MULTIMODAL FAKE NEWS DETECTION

ARTICLES FOR FACULTY MEMBERS

Title/Author	A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks / Song, C., Ning, N., Zhang, Y., & Wu, B.
Source	*Information Processing & Management* Volume 58 Issue 1 (2021) 102437 Pages 1-14 https://doi.org/10.1016/J.IPM.2020.102437 (Database: ScienceDirect)

Title/Author	An inter-modal attention-based deep learning framework using unified modality for multimodal fake news, hate speech and offensive language detection / Ayetiran, E. F., & Özgöbek, Ö.
Source	*Information Systems* Volume 123 (2024) 102378 Pages 1-11 https://doi.org/10.1016/j.is.2024.102378 (Database: ScienceDirect)

Title/Author	A novel hybrid multi-modal deep learning for detecting hashtag incongruity on social media / Dadgar, S., & Neshat, M.
Source	*Sensors* Volume 22 Issue 24 (2022) Pages 1-31 https://doi.org/10.3390/s22249870 (Database: MDPI)

MULTIMODAL FAKE NEWS DETECTION

ARTICLES FOR FACULTY MEMBERS

Title/Author	CAF-ODNN: Complementary attention fusion with optimized deep neural network for multimodal fake news detection / Luvembe, A. M., Li, W., Li, S., Liu, F., & Wu, X.
Source	*Information Processing and Management* Volume 61 Issue 3 (2024) 103653 Pages 1-26 https://doi.org/10.1016/j.ipm.2024.103653 (Database: ScienceDirect)

Title/Author	Combating multimodal fake news on social media: Methods, datasets, and future perspective / Hangloo, S., & Arora, B.
Source	*Multimedia Systems* Volume 28 Issue 6 (2022) Pages 2391–2422 https://doi.org/10.1007/S00530-022-00966-Y (Database: SpringerLink)

Title/Author	Feature importance in the age of explainable AI: Case study of detecting fake news & misinformation via a multi-modal framework / Kumar, A., & Taylor, J. W.
Source	*European Journal of Operational Research* Volume 317 Issue 2 (2024) Pages 401–413 https://doi.org/10.1016/J.EJOR.2023.10.003 (Database: ScienceDirect)

MULTIMODAL FAKE NEWS DETECTION

ARTICLES FOR FACULTY MEMBERS

Title/Author	**Multimodal Fake News Detection / Segura-Bedmar, I., & Alonso-Bartolome, S.**
Source	***Information*** **Volume 13 Issue 6 (2022) Pages 1-16** **https://doi.org/10.3390/info13060284** **(Database: MDPI)**

Title/Author	**Positive unlabeled fake news detection via multi-modal masked transformer network / Wang, J., Qian, S., Hu, J., & Hong, R.**
Source	***IEEE Transactions on Multimedia*** **Volume 26 (2024) Pages 234–244** **https://doi.org/10.1109/TMM.2023.3263552** **(Database: IEEE Xplore)**

Title/Author	**QMFND: A quantum multimodal fusion-based fake news detection model for social media / Qu, Z., Meng, Y., Muhammad, G., & Tiwari, P.**
Source	***Information Fusion*** **Volume 104 (2024) 102172 Pages 1-11** **https://doi.org/10.1016/J.INFFUS.2023.102172** **(Database: ScienceDirect)**

ARTICLES FOR FACULTY MEMBERS

MULTIMODAL FAKE NEWS DETECTION

Title/Author	**A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks / Song, C., Ning, N., Zhang, Y., & Wu, B.**
Source	***Information Processing & Management*** **Volume 58 Issue 1 (2021) 102437 Pages 1-14** **https://doi.org/10.1016/J.IPM.2020.102437** **(Database: ScienceDirect)**

Information Processing and Management 58 (2021) 102437

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks

Chenguang Song [a], Nianwen Ning [a], Yunlei Zhang [b], Bin Wu [*,a]

[a] Beijing Key Laboratory of Intelligence Telecommunication Software and Multimedia, Beijing University of Posts and Telecommunications, Beijing 100876, PR China
[b] North China Institute of Science and Technology, Hebei 065201, PR China

ARTICLE INFO

Keywords:
Fake news detection
Crossmodal attention
Residual network
Convolutional neural network

ABSTRACT

In recent years, social media has increasingly become one of the popular ways for people to consume news. As proliferation of fake news on social media has the negative impacts on individuals and society, automatic fake news detection has been explored by different research communities for combating fake news. With the development of multimedia technology, there is a phenomenon that cannot be ignored is that more and more social media news contains information with different modalities, e.g., texts, pictures and videos. The multiple information modalities show more evidence of the happening of news events and present new opportunities to detect features in fake news. First, for multimodal fake news detection task, it is a challenge of keeping the unique properties for each modality while fusing the relevant information between different modalities. Second, for some news, the information fusion between different modalities may produce the noise information which affects model's performance. Unfortunately, existing methods fail to handle these challenges. To address these problems, we propose a multimodal fake news detection framework based on **C**rossmodal **A**ttention **R**esidual and **M**ultichannel convolutional neural **N**etworks (CARMN). The **C**rossmodal **A**ttention **R**esidual **N**etwork (CARN) can selectively extract the relevant information related to a target modality from another source modality while maintaining the unique information of the target modality. The **M**ultichannel **C**onvolutional neural **N**etwork (MCN) can mitigate the influence of noise information which may be generated by crossmodal fusion component by extracting textual feature representation from original and fused textual information simultaneously. We conduct extensive experiments on four real-world datasets and demonstrate that the proposed model outperforms the state-of-the-art methods and learns more discriminable feature representations.

1. Introduction

As its low cost, convenience, and rapid propagation of information, social media has gradually become one of the important platforms for people to seek out and consume news in recent years Shu, Cui, Wang, Lee, and Liu (2019); Shu, Sliva, Wang, Tang, and Liu (2017); Zhang and Ghorbani (2020). Compared with traditional news media, the lack of effective supervision measures for social

* Corresponding author.
E-mail address: wubin@bupt.edu.cn (B. Wu).

https://doi.org/10.1016/j.ipm.2020.102437
Received 29 February 2020; Received in revised form 27 September 2020; Accepted 8 November 2020
Available online 16 November 2020

media weakens the journalistic norms of objectivity. One can publish and spread fake news on social media at a very low cost. The proliferation of fake news on social media will bring negative impacts on both individuals and society. It may undermine the traditional news sources that have enjoyed high levels of public trust and credibility and harm stability and harmony of society Lazer et al. (2018).

One of the methods to mitigate the serious negative effects caused by the fake news is manual fact-checking Zhou, Zafarani, Shu, and Liu (2019b), which includes expert-based fact-checking and crowd-sourced manual fact-checking. Expert-based fact-checking can obtain high accuracy but needs intensive labor and cost of time and has difficulty in scaling with emerging fake news. Crowd-sourced manual fact-checking do well in scalability but will get a relatively less credible label, which fails to meet the qualification of accurate fake news detection. As the limitations of manual fact-checking approaches, automatic fake news detection techniques have been developed to solve the problem Zhou et al. (2019b). Some early researchers try to manually design a series of features which are fed into a machine learning model to identify fake news Castillo, Mendoza, and Poblete (2011); Sejeong, Meeyoung, Kyomin, Wei, and Yajun (2013); Yang, Liu, Yu, and Yang (2012), but these methods are still time-consuming and poor in generalizability.

As the powerful ability of the deep neural networks (DNN) to automatic capture complex patterns, it was introduced to alleviate the shortcomings of traditional methods. Most of existing studies are mainly focus on using the textual features to detect fake news Oshikawa, Qian, and Wang (2018); Su, Macdonald, and Ounis (2019). However, there is a phenomenon that cannot be ignored is that more and more social media news contains information with different modalities such as texts, pictures, and videos. There are complementary and enhanced relationships between different modalities Cao et al. (2018); Cui, Wang, and Lee (2019); Zhao et al. (2019). More importantly, news with visual information is likely to attract much more attention from users and thus gains a larger propagation range Jin, Cao, Guo, Zhang, and Luo (2017a); Qi, Cao, Yang, Guo, and Li (2019). But limited work has been performed on verifying the credibility of news by exploiting visual information. Jin et al. first proposed a Recurrent Neural Networks (RNN)-based automatic multimodal fake news detection model, in which the multimodal features are fused via attention mechanism Jin et al. (2017a). Wang et al. proposed a multi-task learning framework to learn both textual and visual transferable feature representations among all the posts by leveraging an additional event discriminator Wang et al. (2018b). A similar idea is that Zhang et al. proposed an event memory network module to learn invariant features among different events Zhang, Fang, Qian, and Xu (2019). Khattar et al. proposed a multimodal fusion fake news detection framework based on Variational Autoencoder (VAE) Khattar, Goud, Gupta, and Varma (2019).

First, despite great progress has been made in previous research, an important problem is ignored—how to keep the unique properties for each modality while fusing the relevant information between different modalities. Textual and visual feature representations are learned by different ways and should have their own unique characteristic. It is not a good choice to fuse different modal feature representations to one. Different modal feature separately representations will fail to fuse the correlative and complementary information between different modalities. Second, it should be noted that multimodal fake news detection task usually uses high-level image embeddings and low-level sentence embeddings Jin et al. (2017a) and the visual feature representation is extracted from the model pretrained on Imagenet set Simonyan and Zisserman (2015), which means that it is impossible to accurately match text and image information. And fake news images from social media have more complex patterns at both physical and semantic levels Cao et al. (2020); Qi et al. (2019). For some posts and their attached images, the visual feature representations extracted from pretrained model are not always what we expect. The fusion between textual and visual information may produce noise information which may affect model's performance. Thus, we should consider both original and fused text information simultaneously. Existing multimodal fake news detection methods fail to meet these requirements.

To overcome the limitations of existing approaches, a multimodal fake news detection model based on **C**rossmodal **A**ttention **R**esidual and **M**ultichannel convolutional neural **N**etworks (CARMN) is proposed in this paper. The **C**rossmodal **A**ttention **R**esidual **N**etwork (CARN) can selectively extract the information related to a target modality from another source modality while maintaining the unique information of the target modality. The **M**ultichannel **C**onvolutional neural **N**etwork (MCN) can extract textual feature representation from original and fused textual information simultaneously and mitigate the influence of noise information which may be generated by crossmodal fusion component Wang, Zhang, Xie, and Guo (2018a). At present, there are only a few reliable social-media-oriented multimodal fake news detection datasets. Thus, we collected a large number of reliable fake and real news from the Weibo platform[1]. Our main contributions can be summarized as follows.

- We present a novel multimodal fake news detection model based on CARN and MCN.
- The CARN is introduced to fuse the relevant information between different modalities and keep the unique properties for each modality.
- To mitigate the influence of noise information which may be generated by crossmodal fusion, the MCN is introduced to extract feature representations from original and fused textual information simultaneously.
- We conduct extensive experiments on four real-world datasets and demonstrate that the proposed model outperforms state-of-the-art methods and learn more discriminable feature representations.
- We contribute a large scale multimodal fake news dataset from Weibo platform and will make it available to the public[2].

[1] http://www.weibo.com/
[2] https://github.com/lumen2018/dataset

2. Related work

2.1. The definition of fake news

Fake news overlaps with other concepts and terms such as false news, rumor, and disinformation Ajao, Bhowmik, and Zargari (2018); Bondielli and Marcelloni (2019); Lazer et al. (2018). An universal definition for fake news is still missing so far Zhou et al. (2019b). Similar to previous work Khattar et al. (2019); Wang et al. (2018b), we define fake news to be verifiable false news.

2.2. Fake news detection

The fake news detection task. The fake news detection task aims to assess the authenticity for the given news Kakol, Nielek, and Wierzbicki (2017); Zhou et al. (2019b). Most existing approaches formulate the fake news detection problem as a binary classification problem (fake or real) Shu et al. (2017). In some cases, it also is considered as multi-classification Karimi, Roy, Saba-Sadiya, and Tang (2018), regression, or clustering problems Oshikawa et al. (2018). The way of binary classification is adopted in this paper. The literature on fake news detection is extensive. We will provide a brief review of the work from the following categories: text-based, user-based, propagation-based and multimodal fake news detection.

Text-based fake news detection. The early studies obtain text-based features by manual linguistic cues selection Rubin, Chen, and Conroy (2015); Ruchansky, Seo, and Liu (2017). For fake news detection task, it is difficult to generalize hand-crafted linguistic features across topics and domains Sharma et al. (2019). A RNN-based model was introduced to automatically learn the hidden representation of temporal textual feature, which outperforms the methods leveraging hand-crafted features Ma et al. (2016). In order to capture the long-range dependency among variable length sequential information, Chen et al. adopted soft-attention and RNN to learn selectively temporal feature representation of post series Chen, Li, Yin, and Zhang (2018). Similarly, Yu et al. proposed a convolutional neural networks (CNN)-based model which is used to extract low-level local-global features from the input sequences and then construct high-level interactions among important features Yu, Liu, Wu, Wang, and Tan (2017). By exploiting the users' feedback towards a target claim, stance information was proved to be a strong indicator for classification Dungs, Aker, Fuhr, and Bontcheva (2018); Kochkina, Liakata, and Zubiaga (2018); Ma, Gao, and Wong (2018a), but each response has to be given a special stance label, which is laborious. Inspired by Generative Adversarial Networks (GAN), Ma et al. proposed a GAN-style fake news detection model Ma, Gao, and Wong (2019). Textual feature representation is improved by adversarial learning between text generator and fake news discriminator. Scholars also explored text-based fake news detection with various way such as user response generating Qian, Gong, Sharma, and Liu (2018), text generation Vo and Lee (2019), reinforcement learning Zhou, Shu, Li, and Lau (2019a), fact-checking url recommendation Vo and Lee (2018) and attention-residual network Chen, Sui, Hu, and Gong (2019).

User-based and propagation-based fake news detection. Apart from textual features, user profiles and propagation-based features as auxiliary information are also used to help differentiate fake news. Shu et al. provided a systematic research about the relationship between user profiles and the credibility of news Shu, Wang, and Liu (2018). Guo et al. fused the propagation features and user profiles with textual features via attention mechanism Guo, Cao, Zhang, Guo, and Li (2018). In addition, diffusion-based models have been introduced to solve this problem. Vosoughi et al. claimed that fake news tend to spreads faster, farther and more broadly than the truth on social network Vosoughi, Roy, and Aral (2018). According to supporting and opposing relations among posts, Jin et al. designed a homogeneous stance signed network to evaluate news credibility Jin, Cao, Zhang, and Luo (2016). Similarly, by exploiting post-repost relationships, Ma et al. proposed two kinds of recursive neural network models based on bottom-up and top-down tree-structured Ma, Gao, and Wong (2018b).

Multimodal fake news detection. Different from all the aforementioned work, visual information, as auxiliary information, also has been adopted to infer the veracity of news articles Gupta, Lamba, Kumaraguru, and Joshi (2013); Gupta, Zhao, and Han (2012); Ke, Song, and Kenny Q (2015). There only a few studies that focus on the correlation between image and credibility of tweets Cao et al. (2018). By introducing some features from the field of image retrieval, Jin et al. first provided a systematic research on image features between fake and real news Jin, Cao, Zhang, Zhou, and Tian (2017b). However, these features are still hand-crafted and do not capture the complex visual content information. Inspired by DNN that achieved impressive results for image and text feature representation task, Jin et al. proposed a RNN-based multimodal fusion fake news detection framework Jin et al. (2017a). The high-level visual features and high-level textual and social features are fused by attention mechanism. Wang et al. proposed a multi-task learning model to learn textual and visual transferable feature representations among all the posts by removing textual and visual event-specific information Wang et al. (2018b). Similarly, for the event-level fake news detection task, Zhang et al. used the memory network to learn event invariant features and obtained better generalizability for newly emerged events Zhang et al. (2019). In order to learn a shared latent representation across modalities, Khattar et al. proposed a multi-modal fusion framework based on VAE Khattar et al. (2019). Recently, transfer learning-based methods also have been introduced to verify the authenticity of news Singhal et al. (2020); Singhal, Shah, Chakraborty, Kumaraguru, and Satoh (2019).

3. Problem formulation

There are two ways to detect fake news: post-level or tweet-level (to identify a single post is fake/real news) and event-level (to identify a news which include a group of posts is fake/real). Our research falls in the former. Let T be a post, $T = [w_1, w_2, \cdots w_n]$, where n is the number of words w and P is an attached image of the post T. Given a post T and an attached image P, the task of this paper is to

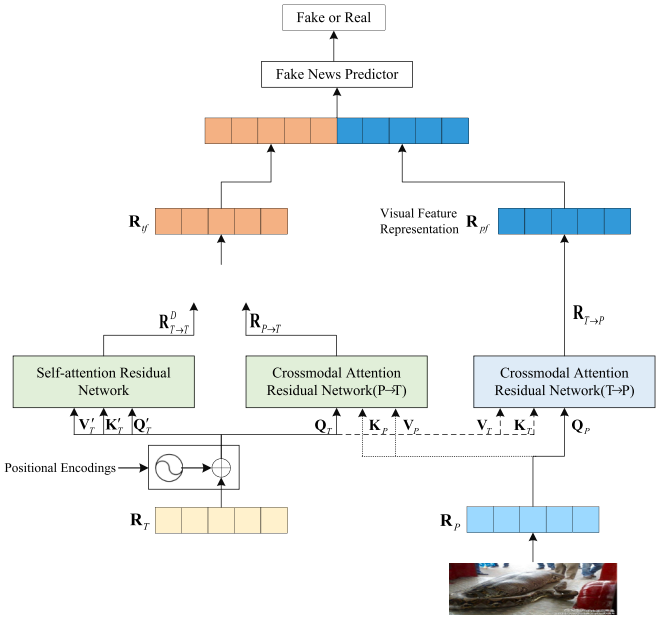


Fig. 1. The proposed model framework.

identify the post T is real ($y = 0$) or fake ($y = 1$) news by learning a fake news detection function $F : F(T, P) \rightarrow (\hat{y})$.

4. Model

4.1. Model framework

The overall structure of the proposed model is shown in Fig. 1. Our model consists of: (1) input embedding layer (to get word embedding matrix $\mathbf{R}_T$ and image embedding matrix $\mathbf{R}_P$), (2) CARN layer (to reinforce the target modality feature representation by selectively extracting information from another source modality) and self-attention residual network layer (to capture the interactions between different sequence element pairs and transmit original textual information to MCN), (3) MCN layer (to alleviate the effect of noise information which may be generated by CARN layer and extract the final textual feature representation $\mathbf{R}_{tf}$), (4) fake news prediction component (to predict a post is real or fake news). Next, we will present the details of the proposed fake news detection model.

4.2. Input embeddings

4.2.1. Word-Level sentence embeddings

For a sentence $T = \{w_1, w_2, \cdots w_n\}$, each w_i represent $i - th$ word of the sentence T and n is the length of the sentence T. Then, we convert each word of sentence T to a pretrained word embedding $\mathbf{e}_i$:

$$\mathbf{e}_i = \text{WordEmbed}(w_i) \tag{1}$$

where $\mathbf{e}_i \in \mathbb{R}^{d_T}$, d_T is is the dimension of word embeddings. The word-level sentence embeddings (i.e., word embedding matrix) of the sentence T can be denoted as:

$$\mathbf{R}_T = \{\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_n\} \tag{2}$$

where $\mathbf{R}_T \in \mathbb{R}^{L_T \times d_T}$, L_T equals the length of the sentence T.

4.2.2. Image embeddings

Given m attached images $P = \{p_1, p_2, \cdots p_m\}$ of the sentence T, we extract the initial image embeddings $\mathbf{R}_{vgg}$ from the VGG-19 net pretrained on Imagenet set Simonyan and Zisserman (2015), which followed by a fully connected layer to transform initial image embeddings $\mathbf{R}_{vgg}$ to the image embedding matrix $\mathbf{R}_P$ with same dimension of word embeddings. Note that $m = 1$ in this paper. The image embedding matrix $\mathbf{R}_P$ of the sentence T can be denoted as:

$$\mathbf{R}_P = \sigma\left(\mathbf{R}_{vgg} \times \mathbf{W}_{bf} + b_P\right) \tag{3}$$

where $\mathbf{R}_{vgg} \in \mathbb{R}^{L_P \times d_{vgg}}$, $\mathbf{R}_P \in \mathbb{R}^{L_P \times d_P}$, $\mathbf{W}_{bf} \in \mathbb{R}^{d_{vgg} \times d_P}$ is the weight matrix of the fully connected layer, b_P is bias term, L_P equals to the number of the attached images m, d_{vgg} denotes the output embedding dimension of the VGG-19 ($d_{vgg} = 1000$), d_P represents the dimension of image embeddings and σ is the Leaky ReLU activation function. It should be noted that $d_P = d_T$ and $L_P = m = 1$ in this paper.

4.3. Attention residual network

4.3.1. Positional encoding

Compared to RNN, the attention-based neural networks can improve the training speed and capture longer dependencies in a sentence. However, there is no essential position information in an attention-based network. In order to enable each token of a sequence carries unique position information, the sentence embeddings are added with the positional encoding Vaswani et al. (2017). Given the embedding matrix of sentence T $\mathbf{R}_T \in \mathbb{R}^{L_T \times d_T}$, its positional encoding (PE) can be computed by:

$$\mathrm{PE}(pos, 2j) = \sin\left(pos \,/\, 10000^{2j/d_T}\right) \tag{4}$$

$$\mathrm{PE}(pos, 2j+1) = \cos\left(pos \,/\, 10000^{2j/d_T}\right) \tag{5}$$

where $pos \in [0, \cdots, L_T]$ is position, $j \in [0, d_T/2)$ is the dimension. Each dimension of the PE corresponds to a sinusoid. Then, the position information is added to a sentence representation by summing token embeddings and corresponding PE (i.e., $\mathbf{R}_T + \mathrm{PE}(\mathbf{R}_T)$).

4.3.2. Crossmodal and unimodal attention

As scaled dot-product attention is the core component of our model, we will provide the definitions of single head crossmodal attention, single head unimodal self-attention and multi-head unimodal self-attention, respectively Tsai et al. (2019); Vaswani et al. (2017). The task of crossmodal attention is to capture the relevant and complementary information between textual and visual information. When pass information from the sentence T to its attached image P (i.e., $T \rightarrow P$), the Queries, Keys and Values are defined as $\mathbf{Q}_P = \mathbf{R}_P \times \mathbf{W}_{Q_P}$, $\mathbf{K}_T = \mathbf{R}_T \times \mathbf{W}_{K_T}$ and $\mathbf{V}_T = \mathbf{R}_T \times \mathbf{W}_{V_T}$, where $\mathbf{W}_{Q_P} \in \mathbb{R}^{d_P \times d_k}$, $\mathbf{W}_{K_T} \in \mathbb{R}^{d_T \times d_k}$ and $\mathbf{W}_{V_T} \in \mathbb{R}^{d_T \times d_v}$. Note that $d_k = d_v = d_T$. The single head crossmodal attention function $Att_{T \rightarrow P} \in \mathbb{R}^{L_P \times d_v}$ is defined as follows.

$$\begin{aligned} Att_{T \rightarrow P} &= \mathrm{softmax}\left(\mathbf{Q}_P \times \mathbf{K}_T^{\top} \Big/ \sqrt{d_k}\right) \times \mathbf{V}_T \\ &= \mathrm{softmax}\left(\mathbf{R}_P \times \mathbf{W}_{Q_P} \times \mathbf{W}_{K_T}^{\top} \times \mathbf{R}_T^{\top} \Big/ \sqrt{d_k}\right) \times \mathbf{R}_T \times \mathbf{W}_{V_T} \end{aligned} \tag{6}$$

when the information from modality P is passed to modality T:

$$Att_{P \rightarrow T} = \left[\mathrm{softmax}\left(\mathbf{Q}_T \times \mathbf{K}_P^{\top} \Big/ \sqrt{d_k}\right)^{\top}\right]^{\top} \times \mathbf{V}_P \tag{7}$$

where $\mathbf{Q}_T = \mathbf{R}_T \times \mathbf{W}_{Q_T}$, $\mathbf{K}_P = \mathbf{R}_P \times \mathbf{W}_{K_P}$, $\mathbf{V}_P = \mathbf{R}_P \times \mathbf{W}_{V_P}$, $\mathbf{W}_{Q_T} \in \mathbb{R}^{d_T \times d_k}$, $\mathbf{W}_{K_P} \in \mathbb{R}^{d_P \times d_k}$, $\mathbf{W}_{V_P} \in \mathbb{R}^{d_P \times d_v}$ and $Att_{P \rightarrow T} \in \mathbb{R}^{L_T \times d_v}$. Similarly, the single head unimodal self-attention function $Att_{T \rightarrow T} \in \mathbb{R}^{L_T \times d_v}$ can be represented as:

$$Att_{T \rightarrow T} = \mathrm{softmax}\left[\mathbf{Q}'_T \times (\mathbf{K}'_T)^{\top} \Big/ \sqrt{d_k}\right] \times \mathbf{V}'_T \tag{8}$$

where $\mathbf{Q}'_T = \mathbf{R}_T \times \mathbf{W}'_{Q_T}$, $\mathbf{K}'_T = \mathbf{R}_T \times \mathbf{W}'_{K_T}$, $\mathbf{V}'_T = \mathbf{R}_T \times \mathbf{W}'_{V_T}$, $\mathbf{W}'_{Q_T} \in \mathbb{R}^{d_T \times d_k}$, $\mathbf{W}'_{K_T} \in \mathbb{R}^{d_T \times d_k}$ and $\mathbf{W}'_{V_T} \in \mathbb{R}^{d_T \times d_v}$.

Compared to single head attention, previous work has shown that multi-head attention can make more efficient use of context information Vaswani et al. (2017). The $\mathbf{Q}'_T$, $\mathbf{K}'_T$, and $\mathbf{V}'_T$ are divided into H different subspaces by exploiting H different, learnable linear projections. The Queries, Keys, and Values of the $h-th$ head can be represented as $\mathbf{Q}'_{T,h} = \mathbf{Q}'_T \times \mathbf{W}'_{Q_T,h}$, $\mathbf{K}'_{T,h} = \mathbf{K}'_T \times \mathbf{W}'_{K_T,h}$ and $\mathbf{V}'_{T,h} = \mathbf{V}'_T \times \mathbf{W}'_{V_T,h}$, respectively. Note that $\mathbf{W}'_{Q_T,h} \in \mathbb{R}^{d_T \times \frac{d_k}{H}}$, $\mathbf{W}'_{K_T,h} \in \mathbb{R}^{d_T \times \frac{d_k}{H}}$, $\mathbf{W}'_{V_T,h} \in \mathbb{R}^{d_T \times \frac{d_v}{H}}$. The unimodal self-attention function of $h-th$ head $Att_{T \rightarrow T,h} \in \mathbb{R}^{L_T \times \frac{d_v}{H}}$ is defined as follows.

$$Att_{T \rightarrow T,h} = \mathrm{softmax}\left[\mathbf{Q}'_{T,h} \times \left(\mathbf{K}'_{T,h}\right)^{\top} \Big/ \sqrt{d_k/H}\right] \times \mathbf{V}'_{T,h} \tag{9}$$

The outputs of all the heads are concatenated together and then are linearly transformed to form multi-head unimodal self-attention function:

$$Att^{mul}_{T \rightarrow T} = \mathrm{Concat}\left[Att_{T,0}, Att_{T,1}, \cdots, Att_{T,H}\right] \times \mathbf{W}_{mul} \tag{10}$$

where $Att^{mul}_{T \rightarrow T} \in \mathbb{R}^{L_T \times d_v}$, $\mathbf{W}_{mul} \in \mathbb{R}^{d_v \times d_v}$.

4.3.3. Crossmodal and unimodal attention residual network

After introducing some preliminary definitions, we will present the structure of CARN module in detail. The target modality

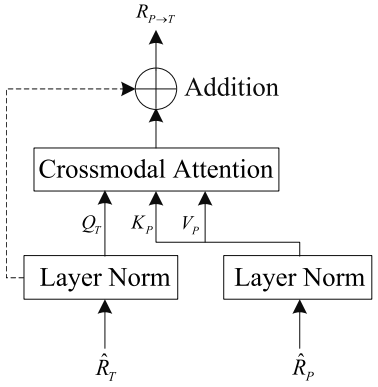


Fig. 2. An illustration for CARN.

selectively extract information from another source modality by exploiting crossmodal attention network. Then, the information is added to the target modality with residual connection. We take pass information from an attached image of the sentence T to the sentence T (i.e., $P \to T$) as an example to make introductions. The overall architecture of the CARN is shown in Fig. 2. To make use of the order of the sequence T, the temporal information is added to the sentence by using PE. It can be computed by:

$$\begin{cases} \widehat{\mathbf{R}}_P = \mathbf{R}_P \\ \widehat{\mathbf{R}}_T = \mathrm{PE}(\mathbf{R}_T) + \mathbf{R}_T \end{cases} \tag{11}$$

Next, the CARN module can be computed by:

$$\mathbf{R}_{P\to T} = Att_{P\to T}\left(\mathrm{LN}\left(\widehat{\mathbf{R}}_P\right), \mathrm{LN}\left(\widehat{\mathbf{R}}_T\right)\right) + \mathrm{LN}\left(\widehat{\mathbf{R}}_T\right) \tag{12}$$

where $\mathbf{R}_{P\to T} \in \mathbb{R}^{L_T \times d_v}$ and LN represents layer normalization Ba, Kiros, and Hinton (2016). Similar to CARN, for a D layers unimodal self-attention residual network (UARN) block, each layer i can be computed by:

$$\begin{aligned} \mathbf{R}^{\mathbf{0}}_{\mathbf{T}\to\mathbf{T}} &= \widehat{\mathbf{R}}_{\mathbf{T}} \\ \mathbf{R}^{i}_{T\to T} &= Att^{mul}_{T\to T}\left(\mathrm{LN}\left(\mathbf{R}^{(i-1)}_{T\to T}\right), \mathrm{LN}\left(\mathbf{R}^{(i-1)}_{T\to T}\right)\right) + \mathrm{LN}\left(\mathbf{R}^{(i-1)}_{T\to T}\right) \end{aligned} \tag{13}$$

where $\mathbf{R}^{i}_{T\to T} \in \mathbb{R}^{L_T \times d_v}$. For simplicity, $\mathbf{R}_{pt} = \mathbf{R}_{P\to T}$, $\mathbf{R}_{tt} = \mathbf{R}^{D}_{T\to T}$. When the target modality is visual information, $\mathbf{R}_{tp} = \mathbf{R}_{T\to P}$.

4.4. Feature extractor

4.4.1. Textual feature extractor

We employed a multi-channel and word-word-aligned CNN-based architecture network (i.e., MCN) to extract the key features from textual information (i.e., $\mathbf{R}_{pt}$ and $\mathbf{R}_{tt}$) processed by CARN and UARN module Kim (2014); Wang et al. (2018a). We first align and stack the embedding matrices $\mathbf{R}_{pt}$ and $\mathbf{R}_{tt}$ as

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}_{pt} \\ \mathbf{R}_{tt} \end{bmatrix} \tag{14}$$

where $\mathbf{R} \in \mathbb{R}^{2\times L_T \times d_v}$. The convolutional filters $\mathbf{W}_c \in \mathbb{R}^{2\times l\times d}$ with various windows size l are used to extract information from embedding matrix $\mathbf{R}$ and produce $L_T - l + 1$ new features. When a filter start with $i - th$ word, the new feature can be denoted as:

$$r_i = \sigma(\mathbf{W}_c \cdot \mathbf{R}_{i:i+l-1} + b_W) \tag{15}$$

where σ is Leaky ReLU activation function and b_W is bias term. The same convolutional operation is performed on each possible window of words in this sentence which generates a feature vector.

$$\mathbf{r} = [r_1, r_2, \cdots, r_{L_T-l+1}] \tag{16}$$

Next, we extract the maximum $\tilde{\mathbf{r}} = \max(\mathbf{r})$ by performing the max-over-time pooling operation on the feature vector $\mathbf{r} \in \mathbb{R}^{L_T-l+1}$. Suppose there being n_W different filters with window size l for the sentence, its feature representation $\tilde{\mathbf{r}}^l \in \mathbb{R}^{n_W}$ can be denoted as:

$$\tilde{\mathbf{r}}^l = \left[\tilde{\mathbf{r}}_1, \tilde{\mathbf{r}}_2 \cdots, \tilde{\mathbf{r}}_{n_W}\right] \tag{17}$$

Table 1
The statistics of datasets

Dataset	# of fake news	# of real news	# of images
Tweet	7,898	6,026	514
Weibo A	4,103	3,605	7,708
Weibo B	5,076	5,008	10,084
Weibo C	5,065	5,065	10,130

Suppose there being n_l filters with different size (i.e., $l \in [1, \cdots, n_l]$),

$$\tilde{\mathbf{R}} = \text{Concat}\left[\tilde{\mathbf{r}}^1, \cdots, \tilde{\mathbf{r}}^{n_l}\right] \tag{18}$$

where $\tilde{\mathbf{R}} \in \mathbb{R}^{d_R}$ and $d_R = n_W \times n_l$. For the sentence T, the final textual feature representation $\mathbf{R}_{tf} \in \mathbb{R}^{d_T}$ is:

$$\mathbf{R}_{tf} = \sigma\left(\mathbf{W}_T \times \tilde{\mathbf{R}} + b_T\right) \tag{19}$$

where σ is Leaky ReLU activation function, $\mathbf{W}_T$ is weight matrix and b_T is bias term.

4.4.2. Visual feature extractor

As a post only with an attached image in our model, we adopted the output (i.e., $\mathbf{R}_{ip}$) of CARN module as the final visual feature representation $\mathbf{R}_{pf} \in \mathbb{R}^{d_P}$.

4.5. Fake news predictor and model learning

We have introduced the mainly modules of this paper. $\mathbf{R}_{tf}$ and $\mathbf{R}_{pf}$ are concatenated together and then are fed to a softmax layer to make the final prediction. The fake news predictor is defined as:

$$\hat{y} = \text{softmax}\left(\mathbf{W} \times \left[\mathbf{R}_{tf}, \mathbf{R}_{pf}\right] + b\right) \tag{20}$$

where $\mathbf{W}$ is the parameters of softmax layer, b is bias term and $\hat{y} = [\hat{y}_0, \hat{y}_1]$. $\hat{y}_0$ and $\hat{y}_1$ denote the probability of a given news is real(0) or fake(1), respectively. We adopt cross-entropy to define the loss function $L(\theta)$ as follows.

$$L(\theta) = -y\log\left(\hat{y}_1\right) - (1-y)\log\left(\hat{y}_0\right) \tag{21}$$

where θ is model parameters. The model aim at minimizing the loss function L_θ for each news by learning θ through back-propagation. We use stochastic gradient decent to train the model and choose Adam as the optimizer with learning rate decay.

5. Experiments

In this section, first, we present the information of four large social media datasets. Second, we provide an introduction about model settings and baseline methods. Third, we make comparisons between the model and baseline methods on four datasets and then bring an detail analysis for experimental results.

5.1. Datasets

The Twitter and Weibo A multimodal dataset are widely adopted by previous work. In addition, in this work, we introduce two new multi-modal fake news datasets for the first time. The detailed statistics information of four datasets is shown in Table 1.

- Twitter Dataset. The Twitter dataset derives from the Verifying Multimedia Use task, the goal of which is to distinguish fake/real news on Twitter with automatic method Boididou et al. (2016). It consists of development set and test set. There is no overlapping events between development set and test set. For each piece of data, it contains text component, an associated image/video and additional user profile information. We only keep the data with text content and attached images.
- Weibo A Dataset. The Weibo A datatset is first presented in Jin et al. (2017a) for fake news detection task. Each post contains text content, user profiles and attached images. The verified fake news is collected from the official fake news debunking system of the Sina Weibo, a website very similar to Twitter. The time span of the data is from May 2012 to January 2016. Jin et al. adopted the news verified by the Xinhua News Agency as real news Jin et al. (2017a). Following the data preprocessing methods of previous research, the low quality and duplicated images are taken away Slaney and Casey (2008). To avoid the same events among training, validation and testing set, we find the same events by exploiting a single-pass clustering method Yang, Pierce, and Carbonell (1998).

- Weibo B Dataset. The Weibo B dataset, a benchmark dataset for internet fake news detection challenger[3], is released by Cao et al. Cao et al. (2018). For each post, it contains text content, attached images, user profile information, news category and corresponding ground-truth label. We take the way same to Weibo A dataset to preprocess and split the Weibo B dataset.
- Weibo C Dataset. To promote fake news detection task, we build a new multi-modal fake news detection dataset. The fake news is collected from the Weibo community management center[4], an official fake news debunking system. The time span of this data is from May 2012 to November 2019. The real news is collected from the People's Daily, an authoritative news source similar to the Xinhua News Agency. Each post contains the original post text, associated images/video and additional user profile information. Following previous research, we preprocess and split this dataset with the way same to Weibo A and B datasets.

5.2. Experimental setup

Following previous work Wang et al. (2018b), development set and test set of the Tweet dataset is used as training set and test set, respectively. For each Weibo datasets, we choose 70%, 10% and 20% of news for training, validation and testing set, respectively. Same to previous research, we obtain 32-dimensional word embedding (i.e., $d_T = 32$) by exploiting Word2Vec model Mikolov, Sutskever, Chen, Corrado, and Dean (2013). For textual feature extractor, the window size of the filter is $l \in [1, 2, 3, 4]$ (i.e., $n_l = 4$) and for each size l, the number of filters is $n_W = 25$. For CARN, the number of layers and heads has little effect on the experimental results, so we choose single head and layer attention residual network. For UARN, we choose the attention residual network with 4 heads and 3 layers (i.e., $H = 4$ and $D = 3$), which achieves the best performance. In the process of training, the batch size and the number of epochs is set to 150. We choose Accuracy, Precision, Recall and F_1 score as evaluation metrics which are widely adopted by related areas Shu et al. (2017).

5.3. Baselines

We make comparisons with a series of baseline fake news detection methods as follows.

5.3.1. Single modality models

- Textual. As the input of model is only post, the CARN module is removed. The output of sentence embedding layer is fed into single channel CNN-based textual feature extractor Kim (2014), which followed by a fully connected layer and softmax layer.
- Visual. The visual features are obtained from the VGG-19 net. After processed by input embedding layer, The visual information is fed into a fully connected layer and softmax layer for making final prediction.

5.3.2. Multimodal models

- VQA Antol et al. (2015). The goal of Visual Question Answering (VQA) is to provide an answer to a question about a given image. As VQA is a multi-classification task, we have to replace the multi-class classifier with a binary classifier. For a fair comparison, we choose a single layer LSTM with hidden layer size 32.
- NeuralTalk Vinyals, Toshev, Bengio, and Erhan (2015). The NeuralTalk model is proposed to generate natural language descriptions from visual information. To adapt the model to fake news detection task, its feature representation is defined as the average of the output of RNN at each time step. Then, the feature representations are fed into a fully connected layer to make prediction. For a fair comparison, we choose both LSTM and fully connected with the hidden layer size 32.
- att-RNN Jin et al. (2017a). att-RNN is a RNN-based automatic multimodal fake news detection model which fuses joint representation of textual features and user profile features and visual features via attention mechanism. For a fair comparison, the user profile information is removed and the hidden layer size of LSTM is 32.
- EANN Wang et al. (2018b). The Event Adversarial Neural Networks (EANN) is a multi-task learning fake news detection model, which aims at learning shared feature representations among all the posts by leveraging an additional adversarial component. Textual and visual feature representations are obtain by exploiting a CNN-based textual features extractor Kim (2014) and VGG-19 network, respectively. For a fair comparison, we remove the adversarial component.
- MVAE Khattar et al. (2019). The state-of-the-art method, the Multimodal Variational Autoencoder (MVAE), is a multi-task learning multimodal fusion fake news detection framework. The modal aims at discovering correlations across modalities by exploiting VAE to reconstructs the textual and visual feature representations from the shared latent representation.
- MKN Zhang et al. (2019). Multi-modal Knowledge-aware Event Memory Network (MKEMN) is event-level multi-modal fake news detection framework, which use the visual information and the external knowledge to assist fake news detection task. The authors adopted an event memory network to learn event invariant features. Considering the differences between event-level and post-level fake news detection and the fairness of comparison, we remove the external knowledge component and event memory network. The modified method is denoted as MKN.

[3] https://biendata.com/competition/falsenews/
[4] http://service.account.weibo.com/

Table 2
The experimental results of different methods on Twitter dataset.

Dataset	Method	Accuracy	Fake News			Real News		
			Precision	Recall	F_1	Precision	Recall	F_1
Twitter	Textual	0.568	0.655	0.379	0.480	0.531	0.778	0.631
	Visual	0.664	0.733	0.568	0.640	0.617	0.770	0.685
	VQA	0.631	0.765	0.509	0.611	0.550	0.794	0.650
	Neural Talk	0.610	0.728	0.504	0.595	0.534	0.752	0.625
	att-RNN	0.681	0.769	0.561	0.650	0.626	0.813	0.707
	EANN	0.677	0.750	0.579	0.653	0.627	0.786	0.699
	MVAE	0.578	0.626	0.488	0.548	0.544	0.677	0.603
	MKN	0.664	0.753	0.537	0.627	0.611	0.805	0.695
	CARMN	**0.741**	**0.854**	**0.619**	**0.718**	**0.670**	**0.880**	**0.760**

Table 3
The experimental results of different methods on three Weibo datasets.

Dataset	Method	Accuracy	Fake News			Real News		
			Precision	Recall	F_1	Precision	Recall	F_1
Weibo A	Textual	0.764	0.776	0.721	0.747	0.755	0.805	0.779
	Visual	0.594	0.583	0.752	0.657	0.615	0.424	0.502
	VQA	0.579	0.581	0.665	0.620	0.576	0.487	0.527
	Neural Talk	0.748	0.739	0.790	0.764	0.758	0.702	0.730
	att-RNN	0.784	0.797	0.781	0.789	0.771	0.787	0.779
	EANN	0.807	0.831	0.788	0.809	0.785	0.828	0.806
	MVAE	0.681	0.756	0.589	0.662	0.630	0.785	0.698
	MKN	0.792	0.805	0.788	0.796	0.778	0.796	0.787
	CARMN	**0.853**	**0.891**	**0.814**	**0.851**	**0.818**	**0.894**	**0.854**
Weibo B	Textual	0.762	0.861	0.623	0.723	0.706	0.900	0.791
	Visual	0.702	0.734	0.630	0.678	0.678	0.773	0.722
	VQA	0.704	0.706	0.695	0.701	0.702	0.713	0.707
	Neural Talk	0.735	0.778	0.652	0.709	0.704	0.817	0.756
	att-RNN	0.780	0.853	0.675	0.753	0.733	0.884	0.801
	EANN	0.815	0.903	0.703	0.791	0.759	0.925	0.834
	MVAE	0.741	0.779	0.671	0.721	0.713	0.811	0.759
	MKN	0.778	0.880	0.643	0.743	0.720	0.913	0.805
	CARMN	**0.869**	**0.935**	**0.796**	**0.860**	**0.820**	**0.944**	**0.878**
Weibo C	Textual	0.772	0.742	0.844	0.790	0.812	0.697	0.750
	Visual	0.831	0.806	0.882	0.842	0.864	0.779	0.820
	VQA	0.807	0.742	0.953	0.834	0.931	0.657	0.770
	Neural Talk	0.796	0.751	0.897	0.817	0.867	0.691	0.769
	att-RNN	0.834	0.778	0.942	0.852	0.923	0.722	0.810
	EANN	0.858	0.807	0.948	0.872	0.934	0.765	0.841
	MVAE	0.821	0.781	0.901	0.837	0.878	0.737	0.802
	MKN	0.842	0.786	0.947	0.859	0.930	0.733	0.820
	CARMN	**0.922**	**0.890**	**0.965**	**0.926**	**0.961**	**0.876**	**0.917**

5.4. Results and analysis

5.4.1. Comparisons of different models

Table 2 and Table 3 show the performance of the proposed model as well as baseline methods in fake news detection task on Twitter and Weibo dataset, respectively. We can observe that CARMN outperforms all the competitive models on different metrics. In fact, the Tweet dataset is not a good choice for the post-level fake news detection task. There are multiple languages, many irregularly written texts, and the textual features lack diversity, which is the reason that the performance of the method based on textual information is worse than visual information. By fusing the information across modalities via attention mechanism, att-RNN outperforms all baseline methods, which confirms the effectiveness of using multimodal information simultaneously in the fake news detection task. Compared with the Tweet dataset, there is rich semantic context information in the text of the three Chinese datasets.

On Chinese datasets, the method only based on text features shows similar or even better performance than some baseline methods. There is a phenomenon that can not be ignored is that, on Weibo C dataset, the method based on visual features outperforms the method based on textual feature representation, which can be attributed to the high-level of quality of the pictures attached to the real news. For multimodal models, attention-based methods (i.e., att-RNN and MKN) show better performance than VQA and NeuralTalk but are less effective and robustness than EANN. MVAE shows worse performance, the reason of which is that it only adopts shared latent representation between textual and visual information. It suggests that it is important to keep unique characteristics for each modality. By introducing the CARN and MCN, our model can keep unique characteristics for each modality while fusing the correlative

Table 4
The comparison of experimental results among variants of CARMN.

Method	Accuracy	Fake News			Real News		
		Precision	Recall	F_1	Precision	Recall	F_1
CNN*	0.862	0.815	0.943	0.874	0.930	0.777	0.847
CARN*	0.890	0.850	0.953	0.898	0.944	0.825	0.881
CARMN-	0.913	0.878	0.963	0.919	0.959	0.861	0.907
CARMN	**0.922**	**0.890**	**0.965**	**0.926**	**0.961**	**0.876**	**0.917**

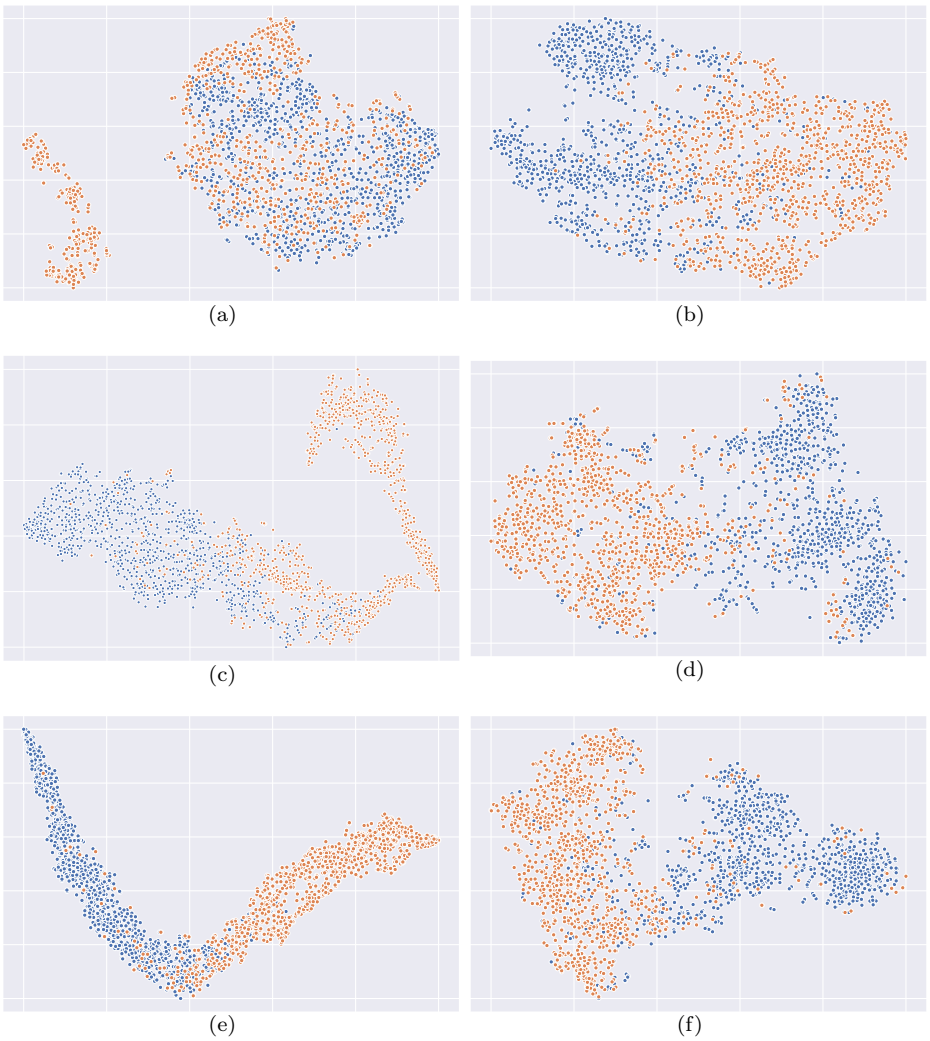


Fig. 3. Visualization of learned latent textual and visual feature representations on the testing data of Weibo C dataset with *t*-SNE, (a) the textual feature representation $\mathbf{R}_{tf}$ learned by CNN*; (b) the visual feature representation $\mathbf{R}_{pf}$ learned by CNN*; (c) the textual feature representation $\mathbf{R}_{tf}$ learned by CARN*; (d) the visual feature representation of $\mathbf{R}_{pf}$ learned by CARN*; (e) the textual feature representation $\mathbf{R}_{tf}$ learned by CARMN; (f) the visual feature representation of $\mathbf{R}_{pf}$ learned by CARMN.

and complementary information between different modalities and alleviate the influence of noise information which may be generated by crossmodal fusion component. The experimental results demonstrate the effectiveness of the proposed model.

5.4.2. Comparisons among variants of CARMN

To further validate the effectiveness of CARMN, we make comparisons with variants of CARMN as follows.

Table 5
The comparison of experimental results between CARMN and SpotFake.

Method	Accuracy	Fake News			Real News		
		Precision	Recall	F_1	Precision	Recall	F_1
SpotFake	0.848	0.804	0.914	0.855	0.904	0.784	0.840
CARMN_Bert	**0.934**	**0.922**	0.952	**0.937**	0.948	**0.916**	**0.932**
CARMN	0.922	0.890	**0.965**	0.926	**0.961**	0.876	0.917

Table 6
The comparison of experimental results between CARMN and SpotFake+.

Method	Accuracy	Fake News			Real News		
		Precision	Recall	F_1	Precision	Recall	F_1
SpotFake+	0.838	0.807	0.882	0.843	0.875	0.796	0.834
CARMN_XLNet	0.922	0.890	**0.965**	**0.926**	**0.961**	0.876	0.917
CARMN	**0.924**	**0.921**	0.930	0.925	0.926	**0.917**	**0.922**

Table 7
The experimental results of different layer number.

# of layer	1	2	3	4	5
Accuracy	0.915	0.917	0.922	0.914	0.906
F1 Score	0.914	0.917	0.921	0.914	0.906

(1) CNN*: The CNN* is the variant of CARMN. It is removed the CARN and UARN. The convolutional filters is changed from $\mathbf{W}_c \in \mathbb{R}^{2\times l\times d}$ to $\mathbf{W}_c \in \mathbb{R}^{1\times l\times d}$.
(2) CARN*: The CARN* is the variant of CARMN. It is removed the UARN. The convolutional filters is changed from $\mathbf{W}_c \in \mathbb{R}^{2\times l\times d}$ to $\mathbf{W}_c \in \mathbb{R}^{1\times l\times d}$.
(3) CARMN-: The CARMN- is the variant of CARMN. It is removed the unimodal attention module.

Due to space limitations, all of the subsequent experimental analyses only focus on the Weibo C dataset. The results are shown in the Table 4. The CARN* is better that CNN* which proves that by fusing the correlative information between different modalities can benefit the model's performance. Compared with CARN*, the usage of the additional residual network and MCN (i.e., CARMN-) can further improve the accuracy and mitigate the influence of noise information which may be generated by crossmodal fusion component. We can achieve the best performance by combining CARMN- with the self-attention module. Then, as shown in the Fig. 3, we visualize the final feature representation $\mathbf{R}_{tf}$ (i.e., a, c, e) and $\mathbf{R}_{pf}$ (i.e., b, d, f) learned by CNN*, CARN* and CARMN with t-SNE van der Maaten and Hinton (2008). The orange and blue color nodes represent fake and real news, respectively. We can observe that our CARMN learns more discriminable feature representations. For textual feature representation, the rank of its discriminability is Fig. 3 (e) > Fig. 3(c) > Fig. 3(a). For visual feature representation, the rank of its discriminability is Fig. 3(f) > Fig. 3(d) > Fig. 3(b). It proves that the CARMN can learn more discriminability feature representations and further validate the effectiveness of the proposed method. The reason why we not to visualize the CARMN- is that the results of CARMN and CARMN- are similar.

5.4.3. Comparisons with transfer learning-based methods

In addition, we also make comparisons with transfer learning-based methods and launch an investigation to find out how the pre-trained Bert Devlin, Chang, Lee, and Toutanova (2018) and XLNet Yang et al. (2019) affect the proposed model's performance. The SpotFake Singhal et al. (2019) and SpotFake+ Singhal et al. (2020), transfer learning-based fake news detection methods, are mainly based on Bert and XLNet model, respectively. The CARMN that takes word embedding representation from Word2Vec model Mikolov et al. (2013) is replaced by CARMN_Bert and CARMN_XLNet that take the representation from pre-trained Bert and XLNet with no fine-tuning. The experimental results are shown in Table 5 and Table 6. Compared with SpotFake and SpotFake+, CARMN_Bert and CARMN_XLNet show better performance. However, the experimental results of CARMN_Bert and CARMN_XLNet are similar to CARMN, which shows that the word embedding representation from pre-trained Bert and XLNet model fail to largely improve the model's performance.

5.4.4. Effects of the number of the heads and layers

In this section, we investigate how the number of the self-attention heads and residual network layers affect the model's performance. Specifically, we set the range of the number of layer to $[1, 2, 3, 4, 5]$. Table 7 shows the performance of CARMN with different layers. The performance of CARMN increases with the number of the layers grows until 3. As the number of the self-attention head

Table 8
The experimental results of different head number.

# of head	1	2	4	8
Accuracy	0.917	0.919	0.922	0.922
F1 Score	0.917	0.919	0.921	0.922

must be divisible by word embedding dimension, we set the range of the number of the head to $[1, 2, 4, 8]$. Table 8 shows the performance of CARMN with different heads. We can observe that the performance of CARMN increases with the number of heads grows until 4. That's the reason why we set the number of the self-attention heads and residual network layers as 4 and 3, respectively.

6. Conclusion

In the field of multimodal fake news detection, there is a challenge of keeping the unique properties for each modality while fusing the relevant information between different modalities. However, for some posts and their attached images, the fusion between textual and visual information may produce noise information which may affect model's performance. To solve these problems, we proposed a multimodal fake news detection model based on CARN and MCN. We conduct extensive experiments on four real-world datasets and demonstrate the effectiveness of the proposed model. As the CARMN is a general model for multimodal fake news detection task, it can be easily expanded to more modalities and the multimodal fusion module can be replaced by other methods. In future work, we will explore event-level multimodal fake news detection by exploiting visual information.

CRediT authorship contribution statement

Chenguang Song: Conceptualization, Methodology, Software, Validation, Resources, Data curation, Writing - original draft, Writing - review & editing, Visualization. **Nianwen Ning:** Resources, Data curation, Software, Writing - review & editing. **Yunlei Zhang:** Writing - original draft, Writing - review & editing. **Bin Wu:** Supervision.

Acknowlgedgments

This work is supported by the National Key Research and Development Program of China (Grant No. 2018YFC0831500), National Natural Science Foundation of China (Grant No. 61972047) and the NSFC-General Technology Basic Research Joint Funds (Grant No. U1936220).

References

Ajao, O., Bhowmik, D., & Zargari, S. (2018). Fake news identification on twitter with hybrid cnn and rnn models, . In *SMSociety '18Proceedings of the 9th international conference on social media and society* (pp. 226–230). ACM. https://doi.org/10.1145/3217804.3217917.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., & Parikh, D. (2015). Vqa: visual question answering. *The ieee international conference on computer vision* (pp. 2425–2433). https://doi.org/10.1007/s11263-016-0966-6.

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization.

Boididou, C., Papadopoulos, S., Dang-Nguyen, D., Boato, G., Riegler, M., Middleton, S. E., … Kompatsiaris, Y. (2016). Verifying multimedia use at mediaeval 2016. *Working notes proceedings of the mediaeval 2016 workshop*. CEUR-WS.org.

Bondielli, A., & Marcelloni, F. (2019). A survey on fake news and rumour detection techniques. *Information Sciences, 497*, 38–55. https://doi.org/10.1016/j.ins.2019.05.035.

Cao, J., Guo, J., Li, X., Jin, Z., Guo, H., & Li, J. (2018). Automatic rumor detection on microblogs: a survey.

Cao, J., Qi, P., Sheng, Q., Yang, T., Guo, J., & Li, J. (2020). *Exploring the role of visual content in fake news detection*. In K. Shu, S. Wang, D. Lee, & H. Liu (Eds.) (pp. 141–161)). Cham: Springer International Publishing.

Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on twitter. *Proceedings of the 20th international conference on world wide web* (pp. 675–684). ACM. https://doi.org/10.1145/1963405.1963500.

Chen, T., Li, X., Yin, H., & Zhang, J. (2018). Call attention to rumors: deep attention based recurrent neural networks for early rumor detection. *Proceedings of the 22nd pacific-asia conference on knowledge discovery and data mining* (pp. 40–52). Springer International Publishing. https://doi.org/10.1007/978-3-030-04503-6_4.

Chen, Y., Sui, J., Hu, L., & Gong, W. (2019). Attention-residual network with cnn for rumor detection. *Proceedings of the 28th acm international conference on information and knowledge management* (pp. 1121–1130). ACM. https://doi.org/10.1145/3357384.3357950.

Cui, L., Wang, S., & Lee, D. (2019). Same: Sentiment-aware multi-modal embedding for detecting fake news. *Proceedings of the 2019 ieee/acm international conference on advances in social networks analysis and mining* (p. 4148). Association for Computing Machinery. https://doi.org/10.1145/3341161.3342894.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding.

Dungs, S., Aker, A., Fuhr, N., & Bontcheva, K. (2018). Can rumour stance alone predict veracity?. *Proceedings of the 27th international conference on computational linguistics* (pp. 3360–3370). Association for Computational Linguistics.

Guo, H., Cao, J., Zhang, Y., Guo, J., & Li, J. (2018). Rumor detection with hierarchical social attention network. *Proceedings of the 27th acm international conference on information and knowledge management* (pp. 943–951). ACM. https://doi.org/10.1145/3269206.3271709.

Gupta, A., Lamba, H., Kumaraguru, P., & Joshi, A. (2013). Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. *Proceedings of the 22nd international conference on world wide web* (pp. 729–736). ACM. https://doi.org/10.1145/2487788.2488033.

Gupta, M., Zhao, P., & Han, J. (2012). Evaluating event credibility on twitter. *Proceedings of the 2012 siam international conference on data mining* (pp. 153–164). Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611972825.14.

Jin, Z., Cao, J., Guo, H., Zhang, Y., & Luo, J. (2017a). Multimodal fusion with recurrent neural networks for rumor detection on microblogs. *Proceedings of the 25th acm international conference on multimedia* (pp. 795–816). ACM. https://doi.org/10.1145/3123266.3123454.

Jin, Z., Cao, J., Zhang, Y., & Luo, J. (2016). News verification by exploiting conflicting social viewpoints in microblogs. *Proceedings of the thirtieth aaai conference on artificial intelligence* (pp. 2972–2978). AAAI Press.

Jin, Z., Cao, J., Zhang, Y., Zhou, J., & Tian, Q. (2017b). Novel visual and statistical image features for microblogs news verification. *IEEE Transactions on Multimedia, 19*(3), 598–608. https://doi.org/10.1109/TMM.2016.2617078.

Kakol, M., Nielek, R., & Wierzbicki, A. (2017). Understanding and predicting web content credibility using the content credibility corpus. *Information Processing & Management, 53*(5), 1043–1061. https://doi.org/10.1016/j.ipm.2017.04.003.

Karimi, H., Roy, P., Saba-Sadiya, S., & Tang, J. (2018). Multi-source multi-class fake news detection. *Proceedings of the 27th international conference on computational linguistics* (pp. 1546–1557). Association for Computational Linguistics.

Ke, W., Song, Y., & Kenny Q, Z. (2015). False rumors detection on sina weibo by propagation structures. *Ieee 31st international conference on data engineering* (pp. 651–662). IEEE. https://doi.org/10.1109/ICDE.2015.7113322.

Khattar, D., Goud, J. S., Gupta, M., & Varma, V. (2019). Mvae: multimodal variational autoencoder for fake news detection. *Proceedings of the 28th international conference on world wide web* (pp. 2915–2921). ACM. https://doi.org/10.1145/3308558.3313552.

Kim, Y. (2014). Convolutional neural networks for sentence classification. *Proceedings of the 2014 conference on empirical methods in natural language processing* (pp. 1746–1751). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1181.

Kochkina, E., Liakata, M., & Zubiaga, A. (2018). All-in-one: multi-task learning for rumour verification. *Proceedings of the 27th international conference on computational linguistics* (pp. 3402–3413). Association for Computational Linguistics.

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., … Zittrain, J. L. (2018). The science of fake news. *Science, 359*(6380), 1094–1096. https://doi.org/10.1126/science.aao2998.

Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B. J., Wong, K.-F., & Cha, M. (2016). Detecting rumors from microblogs with recurrent neural networks. *Proceedings of the twenty-fifth international joint conference on artificial intelligence* (pp. 3818–3824). AAAI Press.

Ma, J., Gao, W., & Wong, K.-F. (2018a). Detect rumor and stance jointly by neural multi-task learning. *Companion proceedings of the the web conference 2018* (pp. 585–593). International World Wide Web Conferences Steering Committee. https://doi.org/10.1145/3184558.3188729.

Ma, J., Gao, W., & Wong, K.-F. (2018b). Rumor detection on twitter with tree-structured recursive neural networks. *Proceedings of the 56th annual meeting of the association for computational linguistics* (pp. 1980–1989). Melbourne, Australia: Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1184.

Ma, J., Gao, W., & Wong, K.-F. (2019). Detect rumors on twitter by promoting information campaigns with generative adversarial learning. *Proceedings of the 28th international conference on world wide web* (pp. 3049–3055). ACM. https://doi.org/10.1145/3308558.3313741.

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research, 9*, 2579–2605.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Proceedings of the 26th international conference on neural information processing systems* (pp. 3111–3119). Curran Associates Inc.. https://doi.org/10.5555/2999792.2999959.

Oshikawa, R., Qian, J., & Wang, W. Y. (2018). A survey on natural language processing for fake news detection.

Qi, P., Cao, J., Yang, T., Guo, J., & Li, J. (2019). Exploiting multi-domain visual information for fake news detection. *2019 ieee international conference on data mining*. IEEE.

Qian, F., Gong, C., Sharma, K., & Liu, Y. (2018). Neural user response generator: Fake news detection with collective user intelligence. *Proceedings of the twenty-seventh international joint conference on artificial intelligence* (pp. 3834–3840). International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2018/533.

Rubin, V. L., Chen, Y., & Conroy, N. J. (2015). Deception detection for news: three types of fakes. *Proceedings of the 78th asis&t annual meeting: Information science with impact: Research in and for the community* (pp. 83:1–83:4). American Society for Information Science.

Ruchansky, N., Seo, S., & Liu, Y. (2017). Csi: A hybrid deep model for fake news detection. *Proceedings of the 2017 acm on conference on information and knowledge management* (pp. 797–806). ACM. https://doi.org/10.1145/3132847.3132877.

Sejeong, K., Meeyoung, C., Kyomin, J., Wei, C., & Yajun, W. (2013). Prominent features of rumor propagation in online social media. *Ieee 13th international conference on data mining* (pp. 1103–1108). IEEE. https://doi.org/10.1109/ICDM.2013.61.

Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M., & Liu, Y. (2019). Combating fake news: a survey on identification and mitigation techniques. *ACM Transactions on Intelligent Systems and Technology, 10*(3), 21:1–21:42. https://doi.org/10.1145/3305260.

Shu, K., Cui, L., Wang, S., Lee, D., & Liu, H. (2019). defend: explainable fake news detection. *Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining* (pp. 395–405). ACM. https://doi.org/10.1145/3292500.3330935.

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: a data mining perspective. *ACM SIGKDD Explorations Newsletter, 19*(1), 2236. https://doi.org/10.1145/3137597.3137600.

Shu, K., Wang, S., & Liu, H. (2018). Understanding user profiles on social media for fake news detection. *Proceedings - ieee 1st conference on multimedia information processing and retrieval* (pp. 430–435). Institute of Electrical and Electronics Engineers Inc.. https://doi.org/10.1109/MIPR.2018.00092.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *International conference on learning representations*.

Singhal, S., Kabra, A., Sharma, M., Shah, R. R., Chakraborty, T., & Kumaraguru, P. (2020). Spotfake+: A multimodal framework for fake news detection via transfer learning (student abstract). *Aaai* (pp. 13915–13916).

Singhal, S., Shah, R. R., Chakraborty, T., Kumaraguru, P., & Satoh, S. (2019). Spotfake: A multi-modal framework for fake news detection. *2019 ieee fifth international conference on multimedia big data (bigmm)* (pp. 39–47).

Slaney, M., & Casey, M. (2008). Locality-sensitive hashing for finding nearest neighbors. *IEEE Signal processing magazine, 25*(2), 128–131.

Su, T., Macdonald, C., & Ounis, I. (2019). Ensembles of recurrent networks for classifying the relationship of fake news titles, . In *SIGIR19Proceedings of the 42nd international acm sigir conference on research and development in information retrieval* (p. 893896). ACM. https://doi.org/10.1145/3331184.3331305.

Tsai, Y.-H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L.-P., & Salakhutdinov, R. (2019). Multimodal transformer for unaligned multimodal language sequences. *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 6558–6569). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1656.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems* (pp. 5998–6008). Curran Associates, Inc..

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: a neural image caption generator. *The ieee conference on computer vision and pattern recognition* (pp. 3156–3164). https://doi.org/10.1109/CVPR.2015.7298935.

Vo, N., & Lee, K. (2018). The rise of guardians: Fact-checking url recommendation to combat fake news. *The 41st international acm sigir conference on research & development in information retrieval* (pp. 275–284). ACM. https://doi.org/10.1145/3209978.3210037.

Vo, N., & Lee, K. (2019). Learning from fact-checkers: analysis and generation of fact-checking language. *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval* (pp. 335–344). ACM. https://doi.org/10.1145/3331184.3331248.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science, 359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559.

Wang, H., Zhang, F., Xie, X., & Guo, M. (2018a). Dkn: Deep knowledge-aware network for news recommendation. *Proceedings of the 2018 world wide web conference* (pp. 1835–1844). International World Wide Web Conferences Steering Committee. https://doi.org/10.1145/3178876.3186175.

Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., … Gao, J. (2018b). Eann: event adversarial neural networks for multi-modal fake news detection. *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining* (pp. 849–857). ACM. https://doi.org/10.1145/3219819.3219903.

Yang, F., Liu, Y., Yu, X., & Yang, M. (2012). Automatic detection of rumor on sina weibo. *Proceedings of the acm sigkdd workshop on mining data semantics* (pp. 13:1–13:7). ACM. https://doi.org/10.1145/2350190.2350203.

Yang, Y., Pierce, T., & Carbonell, J. (1998). A study of retrospective and on-line event detection, . In *SIGIR 98Proceedings of the 21st annual international acm sigir conference on research and development in information retrieval* (pp. 28–36). ACM. https://doi.org/10.1145/290941.290953.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems 32* (pp. 5753–5763). Curran Associates, Inc.

Yu, F., Liu, Q., Wu, S., Wang, L., & Tan, T. (2017). A convolutional approach for misinformation identification. *Proceedings of the 26th international joint conference on artificial intelligence* (pp. 3901–3907). AAAI Press. https://doi.org/10.24963/ijcai.2017/545.

Zhang, H., Fang, Q., Qian, S., & Xu, C. (2019). Multi-modal knowledge-aware event memory network for social media rumor detection. *Proceedings of the 27th acm international conference on multimedia* (pp. 1942–1951). ACM. https://doi.org/10.1145/3343031.3350850.

Zhang, X., & Ghorbani, A. A. (2020). An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management, 57*(2), 102025. https://doi.org/10.1016/j.ipm.2019.03.004.

Zhao, Z., Zhu, H., Xue, Z., Liu, Z., Tian, J., Chua, M. C. H., & Liu, M. (2019). An image-text consistency driven multimodal sentiment analysis approach for social media. *Information Processing & Management, 56*(6), 102097. https://doi.org/10.1016/j.ipm.2019.102097.

Zhou, K., Shu, C., Li, B., & Lau, J. H. (2019a). Early rumour detection. *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies* (pp. 1614–1623). Minneapolis, Minnesota: Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1163.

Zhou, X., Zafarani, R., Shu, K., & Liu, H. (2019b). Fake news: fundamental theories, detection strategies and challenges. *Proceedings of the twelfth acm international conference on web search and data mining* (pp. 836–837). ACM. https://doi.org/10.1145/3289600.3291382.

ARTICLES FOR FACULTY MEMBERS

MULTIMODAL FAKE NEWS DETECTION

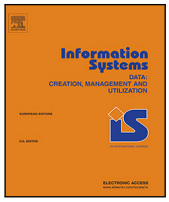
Title/Author	**An inter-modal attention-based deep learning framework using unified modality for multimodal fake news, hate speech and offensive language detection / Ayetiran, E. F., & Özgöbek, Ö.**
Source	***Information Systems*** **Volume 123 (2024) 102378 Pages 1-11** **https://doi.org/10.1016/j.is.2024.102378** **(Database: ScienceDirect)**

Information Systems 123 (2024) 102378

Contents lists available at ScienceDirect

Information Systems

journal homepage: www.elsevier.com/locate/is



An inter-modal attention-based deep learning framework using unified modality for multimodal fake news, hate speech and offensive language detection

Eniafe Festus Ayetiran [a,b,*], Özlem Özgöbek [a]

[a] Department of Computer Science, Norwegian University of Science and Technology, Trondheim, Norway
[b] Department of Computer Science, Achievers University, Owo

ARTICLE INFO

Keywords:
Inter-modal attention
Unified modality
Multimodal fusion
BiLSTM-CNN
Multimodal content understanding
Fake news
Hate speech
Offensive language

ABSTRACT

Fake news, hate speech and offensive language are related evil triplets currently affecting modern societies. Text modality for the computational detection of these phenomena has been widely used. In recent times, multimodal studies in this direction are attracting a lot of interests because of the potentials offered by other modalities in contributing to the detection of these menaces. However, a major problem in multimodal content understanding is how to effectively model the complementarity of the different modalities due to their diverse characteristics and features. From a multimodal point of view, the three tasks have been studied mainly using image and text modalities. Improving the effectiveness of the diverse multimodal approaches is still an open research topic. In addition to the traditional text and image modalities, we consider image–texts which are rarely used in previous studies but which contain useful information for enhancing the effectiveness of a prediction model. In order to ease multimodal content understanding and enhance prediction, we leverage recent advances in computer vision and deep learning for these tasks. First, we unify the modalities by creating a text representation of the images and image–texts, in addition to the main text. Secondly, we propose a multi-layer deep neural network with inter-modal attention mechanism to model the complementarity among these modalities. We conduct extensive experiments involving three standard datasets covering the three tasks. Experimental results show that detection of fake news, hate speech and offensive language can benefit from this approach. Furthermore, we conduct robust ablation experiments to show the effectiveness of our approach. Our model predominantly outperforms prior works across the datasets.

1. Introduction

The exponential growth of world wide web (WWW) and social media have fueled the menaces of fake news, hate speech and offensive language in recent times. Fake news are a form of disinformation, fabricated to deceive readers to believe they are real by imitating mainstream news. The main goal of any form of disinformation is to intentionally mislead people through the creation and spread of false information. In some cases, fake news take genuine part of mainstream news and modify them by injecting some form of falsehood into them. Such modification and injection affect not only the text modality but also images. The main difference between disinformation and misinformation is that in disinformation, the piece of information is deliberately created to mislead people while misinformation is an unintentional propagation of false information. Hate speech is more complex to define because what constitutes hate is relative and differs across jurisdictions. However, in order to provide a general perspective, the United Nations Strategy and Plan of Action on Hate Speech[1] defines hate speech as *"any kind of communication in speech, writing or behavior, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, color, descent, gender or other identity factor"*. Hate speech and offensive language share a common characteristic as a form of abuse or attack but are different in a sense, though some works in literature often use the two terms interchangeably to mean the same. Offensive language is simply a statement which upsets another person. Hence, hate speech is considered more severe as it may lead to extreme action(s) and constitute severe harm to the target. Furthermore, the common effect of the three menaces of fake news, hate speech and offensive language is emotional harm to their targets. Giachanou and Rosso [1] identifies this nexus among the phenomena and refer to

* Corresponding author.
E-mail addresses: eniafe.ayetiran@ntnu.no (E.F. Ayetiran), ozlem.ozgobek@ntnu.no (Ö. Özgöbek).

[1] https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech (accessed 14th March, 2023).

https://doi.org/10.1016/j.is.2024.102378
Received 9 June 2023; Received in revised form 29 November 2023; Accepted 14 March 2024
Available online 16 March 2024

them as harmful information. A greater percentage of fake news, hate speech and offensive language happen online particularly on social media. These three tasks have been studied individually. However, there are connections among them as one may lead to the other. For instance, fake news may lead to hate speech or offensive language. On the other hand, hatred or animosity towards someone or a community of people may make them a target of fake news. In most cases, this is especially applicable to public figures or celebrities. In fact, some fake news doubles as hate speech and/or offensive language. The European Foundation for South Asian Studies identifies and discusses some roles played by fake news in promoting hate speech and extremism online [2]. Both the actions and reactions pertaining to fake news, hate speech and offensive language are strongly tied to intention and harm. Till now, the detection of the three phenomena in digital contents have mostly been studied individually with different approaches. Early studies focused mainly on text modality but as the digital media and repositories continue to grow in multimedia contents (i.e. text, videos, images and audio), the interest to research the prospects of these modalities has risen over the years. Besides text, image is the most commonly used because it is the next most readily available in online digital contents. With the aim to explore the prospects of other modalities, studies based on multimodality are gaining ground [3–8] but still under-explored [9]. Multimodal content understanding aims at recognizing and localizing objects, determining the attributes of entities, characterizing the relationships among entities and describing the common semantic features among different modalities [10]. Chen et al. [10] specifically identified two major gaps in deep multimodal content understanding. First is the "heterogeneity gap", which arises as a result of differences and uniqueness of features of images and texts. This characteristics are directly related the second challenge of "semantic gap", caused by the peculiarities of individual modalities thereby leading to different abstract representations. Extracted features from individual modalities are not directly comparable and are inconsistently distributed. Techniques to mitigate these problems rely mainly on embeddings of individual unimodal features into a common latent space with the help of mapping functions in order to make them comparable and consistent. At present, these techniques still do not fully resolve these problems and advances are still being explored. In view of these current challenges and inspired by advances in computer vision precisely image captioning and Optical Character Recognition (OCR), we develop a unified modality-based deep learning framework which presents the advantage of direct comparison and consistency across modalities. Image captioning and OCR enable the unification of the modalities for comparison and consistency of the modalities. Our deep learning framework comprises a Bidirectional Long Short-Term Memory (BiLSTM) layer [11], a Convolutional Neural Network (CNN) layer with an inter-modal attention mechanism among other layer/modules. The model can be trained on datasets involving any language with appropriate preprocessing. The important contributions of this paper are as follow:

- We develop a unified modality for multimedia contents to resolve the barriers in multimodal content understanding
- We propose an inter-modal attention mechanism for complementarity among modalities in order to improve multimodal content understanding
- We develop a deep learning framework based on the inter-modal attention mechanism for fake news, hate speech and offensive language detection
- Our deep learning framework achieves state-of-the-art performance on three benchmark datasets covering the three tasks.

The rest of the paper is structured as follows: Section 2 reviews the relevant related works covering the three tasks. Section 3 describes the deep learning framework while Section 4 discusses the experiments and model implementation. In Section 5, we discuss the model evaluation and results. Section 6 concludes the paper.

2. Related works

Unimodal content understanding has been widely studied for a wide range of tasks. On the other hand, studies on multimodal content understanding is currently limited [9,12], with some inherent challenges [13,14]. Some of the challenges as discussed in Section 1 include heterogeneity and semantic gaps. In the following subsections, we discuss prior multimodal works on fake news, hate speech and offensive language detection.

2.1. Multimodal fake news detection

One of the earliest works on multimodal fake news detection is the work of Wang et al. [15]. They proposed an end-to-end framework based on neural networks named EANN comprising three primary components; a multimodal feature extractor, a fake news detector and an event discriminator. The multimodal feature extractor comprises two sub-components; a text features extractor and a visual features extractor. Each of the components cooperate for the task of multimodal fake news detection. Experiments on two benchmark datasets show performance improvements over baselines. Khattar et al. [3] proposed MVAE, a similar work to EANN and which as the name suggests used variational autoencoder for classifying multimodal news contents as real or fake. The components of MVAE are an encoder, a decoder and a fake news detector. Both the encoder and the decoder each comprises a text and visual extractor. While the encoder basically encodes the multimodal inputs and outputs a shared representation of learnt features as latent vectors, the decoder reconstructs the latent vectors. The encoded representations serve as inputs to the decoder and the fake news detector component. The fake news detector classifies news content based on the encoded representations, sum of reconstructed and Kullback–Leibler divergence losses. The evaluation of MVAE was carried out on the same datasets as EANN with significant performance improvements over EANN and other baselines. With the goal of achieving a pure classifier without any subtask, SpotFake [16] employs pretrained transformers to incorporate contextualized information and image recognition into multimodal fake news classification. Precisely, they employ Bidirectional Encoder Representations from Transformers (BERT) [17] and VGG-19 [18] to extract textual and visual features which were fused for the classification. The evaluation of SpotFake on the same datasets as EANN and MVAE shows that it outperforms both on only one of the datasets. In order to help identify fake news based on irrelevant images in news content, Zhou et al. [5] introduced SAFE. For textual and visual features extraction, SAFE extended a method based on convolutional neural network (CNN). Images are first processed as text using image captioning. The main crux of SAFE is the computation of similarity between text and image features which is used to optimize model learning parameters. Giachanou et al. [19] combines textual, visual and semantic information for fake news detection using neural network classifier. Textual features include embeddings of posts and sentiments while visual features comprise image tags and local binary patterns (LBP). The model was evaluated on three datasets. In a follow-up work, Giachanou et al. [20] extended the visual features to include multi-image information and in contrast to their earlier work uses BERT [17] and VGG-16 [18] for the extraction of text and image features respectively. The main underlying idea in both works is the computation of semantic similarity between textual and visual features. Multimodal Consistency Neural Network (MCNN) [21] is a network-based approach which consists of five subnetworks namely: a text feature extraction module, a visual semantic feature extraction module, a visual tampering feature extraction module, a similarity measurement module and a multimodal fusion module. MCNN experiments on four datasets show improvements over baselines. In another work, Multimodal Fusion with Co-attention Networks (MCAN) [22] was proposed. MCAN includes a co-attention block, a co-attention layer and multiple co-attention stacking on spatial-domain, frequency-domain

and textual features. Experimental evaluation of MCAN on two domain datasets improves baselines. A cross-modal ambiguity learning model (named CAFE) was proposed by Chen et al. [23]. CAFE comprises three modules namely: a cross-modal alignment module, a cross-modal ambiguity learning module and a cross-modal fusion module. The main goal of CAFE is adaptive aggregation of unimodal features and cross-modal correlations. The evaluation of CAFE was carried out on two benchmark datasets with improvements over baselines. Zhang et al. [24] proposed a model (named SceneFND) with a different approach from prior works by incorporating contextual scene information in addition to textual and visual information. The scene features were obtained from the images by calculating the probabilities of each scene category with different scene recognition methods. They presented results for several variants of the model and SpotFake [16]. Similar to some of the previous works, TRIMOON [25] uses BERT and VGG-19 to extract text and image features respectively followed by a fusion module. The fusion module consists of two co-attention blocks and gate-based fusion component. Experiments on two real-world datasets show improvements over baselines.

2.2. Multimodal hate speech detection

The work of Hosseinmardi et al. [26] is one of the earliest works on multimodal hate speech detection in which they focussed on cyberbullying using both textual and image features. Their model is based logistic regression classifier trained with a forward feature selection method. They experimented on a dataset collected from Instagram for the purpose of validating the model. Automated hate speech detection was explored by Yang et al. [27] with multimodal modalities involving text and image. They experimented with quite a number of multimodal fusion approaches including concatenation and addition with attention mechanism. Evaluation reports on the experiments did not show any tangible gain in fusing the two modalities. As part of the hateful memes challenge competition, Kiela et al. [28] developed a dataset of multimodal memes for the task of identifying whether the memes are hateful or not. They presented a number of models evaluated based on defined benchmarks. What can be referred to as a truly standard benchmark dataset for multimodal hate speech classification was developed by Gomez et al. [6] which they named MMHS150K. It was collected from Twitter and annotated on a large scale. In contrast to other datasets, it also leverages image–texts in addition to the main text and image modalities. They experimented widely with diverse models and similar to Kiela et al. [28], also reported that multimodality did not result in tangible gain when compared with using a single or two modalities. Maity et al. [29] introduces a model for detecting cyberbullying in multimodal memes taking into account sentiment, emotion and sarcasm. This led to the development of a dataset on which the model was evaluated. A recent work Yang et al. [30] explores transfer learning for hate speech detection. The authors opine that there is a high correlation between hate speech and sarcasm and therefore designate them as primary and auxiliary tasks for the purpose of cross-task transfer learning. The model consists mainly of adaptation modules namely: semantic, definition and domain adaptation modules. A joint objective for the modules is optimized for learning the parameters. Experiments show efficacy of the approach across benchmark datasets.

Besides the traditional hate speech, research on misogyny detection using multimodal contents is now generating interests [31,32]. Misogyny is a type of hate targeted at women. Fersini et al. [31] specifically organized a task on this problem using multimedia contents while Rizzi et al. [32] proposed to answer some open questions on the topic which include but not limited to determining which modality contributes most to misogyny detection.

2.3. Multimodal offensive language detection

To the best of our knowledge, the work of [4] is the first work to experiment on a truly multimedia contents for offensive language detection. They developed a dataset (MultiOFF) for this purpose using existing meme data collection and experimented with different known classifiers. In their study, multimodal experiments show very little improvements over unimodal experiments when the same algorithms are used. Curiously, some unimodal experiments with different algorithms outperform multimodal experiments. Lee et al. [33] proposed a method called DisMultiHate to disentangle target entities in multimodal memes for hate detection. Their proposed method consists of three modules; data pre-processing, text representation learning and visual representation learning modules. DisMultiHate uses a regression layer to generate the probability of a multimedia content being hate or not and experimental evaluation of the method on MultiOFF improved performance over compared baselines. Pramanick et al. [34] developed a framework (MOMENTA) for detection of harmful memes and the target entities. It uses Google's Vision API to extract image–texts. The extracted text and images are then encoded with a pre-trained visual-linguistic model and VGG-19 respectively. A key component of MOMENTA is intra-modal and cross-modal attention fusion. It outperforms majority of the baselines. MeBERT is another work [35] that uses external knowledge-base to enhance semantic representation for meme classification. It fuses texts and images based on attention mechanism for the classification task. Experiments on two public datasets show the effectiveness of the method. A recent work on multimodal offensive language is MemeFier [8], a deep learning framework for classifying offensive memes. It incorporates external knowledge into features encoding. A key component of MemeFier is alignment-aware fusion of modalities. Experiments on three datasets reveal MemeFier outperforms baselines on two of the three datasets.

3. Methodology

We define the problem and describe the unified framework for multimodal content classification for fake news, hate speech and offensive language. The general architecture of the unified deep learning framework is presented in Fig. 1. It consists of a modality unification module, an embedding layer, a BiLSTM layer, a CNN layer, an inter-modal attention module, a fusion module and a prediction module (dense layer with sigmoid activation).

3.1. Problem formulation

Let M denote a multimedia data comprising a text T, an image X and an image–text Y, belonging to a binary class C. Given a set of multimedia data M_i^k, for each $m_i \in M_i^k$ comprising text t_i, an image x_i and an image–text y_i, the problem is to determine the class $c_i \in C_i^{n=2}$ to which m_i belongs, where $C_i^{n=2}$ is a set of predefined binary classes. We adapt this formulation to fake news, hate speech and offensive language detection where m_i is either a news article, hateful or non-hateful and offensive or non-offensive contents respectively. Furthermore, c_i represents fake or real, hateful or non-hateful and offensive or non-offensive classes for the three tasks respectively.

3.2. Modality unification module

For each sample m_i in a multimedia content, we obtain the caption x, of the image and the image–text y. We use LAVIS [36], a deep learning library for LAnguage-and-VISion intelligence research and applications to retrieve image captions. LAVIS consists over thirty state-of-the-art language vision models including but not limited to Contrastive Language-Image Pre-training (CLIP) [37] and Bootstrapping Language-Image Pre-training (BLIP) for Unified Vision-Language

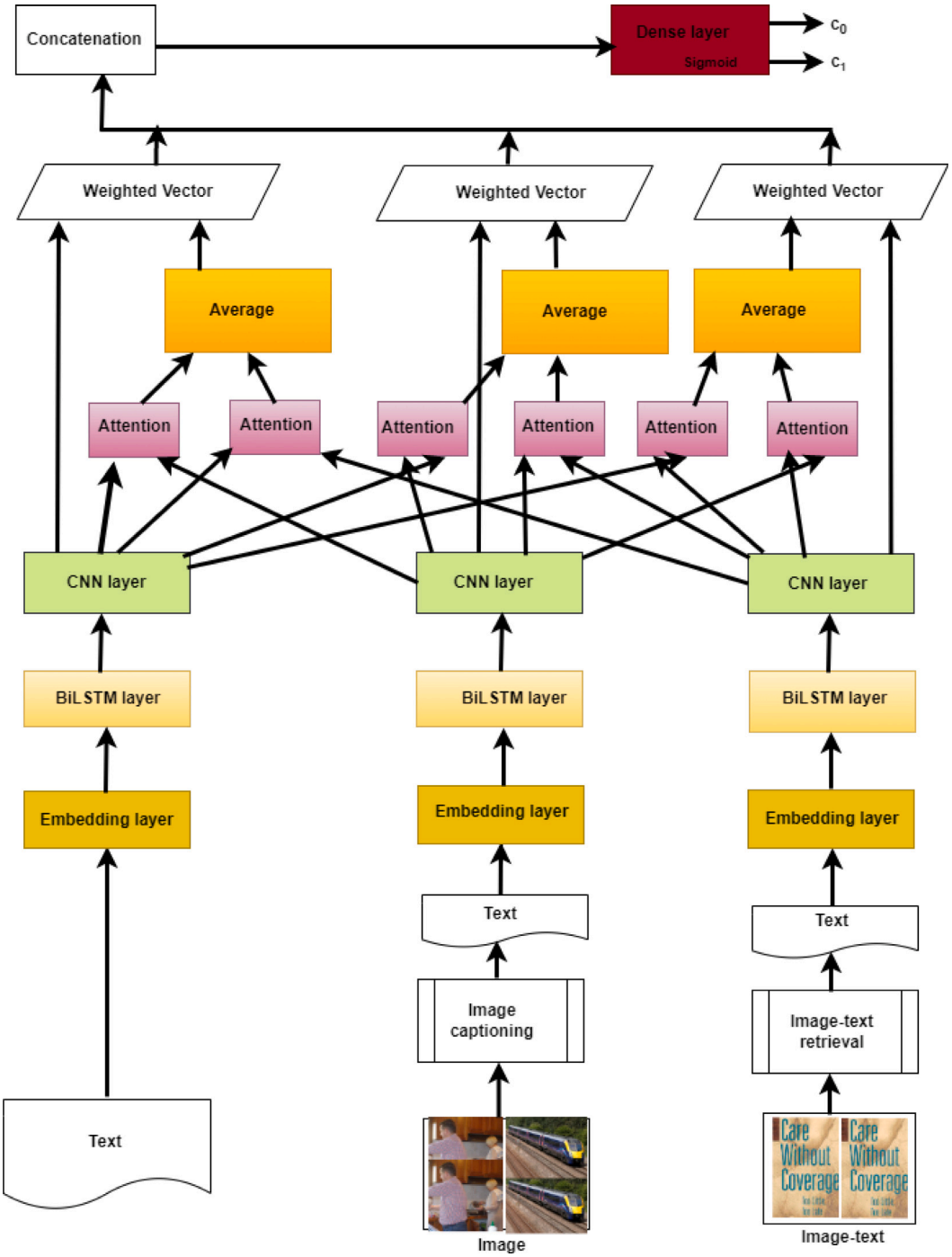


Fig. 1. Architecture of the unified inter-modal attention framework.

Understanding and Generation [38]. Sample images with captions on top are presented in Fig. 2.

We use EasyOCR[2] to retrieve texts inserted within the images. Sample images with inserted texts are shown in Fig. 3. We therefore have the text representations for the original texts, images and image–texts denoted t, x and y respectively. EasyOCR and LAVIS have been chosen for OCR and image captioning respectively because of their state-of-the-art efficacy and easy-to-use Application Programming Interface (API). In few cases where the outputs of OCR are not 100% perfect, the results are still useful as some of the recognized texts are still accurate. The same situation applies to the generated image captions. When the captions do not fully describe the scene, they still capture them to reasonable extents useful for understanding the contents.

3.3. Embedding module

For each word in the unified modality, that is $w \in \{t, x, y\}$, we obtain their embeddings e_w from an embedding matrix $E \in \mathbb{R}^{V \times d}$, where V is the vocabulary size of the embedding matrix and d, the dimension. Specifically, Eqs. (1) to (3) present the word embeddings for a text in each modality as follow:

$$e_t = \{e_{w_1}, e_{w_2}, \ldots, e_{w_n}\} \tag{1}$$

[2] https://github.com/JaidedAI/EasyOCR

Fig. 2. Sample images with captions on top.

Fig. 3. Sample images with inserted texts.

$$e_x = \{e_{w_1}, e_{w_2}, \ldots., e_{w_n}\} \tag{2}$$

$$e_y = \{e_{w_1}, e_{w_2}, \ldots., e_{w_n}\} \tag{3}$$

where n is the number the words in each of t, x and y. The embedded texts are fed to the Bidirectional Long Short-Term Memory (BiLSTM) layer.

3.4. Bidirectional Long Short-Term Memory (BiLSTM) layer

The original Long Short-Term Memory (LSTM) [39] was developed to address the exploding and vanishing gradient problems in feed-forward neural networks. The LSTM architecture comprises three gates; an input gate i_t, a forget gate f_t and an output gate o_t. It also has a memory cell c_t with capability to learn long-term dependencies in sequences and a hidden state h_t. The transition equations of the LSTM are presented in Eqs. (4) to (8):

$$i_t = \sigma\left(w_i \left[h_{t-1}, x_t\right] + b_i\right) \tag{4}$$

$$f_t = \sigma\left(w_f \left[h_{t-1}, x_t\right] + b_f\right) \tag{5}$$

$$o_t = \sigma\left(w_o \left[h_{t-1}, x_t\right] + b_o\right) \tag{6}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh\left(w_c \left[h_{t-1}, x_t\right] + b_c\right) \tag{7}$$

$$h_t = o_t \odot \tanh\left(c_t\right) \tag{8}$$

where w_i, w_f and w_o are the weights of the neurons. b_i, b_f and b_o are the biases to be learned during training. σ denotes a logistic sigmoid function and $\odot$ denotes element-wise multiplication. tanh() is a hyperbolic tangent function. Bidirectional LSTM [11] is a variant of the conventional LSTM which consists of two LSTMs that are run forward and backward simultaneously on the input sequence. The backward LSTM is used to capture the past contextual information while the forward LSTM is used to capture future contextual information. The BiLSTM is used to capture the sequential and contextual information in the input sequences. The outputs are hidden representations of the inputs as presented in Eq. (9):

$$[h_1, \ldots., h_n] = BiLSTM([e_{w_1}, \ldots., e_{w_n}], \theta_{bilstm}) \tag{9}$$

where the input sequences e_w are the embeddings from the embedding layer. θ_{bilstm} is a trainable parameter. Therefore, the final hidden state is obtained using Eq. (10):

$$h_t = \mu\left(\overrightarrow{h_t}, \overleftarrow{h_t}\right) \tag{10}$$

where μ is the average of the hidden states of the forward and the backward LSTMs.

3.5. Convolutional Neural Network (CNN) layer

Convolutional Neural Network (CNN) [40] has become one of the most popular choice in the field of deep learning for image classification and feature extraction. Convolutional Neural Network is a kind of feedforward neural network that is able to extract features from data with convolution structures. In contrast to the traditional feature extraction methods, CNN does not need to extract features manually. Computer vision based on Convolutional Neural Networks has enabled accomplishments that had been considered impossible in the past decades. These areas include face recognition, autonomous vehicles and intelligent medical treatment. They have also recently found applications in sequence modeling problems such as text classification, sentiment analysis, prediction tasks among others. Therefore, this layer is meant to extract important features from the BiLSTM hidden vectors. The CNN component of our architecture uses a one-dimensional convolutional layer with f filters and k kernels. The output hidden state representations from the BiLSTM layer is fed into this layer. Similar to TextCNN [41], for each BiLSTM hidden vector $h_i \in \mathbb{R}^d$ (d is the dimension of the vector), a convolution operation which involves a filter $f \in \mathbb{R}^{jd}$ is applied to a window of j hidden vectors of words to produce a new feature map. A feature map g is extracted from a window of j word hidden vectors as given by Eq. (11):

$$g = tanh\left(f \cdot h_{i:i+j-1} + b\right) \tag{11}$$

where tanh() is a hyperbolic tangent and $b \in \mathbb{R}$ is a bias term. Therefore, for each possible window of words j in a text (t_i, x_i and y_i for the three modalities), the filter is applied to produce a set of feature vector for a text in each modality as presented in Eqs. (12) to (14):

$$g_t = \{g_1, g_2, \ldots., g_{n-j+1}\} \tag{12}$$

$$g_x = \{g_1, g_2, \ldots., g_{n-j+1}\} \tag{13}$$

$$g_y = \{g_1, g_2, \ldots., g_{n-j+1}\} \tag{14}$$

3.6. Inter-modal attention module

Given the set of feature maps of words for each modal texts t_i, x_i and y_i, we propose an inter-modal attention layer to assign weights to each word in the text. This attention mechanism is based on the variant proposed by Luong et al. [42] and applied in [43]. At a time, we take either feature map g_t, g_x or g_y as the source and another text of different modality as the target to produce an alignment vector a. For instance, taking g_x as the source and g_t as the target, the alignment vector $a_t(x)$ is given by Eq. (15):

$$a_t(x) = \frac{exp\left(score\left(g_t^{\mathbf{T}}, g_x\right)\right)}{\sum_{x'} exp\left(score\left(g_t^{\mathbf{T}}, g_{x'}\right)\right)} \tag{15}$$

where $score$ is a function which computes the semantic relationship among words in the source and target; dot product in this case. The resulting alignment vector is passed through a softmax activation to predict the probabilities (attention weights) α of each word. The weights are given by Eq. (16):

$$\alpha_t(x) = softmax\left(a_t(x)\right) \tag{16}$$

When the three modalities are considered, the final attention weights for a modality is obtained by computing the average [44] of the resulting attention weights obtained from using the other two modalities as targets. For instance, the weight $\alpha_t(x, y)$ is the average of $\alpha_t(x)$ and $\alpha_t(y)$ given by Eq. (17):

$$\alpha_t(x, y) = \frac{1}{2} \sum \alpha_t(x), \alpha_t(y) \tag{17}$$

If only two modalities are involved or considered, the weighted representation of a modality is computed by simply using one of the two modalities as source and the other as target and vice versa.

3.7. Fusion module

The final attention-weighted vector representation of a text t is presented in Eq. (18):

$$v_t = \alpha_t g_t \tag{18}$$

The same weighted representation is applicable to image and image–text, denoted v_x and v_y respectively. To obtain a multimodal representation of a multimedia content, the weighted multimodal representation v_m is derived through concatenation of individual weighted representations v_t, v_x and v_y as shown in Eq. (19):

$$v_m = v_t \oplus v_x \oplus v_y \tag{19}$$

where v_t, v_x and v_y are the weighted vectors for original text, caption and image–text respectively.

3.8. Prediction module

The resulting multimodal fused representation from the fusion module is fed into an output layer to predict the probability P of a data sample of a multimedia content belonging to a particular class as shown in Eq. (20):

$$p = \sigma(v_m) \tag{20}$$

where σ is a sigmoid activation function. The objective loss function which the model seek to minimize is a cross-entropy function given by Eq. (21), with specific application to our binary classification problem as given in Eq. (22):

$$\mathcal{L} = -\sum_{i=1}^{c} y_i \log p_i \tag{21}$$

$$\mathcal{L}_b = -(y_i \log(p_i) + (1 - y_i) \log(1 - p_i)) \tag{22}$$

where y is the true class label and p is the predicted probability of a data sample belonging to a particular category between the binary categories.

4. Experiments and model implementation

In this section, we discuss the experiments and model implementation details.

4.1. Datasets description

We describe the datasets used for the experiments as per the tasks. We however present a summary of the statistics of the datasets for fake news, hate speech and offensive language in Tables 1 to 3 respectively.

Table 1
Statistics of PolitiFact dataset.

	Real	Fake	Total
News contents	624	432	1056
- with text	528	420	948
- with image	447	336	783

Table 2
Statistics of MHS150K dataset.

	Train	Validation	Test
Hate	29,447	2,500	5,001
NotHate	105,346	2,500	4,999
Total	134,823	5,000	10,000

4.1.1. Fake news dataset

- **PolitiFact**: PolitiFact dataset is part of FakeNewsNet [45], a repository of news contents which fact-checks political reports and issues. It has been collected from the website[3] of the organization. It consists of three contexts; news content, social context and spatio-temporal information. Like most prior works, we used the news content. Annotations were done by human annotators as part of the dataset development. The news content comprises mainly the news headline and body. PolitiFact consists of news articles that were published from May, 2002 to July, 2018. It comprises 1,056 news articles with 624 real news and 432 fake news. Other statistics on textual and visual contents are shown in Table 1. In our experiment, we have treated the headline and body as original text modality but applied separately when dealing with other modalities in the model. For instance, in computing attention weights, both are weighted separately using captions and image–texts.

4.1.2. Hate speech dataset

- **MMHS150K**: MMHS150K [6] is a large-scale collection of tweets for hate speech classification task. The raw MMHS150K consists of 150,000 samples, on which annotations were done by human annotators based on majority voting. The annotated data consists of 112,845 *NotHate* samples and 36,978 *Hate* samples. The annotations fall into six classes including *"No attacks to any community", "racist", "sexist", "homophobic", "religion based attacks" and "attacks to other communities"*. The other five labels apart from "No attacks to any community" are hate categories. The dataset was further split into a test set consisting of 10,000 samples, a validation set consisting of 5,000 samples while the remaining were set aside as training set. Each data sample has a text and associated image and majority of the images have texts inserted within them. Following the authors of MHS150K, in our experiment, the dataset was treated as binary by taking all hateful categories as *"Hate"* label and *"No attacks to any community"* as *"NotHate"* label. Full statistical details about the dataset are presented in Table 2.

4.1.3. Offensive language dataset

- **MultiOFF**: MultiOFF [4] was developed from a collection of memes from social media such as Facebook, Twitter etc. and annotated for offensiveness or otherwise. It is an extension of an existing dataset about 2016 U.S. Presidential Election. In all, MultiOFF consists of 743 samples split into training, validation and test sets. It consists only of the memes (images) and texts

[3] https://www.politifact.com/

Table 3
Statistics of MultiOFF dataset.

	Train	Validation	Test
Offensive	187	59	59
Non-offensive	258	90	90
Total	445	149	149

(full details in Table 3). The text modality constitutes the image–texts already extracted by the authors. Therefore, the dataset does not have separate text and image–text modalities. Experiments on MultiOFF are carried out using the two modalities.

4.2. Data pre-processing

Traditional pre-processing on text data such data cleaning, removing noise, lower casing among others were performed on the data according to the peculiarity of the data. For instance, the hate speech dataset was collected from Twitter which contains slangs and some informal writing styles. All numeric tokens are represented as "<number>", all uniform resource locators (URLs) as "<url>", emojis' interpretation among other pre-processing activities.

4.3. Model implementation

In the following subsections, we present a brief description of task-specific details of the experiments and present a summary of the hyperparameters for all models in Table 4. In all experiments, a 300-dimension GloVe embeddings [46] (trained on 840 billions word tokens) was used to initialize the Embedding layer. Furthermore, for datasets which suffer from class imbalance, we adopt sample weights from each class in computing the loss.

4.3.1. Fake news detection experiments

We split the fake news datasets in the ratio 7:1:2 for training, validation and test respectively. In training the model, we use a batch size of 32 using ADAM [47] as the optimization algorithm with a learning rate of $1e{-}3$. The model was regularized with a Dropout [48] probability of 0.2, applied after the concatenation layer. The number of LSTM neurons is 300 while CNN filters and kernel windows are 300 and 3 respectively.

4.3.2. Hate speech detection experiments

The MMHS150K dataset used for experiment is already split into training, validation and test sets as shown in Table 2. Batch normalization [49] was applied to standardize inputs to the BiLSTM layer. In training the model, we use a batch size of 128 using NADAM [50] as the optimization algorithm with a learning rate of $1e{-}3$. A Dropout [48] probability of 0.2 was applied after the concatenation layer to regularize the model. The number of LSTM neurons and CNN filters are 100 while the kernel window size is 4.

4.3.3. Offensive language detection experiments

Like MMHS150K, MultiOFF dataset is also already split into training, validation and test sets as shown in Table 3. A batch size of 32 is adopted in training the model while Adam [47] is used as the optimization algorithm with a learning rate of $1e{-}3$. A Dropout [48] probability of 0.2, applied after the concatenation layer is used to regularize the model. The number of LSTM neurons is 300 while CNN filters and kernel windows are 300 and 3 respectively.

5. Model evaluation and results discussion

In the following subsections, we evaluate each of the models, report their performances and that of ablation studies and then compare with state-of-the-art models.

Table 4
Hyperparameters for the task-specific models.

Task	Fake news	Hate speech	Offens lang
LSTM neurons	300	100	300
CNN filters	300	100	300
Batch size	32	128	32
Embedding size	300	300	300
Optimizer	ADAM	NADAM	ADAM
Learning rate	$1e{-}3$	$1e{-}3$	$1e{-}3$
Dropout	0.2	0.2	0.2
Batch norm	NA	Yes	NA

Keys: NA — Not applied, Offens lang — Offensive language.

Table 5
Performance of inter-modal attention models with different combination of modalities on PolitiFact dataset. Best results in bold.

Modality	Accuracy	Precision	Recall	F1
TT & IM	0.893	0.898	0.894	0.896
TT & IT	0.893	0.893	0.903	0.898
IM & IT	0.570	0.568	0.546	0.557
TT & IM & IT	**0.940**	**0.939**	**0.940**	**0.939**

Keys: TT — text, IM — Images and IT — Image–text.

Table 6
Performance of BiLSTM-CNN models with different combination of modalities on PolitiFact dataset. Best results in bold.

Modality	Accuracy	Precision	Recall	F1
TT	0.893	0.898	0.895	0.897
IM	0.631	0.744	0.658	0.698
IT	0.557	0.776	0.515	0.619
TT & IM	0.893	0.892	0.895	0.894
TT & IT	0.866	0.867	0.876	0.871
IM & IT	0.651	0.694	0.626	0.658
TT & IM & IT	**0.879**	**0.882**	**0.884**	**0.883**

Keys: TT — Text, IM — Images and IT — Image–text.

5.1. Evaluation results on fake news detection task

Table 5 shows the results of the inter-modal attention model on the PolitiFact dataset with different modalities in terms of Accuracy, Precision, Recall and F1. Table 6 shows the results of the ablation experiments which use BiLSTM-CNN directly without the inter-modal attention module for all combination of modalities.

The result in Table 5 shows that the inter-modal attention model produces the best result when the three modalities are used. Text used with image and text used with image–text have comparable performance. Image used with image–text does not seem to be helpful in detecting fake news as the combination performs way below other combinations. The result of the ablation studies which uses direct concatenation of BiLSTM-CNN features without attention layer is presented in Table 6. Analysis of the ablation results also reveals that combination of the three modalities best detects fake news. The performances of the different combination of modalities follows the same pattern as the main inter-modal attention model. However, the impact of the inter-modal attention mechanism is noteworthy. It enhances performance by 6.1%, 5.7%, 5.6% and 5.6% for Accuracy, Precision, Recall and F1 respectively.

Fig. 4 shows the Receiver Operating Characteristic (ROC) curve of our inter-attention model. The ROC curve tilts towards the True Positive Rate and farther away from the random curve which confirms the quality of the model.

5.1.1. Performance comparison with baselines on fake news detection

Performance comparison of the inter-modal attention model with state-of-the-art models is presented in Table 7. Brief descriptions of the baselines are as follow:

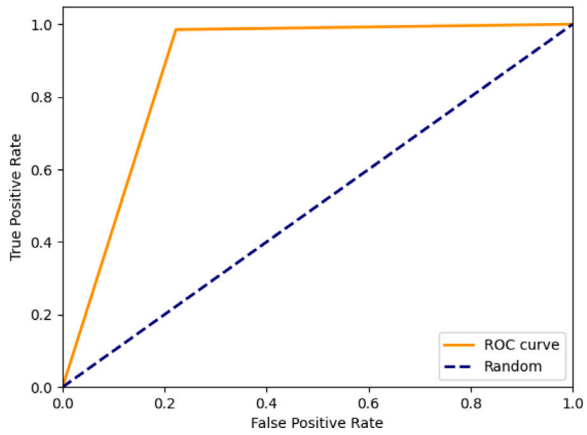


Fig. 4. The Receiver Operating Characteristic (ROC) curve for fake news detection task.

Table 7
Performance comparison of the inter-modal attention model with state-of-the-art models on PolitiFact dataset. Best results in bold.

Model	Accuracy	Precision	Recall	F1
LBP-sim [19]	–	–	–	0.925
SAFE [5]	0.874	0.889	0.903	0.896
MCNN [21]	0.884	**0.973**	0.867	0.917
Our model	**0.940**	0.939	**0.940**	**0.939**

Table 8
Performance of the inter-modal attention models with different combination of modalities on MMHS150K dataset. Best results in bold.

Modality	Acc	Prec	Rec	F1	AUC
TT & IM	0.684	0.656	0.763	0.705	0.684
TT & IT	0.680	0.653	0.764	0.704	0.680
IM & IT	0.514	0.516	**0.934**	0.665	0.514
TT & IM & IT	**0.687**	**0.659**	0.761	**0.706**	**0.687**

Keys: Acc — Accuracy, Prec — Precision, Rec — Recall, TT — text, IM — Images and IT — Image–text.

- **Text-tags-LBP-similarity**: Text-tags-LBP-similarity [19] is a model based on neural network classifier which computes the similarity of text tags and LBP features.
- **SAFE**: SAFE [5] uses an extension of a method based on Convolutional Neural Network (CNN) for textual and visual features extraction. The main component of SAFE is the computation of similarity between text and image features which was used to optimize model learning parameters.
- **MCNN**: MCNN [21] is a network-based approach which consists of five sub-networks namely: a text feature extraction module, a visual semantic feature extraction module, a visual tampering feature extraction module, a similarity measurement module and a multimodal fusion module

Comparison of the performance of our model with baselines shows that our inter-modal attention model outperforms the other baselines across all the metrics except precision where MCNN has a better performance.

5.2. Evaluation results on hate speech detection task

Table 8 shows the results of the inter-modal attention model on the MMHS150K dataset with different modalities in terms of Accuracy, Precision, Recall, F1 and Area Under Curve (AUC). Table 9 shows the results of ablation experiments which use the BiLSTM-CNN directly without the inter-modal attention module for all combination of modalities.

Table 9
Performance of BiLSTM-CNN model with different combination of modalities on MMHS150K dataset. Best results in bold.

Modality	Acc	Prec	Rec	F1	AUC
TT	**0.678**	**0.653**	0.762	**0.703**	**0.678**
IM	0.524	0.516	0.892	0.654	0.524
IT	0.506	0.508	**0.965**	0.666	0.506
TT & IM	0.677	**0.653**	0.762	**0.703**	0.677
TT & IT	0.677	0.652	0.763	**0.703**	0.677
IM & IT	0.524	0.520	0.883	0.654	0.524
TT & IM & IT	0.675	0.652.	0.761	0.702	0.675

Keys: Acc — Accuracy, Prec — Precision, Rec — Recall, TT — Text, IM — Images and IT — Image–text.

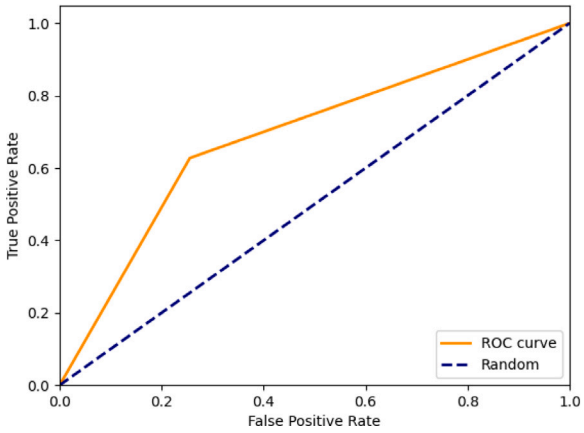


Fig. 5. The Receiver Operating Characteristic (ROC) curve for hate speech detection task.

Combination of the three modalities results in better performance across all the metrics except Recall, where utilization of image and image–text results in better performance with a score of 93.4%. Combination of text and image compared with text and image–text produces comparable performance. Performance of image and image–text results in considerably general lower performance when compared to other combinations except in Recall where it produces a significantly higher performance.

In contrast to the results produced by the inter-modal attention model, combination of text with image or image–text performs best using the BiLSTM-CNN model. Still on the BiLSTM-CNN models, the combination of text and either of image or image–text produce better performance than using the three modalities. More surprising is the fact that text only outperforms any other combination of modalities except in Recall where image–text only has the best performance.

We also qualitatively evaluate the inter-attention model on hate speech detection using ROC curve as presented in Fig. 5. The position of the ROC curve as shown in the figure confirms the quantitative performance.

5.2.1. Performance comparison with baselines on hate speech detection

Table 10 depicts the performance of the inter-modal attention model with state-of-the-art models. The descriptions of the baseline models are as follow:

- **FCM**: Features Concatenation Model (FCM) [6] is a concatenation of features from each of the modalities in which a CNN-based pretrained model (Inception v3) was used for image features representation with Average Pooling while LSTM was used for the representation of text and image–text features.
- **SCM**: Spatial Concatenation Model (SCM) [6] is the same as FCM but with a change in the feature vectors of Inception v3.

Table 10
Performance comparison of the inter-modal attention model with state-of-the-art models on MMHS150K dataset. Best results in bold.

Model	Acc	Precision	Recall	F1	AUC
FCM [6]	0.684	–	–	0.704	**0.734**
SCM [6]	0.685	–	–	0.702	0.732
TKM [6]	0.682	–	–	0.701	0.731
Our model	**0.687**	**0.659**	**0.761**	**0.706**	0.687

Keys: Acc — Accuracy.

Table 11
Performance of inter-modal attention and BiLSTM-CNN models with individual and combination of modalities on MultiOFF dataset. Best results in bold.

Modality	Acc	Prec	Rec	F1
BiLSTM-CNN(TT)	0.658	0.641	0.598	0.619
BiLSTM-CNN(IM)	0.537	0.614	0.635	0.624
BiLSTM-CNN(TT & IM)	0.664	0.655	0.643	0.649
Inter-att(TT & IM)	**0.718**	**0.703**	**0.700**	**0.702**

Keys: Acc — Accuracy, Prec — Precision, Rec — Recall, TT — text, IM — Images and IT — Image–text. Key: Inter-att — Inter-modal attention.

- **TKM**: Textual Kernels Model (TKM) [6] aims to boost interactions among modalities by learning dependent text kernels for texts and image–texts.

Comparison of the performance of our inter-modal attention model with baseline models as presented in Table 10 shows that it outperforms the baselines on Accuracy and F1 with a score of 68.7% and 70.6% respectively. FCM is the best on AUC metric. Our model achieves 65.9% and 76.1% in Precision and Recall respectively. The baselines do not present evaluation results for Precision and Recall.

5.3. Evaluation results on offensive language detection task

Since the MultiOFF dataset consists of two modalities, we present the results of the inter-modal attention model with those of BiLSTM-CNN models in Table 11.

The results in Table 11 confirms that combined usage of text and image is beneficial for offensive language detection. On all the evaluation metrics, multimodality improves performance. For unimodal approaches, usage of text only is better for Accuracy and Precision while image modality is better in Recall and F1. The inter-modal attention model outperforms the BiLSTM-CNN model with a significant margin which therefore confirms its effectiveness.

The ROC curve for our inter-attention model on offensive language is presented in Fig. 6. The positioning of the curve in relation to True Positive Rate and random curve confirm the quality of the quantitative performance of the model.

5.3.1. Performance comparison with baselines on offensive language detection

Performance comparison of our inter-modal attention model with state-of-the-art models is presented in Table 12. Brief descriptions of the baselines are as follow:

- **Stacked LSTM + VGG16**: Stacked LSTM + VGG16 [4] uses stacked LSTM and VGG16 to for text and image representation respectively in a neural classifier
- **BiLSTM + VGG16**: BiLSTM + VGG16 [4] uses BiLSTM to represent the texts combined with image features extracted with VGG16
- **CNNText + VGG16**: CNNText + VGG16 [4] employs VGG16 for image features extraction while traditional CNN was used for textual features representation
- **DisMultiHate**: DisMultiHate [33] extracts target entities for hate detection in multimodal memes

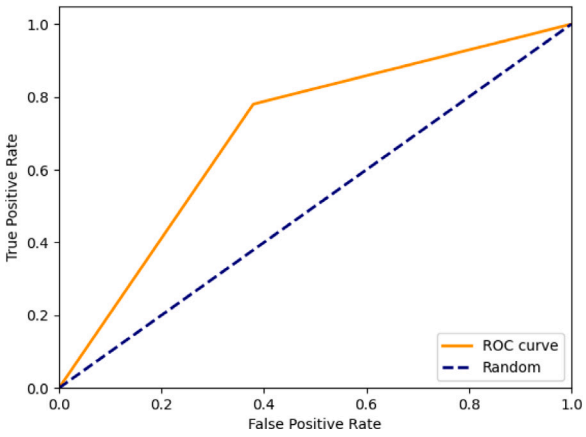


Fig. 6. The Receiver Operating Characteristic (ROC) curve for offensive language detection task.

Table 12
Performance comparison of the inter-modal attention model with state-of-the-art models on MultiOFF dataset. Best results in bold.

Model	Acc	Prec	Rec	F1
Stacked LSTM+VGG16 [4]	–	0.400	0.660	0.500
BiLSTM+VGG16 [4]	–	0.400	0.440	0.410
CNNText+VGG16 [4]	–	0.380	0.670	0.480
DisMultiHate [33]	–	0.645	0.651	0.646
MeBERT [35]	–	0.670	0.671	0.671
MemeFier [8]	0.685	–	–	0.625
Our model	**0.718**	**0.703**	**0.700**	**0.702**

Key: Acc — Accuracy, Prec — Precision, Rec — Recall.

- **MeBERT**: MeBERT [35] fuses texts and images enhanced with external knowledge for semantic representation
- **MemeFier**: MemeFier [8] is a deep learning framework for classifying memes. It also incorporates external knowledge into features encoding. The fusion of modalities is based on alignment among the multimodal features.

Our inter-modal attention model with 70.3%, 70.0% and 70.2% on Precision, Recall and F1 respectively is the best performing model among the compared baselines. It also achieves better Accuracy when compared with MemeFier; the only baseline which evaluated on Accuracy.

5.4. Further note

Result analysis across the tasks reveals that combination of the three modalities mostly lead to the best performance on both inter-modal attention and BiLSTM-CNN models with the exception of the hate speech dataset (MMHS150K) where only text modality leads to the best performance on the BiLSTM-CNN model. In general, the efficacy of the inter-modal attention model is evident across all the tasks.

6. Conclusion

Multimodal content understanding is a challenging and still an open research area due to the *heterogeneity* and *semantic* gaps in the modalities involved. Majority of the prior works in multimodal content understanding for fake news, hate speech and offensive language detection do not take into account how modalities involved complements one another due to issues caused by the aforementioned gaps. In this work, we introduce an additional modality and filled the gaps by leveraging on advances in computer vision to unify the diverse modalities. We further develop a unified deep learning framework based on inter-modal attention mechanism on the unified modalities. Our

framework consists of several modules/layers based mainly on neural networks. We conduct extensive experiments on three public benchmark datasets covering fake news, hate speech and offensive language. Our model significantly enhance prediction and achieves state-of-the-art performance on most of the datasets. We further conduct ablation experiments covering the three tasks to show the effectiveness of our unified inter-modal attention approach.

CRediT authorship contribution statement

Eniafe Festus Ayetiran: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Özlem Özgöbek:** Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data are publicly available online.

Acknowledgment

This work was carried out during the tenure of the first author as an ERCIM “Alain Bensoussan” fellow.

References

[1] A. Giachanou, P. Rosso, The battle against online harmful information: The cases of fake news and hate speech, in: M. d’Aquin, S. Dietze, C. Hauff, E. Curry, P. Cudré-Mauroux (Eds.), CIKM ’20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020, ACM, 2020, pp. 3503–3504, http://dx.doi.org/10.1145/3340531.3412169.

[2] European Foundation for South Asian Studies, The role of fake news in fueling hate speech and extremism online; promoting adequate measures for tackling the phenomenon, 2021, https://www.efsas.org/publications/study-papers/the-role-of-fake-news-in-fueling-hate-speech-and-extremism-online/.

[3] D. Khattar, J.S. Goud, M. Gupta, V. Varma, MVAE: multimodal variational autoencoder for fake news detection, in: L. Liu, R.W. White, A. Mantrach, F. Silvestri, J.J. McAuley, R. Baeza-Yates, L. Zia (Eds.), The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019, ACM, 2019, pp. 2915–2921, http://dx.doi.org/10.1145/3308558.3313552.

[4] S. Suryawanshi, B.R. Chakravarthi, M. Arcan, P. Buitelaar, Multimodal meme dataset (multiOFF) for identifying offensive content in image and text, in: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, European Language Resources Association (ELRA), Marseille, France, 2020, pp. 32–41, URL https://aclanthology.org/2020.trac-1.6.

[5] X. Zhou, J. Wu, R. Zafarani, SAFE: similarity-aware multi-modal fake news detection, in: H.W. Lauw, R.C. Wong, A. Ntoulas, E. Lim, S. Ng, S.J. Pan (Eds.), Advances in Knowledge Discovery and Data Mining - 24th Pacific-Asia Conference, PAKDD 2020, Singapore, May 11-14, 2020, Proceedings, Part II, in: Lecture Notes in Computer Science, 12085, Springer, 2020, pp. 354–367, http://dx.doi.org/10.1007/978-3-030-47436-2_27.

[6] R. Gomez, J. Gibert, L. Gómez, D. Karatzas, Exploring hate speech detection in multimodal publications, in: IEEE Winter Conference on Applications of Computer Vision, WACV 2020, Snowmass Village, CO, USA, March 1-5, 2020, IEEE, 2020, pp. 1459–1467, http://dx.doi.org/10.1109/WACV45572.2020.9093414.

[7] C. Yang, F. Zhu, G. Liu, J. Han, S. Hu, Multimodal hate speech detection via cross-domain knowledge transfer, in: J. ao Magalhães, A.D. Bimbo, S. Satoh, N. Sebe, X. Alameda-Pineda, Q. Jin, V. Oria, L. Toni (Eds.), MM ’22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022, ACM, 2022, pp. 4505–4514, http://dx.doi.org/10.1145/3503161.3548255.

[8] C. Koutlis, M. Schinas, S. Papadopoulos, MemeFier: Dual-stage modality fusion for image meme classification, in: I. Kompatsiaris, J. Luo, N. Sebe, A. Yao, V. Mazaris, S. Papadopoulos, A. Popescu, Z.H. Huang (Eds.), Proceedings of the 2023 ACM International Conference on Multimedia Retrieval, ICMR 2023, Thessaloniki, Greece, June 12-15, 2023, ACM, 2023, pp. 586–591, http://dx.doi.org/10.1145/3591106.3592254.

[9] L. Bozarth, C. Budak, Toward a better performance evaluation framework for fake news classification, in: M.D. Choudhury, R. Chunara, A. Culotta, B.F. Welles (Eds.), Proceedings of the Fourteenth International AAAI Conference on Web and Social Media, ICWSM 2020, Held Virtually, Original Venue: Atlanta, Georgia, USA, June 8-11, 2020, AAAI Press, 2020, pp. 60–71, URL https://ojs.aaai.org/index.php/ICWSM/article/view/7279.

[10] W. Chen, W. Wang, L. Liu, M.S. Lew, New ideas and trends in deep multimodal content understanding: A review, Neurocomputing 426 (2021) 195–215, http://dx.doi.org/10.1016/j.neucom.2020.10.042.

[11] A. Graves, N. Jaitly, A. Mohamed, Hybrid speech recognition with deep bidirectional LSTM, in: 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, Olomouc, Czech Republic, December 8-12, 2013, IEEE, 2013, pp. 273–278, http://dx.doi.org/10.1109/ASRU.2013.6707742.

[12] I. Segura-Bedmar, S. Alonso-Bartolome, Multimodal fake news detection, Inf. 13 (6) (2022) 284, http://dx.doi.org/10.3390/info13060284.

[13] D. Lahat, T. Adali, C. Jutten, Challenges in multimodal data fusion, in: 22nd European Signal Processing Conference, EUSIPCO 2014, Lisbon, Portugal, September 1-5, 2014, IEEE, 2014, pp. 101–105, URL https://ieeexplore.ieee.org/document/6951999/.

[14] F. Alam, S. Cresci, T. Chakraborty, F. Silvestri, D. Dimitrov, G.D.S. Martino, S. Shaar, H. Firooz, P. Nakov, A survey on multimodal disinformation detection, in: Proceedings of the 29th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 6625–6643, URL https://aclanthology.org/2022.coling-1.576.

[15] Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su, J. Gao, EANN: event adversarial neural networks for multi-modal fake news detection, in: Y. Guo, F. Farooq (Eds.), Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018, ACM, 2018, pp. 849–857, http://dx.doi.org/10.1145/3219819.3219903.

[16] S. Singhal, R.R. Shah, T. Chakraborty, P. Kumaraguru, S. Satoh, SpotFake: A multi-modal framework for fake news detection, in: Fifth IEEE International Conference on Multimedia Big Data, BigMM 2019, Singapore, September 11-13, 2019, IEEE, 2019, pp. 39–47, http://dx.doi.org/10.1109/BigMM.2019.00-44.

[17] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 4171–4186, http://dx.doi.org/10.18653/v1/n19-1423.

[18] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: Y. Bengio, Y. LeCun (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015.

[19] A. Giachanou, G. Zhang, P. Rosso, Multimodal fake news detection with textual, visual and semantic information, in: P. Sojka, I. Kopecek, K. Pala, A. Horák (Eds.), Text, Speech, and Dialogue - 23rd International Conference, TSD 2020, Brno, Czech Republic, September 8-11, 2020, Proceedings, in: Lecture Notes in Computer Science, vol. 12284, Springer, 2020, pp. 30–38, http://dx.doi.org/10.1007/978-3-030-58323-1_3.

[20] A. Giachanou, G. Zhang, P. Rosso, Multimodal multi-image fake news detection, in: G.I. Webb, Z. Zhang, V.S. Tseng, G. Williams, M. Vlachos, L. Cao (Eds.), 7th IEEE International Conference on Data Science and Advanced Analytics, DSAA 2020, Sydney, Australia, October 6-9, 2020, IEEE, 2020, pp. 647–654, http://dx.doi.org/10.1109/DSAA49011.2020.00091.

[21] J. Xue, Y. Wang, Y. Tian, Y. Li, L. Shi, L. Wei, Detecting fake news by exploring the consistency of multimodal data, Inf. Process. Manag. 58 (5) (2021) 102610, http://dx.doi.org/10.1016/j.ipm.2021.102610.

[22] Y. Wu, P. Zhan, Y. Zhang, L. Wang, Z. Xu, Multimodal fusion with co-attention networks for fake news detection, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021, in: Findings of ACL, ACL/IJCNLP 2021, Association for Computational Linguistics, 2021, pp. 2560–2569, http://dx.doi.org/10.18653/v1/2021.findings-acl.226.

[23] Y. Chen, D. Li, P. Zhang, J. Sui, Q. Lv, T. Lu, L. Shang, Cross-modal ambiguity learning for multimodal fake news detection, in: F. Laforest, R. Troncy, E. Simperl, D. Agarwal, A. Gionis, I. Herman, L. Médini (Eds.), WWW ’22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022, ACM, 2022, pp. 2897–2905, http://dx.doi.org/10.1145/3485447.3511968.

[24] G. Zhang, A. Giachanou, P. Rosso, Scenefnd: Multimodal fake news detection by modelling scene context information, J. Inf. Sci. (2024) 01655515221087683, http://dx.doi.org/10.1177/01655515221087683, arXiv:10.1177/01655515221087683.

[25] S. Xiong, G. Zhang, V. Batra, L. Xi, L. Shi, L. Liu, TRIMOON: two-round inconsistency-based multi-modal fusion network for fake news detection, Inf. Fusion 93 (2023) 150–158, http://dx.doi.org/10.1016/j.inffus.2022.12.016.

[26] H. Hosseinmardi, R.I. Rafiq, R. Han, Q. Lv, S. Mishra, Prediction of cyberbullying incidents in a media-based social network, in: R. Kumar, J. Caverlee, H. Tong (Eds.), 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2016, San Francisco, CA, USA, August 18-21, 2016, IEEE Computer Society, 2016, pp. 186–192, http://dx.doi.org/10.1109/ASONAM.2016.7752233.

[27] F. Yang, X. Peng, G. Ghosh, R. Shilon, H. Ma, E. Moore, G. Predovic, Exploring deep multimodal fusion of text and photo for hate speech classification, in: Proceedings of the Third Workshop on Abusive Language Online, Association for Computational Linguistics, Florence, Italy, 2019, pp. 11–18, http://dx.doi.org/10.18653/v1/W19-3502, URL https://aclanthology.org/W19-3502.

[28] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, D. Testuggine, The hateful memes challenge: Detecting hate speech in multimodal memes, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, Virtual, 2020.

[29] K. Maity, P. Jha, S. Saha, P. Bhattacharyya, A multitask framework for sentiment, emotion and sarcasm aware cyberbullying detection from multi-modal code-mixed memes, in: E. Amigó, P. Castells, J. Gonzalo, B. Carterette, J.S. Culpepper, G. Kazai (Eds.), SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022, ACM, 2022, pp. 1739–1749, http://dx.doi.org/10.1145/3477495.3531925.

[30] C. Yang, F. Zhu, G. Liu, J. Han, S. Hu, Multimodal hate speech detection via cross-domain knowledge transfer, in: J. ao Magalhães, A.D. Bimbo, S. Satoh, N. Sebe, X. Alameda-Pineda, Q. Jin, V. Oria, L. Toni (Eds.), MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022, ACM, 2022, pp. 4505–4514, http://dx.doi.org/10.1145/3503161.3548255.

[31] E. Fersini, F. Gasparini, G. Rizzi, A. Saibene, B. Chulvi, P. Rosso, A. Lees, J. Sorensen, SemEval-2022 task 5: Multimedia automatic misogyny identification, in: G. Emerson, N. Schluter, G. Stanovsky, R. Kumar, A. Palmer, N. Schneider, S. Singh, S. Ratan (Eds.), Proceedings of the 16th International Workshop on Semantic Evaluation, SemEval@NAACL 2022, Seattle, Washington, United States, July 14-15, 2022, Association for Computational Linguistics, 2022, pp. 533–549, http://dx.doi.org/10.18653/V1/2022.SEMEVAL-1.74.

[32] G. Rizzi, F. Gasparini, A. Saibene, P. Rosso, E. Fersini, Recognizing misogynous memes: Biased models and tricky archetypes, Inf. Process. Manag. 60 (5) (2023) 103474, http://dx.doi.org/10.1016/J.IPM.2023.103474.

[33] R.K. Lee, R. Cao, Z. Fan, J. Jiang, W. Chong, Disentangling hate in online memes, in: MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021, ACM, 2021, pp. 5138–5147, http://dx.doi.org/10.1145/3474085.3475625.

[34] S. Pramanick, S. Sharma, D. Dimitrov, M.S. Akhtar, P. Nakov, T. Chakraborty, MOMENTA: A multimodal framework for detecting harmful memes and their targets, in: M. Moens, X. Huang, L. Specia, S.W. Yih (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021, Association for Computational Linguistics, 2021, pp. 4439–4455, http://dx.doi.org/10.18653/v1/2021.findings-emnlp.379.

[35] Q. Zhong, Q. Wang, J. Liu, Combining knowledge and multi-modal fusion for meme classification, in: MultiMedia Modeling - 28th International Conference, MMM 2022, Phu Quoc, Vietnam, June 6-10, 2022, Proceedings, Part I, in: Lecture Notes in Computer Science, vol. 13141, Springer, 2022, pp. 599–611, http://dx.doi.org/10.1007/978-3-030-98358-1_47.

[36] D. Li, J. Li, H. Le, G. Wang, S. Savarese, S.C.H. Hoi, LAVIS: A library for language-vision intelligence, 2022, arXiv:2209.09019.

[37] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, in: Proceedings of Machine Learning Research, vol. 139, PMLR, 2021, pp. 8748–8763, URL http://proceedings.mlr.press/v139/radford21a.html.

[38] J. Li, D. Li, C. Xiong, S.C.H. Hoi, BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation, in: K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, S. Sabato (Eds.), International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA, in: Proceedings of Machine Learning Research, vol. 162, PMLR, 2022, pp. 12888–12900, URL https://proceedings.mlr.press/v162/li22n.html.

[39] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780, http://dx.doi.org/10.1162/neco.1997.9.8.1735.

[40] K. Fukushima, S. Miyake, Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position, Pattern Recognit. 15 (6) (1982) 455–469, http://dx.doi.org/10.1016/0031-3203(82)90024-3.

[41] Y. Kim, Convolutional neural networks for sentence classification, in: A. Moschitti, B. Pang, W. Daelemans (Eds.), Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, a Meeting of SIGDAT, a Special Interest Group of the ACL, ACL, 2014, pp. 1746–1751, http://dx.doi.org/10.3115/v1/d14-1181.

[42] T. Luong, H. Pham, C.D. Manning, Effective approaches to attention-based neural machine translation, in: L. Màrquez, C. Callison-Burch, J. Su, D. Pighin, Y. Marton (Eds.), Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015, The Association for Computational Linguistics, 2015, pp. 1412–1421, http://dx.doi.org/10.18653/v1/d15-1166.

[43] E.F. Ayetiran, Attention-based aspect sentiment classification using enhanced learning through CNN-BiLSTM networks, Knowl.-Based Syst. 252 (2022) 109409, http://dx.doi.org/10.1016/j.knosys.2022.109409.

[44] E.F. Ayetiran, P. Sojka, V. Novotný, EDS-MEMBED: multi-sense embeddings based on enhanced distributional semantic structures via a graph walk over word senses, Knowl.-Based Syst. 219 (2021) 106902, http://dx.doi.org/10.1016/j.knosys.2021.106902.

[45] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, H. Liu, FakeNewsNet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media, Big Data 8 (3) (2020) 171–188, http://dx.doi.org/10.1089/big.2020.0062.

[46] J. Pennington, R. Socher, C.D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, a Meeting of SIGDAT, a Special Interest Group of the ACL, ACL, 2014, pp. 1532–1543, http://dx.doi.org/10.3115/v1/d14-1162.

[47] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: Y. Bengio, Y. LeCun (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015.

[48] N. Srivastava, G.E. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (1) (2014) 1929–1958, URL http://dl.acm.org/citation.cfm?id=2670313.

[49] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: F.R. Bach, D.M. Blei (Eds.), Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015, in: JMLR Workshop and Conference Proceedings, vol. 37, JMLR.org, 2015, pp. 448–456, URL http://proceedings.mlr.press/v37/ioffe15.html.

[50] T. Dozat, Incorporating nesterov momentum into adam, in: 4th International Conference on Learning Representations, ICLR 2016 Workshop Track, San Juan, Puerto Rico, USA, May 2-4, 2016, Conference Track Proceedings, 2016.

ARTICLES FOR FACULTY MEMBERS

MULTIMODAL FAKE NEWS DETECTION

Title/Author	**A novel hybrid multi-modal deep learning for detecting hashtag incongruity on social media / Dadgar, S., & Neshat, M.**
Source	***Sensors*** **Volume 22 Issue 24 (2022) Pages 1-31** **https://doi.org/10.3390/s22249870** **(Database: MDPI)**

sensors

MDPI

Article

A Novel Hybrid Multi-Modal Deep Learning for Detecting Hashtag Incongruity on Social Media

Sajad Dadgar [1,*] and Mehdi Neshat [2,3,*]

[1] Department of Mathematics and Computer Science, Amirkabir University of Technology, Tehran 15875-4413, Iran
[2] Adjunct Research Fellow at Center for Artificial Intelligence Research and Optimization, Torrens University Australia, Brisbane, QLD 4006, Australia
[3] Faculty of Engineering and Information Technology, University of Technology Sydney, Ultimo, NSW 2007, Australia
* Correspondence: s.dadgar@aut.ac.ir (S.D.); mehdi.neshat@torrens.edu.au (M.N.)

Abstract: Hashtags have been an integral element of social media platforms over the years and are widely used by users to promote, organize and connect users. Despite the intensive use of hashtags, there is no basis for using congruous tags, which causes the creation of many unrelated contents in hashtag searches. The presence of mismatched content in the hashtag creates many problems for individuals and brands. Although several methods have been presented to solve the problem by recommending hashtags based on the users' interest, the detection and analysis of the characteristics of these repetitive contents with irrelevant hashtags have rarely been addressed. To this end, we propose a novel hybrid deep learning hashtag incongruity detection by fusing visual and textual modality. We fine-tune BERT and ResNet50 pre-trained models to encode textual and visual information to encode textual and visual data simultaneously. We further attempt to show the capability of logo detection and face recognition in discriminating images. To extract faces, we introduce a pipeline that ranks faces based on the number of times they appear on Instagram accounts using face clustering. Moreover, we conduct our analysis and experiments on a dataset of Instagram posts that we collect from hashtags related to brands and celebrities. Unlike the existing works, we analyze these contents from both content and user perspectives and show a significant difference between data. In light of our results, we show that our multimodal model outperforms other models and the effectiveness of object detection in detecting mismatched information.

Keywords: hybrid deep learning models; machine learning models; stacking ensemble; XGBoost; fine–tuning; image–text multimodal classification; object detection; hashtags; social media analysis

Citation: Dadgar, S.; Neshat, M. A Novel Hybrid Multi-Modal Deep Learning for Detecting Hashtag Incongruity on Social Media. *Sensors* **2022**, *22*, 9870. https://doi.org/10.3390/s22249870

Academic Editors: Barbara Guidi and Michienzi Andrea

Received: 6 November 2022
Accepted: 12 December 2022
Published: 15 December 2022

1. Introduction

Over the past decade, with the drastic rise in the popularity of social media, these platforms have played an indispensable role in users' social lives. They have become more than just a tool for communication and sharing information privately and publicly. Social media platforms provide services allowing users to maintain direct relationships with their followers. Thus, it is a great opportunity for commercial brands and celebrities that encourage them to share visual and textual posts. In other words, social media nowadays can be considered a two-way channel where brands can amplify their marketing strategies, take customer services to the next level and enhance their knowledge about their customers by monitoring activities taking place on social networks [1] and celebrities and public figures can reach a wider audience and monetize with the content they produce. Despite the advantages of services offered to users, some problems may still transpire due to public access to these platforms.

This paper focuses on incongruent Instagram content that can also be referred to as spam, which are posts that do not match users' expectations [2]. These contents can be

found in a variety of cases on Instagram. However, we specifically assess them in hashtag searches. Hashtags are searchable keywords preceded by the hash sign # that are used on social media platforms to categorize information in different contexts. The increase in the number of social searches using hashtags sometimes makes it a substitution for conventional search engines on social media [3]. In addition, by tagging a post, users can contribute and link their posts to other related images and videos. Since using hashtags has no restriction, some users may choose the wrong hashtags inadvertently or intentionally to get more followers or monetize by advertising, which can be annoying for consumers who have a common interest through the hashtag, and also prevent creating a collection of related content that may have valuable insights. Therefore, as illustrated in Figure 1, many irrelevant and incongruent contents on hashtags can be troublesome. Not only are they disturbing and confusing for users searching through the hashtags, but they also impede brands from analyzing their customers in the best way. Additionally, post-hashtag mismatches on Instagram, which are in text and visual formats, have the potential to convey misinformation which is a significant threat [4]. Accordingly, utilizing incongruous hashtags on Instagram posts that comprise both visual and textual data may have adverse consequences. To avoid the potential threats and improve visual-sharing social network performance for any user who spends time on these platforms, there is a need to detect such posts automatically via their visual and textual information.

Computer Vision and Natural Language Processing (NLP) are two fields of Artificial Intelligence (AI) that contain many techniques that can be applied, respectively, to visuals and texts and even combined to solve several real-world problems, such as images with tags on social media [5]. As a result, our primary objective is to find a solution using these techniques that could be used to detect posts with irrelevant hashtags, whether a tag link to an individual or a prominent brand. To this end, we crawl Instagram posts with various information (e.g., images, texts, metadata) from a few hashtags with a large number of contents. In addition, we propose detection models to identify incongruity between posts and hashtags using visual and textual features. Since related objects in images can indicate whether an image is associated with a hashtag or not, we further develop object detection models to detect logos and faces and analyze their performance. Moreover, we analyze the characteristics of mismatched information and users who contribute prevalence of these contents. The key contributions are summarized as follows:

1. We introduce a dataset for Instagram that consists of metadata, visual and textual information collected from different hashtags pertinent to brands and celebrities with additional generic features related to images and texts.
2. We develop machine learning and deep learning models based on metadata, text and images for incongruity detection. We also propose a multimodal model by fusing text and image classifiers. Further, by comparing the experimental results of models, we show that our proposed multimodal models outperform other models.
3. We apply object detection to the two categories of images. First, we use brand-related images to detect a brand's logos. Second, we employ celebrity-related images to recognize the faces of the celebrity and other people who are somehow connected to them by performing clustering on their Instagram accounts and show the effectiveness of object detection to discriminate incongruent information from other relevant information.
4. We conduct an explorative analysis and empirical study of our dataset from different perspectives to categorize the type of incongruity in posts and examine the characteristic of social media users who share such posts.

The overview of the rest of the paper is as follows. We first review the related literature in Section 2. Then, the proposed approach and our methods are presented in Section 3. We conduct experimental results and explicate data analysis in Section 4. Next, we discuss limitations and provide insight into future works in Section 5. Finally, we conclude and summarize our work in Section 6.

Figure 1. An example of incongruent content that users have shared with irrelevant hashtags on Instagram.

2. Related Works

2.1. Brand Marketing and Advertising on Social Media

The previous findings reveal that different factors motivate users to use Instagram, including establishing social interaction, recording events and peeking at celebrities [6], who are public figures that have been shown to affect human behaviour strongly [7]. Thus, brands should comprehend these motivations to establish a reinforced relationship with their customers by identifying customer needs and utilizing various advertising techniques that help them obtain a desirable outcome. In recent years, marketers and brands have taken advantage of visual-sharing social networks because visual information is more memorable and provokes stronger emotional reactions than textual information [8]. Consequently, visual brand-related content is a persuasive tool that significantly influences consumers' buying intentions [9]. Instagram is one of the most popular visual-based social media platforms with a large number of active users. Along with allowing users to share visual information, Instagram, by adding new features in recent years (e.g., product tags on images), created a good marketing atmosphere that can develop more trust between users and companies [10]. Moreover, brands on this platform can implement image strategies that express their concepts and promote the public's cognitive efficiency [11]. In addition to these benefits, Instagram APIs grant users access to data from business and creator accounts, allowing brands to mine perceptual and semantic features to yield promising results. Many prior studies, therefore, were conducted about marketing and advertising on Instagram. For instance, Liu et al. [12] estimate how brands are represented on Instagram by studying consumer-created images.

In the latest works, employing Machine Learning (ML) and Deep Learning (DL) approaches also allows brands to gain more valuable marketing insight by leveraging disparate information. For instance, Paolanti et al. [13] measured the overall sentiment of brand-related images by proposing a deep Convolutional Neural Network (CNN) model. A Support Vector Machine (SVM) model was presented by Apostolova and Tomuro [14] for extracting named entities in online marketing materials. Wijenayake et al. [15] studied users' expressions and opinions toward brands and developed a Long Short-Term Memory (LSTM) model to generate and monitor brand personalities. Nakayama and Baier [16] introduced an approach to predict and prevent confusion in a brand's visual advertisements using CNN. Tous et al. [17] proposed a CNN model to efficiently filter and curate

brand-related images from Instagram and Twitter by applying object detection to discover brands' logos.

In the context of advertising, the privilege of social media in exchanging information at a high level enables brands to create positive attitudes toward their products more than traditional advertising channels. Correspondingly, advertisements have appeared in a variety of forms on social media. Both celebrities and social media influencers who accumulate a high number of followers have the capability to affect a brand's preferences [18]. One way to advertise, thus, is via exploiting influencers and celebrities, and finding appropriate ones is a challenging task. The most straightforward and relevant task of identifying proper influencers comes from research that suggests a DL algorithm to classify these influencers and disclose the impact of visual congruence on consumers' brand engagement by analyzing their interaction with their followers [19]. Another way is to customize advertising based on a user's interests, preferences and personal characteristics to boost engagement [20]. To personalize the advertisement, Hong et al. [21] have also proposed a hybrid interest classification system using Recurrent Neural Network (RNN) and CNN to classify text and images, respectively. Therefore, visual-based platforms are predominantly helpful for brand analysis.

Although social media platforms provide brands with a way to better advertise and exhibit their products that the above works have focused on, some problems can stop brands from completely taking advantage of these platforms. Unlike these works, we attempt to increase the efficiency of hashtag searches in social media, which are crucial for brand marketing and advertising, by identifying incongruent content from congruent content.

2.2. Incongruent Content, Misinformation and Spam

Despite the benefits brought by social media to brands, a huge number of unwanted information has been found on these platforms, such as incongruent content [22], spam [23] and misinformation [24] and due to the emergence of new challenges, they have been a top priority in the field of research for the past decade. This content can be broadcasted accidentally or deliberately in a fraction of a second due to the broad audience [25]. Nonetheless, we do not scrutinize the intention of such content, and we examine incongruent content regardless of intent in this study. Incongruence can be defined as information about a specific topic presented in an unrelated context. We investigate a type of incongruent information on Instagram where posts are not associated with their hashtag. Ha et al. [22] worked on the contradiction between brand-related visual data and hashtags by leveraging Computer Vision to analyze and detect these data on Instagram with images, text and meta-data cues. Basically, hashtags are one way to label content and assign it to other related content. When a hashtag is used in a post, the post will emerge on the hashtag's page. They are beneficial for garnering opinions, surveys and engagement across events. Apart from differentiating between hashtags with visual and textual information, hashtag recommendation methods have also been proven to prevent mismatched information. Alsini et al. [26] reviewed these methods on Twitter and divided them into three categories: first, methods that employ text-based [27], graph-based [28] and classification models [29,30]; next, hybrid user-based methods recommend hashtags based on similarities among users' interactions and behaviour [31,32]; lastly, hybrid miscellaneous methods whose recommendations are conducted with multimodal features [33]. In contrast to the studies that analyzed irrelevant content, other papers have concentrated on detecting and classifying the images relevant to a company with real-time object detection systems and deep learning techniques on Instagram [34]. Moreover, incongruent content and turmoil and befuddlement caused by the exposure to such information have demonstrated that users were required to make more efforts to process information, which, in the marketing domain, is a negative aspect of a brand [35–37].

The discrepancy between information can also be considered an indicator of finding fake news and misinformation [38,39], for example, misinformation detection pertinent to headlines and news [40–42]. Misinformation is a type of misleading information that

has been disseminated unintentionally and goes through a variety of labels, such as fake news, clickbait and rumors [43]. It is widely accepted that misinformation is a serious menace to societies [44], and the multiple negative impacts of such false information have led researchers to focus on this issue in several areas, ranging from health to marketing. Consequently, previous studies have addressed the task of detecting misinformation, mainly in social media as a source of information. Some have explored the concept from a verbal perspective on text-based platforms such as Twitter, while others presented works from a visual perspective due to the greater deceiving influence [45] and to avoid the dangerous use of social media and technologies that can produce misinformation, such as Deepfake technology [46,47]. Furthermore, with the positive nature of Instagram and the presence of images and video accompanied by textual information, it has provided a place for researchers to work with multimodality by fusing data from several dimensions that have been shown to perform significantly better [48,49]. Amid a wide range of models for detecting misinformation, the dissemination of this content can be amplified by automated fake accounts [50]. In consideration of that, studies were also conducted for Instagram platforms using machine-learning algorithms to discover fake accounts [51,52].

Regarding marketing, brands and companies can also be affected by misinformation in consumer reviews and fake news alongside advertising which undermine consumers' trust in them and damage brands' reputations and consumers' overall attitude toward brands [53]. Among works that assessed this issue, Vidanagama et al. [54] provided a comprehensive analysis of previous research that proposed approaches to detect deceptive consumer reviews. On the other hand, some researchers used consumer reviews for fact-checking. For instance, Zhang et al. [55] developed a model to predict the integrity of answers to consumers' Question-Answering related to products on Amazon by retrieving evidence from consumer reviews and product descriptions.

The other type of unwanted information that users encounter is spam, which is considered one form of misinformation [56]. Spam is defined as irrelevant and worthless texts and images with a high rate of repetition that is proliferated in any media, like social network platforms and emails. Their form classifies spam into social network spam, image spam, spam links, email spam and advertisement spam [57]. However, allegedly, spam appeared and increased quickly, firstly in emails [58]. The majority of previous work on the processing of spam has been conducted on email text [59], images attached to emails [60,61] and multimodal approaches to eliminate spam from emails [62–64]. Even though spam content has become part of the human experience on emails and web, the development of social media platforms and the appearance of spam content brought new challenges to this issue. They lead to a stream of research investigating how to identify and analyze spam on social media. Regarding Instagram, spam content is tied between visual and textual information [65]. Processing of spam content, therefore, is associated with images/videos along with captions and comments. For example, CNN models have been presented with different architectures to detect spam images [2] and spam comments; Complementary Naïve Bayes and SVM models have been developed on balanced and imbalanced datasets of Instagram comments [66]. Other studies focused on extracting texts from spam images by leveraging optical character recognition (OCR), which has shown that its combination with NLP and ML outperforms other ML models trained without OCR [67]. Moreover, the challenges involved in detecting spam content on a large scale have led researchers to look at spam profiles and identify spammers to prevent these accounts from generating spam and remove spam before the user falls for it [68].

Although several methods have been proposed for misinformation and spam, very few works have addressed the problem of incongruent information detection that only focuses on brand-related images. Therefore, we focused on incongruent information detection for brands and celebrities with a large number of followers on social media. Moreover, to the best of our knowledge, very few articles have explored the semantic aspect and the relationship between spam and its context [69]. In this article, we are looking for posts with irrelevant tags in the hashtag search. At first glance and regardless of other posts, a post

with irrelevant hashtags might not be spam. However, in the wrong context, when that post is found among other posts that have nothing to do with it, it becomes a worthless post that users do not expect to see among other posts, just like spam.

2.3. Machine Learning, NLP and Computer Vision

Promising results over the years obtained by different AI techniques and state-of-the-art methods that have been proposed have led researchers to deal with various problems using these technologies. Although AI technologies are broad and cannot all be mentioned in detail, three areas used in this study should be considered while examining related work; ML, NLP and Computer Vision.

ML is a component of AI that relies on algorithms and data to provide models that are able to perform a specific task automatically with accurate results. Generally speaking, according to the scale of the data and the type of tasks to be performed, different ML algorithms, such as traditional ML and DL, derive benefits from two main learning methods: supervised learning and unsupervised learning. Traditional ML techniques are predominantly supervised learning methods that include a wide variety of algorithms, such as Logistic Regression (LR), SVM, Decision Tree (DT), Naïve Bayes (NB), Random Forest (RF), K-Nearest Neighbor (KNN), etc., each of which has its own advantages [70]. Even though these algorithms are still suited for many tasks independently or as a component of ensemble models, most of them are unsuitable to be used directly in high-dimensional vector information, such as images [71]. Moreover, they cannot work without predefined data. As a result, unsupervised learning methods were performed that allowed models to recognize patterns by themselves and even deal with data that showed up in a matrix form, including images. For example, Zhang [72] proposed unsupervised image clustering algorithms on two datasets to group images into meaningful categories. The drawbacks have also caused traditional ML algorithms to be significantly overshadowed by DL, which can be considered mathematically sophisticated algorithms with a spectrum of architectures capable of solving problems using high-dimensional data [73]. From CNN with its typical layers (e.g., convolution, pooling, fully connected) that are mainly used in image processing and Computer Vision [74] to RNN, LSTM and Gated Recurrent Unit (GRU) with their ability to memorize and recognize sequential patterns in sequential data such as natural language [75]. In addition to these DL algorithms to process these forms of data, it is better to refer to other components of AI that make it possible for computer systems to perceive and process texts and images.

NLP is associated with the capability of a computer system to understand natural language, just as humans, from written or spoken communication, concerning which great strides have been made in performing various tasks such as text classification, sentiment analysis, topic modelling, translation, etc. [76]. Another component of AI is Computer Vision, which is very different to NLP, which includes processing, analyzing and comprehending digital images and videos [5]. In this area, when computer systems' functions integrate with intelligence, they can fulfil various tasks, such as image classification, object detection, semantic segmentation and so on. Though NLP and Computer Vision are two distinct and active research areas, their combination gives rise to a new interdisciplinary field with an assortment of applications in industries. Some application domains that intersect NLP and Computer Vision include image and video captioning [77], document image classification [78] and visual question answering [79]. Additionally, many studies proposed image–text multimodal and multi-view for classification models that used NLP and Computer Vision techniques to extract textual and visual features [80,81].

A comprehensive summary of the main ML and DL models applied in this field can be found in Table 1.

Table 1. A summary and comparison of the main characteristics of related ML models in the literature.

Ref.	Year	Task	Source of Information	Dataset	Data Type	Models	Main Focus
[22]	2020	Mismatch Detection	Instagram	7769 labeled posts and 444,491 unlabeled posts	Textual, Visual, Metadata	LR, SVM, RF	Detecting brand-irrelevant posts in brand-relevant hashtags
[40]	2020	News-Headlines Incongruent Detection	News	1.7 million news articles	Textual	SVM, DL	Introducing a web interface for predicting the incongruence between news-headline
[41]	2022	News-Headlines Incongruent Detection	News	Incongruent News Headline Dataset	Textual	Deep Learning (GRU)	Proposing a method to detect incongruence of news headlines using the lexical and contextual connection between news body and its headline
[42]	2020	News Headlines Incongruent Detection	News	NELA17: 91,042 news, Clickbait Challenge: 21,033 social media posts	Textual	SVM, LSTM	Incongruence detection using inter-mutual attention-based semantic matching
[27]	2021	Hashtag Recommendation	Twitter	30 million news tweets	Textual	SVM	Proposing a novel method to recommend hashtags of tweets using lexical, topical, semantic and user influence features
[29]	2020	Hashtag Recommendation	-	Two public datasets: 18,464 articles with five tags and 127,600 articles with four tags	Textual	AdaBoost, RF, LSTM, Bi-LSTM, CNN	An approach to recommend hashtags using text classification.
[30]	2020	Hashtag Recommendation	Instagram	HARRISON dataset: 57,383 multi-labeled images	Visual	Voting Deep Neural Network with Associative Rules Mining	Recommending one to ten hashtags for images
[31]	2016	Hashtag Recommendation	Twitter	1,674,789 tweets with 28,526 hashtags	Textual	Latent Dirichlet Allocation (LDA)	Hashtag recommendation using the latent relationship between words and hashtags
[32]	2020	Hashtag Recommendation	Twitter	Dataset-UDI-TwitterCrawl-Aug2012	Textual	Clique percolation method (CPM)	A community-based approach to recommend hashtags using tweet similarity
[33]	2022	Hashtag Recommendation	Social media	-	Textual, Visual, User information	Hybrid deep neural network	Proposing a multimodal personalized hashtag recommendation
[49]	2020	Misinformation Detection	Instagram	30,000 posts	Textual, Visual	LSTM, GRU, VGG16, VGG19, ResNet50, ResNet101, DenseNet121, DenseNet169, Ensemble model	Detecting medical misinformation with semantic level and task-level attention to focus on important contents
[39]	2020	Fake News Detection	-	George McIntires dataset: 10,558 texts	Textual	LSTM, FNN	Fake news detection using NLP and deep learning by including auxiliary features from live data mining
[51]	2020	Fake account Detection	Instagram	10,000 accounts	Metadata	SVM, RF, NB, DT, MLP	Introducing a method to identify fake accounts efficiently

Table 1. *Cont.*

Ref.	Year	Task	Source of Information	Dataset	Data Type	Models	Main Focus
[52]	2019	Fake Account Detection	Instagram	Two datasets: 1002 real and 201 fake accounts; 700 real and 700 automated accounts	Metadata	NB, LR, SVM, NN	Detecting fake and automated accounts
[55]	2020	Fact-Checking	Product Q&A forums on Amazon	60,864 answer claims about products	Textual	CNN, LSTM, AVER (proposed model)	Proposing AVER, a model to predict the veracity of answers based on evidence
[25]	2019	Spam Detection	Twitter	2000 Dialectical Arabic tweets	Textual	SVM, NB	Detecting malicious and spam content on Twitter written in Dialectical Arabic
[59]	2019	Spam Detection	Email	962 emails	Textual	NB, SVM	Effect of preprocessing of text on the performance of models
[60]	2016	Spam Detection	Email	52,934 images in 7 categories	Visual	CNN and SVM instead of the Softmax layer	Classifying spam images into seven categories
[61]	2022	Spam Detection	Email	1,725,928 spam images extracted from real spam emails	Visual	RF, DT, KNN, SVM, NB, CNN	Classifying spam images and analyzing the performance of ML models
[62]	2008	Spam Detection	Email	14,723 emails	Textual, Visual	DT	Proposing a system to filter out spam emails using different sets of features
[63]	2019	Spam Detection	Email	Text dataset: 2893 message, Image dataset: 2359 images	Textual, Visual	SVM	Propose a method to improve spam classification using a dataset with a small number of data
[64]	2017	Spam Detection	Email	1251 spam images from emails, Enron Spam Dataset: 33,645 texts	Textual, Visual	CNN	Detecting spam emails with hybrid architecture
[2]	2019	Spam Detection	Instagram	8000 images	Visual	CNN	Detecting spam images and comparing five different CNN architectures
[66]	2019	Spam Detection	Instagram	2600 comments	Textual	SVM, Complementary NB	Detecting spam using a balanced and an imbalanced dataset
[67]	2020	Spam Detection	-	Mark Dredze spam images: 10,000 images	Textual, Visual	DL	Extracting text from images using OCR to improve spam classification
[68]	2021	Spam Profile Detection	Instagram	916 user profiles	Metadata	MLP, RF, KNN, SVM	Detecting spammers by extracting additional features
[12]	2020	Image Classification	Flickr, Instagram	13 features about color, shape, and texture from 16,368 images	Visual	SVM, CNN	Measuring how brands are portrayed on social media
[19]	2020	Image Classification	Instagram	More than 45,000 images	Visual	CNN	Classifying images' themes and analyzing to reveal the hidden relationship between visual content and brand engagement
[17]	2018	Image Recognition, Object Detection	Instagram, Twitter	More than 50,000 images in 100 categories	Visual	CNN	Minimizing manual curation of brand-related images

Table 1. *Cont.*

Ref.	Year	Task	Source of Information	Dataset	Data Type	Models	Main Focus
[34]	2021	Image Recognition, Object Detection	Instagram	Starbucks Instagram images	Visual	Mask R-CNN, Faster R-CNN, YOLO, SSD	Proposing a model to recognize the identity of a brand using object detection
[72]	2021	Image Clustering	Chinese social media, Instagram	Images of protests, images related to climate change	Visual	K-means, Deep-Cluster	Developing three image clustering algorithms on two datasets
[13]	2017	Sentiment Analysis	Instagram	GfK Verein Dataset: 4200 positive, negative and neutral images	Textual, Visual	Deep CNN, KNN, SVM, DT, RF, NB, ANN	Estimate the overall sentiment of brand-related pictures from social media.
[21]	2020	User Interest Classification	Instagram, Twitter, Facebook, Flickr, Google	33,647 images and 21,022 texts	Textual, Visual	CNN, RNN	Improving personalized advertising based on users' interests
[14]	2014	Named Entity Recognition (NER)	Online sources	1920 online flyers	Textual, Visual	SVM	Recognizing 12 types of named entities in online marketing materials
[16]	2020	Predicting Brand Confusion	All channels	Image and video advertising	Visual	CNN	Proposed an approach to predict the uniqueness of brand positionings
[15]	2021	Generate Brand Personalities	Social media data	1.2 million posts	Textual	Deep LSTM	Investigating how users' opinions can be used to generate and monitor brand personalities

3. Materials and Methods

3.1. Approach Overview

In this section, we describe the general approach of our study on incongruent information. The proposed approach to detect this content consists of two primary modules: classification and object detection. This study starts with the classification module, which employed ML and DL methods to identify incongruity between Instagram posts and hashtags. We investigate different models to detect and classify posts based on extracted features from metadata, text and images. We further propose a hybrid multimodal model by fusing image–text classifiers. In the object detection module, we apply an object detection model to brand and celebrity-related images that enables us to identify and recognize related objects from images in our dataset. Accordingly, the first and foremost step is to construct a dataset containing Instagram posts with applicable information that allows us to research the concept of incongruent data.

3.2. Data Collection

Before dealing with the main modules, there is a need for a high-quality dataset with multimodal information for the processing to be carried out with the highest success. Hence, we used Instagram, a visual-based social media platform that provides multimodal information. Although there are different types of users on Instagram, the most followed user accounts can be divided into brands and celebrities, which were earlier found to have a higher number of spam comments [82]. Likewise, due to having more visits by other users, it is expected that more incongruent information will be found on hashtags related to prominent brands and celebrities. Therefore, we identified a set of hashtags related to brands/celebrities that have been used frequently in users' posts and created a dataset of Instagram posts by searching through these hashtags. For retrieving information, we used Instaloader, which is a python library to crawl images and videos along with JSON

files containing captions, post engagements (e.g., like count, comment count) and user information (e.g., follower count, post count, profile picture). However, we excluded videos from our study. Instaloader further allows us to garner data based on time intervals. The dataset is aggregated from posts that were published at least 30 days ago to get enough feedback. In total, we were able to gather 12,119 data objects from 8014 users, which we collected from four hashtags.

3.3. Data Annotation

Our approach for detecting incongruent information is based on supervised learning that learns from labelled training data. Consequently, the significant challenge is the availability of a dataset with reliable labels, so we annotated the data manually, which is a cumbersome and demanding task. For annotation of our dataset, we created a team of ten people, including seven master's degree students and three bachelor's degree students, all with background knowledge of computer science. Based on our experience, finding incongruent information, especially on most-used hashtags, can be identified usually by observation of images and captions. However, detailed discussions were conducted for uniformity of data labelling and acquainting annotators with the task and sample data were provided as a guideline throughout the annotation process. At the beginning of the task, we distributed our dataset among annotators in a way that three annotators evaluated each post. To simplify and accelerate the task, we developed a website that enables annotators to upload their data and evaluate each post by observing textual and visual information. After labelling each post three times, if there was even a single disagreement between the labels, we evaluated them for the last time. According to the obtained results, 76.2% of the data reached a full agreement, and the rest had at least one disagreement, which shows the effectiveness of the guideline and website. Once the data annotation task was completed, we segregated data into "match" and "mismatch" labels. Table 2 presents a list of hashtags used in the dataset with their statistical information regarding the hashtags' details and distribution of the match and mismatch content in the dataset. According to the table, among 6494 brand-related posts, 39.57% and 60.43% of them were annotated as match and mismatch, respectively. In celebrity-related posts, 38.36% of the samples belong to matched data, and 61.63% belong to mismatched data. As a result, irrelevant hashtags were used in more than half of the collected posts.

Table 2. The distribution of the collected data with additional statistical information for each hashtag.

Type	Hashtag	Total Number of Posts	No. of Collected Posts	No. of Matches	No. of Mismatches	No. of Users
Brands	#Nike	125.6 million	3151	1531	1620	2266
	#Gucci	69.4 million	3343	1039	2304	1940
Celebrities	#CristianoRonaldo	12.7 million	3481	1024	2457	2405
	#EdSheeran	5.6 million	2144	1134	1010	1403

3.4. Classification Module

This section briefly clarifies the classification module that relies on different ML and DL models with a collaboration of NLP and Computer Vision techniques to address the issue of detecting post-hashtag incongruence. The classification module adopts a four-stage approach. Each stage is explained in detail below:

3.4.1. Metadata Classification

The first stage is carried out by extracting metadata and generic features related to texts and images to quantify the characteristics of posts. Early work in various domains leverages supervised-learning methods that have specifically focused on manually curated features to solve various tasks. To determine the distinctive features between incongruence and congruence information and better analyze their characteristics, we initially explored features that have been analyzed and used for classifying other unwanted information,

such as misinformation [83], rumors [84] and spam [85] and also other works on social media that can be seen in Table 1. Then, we selected and mined the most useful features for predictive models and showed their potential to distinguish information in other tasks.

These features are mainly extracted from metadata generated by user engagement and interaction. Instagram generally makes these features available, and we extracted them straightforwardly from the collected data. Besides these features, we also extracted additional features with a bit of computing that indicates a series of attributes related to captions (e.g., word count, sentiment) and images (e.g., size, dominant colors). The overall extracted features from our datasets are listed in Table 3.

Once the features are extracted from raw data, the feature selection takes place to discover the features' importance. The selected feature vectors representing each post are then fed into different ML and DL classifiers to learn from the metadata features. Analyzing each feature's characteristic and the classifiers' performance is discussed further in Section 4.

Table 3. List of extracted features with their descriptions.

Type	Feature	Description
User	user_follower_count	Number of followers
	user_following_count	Number of followings
	user_post_count	Number of posts published by the user
	user_business_category	Type of accounts business
	user_is_business_account	Type of the account
	user_is_verified	Whether the user is verified by Instagram
Post	like_count	Number of the post's like
	comment_count	Number of comments
	has_location	Whether the location has been specified by the user in the post
	mention_num	Number of mentions (@)
	hashtags_num	Number of hashtags (#)
Text	sentiment	Sentiment of captions
	text_word_count	Number of words in captions
	capital_char_num	Number of capital characters
	digit_num	Number of digits
	hashtag_sequence_num	Index of a target hashtag in a sequence of hashtags
	is_comment_hashtags	Whether hashtags are used in the caption or comments
	mention_target_account	Whether the user also mentions the target account
	hashtag_other_related_brands	Whether the user tagged other famous brands
	tag_other_related_brands	Whether the user mentions other famous brands
	is_bio_related_brand	Whether the bio of the account relates to the target account
	is_username_related_brand	Whether the username relates to the target account
Image	dominant_color	The dominant color of the image
	image_original_size	Size of the image

3.4.2. Text Classification

This stage aims to adopt a DL approach to classify textual information. Text classification is a fundamental task in NLP that has been proven to be solved using DL models by differentiating verbal patterns and distinguishing semantic relations. Here, we developed a text classification model based on a transfer learning approach using the pre-trained Bidirectional Encoder Representations from Transformers (BERT) model to detect incongruence textual information automatically. Textual information generated by the user who publishes an Instagram post can be found in captions, which are titles or brief descriptions of images or videos and also can be found within digital images that refer to text embedded inside of the images. Therefore, our input data are divided into these two corpora. Apart from captions, which are available through the dataset, we need to extract texts from images.

Hence, we used OCR to identify and recognize alphanumeric characters automatically and symbols from digital images, including printed, typewritten and handwritten texts, and convert them into a machine-readable text format [86]. With OCR, all words and sentences overlaying the images can be recognized character by character with a confidence rate. Images relate to a scene and texts placed in the background or they are associated with advertising and texts written on objects, as shown in Figure 2. In the dataset, by performing OCR with a confidence rate of 0.7 on the original-sized images, 8159 of the images contained text, and the rest were without any text inside the images. In this study, after preparing and preprocessing the corpus, we fine-tuned the BERT pre-trained model with captions and texts extracted from OCR as two input data.

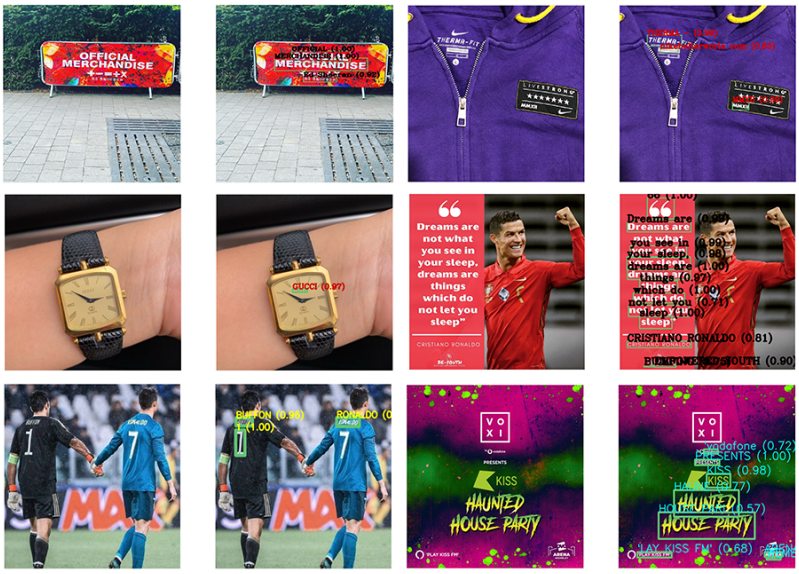


Figure 2. Sample images and recognition of corresponding overlaying text in different hashtags.

BERT [87] is a transformer-based model that consists of two steps, i.e., unsupervised pre-training and supervised fine-tuning. The pre-training step was performed on a large amount of unlabeled textual data related to a variety of domains gathered from BooksCorpus and Wikipedia. The fine-tuning step refers to the procedure of retraining the pre-trained model to adapt and perform on a custom dataset. It has been shown that such fine-tuned models outperform traditional models that require large training data sets. In general, BERT serves both as an encoder to process and extract features from input texts and a decoder to employ the features to generate outputs. However, in this study, we used BERT as an encoder to preprocess raw input data and transform them into BERT readable features, i.e., Token IDs, Input Mask and Type IDs that contain 0 or 1, indicating the padding and the token's sentence, respectively. For encoding, the input data are first tokenized and moved through an embedding layer that can transform each input token into a 768-dimensional vector. Special classification (CLS) and separator (SEP) tokens are then added correspondingly at the first and last of the sentences to clarify each sentence. Finally, positional encoding took place to learn the positions and assign a token to a unique representation based on its context (contextualized embeddings).

3.4.3. Image Classification

Another type of information that can help us to classify our data is images. Images and the features that can be obtained from them give us an advantage that might perform better than the other types of information we mentioned in previous sections. CNN is one of the DL methods that is widely used for image classification. We refer readers to a survey by Rawat and Wang [88] about a comprehensive overview of CNN in the image classification task. Although a wide variety of CNN architectures have been proposed, each

depending on the task, Mascarenhas and Agarwal [89] concluded the better performance of Resnet50 on the image classification task by comparing other pre-trained models, including VGG16 and VGG19. Thus, In this stage, we encoded the images using the Resnet50 model, which had been pre-trained on the ImageNet dataset [90] to detect image mismatches.

ResNet [91] is a CNN architecture that enables the construction of networks with thousands of convolutional layers by overcoming the vanishing gradient and exploding gradient problems. In DL models, the deeper the networks get, the less the gradient value changes. Therefore, weights are barely updated during backpropagation. However, Resnet, with its stacked (two-layer) residual blocks that have additional shortcut connections, allows the network to reduce computation and improve performance by skipping some layers and learning with deeper models. Resnet50 is a version of residual networks that consists of 48 convolutional layers along with two pooling layers, i.e., max pooling and average pooling layers. Each two-layer block is replaced with a three-layer block called a bottleneck in ResNet50.

3.4.4. Hybrid Multimodal Deep Learning Model

In real-world problems, it is often the case that the information comes not just from a single modal, but from a multimodal combination of information, just like our tasks. It has been found that multimodal classification can be most effective when text and image diverge semiotically [92]. Therefore, multimodal classification, where image and text are fused, gives us the privilege of strengthening the detection of incongruity information. Until now, we developed text classification and image classification models separately using pre-trained models. In the text classification, we extracted texts from images in the OCR and along with captions, we fed them as inputs to the BERT pre-trained model. We fed input images into a pre-trained Resnet50 model in the image classification. In this stage, however, we proposed a multimodal network by fusing these two models that simultaneously learn from images and textual contents. As illustrated in Figure 3, they moved through the learning process after passing information to the model and extracting embeddings for textual and visual information. The learning process contains fully connected layers accompanied by dropout to reduce overfitting and a batch normalization layer to normalize input features across the batch dimension that improve the training time and add a regularization effect on the network. Finally, we used a late fusion process to concatenate the two modalities and fed them to the final classification layer, a one-unit dense layer with a sigmoid activation function to detect match and mismatch content.

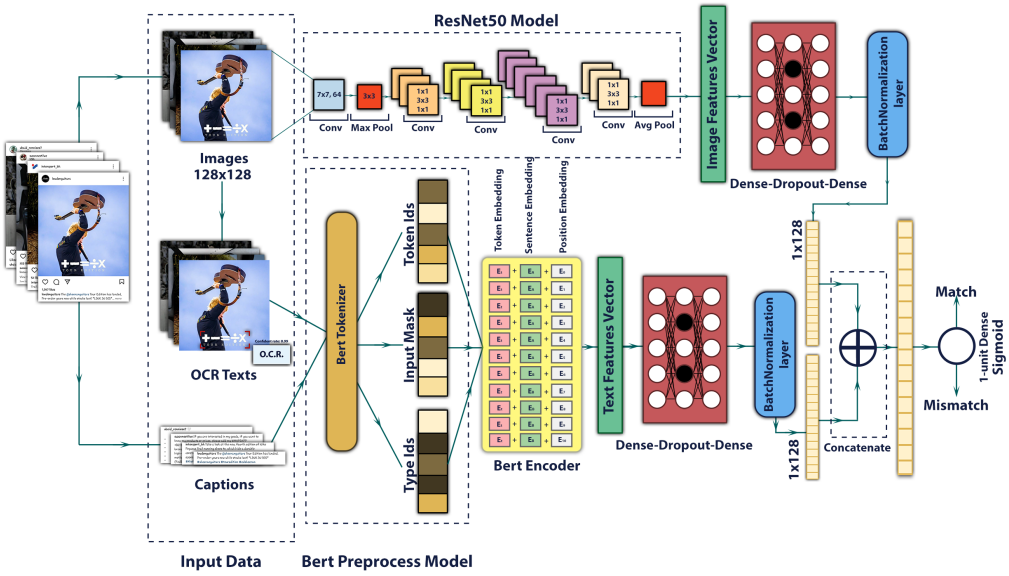


Figure 3. The proposed hybrid multimodal deep learning model.

3.5. Object Detection Module

In the previous module, we discussed different classifiers which were trained on the dataset. The dependencies of these models on the dataset could be a drawback because content on social media is evolving over time, and the classifiers cannot perform well on the other data as much as on the collected dataset. Nonetheless, relevant posts about a hashtag usually contain the same objects representing the hashtags. Therefore, identifying the objects that appear in most images can effectively detect incongruence information. Note that this module aims to show the ability to detect related objects on images and its performance in identifying incongruence data. As an object detection algorithm, we used YOLO [93], a real-time object detection system that applies a single CNN to the entire image in order to divide regions and predict classes and bounding boxes of the detected object with a probability. Moreover, employing YOLO to detect objects from images in the application has been shown to have advantages in terms of speed and accuracy over other CNN architectures [34]. Therefore, YOLO's performance in earlier works is another reason for using YOLO in the training process to recognize faces, products and logos.

As mentioned earlier, we divided the data into two categories: celebrities and brands, which each include different associated objects to detect. For example, faces for celebrities and logos for brands. Thus, this stage is also divided into face recognition and logo detection sections. Figure 4 illustrates the result of face recognition and logo detection.

(**a**) (**b**)

Figure 4. The result of the models in the object detection module. (**a**) face detection, (**b**) logo detection.

3.6. Face Recognition

The most important object, which is an indicator of the presence of celebrities in an image, without any doubt, is their faces because if their faces are present in an image of a post, the post is relevant to a hashtag, and it is right to use the hashtag. The first step for face recognition is to have many images of the target face. Hence, a source of information is required to obtain these images. Instagram is an appropriate place to collect images to recognize celebrities' faces on Instagram. The Instagram platform lets users share images and save them on their accounts. Therefore, we could find and download all images from their accounts. However, we need to detect and extract the faces from each image before face recognition. We used MTCNN (Multi-task Cascaded Convolutional Networks) [94] as a face detection algorithm containing CNNs in three stages. First, a shallow CNN generates windows and locates candidates. Second, another CNN refines the result of the previous stage and eliminates non-face candidates. Finally, a deep CNN is used to refine the candidates again and returns facial landmark positions such as nose, eyes and mouth. Even though we could extract all faces from the accounts using this algorithm, most faces do not belong to the owner of the Instagram accounts. Consequently, clustering the faces is crucial. To be able to cluster faces, we compacted high-dimensional images of faces into 128-dimensional embeddings using Google's FaceNet [95], which in light of the face clustering results, has been shown to be invariant to a variety of cases. Finally, we

performed K-means to cluster all the obtained faces, and the Elbow method was employed to determine the optimal value of clusters (K). Moreover, we sorted clusters by the number of members and filtered out clusters containing poor-quality images because K-means could not work well due to the quality of images, occlusion, image lighting and persons' pose in images. Moreover, we eliminated clusters that resulted from a group of faces of different people. The procedure of face clustering is illustrated in Figure 5.

In this way, by performing the steps on Instagram accounts of users who often share pictures of themselves with their followers, we can extract their faces since they appear in a significant number of images. Celebrities are also aware of the effects that social media accounts have on their fans [96] and publish different images of themselves, more than others, to followers. Consequently, the cluster with the highest number of members belongs to the person that owns these accounts. Furthermore, with the right number of clusters that can obtain from the Elbow method, we can also extract other faces from people who are somehow related to the owner of the Instagram accounts. Regardless of removed clusters, the people in images will be ranked based on their appearance on the image in the account by sorting the clusters based on the cluster's frequency. This means the more a person is present in an image of the respective Instagram account, and the image is placed in higher clusters; thus, the possibility that an image of this person is incongruent with the related hashtag is decreased and vice versa. All faces and corresponding cluster labels were ultimately passed to the training model as inputs.

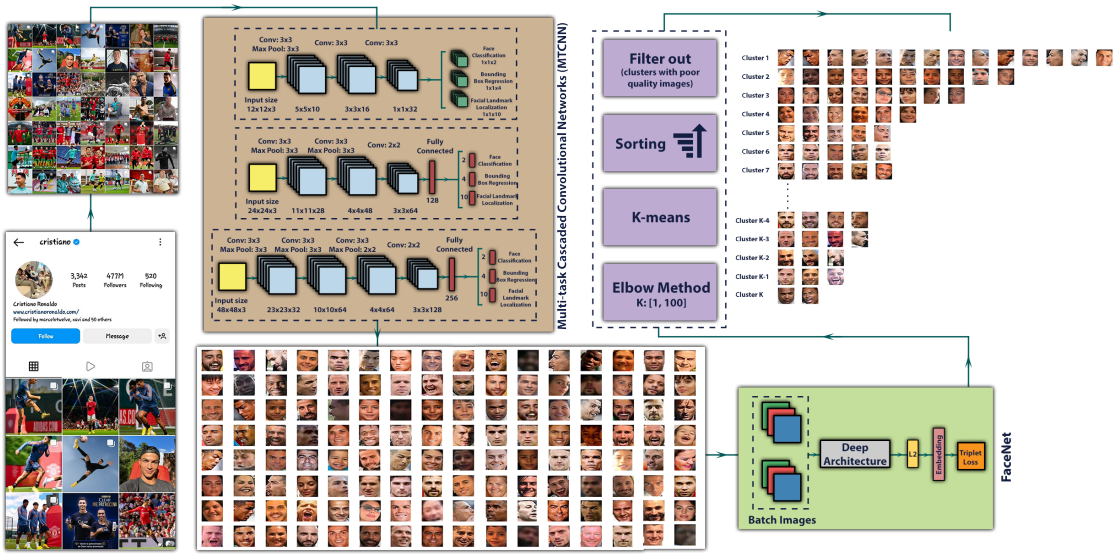


Figure 5. The procedure of face recognition. The images of MTCNN and FaceNet architecture are taken from the corresponding referenced papers.

3.7. Logo Detection

The presence of a logo in brand-related images and their detection can effectively differentiate match images from mismatches. The logo is a key component of branding; it is a brand's emblem, trademark or identity, as it typically joins brand names, products and packaging [97]. As a result, it is a fundamental part of brand-related images shared on Instagram and actually represents the brand. In this section, we used brands' logos to differentiate mismatched from matched content. Hence, we performed the YOLO object detection model to exhibit the potential of object detection by identifying the logos from images. In the training process, the images of a brand's logo were first collected via the brand's Instagram account and Google Images, containing icons and images of logos overlaying on the products. Then, the logos were labelled using LabelImg, a tool for labelling bounding boxes in images. After the labelling process, the images and their labelled bounding box passed through the training model to learn objects from images.

4. Experiment and Analysis

4.1. Data Analysis

In this section, we discuss an extensive analysis of incongruent information and explicate the characteristics of incongruent information and users that use irrelevant tags.

4.1.1. Mismatch Topics

We conducted an empirical study to discover the relation among the hashtags and their various irrelevant topics and attempt to find answers to different questions, which can help us to gain a better comprehension of this content. For example, which topics include incongruent information, which topic constitutes the majority of this information and so on. As mentioned in the discussion on data collection, we collected data for two hashtag categories for which consumers post many photos: brands and celebrities. We found incongruent hashtag-post information in 60.42% of brand data and 61.63% of celebrity data and categorized them into different topics based on observation. Although there are several ways to categorize mismatched information based on their topic, generally, the major topics on Instagram were:

- Personal: selfies and photos of an individual or group without any relation to the hashtag.
- Art: painting, graphic art, musical instruments and artists.
- Sport: sports equipment, pictures of professional sports, athletes.
- Animal: all pictures of animals.
- Food: meals and beverages and simply everything edible and drinkable.
- Cosmetic: hairdressing, makeup, cosmetic treatments, even healthcare.
- Environment: photo of nature, building.
- Quote: images of quotes, memes, tweets, manuscripts.
- Screenshot: photos displayed on the screen of a computer or mobile phone.
- Ads: posters and flyers.
- Economy: images relating to bitcoin and other digital currencies.
- Shop: online sales and products related to other brands.
- Inappropriate: sensitive and sexual pictures that are not suitable for all users. Note that Instagram strictly handles this sort of content, so there is little of them.
- Other: the remaining images do not belong to mentioned topics.

In addition, we categorized the shop category in the brand-related hashtag into related and unrelated products due to the different nature of these two types of hashtags. Figure 6 shows the proportion of topics for each hashtag in the dataset. As illustrated in Figure 6, in both brands' hashtags, the most common type of incongruent information involves brand-related products from other brands and more than half of the posts belong to this topic. As a result, it makes it difficult for our classification models to distinguish between brand-related data. On the other hand, in celebrity-related mismatches, the theme of some data is similar to the corresponding hashtag. For example, in #CristianoRonaldo, which is associated with an athlete, the sports topic has a high percentage. In #EdSheeran, which is related to an artist, the art topic forms a large part of the data. Moreover, it can be inferred that online shops and users who sell products on Instagram use celebrities' tags to attract users who constantly or occasionally visit these hashtags since a significant part of incongruent information in celebrity-related hashtags comprises the shop topic. In addition, the role of personal content is undeniable, which intertwines with mismatches in all hashtags.

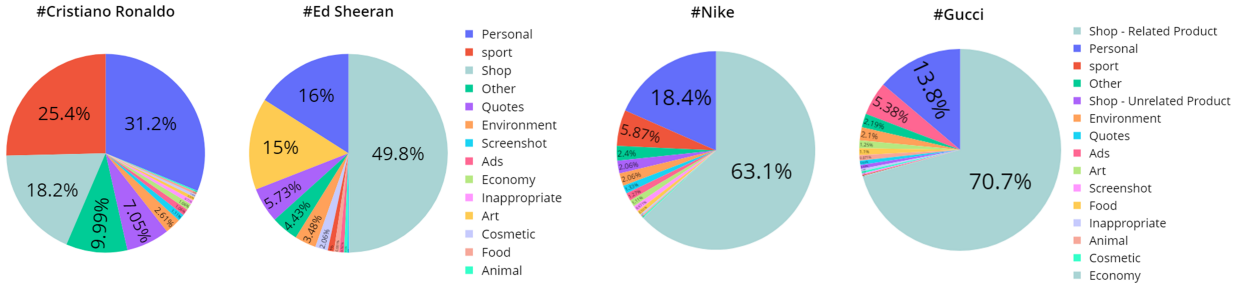


Figure 6. Distribution of the main mismatch topics over posts in the dataset.

4.1.2. Hashtag Analysis

In Figure 7, by ignoring the target hashtags that have been used for data collection and consequently are present in all sample data, we listed the most frequent hashtags in each tag used in the dataset. We found differences between the hashtag distributions in match and mismatch content. While in the congruent content, hashtags mainly refer to the target hashtag's general idea, those in the incongruent content are pertinent to diverse themes with much more repetition throughout all Instagram posts. For instance, in #EdSheeran, the congruent information is about music, concerts and tours. However, the incongruent information includes hashtags about other artists, fashion, love and business. As a result, we conclude that other unrelated hashtags that have nothing to do with each other can be found frequently in this type of content.

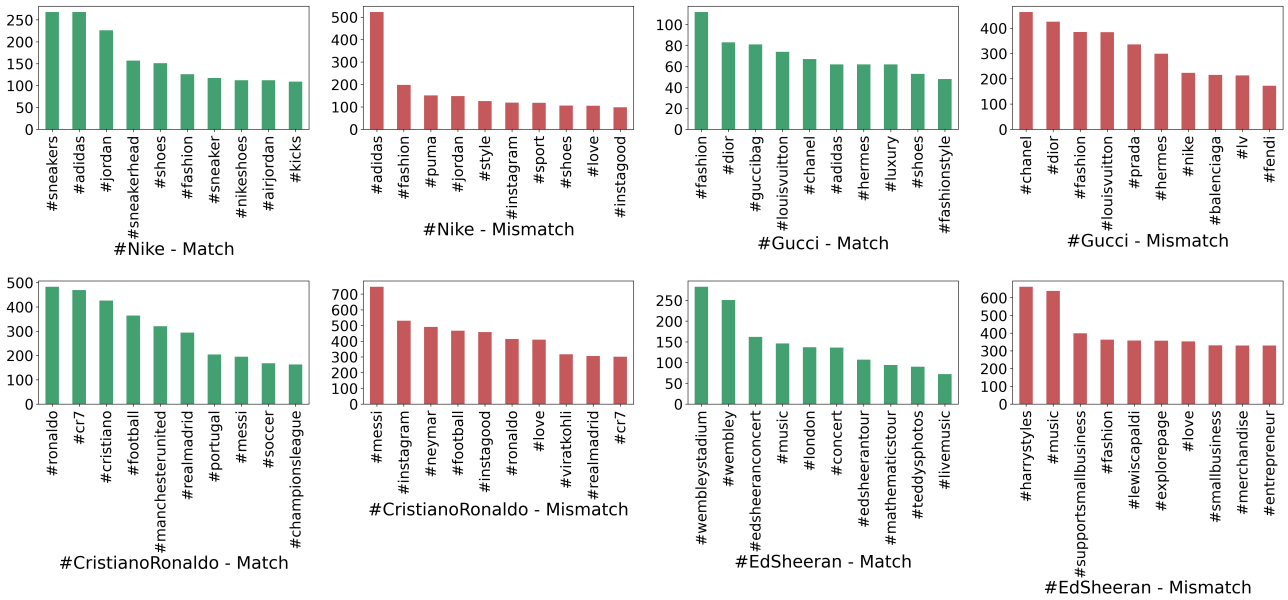


Figure 7. Frequency of the top 10 hashtags about the match and mismatch content in each category on Instagram.

Another analysis that can be pointed out about hashtags is the number of hashtags and the order of their placement in a single caption. As illustrated in Figure 8a, we discovered a significant difference between the number of hashtags in a single post. Users sharing incongruent content tend to use more hashtags in their posts to be seen by more visitors. Furthermore, considering the hashtags of a post as a sequence of tags, we obtained the index of the target hashtag in the sequence in each sample datum. As shown in Figure 8b, the target hashtag usually appears in the congruent information at the beginning of this sequence. In comparison, the hashtag in incongruent information may occur at the end of the sequence.

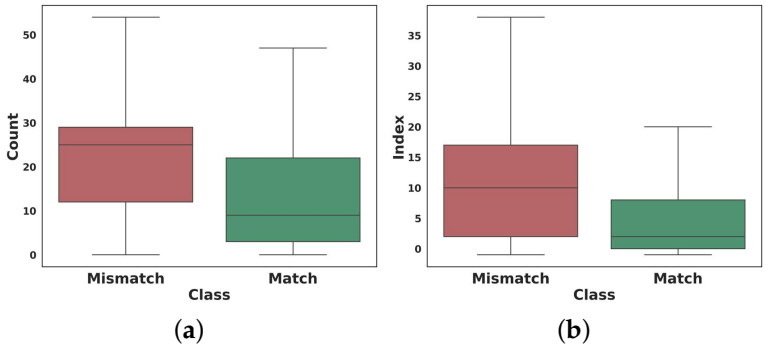


Figure 8. The comparison of hashtags in matched and mismatched content (the outliers are ignored), (**a**) the number of hashtags, (**b**) the Hashtag sequence index.

4.1.3. User Analysis

Other than content analysis, we examine the characteristics of users who are responsible for creating mismatched information by using irrelevant hashtags. As mentioned in the data collection discussion, the data were collected at least a month after their being published on Instagram. So, enough feedback and reactions had been received from others. As shown in Figure 9, based on the user engagement information obtained by measuring the audience's interaction in the sample posts, users who shared incongruent information followed more users and were followed less by others. Moreover, the number of posts of these users is less than users who create congruent information. In addition, although congruent posts received more likes, there was no significant difference in the number of comments.

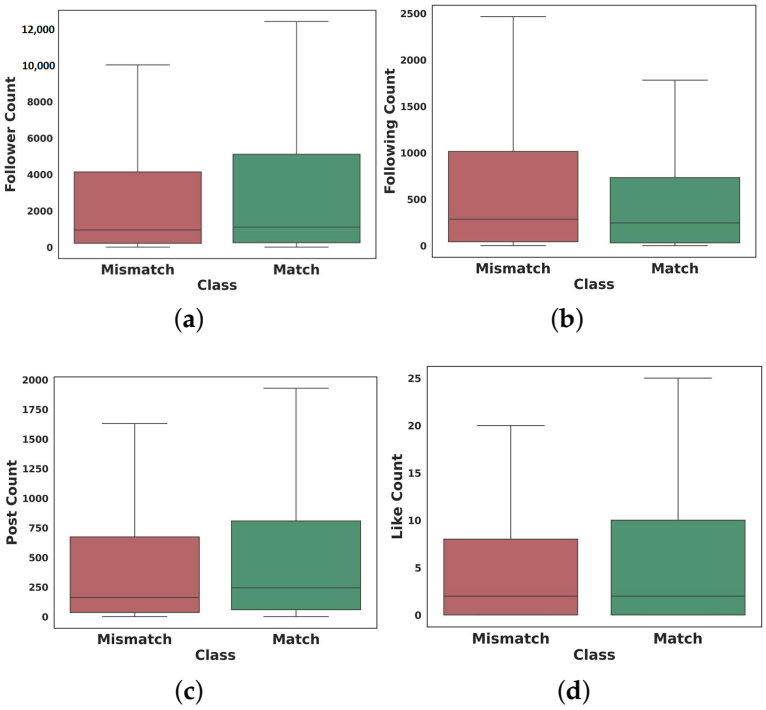


Figure 9. User engagements in the match and mismatch content (the outliers are ignored), (**a**) Follower count, (**b**) Following count, (**c**) Post count, and (**d**) Like count.

Moreover, Instagram enables users to create business accounts that provide additional features that help them to expand their business and improve their strategies, such as the ability to run advertising, access to insights to analyze their profile, posts and more. Moreover, as part of the accounts set up for business, Instagram allows users to select a

business category from hundreds of categories, letting visitors understand their type of business better. In our dataset, business accounts comprise 34.92% of the total data in 20 categories. Among these categories, "Personal Goods & General Merchandise Stores" is the most common category, followed by "Creators & Celebrities" and " Publishers". According to Figure 10, each category has more congruent information, which shows that users who do not have business accounts are more involved in generating incongruent information.

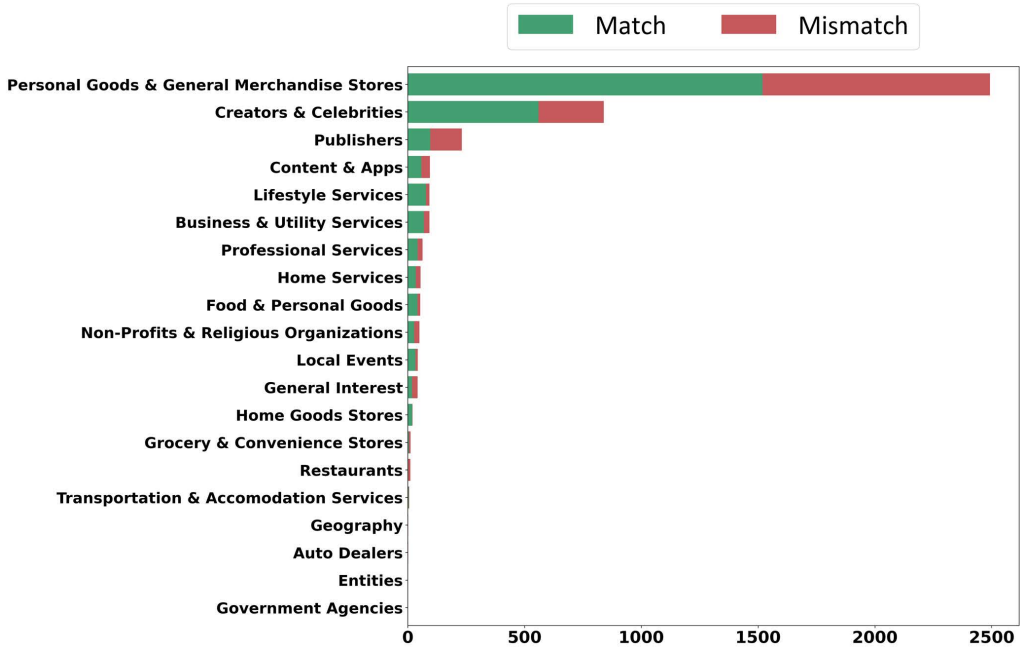


Figure 10. The comparison of business accounts in the dataset.

Finally, we investigated gender to fulfil our analysis of users. Since such information cannot be obtained from Instagram APIs, we carried out another empirical study to examine users' gender. At the time of the data collection process, we stored the profile image of each user along with other features. In this study, we divided the users into four groups based on their profile pictures: business, male, female and unknown. The unknown group includes profile pictures concerning which the gender cannot be determined due to not setting a profile picture or using fake pictures. Moreover, due to the limitation under the API, we could not extract some profile URLs in JSON files. From Figure 11, we discover that the most frequent group belongs to businesses and males share a slightly higher percentage of mismatched content than females.

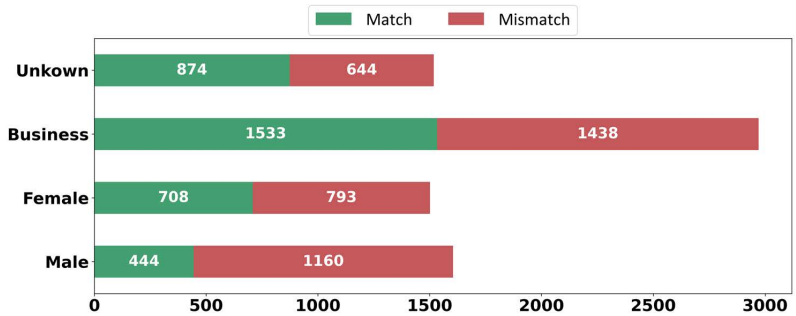


Figure 11. Comparison of the gender of users.

4.2. Experimental Results

In this section, we discuss the experimental setup and the performance evaluation of the classification and object detection models.

4.2.1. Feature Selection

In Section 3.4.1, we extracted features from the collected dataset. To discover those features that contribute most to differentiating incongruent information from congruent information and exclude redundant and irrelevant variables from the model training, we performed Recursive Feature Elimination (RFE). RFE is a recursive feature selection algorithm that identifies the important features based on an estimator's accuracy. In this experiment, we used RF as the estimator of RFE. Moreover, we performed feature importance via RF to better demonstrate the impact of the extracted features and their discriminative power. As shown in Figure 12, the number of hashtags that were also analyzed in Section 4.1.2 has the most impact on classification. In the following, user engagement features are in the next ranks. In contrast, due to the limited number of samples with corresponding characteristics, most features were extracted from the captions and other features (e.g., user_is_verified) have less contribution.

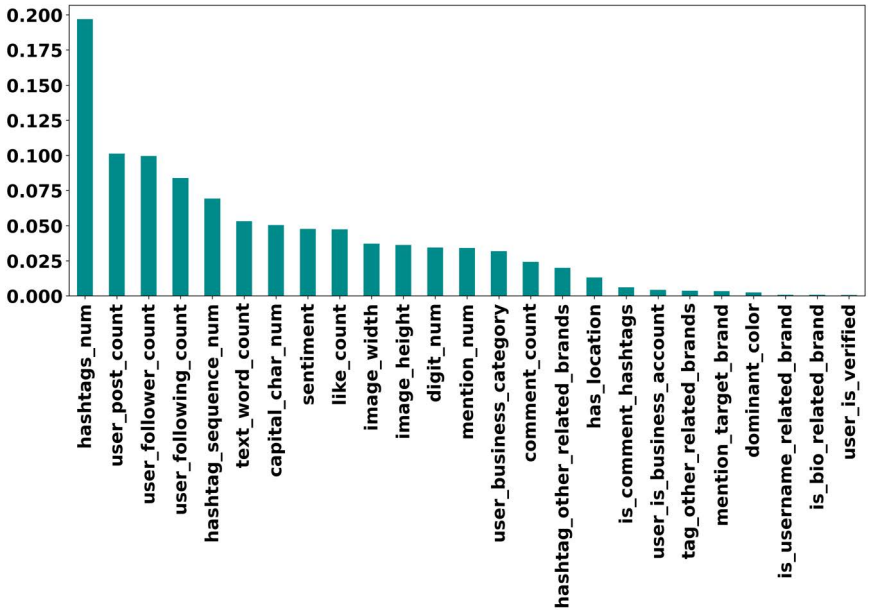


Figure 12. The feature importance of the collected features using Random Forest.

4.2.2. Classification Results

In the classification module, the collected dataset was initially divided into training, test and validation sets with an 80:10:10 ratio. We tested different architectures and tuned the hyperparameters to find the optimal models. All hyperparameters used for each method are enumerated in Appendix A. Then, the models were built using the collected dataset, as described in Section 3. In metadata classification, after performing feature selection, we used ML algorithms, including SVM [98], Stacking Ensemble ML [99], RF [100], XGBoost [101] and Deep Dense layers to train classification models. In addition to the pre-trained models described in Section 3, we also used these ML models for text and image classification to provide a comprehensive experiment. In the text and image classification tasks using these algorithms, the encoded data are obtained from the pre-trained models and used as the inputs of the ML models. The learning process was conducted using scikit-learn and TensorFlow to build the BERT, Resnet50 and VGG19 models and other ML models. The pre-trained models were fine-tuned on the collected dataset. Then, the Adam optimizer was used during the model training with a learning rate of 0.001 and a batch size of 32 for 100 epochs. The model with the best validation performance was used for evaluation. To evaluate the performance of the models in the classification module, we used accuracy and F-score, which are shown in Tables 4 and 5, respectively.

$$\text{F-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (1)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (2)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3)$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (4)$$

Based on the results obtained from the models, the image–text multimodal architecture, as expected, obtained relatively more satisfactory results than text and image classification separately. Moreover, the results indicate that integrating caption with OCR texts yields better results than classification without OCR in most cases. In addition, we observe a slight difference between the performance of the two types of hashtags in some classifiers. These models cannot classify data in brand-related hashtags as much as in celebrity-related hashtags. Our empirical study on the type of mismatched content can justify these differences. As shown in Figure 6, there was a large volume of incongruent information about brand-related products from other brands, which makes it difficult for our models to discriminate between them. Therefore, the object detection model can help in this case due to its ability to detect logos, among other related product images. Moreover, as stated in Section 3, it has been shown that the Resnet50 model outperforms other pre-trained models, such as VGG19. However, we again tested VGG19 for the image classification task and compared it with Resnet50. As a result, Resnet50 is employed in the multimodal model.

Table 4. The accuracy of the models on the test set. The best-performed model shows in bold.

Type	Models	#Nike	#Gucci	#CristianoRonaldo	#EdSheeran	All Hashtags
Metadata Classification	SVM	0.6518	0.6956	0.7106	0.7558	0.7112
	Stacking Ensemble	0.6518	0.6895	0.7220	0.7883	0.7194
	RF	0.6993	0.7492	0.7841	0.7868	0.7568
	XGBoost	0.7267	0.7522	0.7965	0.7930	0.7582
	Deep Dense layers	0.6990	0.7019	0.7177	0.7341	0.7417
Text Classification (Without OCR)	SVM	0.6307	0.6542	0.6760	0.6832	-
	Stacking Ensemble	0.6174	0.6412	0.6497	0.7037	-
	RF	0.6234	0.6955	0.6607	0.7204	-
	XGBoost	0.6429	0.7004	0.6946	0.6981	-
	Fine-tuned BERT	0.7358	0.7448	0.8142	0.8096	-
Text Classification (With OCR)	SVM	0.6317	0.6483	0.6814	0.6722	-
	Stacking Ensemble	0.6192	0.6559	0.6507	0.7305	-
	RF	0.6344	0.6784	0.6637	0.7253	-
	XGBoost	0.6830	0.7067	0.6911	0.7629	-
	Fine-tuned BERT	0.7509	0.7640	0.8172	0.8342	-
Image Classification	SVM	0.5889	0.5972	0.6175	0.6487	-
	Stacking Ensemble	0.6137	0.6214	0.6432	0.6663	-
	RF	0.6964	0.6811	0.7139	0.7100	-
	XGBoost	0.7320	0.7274	0.7548	0.7413	-
	Fine-tuned VGG19	0.7632	0.7119	0.8295	0.8444	-
	Fine-tuned Resnet50	0.7849	0.7988	0.8412	0.8785	-
Image–Text Multimodal Model	**BERT + Resnet50**	**0.8363**	**0.8536**	**0.8762**	**0.9218**	-

Table 5. The F-Score of the models on the test set. The best-performed model shows in bold.

Type	Models	#Nike	#Gucci	#CristianoRonaldo	#EdSheeran	All Hashtags
Metadata Classification	SVM	0.6867	0.7104	0.7191	0.7532	0.6864
	Stacking Ensemble	0.6550	0.7045	0.7265	0.7569	0.7208
	RF	0.7079	0.7103	0.7528	0.7697	0.7528
	XGBoost	0.7282	0.7564	0.7779	0.7638	0.7408
	Deep Dense layers	0.6987	0.6827	0.7211	0.7258	0.7169
Text Classification (Without OCR)	SVM	0.6395	0.6499	0.6993	0.6798	-
	Stacking Ensemble	0.6313	0.6238	0.6807	0.6968	-
	RF	0.5986	0.6704	0.6914	0.6835	-
	XGBoost	0.6250	0.6858	0.6851	0.7058	-
	Fine-tuned BERT	0.7460	0.7410	0.8155	0.8372	-
Text Classification (With OCR)	SVM	0.6493	0.6746	0.6858	0.6930	-
	Stacking Ensemble	0.6234	0.6432	0.6776	0.6625	-
	RF	0.6059	0.6807	0.6411	0.7167	-
	XGBoost	0.6955	0.6873	0.6798	0.7024	-
	Fine-tuned BERT	0.7547	0.7639	0.8317	0.8558	-
Image Classification	SVM	0.5275	0.5664	0.5721	0.5917	-
	Stacking Ensemble	0.5985	0.6067	0.6779	0.6776	-
	RF	0.6552	0.6775	0.7075	0.6922	-
	XGBoost	0.7082	0.7295	0.7633	0.7096	-
	Fine-tuned VGG19	0.7249	0.7198	0.8458	0.8513	-
	Fine-tuned Resnet50	0.7592	0.7743	0.8507	0.8627	-
Image–Text Multimodal Model	**BERT + Resnet50**	**0.8104**	**0.8359**	**0.8860**	**0.9106**	-

4.2.3. Object Detection Results

Additional experiments were performed by focusing on the object detection module. In the second module, we used additional images and fed them into YOLO to detect faces and logos from images in the collected dataset. First, to recognize faces from posts in celebrity-related hashtags (#CristianoRonaldo and #EdSheeran), we downloaded all images from their account (@cristiano and @teddysphotos) and performed the procedure which contains face detection, finding the optimal number of clusters using the Elbow method, clustering the faces with K-means and filtering out clusters with low-quality of faces and clusters with faces belonging to different people (unknown). The statistics of face clustering are shown in Table 6. Afterwards, the related faces with their cluster label were fed as the input to the model. Second, to detect brands' logos, we downloaded 1000 images of each logo, including logos overlayed on products and their icons from the corresponding Instagram accounts and Google Images. Ultimately, the images and their labelled logo bounding boxes pass through to the object detection model.

Table 6. Statistics of face clustering.

Hashtag	@CristianoRonaldo	@EdSheeran
Number of account images	2271	3095
Number of faces	8481	11,353
Number of clusters	65	24
Clusters with the highest number of members	1742 (20.54%)	1213 (12.30%)
Number of poor-quality faces	2778 (32.75%)	4206 (42.66%)
Number of faces in unknown clusters	1375 (16.21%)	5652 (57.33%)

In the learning process, we split the additional visual data into training and test sets with an 80:20 ratio and ran the YOLO model with a batch size of 64 over 1000 epochs. To measure the performance of the YOLO models, which make predictions in terms of bounding boxes and labels, we used Mean Average Precision (mAP). the mAP is obtained from the average of AP, which is calculated by averaging the precision of recall values for each class.

$$\text{mAP} = \frac{1}{n}\sum_{k=1}^{k=n} \text{AP}_k = 1, \quad \text{AP}_k = \text{The AP of class k}, \quad n = \text{the number of classes} \tag{5}$$

Based on our experiments, for models trained on brand-related images, the mAP values are 0.84 and 0.81 for images pertinent to Nike and Gucci, respectively. For celebrity-related images and recognizing their faces, since each face is extracted from images and the bounding box is set to the size of the image, the mAP is not a good metric to evaluate the model. We used the accuracy to measure what percentage of faces could be detected by their correct cluster labels. The result obtained by performing the model on the test datum is 87.43% for CristianoRonaldo and 72.61% for EdSheeran. Finally, to demonstrate the ability of object detection models, we perform the models on the visual data in the collected dataset. In light of the results, Figure 13 illustrates the logos and faces detected by the object detection model on the match and mismatch data in each hashtag. Based on the result obtained from running the trained models on the images of hashtags, we noticed that models were able to recognize faces and logos in many matched images. Nonetheless, it is expected that a better result will be obtained by using more data in the training process.

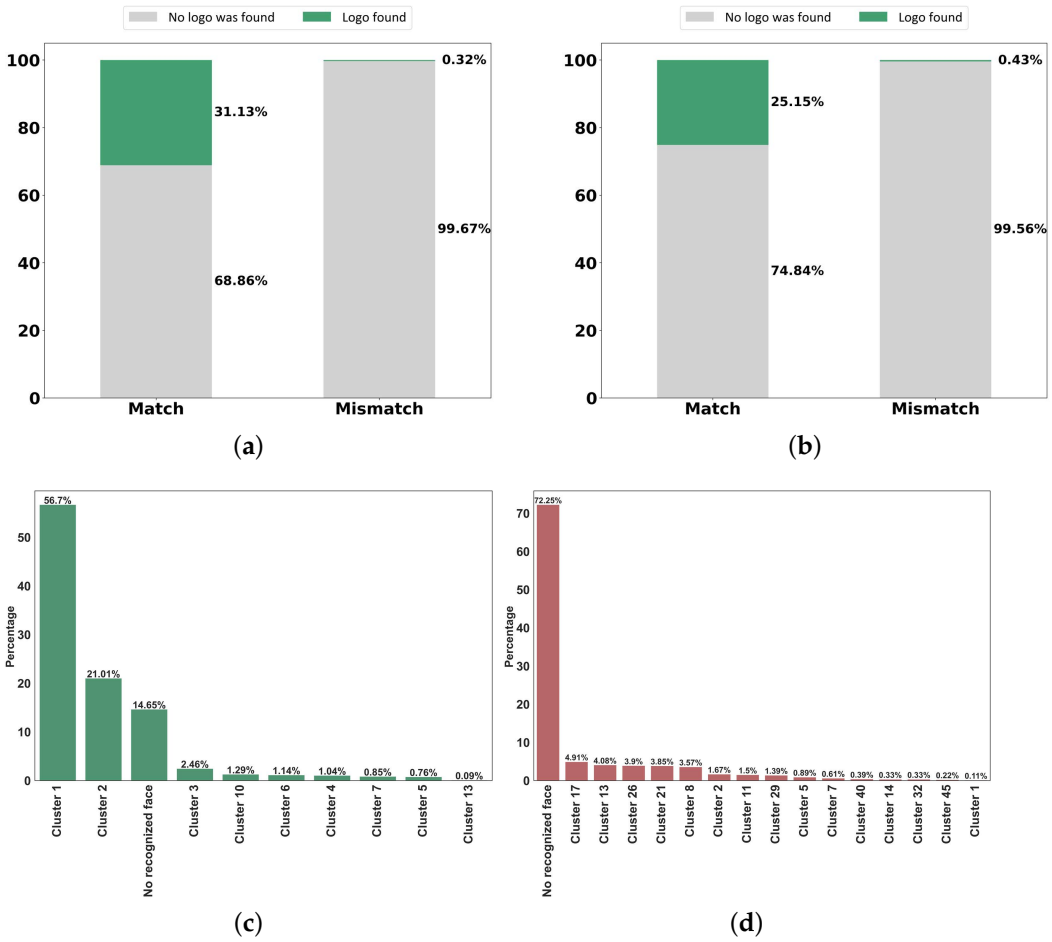


Figure 13. *Cont.*

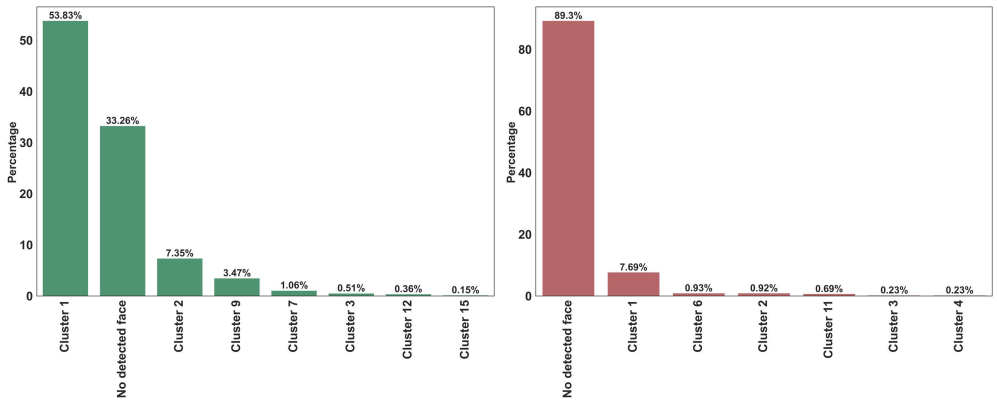


Figure 13. The result of YOLO on the dataset. (**a**) Logo detection on the #Nike hashtag. (**b**) Logo detection on the #Gucci hashtag. (**c**) Face recognition on the #CristianoRonaldo hashtag with a Match label. (**d**) Face recognition on the #CristianoRonaldo hashtag with a Mismatch label. (**e**) Face recognition on the #EdSheeran hashtag with Match label. (**f**) Face recognition on the #EdSheeran hashtag with Mismatch label.

5. Limitations and Future Works

This research explored hashtags pertinent to brands and celebrities to identify and filter out incongruent information. Hashtags also play an essential role among people during critical situations; for example, #COVID-19, is used worldwide for notifying people about the pandemic and other hashtags have helped people to be informed about occurrences and events by using them frequently and becoming a trend. Therefore, future work can concentrate on these topics as a source of information, which contain a high amount of unwanted information. This study also has several limitations that need to be explored in future research. First, we have explored incongruent information regardless of videos. Future work can investigate videos by developing methods that can be applied to audio and video. Second, we extracted several features to analyze and classify data based on content and user characteristics. Nonetheless, more features can still be extracted from text and images in the future. Third, we employed grid search to optimize the hyperparameters in the models. In addition to grid search techniques, many methods could be addressed. Future works could address the application of optimization methods to adjust the hyperparameters and develop faster and more effective auto-tuners, such as methods used in [102,103]. Fourth, as mentioned in Section 4, our classification models depend on the dataset, which brings some limitations. Other than object detection models used in this paper, future studies can focus on real-time methods to overcome these limitations. Finally, although some papers have conducted experiments to investigate the performance of object detection models in terms of speed and consistency, the same as [34], some experiments could still be conducted by applying different object detection models to detect related objects in this task.

6. Conclusions

In this research, we presented work on post-hashtag incongruent information and discussed their prevalence in hashtags searches of brands and celebrities. We initially collected a dataset consisting of Instagram posts and annotated it into match and mismatch labels. Then, we conducted our research in two modules: classification and object detection. In the classification module, we proposed methods that adopt DL, NLP and Computer Vision to detect incongruent contents from different aspects, including metadata, text and image. We also proposed a hybrid multimodal DL model based on transfer learning to learn simultaneously from visual and textual information. In the second module, to illustrate the ability of object detection models to discriminate between matched and mismatched

information, we performed YOLO on the images in the dataset to recognize faces and logos related to the hashtag. For face recognition, we trained the model using faces extracted from Instagram using a novel pipeline that ranks the faces based on the number of their appearances on an Instagram account. We also trained the logo detection model using images of logos collected from the brands' Instagram accounts and Google Images. To demonstrate the potential of our approaches in the two modules and analyze the data, we conducted experiments on the dataset. In particular, the results indicate that leveraging from both image and text simultaneously improves the results compared to other models. Furthermore, the results suggest that detecting related objects, which are the identities that link the posts to the hashtag, particularly helps to differentiate between matched and mismatched information. Finally, we conducted an explorative analysis and empirical study on our dataset. In the data analysis, we investigated characteristics of incongruent content and discussed the differences between topics, hashtags, engagements and user accounts.

Author Contributions: Conceptualization, S.D. and M.N.; Methodology, S.D. and M.N.; Software, S.D.; Validation, S.D.; Formal analysis, S.D.; Investigation, S.D. and M.N.; Resources, S.D.; Data curation, S.D.; Writing—original draft preparation, S.D.; Review and editing, S.D. and M.N.; Visualization, S.D.; Supervision, M.N.; Correspondence, M.N. and S.D. All authors have read and agreed to the published version of the manuscript.

Funding: The authors received no financial support for the research, authorship and/or publication of this article.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset collected and analyzed in this study are publicly available in https://github.com/sajaddadgar/Multi-Modal-Deep-Learning-for-Detecting-Hashtag-Incongruity (accessed on 20 November 2022).

Acknowledgments: We would like to express our gratitude to Yaser Mansouri for his collaboration and helpful comments. Moreover, we would like to thank anyone who contributed to the data annotation task and, by great effort and preparing data in a short amount of time, shed light on our way to improve the quality of our paper.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
ML	Machine Learning
DL	Deep Learning
NLP	Natural Language Processing
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
GRU	Gated Recurrent Unit
SVM	Support Vector Machines
LR	Logistic Regression
DT	Decision Tree
NB	Naïve Bayes
RF	Random Forest
KNN	K-Nearest Neighbor
OCR	Optical Character Recognition
BERT	Bidirectional Encoder Representations from Transformers
ResNet	Residual neural network
MTCNN	Multi-task Cascaded Convolutional Networks

RFE	Recursive Feature Elimination
mAP	Mean Average Precision

Appendix A

While conducting the aforementioned experiment using scikit-learn and TensorFlow, there were various hyperparameters that needed to be optimized before starting the training process. Hence, We tuned hyperparameters to find a set of optimal values that can maximize the performance of the models. As a tuning technique, we performed Grid Search on ML and DL models used in the classification module. Table A1 shows all hyperparameters and their optimal values used for building classifiers. In this table, the text classification models with and without are built with the same hyperparameters for better comparison. Moreover, we developed the multimodal model by fusing text and image classifiers with tuned hyperparameters. So we did not perform hyperparameter tuning for the image–text multimodal model. Moreover, the number of neurons and activation functions used for developing DL models is shown for each layer, respectively, in the optimal values. We also trained the stacking ensemble model with different ML algorithms for estimators and the final estimator and we selected the best combination.

$$\text{Number of Filters} = (\text{Number of classes} + 5) \times 3 \tag{A1}$$

Moreover, configuring the YOLO to train the custom object is required. Thus, we set the various hyperparameters, including the number of classes (i.e., number of clusters in celebrities, one class to detect brands' logo), number of filters to detect three boxes per grid cell that have five variables consisting of classes, width, height, x, y, confidence rate (Formula (A1)).

Table A1. The hyperparameters used in the ML and DL models in the classification module and their optimal value obtained from the grid search technique.

Type	Models	Hyperparameters	Optimal Values
Metadata Classification	SVM	kernel = [linear, poly, rbf, sigmoid], C = [1, 10, 100, 1000]	kernel = rbf, C = 1
	Stacking Ensemble	estimators = [SVM, DT, XGBoost, NB, LR]	estimators = [SVM, DT, NB], final_estimator = LR
	RF	n_estimators = [10, 20, 50, 100, 200, 500], criterion = [gini, entropy, log_loss], max_depth = [None, 2, 5, 10], max_features = [sqrt, log2, None]	n_estimators = 10, criterion = gini, max_depth = None, max_features = sqrt
	XGBoost	loss = [log_loss, deviance, exponential], learning_rate = [0.01, 0.025, 0.05, 0.075, 0.1, 0.15, 0.2], max_depth = [3, 5, 8]	loss = log_loss, learning_rate = 0.15, max_depth = 3
	Deep Dense layers	Optimizer = [SGD, RMSprop, Adagrad, Adadelta, Adam, Adamax, Nadam], Learning rate = [0.0001, 0.001, 0.01, 0.1], Dropout = [0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9], Number of neurons (hidden layers) = [16, 32, 64, 128, 256], Activation functions = [softmax, softplus, softsign, relu, tanh, sigmoid, hard_sigmoid, linear]	optimizer = Adamax, learning_rate = 0.1, Dropout = 0, number of neurons (hidden layers) = [32, 32, 32], Activation functions = [linear, relu, relu]

Table A1. *Cont.*

Type	Models	Hyperparameters	Optimal Values
Text Classification	SVM	kernel = [linear, poly, rbf, sigmoid] C = [1, 10, 100, 1000]	kernel = rbf, C = 100
	Stacking Ensemble	estimators = [SVM, DT, XGBoost, NB, LR]	estimators = [SVM, XGBoost, NB], final_estimator = LR
	RF	n_estimators = [10, 20, 50, 100, 200, 500], criterion = [gini, entropy, log_loss], max_depth = [None, 2, 5, 10], max_features = [sqrt, log2, None]	n_estimators = 100, criterion = gini, max_depth = None, max_features = sqrt
	XGBoost	loss = [log_loss, deviance, exponential], learning_rate = [0.01, 0.025, 0.05, 0.075, 0.1, 0.15, 0.2], max_depth = [3, 5, 8]	loss = log_loss, learning_rate = 0.1, max_depth = 3
	Fine-tuned BERT	Optimizer = [SGD, RMSprop, Adagrad, Adadelta, Adam, Adamax, Nadam], Learning rate = [0.0001, 0.001, 0.01, 0.1], Dropout = [0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9], number of neurons (hidden layers) = [16, 32, 64, 128, 256], Activation functions = [softmax, softplus, softsign, relu, tanh, sigmoid, hard_sigmoid, linear]	Optimizer = Adam, Learning rate = 0.001, Dropout = 0.1, number of neurons (hidden layers) = [256, 32], Activation functions = [relu, relu]
Image Classification	SVM	kernel = [linear, poly, rbf, sigmoid] C = [1, 10, 100, 1000]	kernel = rbf, C = 100
	Stacking Ensemble	estimators = [SVM, DT, XGBoost, NB, LR]	estimators = [SVM, XGBoost], final_estimator = LR
	RF	n_estimators = [10, 20, 50, 100, 200, 500], criterion = [gini, entropy, log_loss], max_depth = [None, 2, 5, 10], max_features = [sqrt, log2, None]	n_estimators = 100, criterion = gini, max_depth = None, max_features = sqrt
	XGBoost	loss = [log_loss, deviance, exponential], learning_rate = [0.01, 0.025, 0.05, 0.075, 0.1, 0.15, 0.2], max_depth = [3, 5, 8]	loss = deviance, learning_rate = 0.1, max_depth = 3
	Fine-tuned VGG19 and Resnet50	Optimizer = [SGD, RMSprop, Adagrad, Adadelta, Adam, Adamax, Nadam], Learning rate = [0.0001, 0.001, 0.01, 0.1], Dropout = [0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9], number of neurons (hidden layers) = [16, 32, 64, 128, 256], Activation functions = [softmax, softplus, softsign, relu, tanh, sigmoid, hard_sigmoid, linear]	Optimizer = Adam, Learning rate = 0.001, Dropout = 0.2, number of neurons (hidden layers) = [256, 256, 16], Activation functions = [relu, relu, relu]

References

1. Maecker, O.; Barrot, C.; Becker, J.U. The effect of social media interactions on customer relationship management. *Bus. Res.* **2016**, *9*, 133–155. [CrossRef]
2. Fatichah, C.; Lazuardi, W.F.; Navastara, D.A.; Suciati, N.; Munif, A. Image spam detection on Instagram using Convolutional Neural Network. In *Intelligent and Interactive Computing*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 295–303.
3. Sung, Y.; Kim, E.; Choi, S.M. # Me and brands: Understanding brand-selfie posters on social media. *Int. J. Advert.* **2018**, *37*, 14–28.
4. Southwell, B.G.; Brennen, J.S.B.; Paquin, R.; Boudewyns, V.; Zeng, J. Defining and measuring scientific misinformation. *Ann. Am. Acad. Political Soc. Sci.* **2022**, *700*, 98–111. [CrossRef]
5. El-Komy, A.; Shahin, O.R.; Abd El-Aziz, R.M.; Taloba, A.I. Integration of Computer Vision and natural language processing in multimedia robotics application. *Inf. Sci.* **2022**, *7*, 6.
6. Lee, E.; Lee, J.A.; Moon, J.H.; Sung, Y. Pictures speak louder than words: Motivations for using Instagram. *Cyberpsychol. Behav. Soc. Netw.* **2015**, *18*, 552–556. [CrossRef]
7. Selkie, E. Influence at the Intersection of Social Media and Celebrity. *JAMA Netw. Open* **2022**, *5*, e2143096. [CrossRef]
8. Casas, A.; Williams, N.W. Images that matter: Online protests and the mobilizing role of pictures. *Political Res. Q.* **2019**, *72*, 360–375. [CrossRef]
9. Jaulkar, M.; Virag, R.; Mohite, D.; Muktadevi, P. Impact of Advertisement on the Development of Brand Image. *SSRN Electron. J.* **2022**, *1*, 1–2. [CrossRef]
10. Broeder, P.; Schouten, M. The Impact of Product Tagging on Trust and Purchase Intention: A cross-cultural perspective in visual e-commerce. *CBR-Consum. Behav. Rev.* **2022**, *6*, 250595. [CrossRef]

11. Fu, L. A Brand Image Design Service Model Using the Visual Communication Technology under the Background of Internationalization. *Wirel. Commun. Mob. Comput.* **2022**, *2022*, 5922967. [CrossRef]
12. Liu, L.; Dzyabura, D.; Mizik, N. Visual listening in: Extracting brand image portrayed on social media. *Mark. Sci.* **2020**, *39*, 669–686. [CrossRef]
13. Paolanti, M.; Kaiser, C.; Schallner, R.; Frontoni, E.; Zingaretti, P. Visual and textual sentiment analysis of brand-related social media pictures using deep Convolutional Neural Networks. In *Image Analysis and Processing—ICIAP 2017*; Springer: Cham, Switzerland, 2017; pp. 402–413.
14. Apostolova, E.; Tomuro, N. Combining visual and textual features for information extraction from online flyers. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1924–1929.
15. Wijenayake, P.; Alahakoon, D.; De Silva, D.; Kirigeeganage, S. Deep LSTM for Generating Brand Personalities Using Social Media: A Case Study from Higher Education Institutions. *Int. J. Comput. Commun. Eng.* **2021**, *10*, 17–27. [CrossRef]
16. Nakayama, A.; Baier, D. Predicting brand confusion in imagery markets based on deep learning of visual advertisement content. *Adv. Data Anal. Classif.* **2020**, *14*, 927–945. [CrossRef]
17. Tous, R.; Gomez, M.; Poveda, J.; Cruz, L.; Wust, O.; Makni, M.; Ayguadé, E. Automated curation of brand-related social media images with deep learning. *Multimed. Tools Appl.* **2018**, *77*, 27123–27142. [CrossRef]
18. Lee, S.S.; Chen, H.; Lee, Y.H. How endorser-product congruity and self-expressiveness affect Instagram micro-celebrities' native advertising effectiveness. *J. Prod. Brand Manag.* **2021**, *31*, 149–162. [CrossRef]
19. Argyris, Y.A.; Wang, Z.; Kim, Y.; Yin, Z. The effects of visual congruence on increasing consumers' brand engagement: An empirical investigation of influencer marketing on Instagram using deep-learning algorithms for automatic image classification. *Comput. Hum. Behav.* **2020**, *112*, 106443. [CrossRef]
20. Strycharz, J.; van Noort, G.; Smit, E.; Helberger, N. Consumer view on personalized advertising: Overview of self-reported benefits and concerns. In *Advances in Advertising Research X*; Springer Gabler: Wiesbaden, Germany, 2019; pp. 53–66.
21. Hong, T.; Choi, J.A.; Lim, K.; Kim, P. Enhancing personalized ads using interest category classification of SNS users based on deep neural networks. *Sensors* **2020**, *21*, 199. [CrossRef]
22. Ha, Y.; Park, K.; Kim, S.J.; Joo, J.; Cha, M. Automatically detecting image–text mismatch on Instagram with deep learning. *J. Advert.* **2020**, *50*, 52–62. [CrossRef]
23. Sirija, M.; Jayashankari, R.; Kalpana, R.; Umamaheswari, B.; Shanthakumari, A. Characteristic based spam detection system to reveal the mock appraise in online social media. *Aip Conf. Proc.* **2022**, *2393*, 020134.
24. Rogers, R. Visual media analysis for Instagram and other online platforms. *Big Data Soc.* **2021**, *8*, 20539517211022370. [CrossRef]
25. Alorini, D.; Rawat, D.B. Automatic spam detection on gulf dialectical Arabic Tweets. In Proceedings of the 2019 International Conference on Computing, Networking and Communications (ICNC), Honolulu, HI, USA, 18–21 February 2019; pp. 448–452.
26. Alsini, A.; Huynh, D.Q.; Datta, A. Hashtag Recommendation Methods for Twitter and Sina Weibo: A Review. *Future Int.* **2021**, *13*, 129. [CrossRef]
27. Kumar, N.; Baskaran, E.; Konjengbam, A.; Singh, M. Hashtag recommendation for short social media texts using word-embeddings and external knowledge. *Knowl. Inf. Syst.* **2021**, *63*, 175–198. [CrossRef]
28. Bhaskar, R.; Bansal, A. Implementing Prioritized-Breadth-First-Search for Instagram Hashtag Recommendation. In Proceedings of the 2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 27–28 January 2022; pp. 66–70.
29. Lei, K.; Fu, Q.; Yang, M.; Liang, Y. Tag recommendation by text classification with attention-based capsule network. *Neurocomputing* **2020**, *391*, 65–73. [CrossRef]
30. Hachaj, T.; Miazga, J. Image hashtag recommendations using a voting deep neural network and associative rules mining approach. *Entropy* **2020**, *22*, 1351. [CrossRef]
31. Zhao, F.; Zhu, Y.; Jin, H.; Yang, L.T. A personalized hashtag recommendation approach using LDA-based topic model in microblog environment. *Future Gener. Comput. Syst.* **2016**, *65*, 196–206. [CrossRef]
32. Alsini, A.; Datta, A.; Huynh, D.Q. On utilizing communities detected from social networks in hashtag recommendation. *IEEE Trans. Comput. Soc. Syst.* **2020**, *7*, 971–982. [CrossRef]
33. Bansal, S.; Gowda, K.; Kumar, N. A Hybrid Deep Neural Network for Multimodal Personalized Hashtag Recommendation. *IEEE Trans. Comput. Soc. Syst.* **2022**. [CrossRef]
34. Fatma, T.; Yüksel, E. Brand Analysis in Social Networks Using Deep Learning Techniques. *Avrupa Bilim Teknol. Derg.* **2021**, *27*, 386–391.
35. Erisen, C.; Redlawsk, D.P.; Erisen, E. Complex thinking as a result of incongruent information exposure. *Am. Politics Res.* **2018**, *46*, 217–245. [CrossRef]
36. Belanche, D.; Casaló, L.V.; Flavián, M.; Ibáñez-Sánchez, S. Building influencers' credibility on Instagram: Effects on followers' attitudes and behavioral responses toward the influencer. *J. Retail. Consum. Serv.* **2021**, *61*, 102585. [CrossRef]
37. De Cicco, R.; Iacobucci, S.; Pagliaro, S. The effect of influencer–product fit on advertising recognition and the role of an enhanced disclosure in increasing sponsorship transparency. *Int. J. Advert.* **2021**, *40*, 733–759. [CrossRef]
38. Tousignant, J.P.; Hall, D.; Loftus, E.F. Discrepancy detection and vulnerability to misleading postevent information. *Mem. Cogn.* **1986**, *14*, 329–338. [CrossRef]

39. Deepak, S.; Chitturi, B. Deep neural approach to Fake-News identification. *Procedia Comput. Sci.* **2020**, *167*, 2236–2243.
40. Park, K.; Kim, T.; Yoon, S.; Cha, M.; Jung, K. BaitWatcher: A lightweight web interface for the detection of incongruent news headlines. In *Disinformation, Misinformation and Fake News in Social Media*; Springer: Cham, Switzerland, 2020; pp. 229–252.
41. Jang, J.; Cho, Y.S.; Kim, M.; Kim, M. Detecting incongruent news headlines with auxiliary textual information. *Expert Syst. Appl.* **2022**, *199*, 116866. [CrossRef]
42. Mishra, R.; Yadav, P.; Calizzano, R.; Leippold, M. MuSeM: Detecting incongruent news headlines using mutual attentive semantic matching. In Proceedings of the 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA), Miami, FL, USA, 14–17 December 2020; pp. 709–716.
43. Zannettou, S.; Sirivianos, M.; Blackburn, J.; Kourtellis, N. The web of false information: Rumors, fake news, hoaxes, clickbait, and various other shenanigans. *J. Data Inf. Qual. (JDIQ)* **2019**, *11*, 1–37. [CrossRef]
44. Ahmed, M.; Bachmann, S.; Martin, C.; Walker, T.; Rooyen, J.; Barkat, A. False Information as a Threat to Modern Society: A Systematic Review of False Information, Its Impact on Society and Current Remedies. *J. Inf. Warf.* **2022**, *21*, 105–120.
45. Mena, P.; Barbe, D.; Chan-Olmsted, S. Misinformation on Instagram: The impact of trusted endorsements on message credibility. *Soc. Media+ Soc.* **2020**, *6*, 2056305120935102. [CrossRef]
46. Shahzad, H.F.; Rustam, F.; Flores, E.S.; Luís Vidal Mazón, J.; de la Torre Diez, I.; Ashraf, I. A Review of Image Processing Techniques for Deepfakes. *Sensors* **2022**, *22*, 4556. [CrossRef]
47. Han, B.; Han, X.; Zhang, H.; Li, J.; Cao, X. Fighting fake news: Two stream network for deepfake detection via learnable SRM. *IEEE Trans. Biom. Behav. Identity Sci.* **2021**, *3*, 320–331. [CrossRef]
48. Xarhoulacos, C.G.; Anagnostopoulou, A.; Stergiopoulos, G.; Gritzalis, D. Misinformation vs. Situational Awareness: The Art of Deception and the Need for Cross-Domain Detection. *Sensors* **2021**, *21*, 5496. [CrossRef]
49. Wang, Z.; Yin, Z.; Argyris, Y.A. Detecting medical misinformation on social media using multimodal deep learning. *IEEE J. Biomed. Health Inform.* **2020**, *25*, 2193–2203. [CrossRef]
50. Miró-Llinares, F.; Aguerri, J.C. Misinformation about fake news: A systematic critical review of empirical studies on the phenomenon and its status as a 'threat'. *Eur. J. Criminol.* **2021**, 1–19. [CrossRef]
51. Sheikhi, S. An Efficient Method for Detection of Fake Accounts on the Instagram Platform. *Rev. D'Intell. Artif.* **2020**, *34*, 429–436. [CrossRef]
52. Akyon, F.C.; Kalfaoglu, M.E. Instagram fake and automated account detection. In Proceedings of the 2019 Innovations in Intelligent Systems and Applications Conference (ASYU), Izmir, Turkey, 31 October–2 November 2019; pp. 1–7.
53. Di Domenico, G.; Sit, J.; Ishizaka, A.; Nunan, D. Fake news, social media and marketing: A systematic review. *J. Bus. Res.* **2021**, *124*, 329–341. [CrossRef]
54. Vidanagama, D.U.; Silva, T.P.; Karunananda, A.S. Deceptive consumer review detection: A survey. *Artif. Intell. Rev.* **2020**, *53*, 1323–1352. [CrossRef]
55. Zhang, W.; Deng, Y.; Ma, J.; Lam, W. AnswerFact: Fact checking in product question answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; pp. 2407–2417.
56. Tainter, J.A.; Taylor, T.G.; Brain, R.; Lobo, J. *Emerging Trends in the Social and Behavioral Sciences: An Interdisciplinary, Searchable and Linkable Resource*; John Wiley & Sons: Hoboken, NI, USA, 2015.
57. Geerthik, S. Survey on internet spam: Classification and analysis. *Int. J. Comput. Technol. Appl.* **2013**, *4*, 384.
58. Ahmed, N.; Amin, R.; Aldabbas, H.; Koundal, D.; Alouffi, B.; Shah, T. Machine learning techniques for spam detection in email and IoT platforms: Analysis and research challenges. *Secur. Commun. Net.* **2022**, *2022*, 1862888. [CrossRef]
59. Ruskanda, F.Z. Study on the effect of preprocessing methods for spam email detection. *Indones. J. Comput. (Indo-JC)* **2019**, *4*, 109–118. [CrossRef]
60. Shang, E.X.; Zhang, H.G. Image spam classification based on Convolutional Neural Network. In Proceedings of the 2016 International Conference on Machine Learning and Cybernetics (ICMLC), Jeju, Korea, 10–13 July 2016; Volume 1, pp. 398–403.
61. Abuzaid, N.N.; Abuhammad, H.Z. Image SPAM Detection Using ML and DL Techniques. *Int. J. Adv. Soft. Compu. Appl.* **2022**, *14*, 226–243. [CrossRef]
62. Gargiulo, F.; Sansone, C. Combining visual and textual features for filtering spam emails. In Proceedings of the 2008 19th International Conference on Pattern Recognition, Tampa, FL, USA, 8–11 December 2008; pp. 1–4.
63. Kumaresan, T.; Saravanakumar, S.; Balamurugan, R. Visual and textual features based email spam classification using S-Cuckoo search and hybrid kernel Support Vector Machine. *Clust. Comput.* **2019**, *22*, 33–46. [CrossRef]
64. Seth, S.; Biswas, S. Multimodal spam classification using deep learning techniques. In Proceedings of the 2017 13th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), Jaipur, India, 4–7 December 2017; pp. 346–349.
65. Chrismanto, A.R.; Sari, K.; Suyanto, Y. Critical evaluation on spam content detection in social media. *J. Theor. Appl. Inf. Technol.* **2022**, *100, 2642–2667*.
66. Haqimi, N.A.; Rokhman, N.; Priyanta, S. Detection of Spam Comments on Instagram Using Complementary Naïve Bayes. *IJCCS (Indones. J. Comput. Cybern. Syst.)* **2019**, *13*, 263–272. [CrossRef]
67. Yaseen, Y.K.; Abbas, A.K.; Sana, A.M. Image spam detection using machine learning and natural language processing. *J. Southwest Jiaotong Univ.* **2020**, *55*, 1–9.
68. Raza, T.; Afsar, S.; Jameel, J.; Mateen, A.; Khalid, A.; Naeem, H. Execution Assessment of Machine Learning Algorithms for Spam Profile Detection on Instagram. *Int. J.* **2021**, *10*, 1889–1894.

69. Fahfouh, A.; Riffi, J.; Mahraz, M.A.; Yahyaouy, A.; Tairi, H. A Contextual Relationship Model for Deceptive Opinion Spam Detection. *IEEE Trans. Neural Net. Learn. Syst.* **2022**, 1–12. [CrossRef]
70. Mahesh, B. Machine learning algorithms-a review. *Int. J. Sci. Res. (IJSR)* **2020**, *9*, 381–386.
71. Lai, Y. A comparison of traditional machine learning and deep learning in image recognition. *J. Phys.* **2019**, *1314*, 012148. [CrossRef]
72. Zhang, H.; Peng, Y. Image clustering: An unsupervised approach to categorize visual data in social science research. *Sociol. Methods Res.* **2021**, 00491241221082603.
73. Mredula, M.S.; Dey, N.; Rahman, M.S.; Mahmud, I.; Cho, Y.-Z. A Review on the Trends in Event Detection by Analyzing Social Media Platforms' Data. *Sensors* **2022**, *22*, 4531. [CrossRef]
74. Li, Z.; Liu, F.; Yang, W.; Peng, S.; Zhou, J. A survey of Convolutional Neural Networks: Analysis, applications, and prospects. *IEEE Trans. Neural Net. Learn. Syst.* **2021**, *33*, 6999–7019. [CrossRef]
75. Van Houdt, G.; Mosquera, C.; Nápoles, G. A review on the long short-term memory model. *Artif. Intell. Rev.* **2020**, *53*, 5929–5955. [CrossRef]
76. Kang, Y.; Cai, Z.; Tan, C.W.; Huang, Q.; Liu, H. Natural language processing (NLP) in management research: A literature review. *J. Manag. Anal.* **2020**, *7*, 139–172. [CrossRef]
77. Hossain, M.Z.; Sohel, F.; Shiratuddin, M.F.; Laga, H. A comprehensive survey of deep learning for image captioning. *ACM Comput. Surv. (CsUR)* **2019**, *51*, 1–36. [CrossRef]
78. Bakkali, S.; Ming, Z.; Coustaty, M.; Rusiñol, M. Visual and textual deep feature fusion for document image classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 562–563.
79. Kafle, K.; Kanan, C. Visual question answering: Datasets, algorithms and future challenges. *Comput. Vis. Image Underst.* **2017**, *163*, 3–20. [CrossRef]
80. Seeland, M.; Mäder, P. Multi-view classification with Convolutional Neural Networks. *PLoS ONE* **2021**, *16*, e0245230. [CrossRef] [PubMed]
81. Guarino, A.; Lettieri, N.; Malandrino, D.; Zaccagnino, R.; Capo, C. Adam or Eve? Automatic users' gender classification via gestures analysis on touch devices. *Neural Comput. Appl.* **2022**, *34*, 18473–18495. [CrossRef]
82. Subyantoro, S.; Apriyanto, S. Impoliteness in Indonesian language hate speech on social media contained in the Instagram account. *J. Adv. Linguist.* **2020**, *11*, 36–46. [CrossRef]
83. Al-Rakhami, M.S.; Al-Amri, A.M. Lies kill, facts save: Detecting COVID-19 misinformation in Twitter. *IEEE Access* **2020**, *8*, 155961–155970. [CrossRef]
84. Liang, G.; He, W.; Xu, C.; Chen, L.; Zeng, J. Rumor identification in microblogging systems based on users' behavior. *IEEE Trans. Comput. Soc. Syst.* **2015**, *2*, 99–108. [CrossRef]
85. Chen, C.; Zhang, J.; Xie, Y.; Xiang, Y.; Zhou, W.; Hassan, M.M.; AlElaiwi, A.; Alrubaian, M. A performance evaluation of machine learning-based streaming spam tweets detection. *IEEE Trans. Comput. Soc. Syst.* **2015**, *2*, 65–76. [CrossRef]
86. Awel, M.A.; Abidi, A.I. Review on optical character recognition. *Int. Res. J. Eng. Technol. (IRJET)* **2019**, *6*, 3666–3669.
87. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
88. Rawat, W.; Wang, Z. Deep Convolutional Neural Networks for image classification: A comprehensive review. *Neural Comput.* **2017**, *29*, 2352–2449. [CrossRef]
89. Mascarenhas, S.; Agarwal, M. A comparison between VGG16, VGG19 and ResNet50 architecture frameworks for Image Classification. In Proceedings of the 2021 International Conference on Disruptive Technologies for Multi-Disciplinary Research and Applications (CENTCON), Bengaluru, India, 19–21 November 2021; Volume 1, pp. 96–99.
90. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
91. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NA, USA, 27–30 June 2016; pp. 770–778.
92. Kruk, J.; Lubin, J.; Sikka, K.; Lin, X.; Jurafsky, D.; Divakaran, A. Integrating text and image: Determining multimodal document intent in Instagram posts. *arXiv* **2019**, arXiv:1904.09073.
93. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NA, USA, 27–30 June 2016; pp. 779–788.
94. Zhang, K.; Zhang, Z.; Li, Z.; Qiao, Y. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process. Lett.* **2016**, *23*, 1499–1503. [CrossRef]
95. Schroff, F.; Kalenichenko, D.; Philbin, J. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 15 October 2015; pp. 815–823.
96. Ward, J. A content analysis of celebrity Instagram posts and parasocial interaction. *Elon J. Undergrad. Res. Commun.* **2016**, *7*, 1–4.
97. Kelly, M. Analysing the complex relationship between logo and brand. *Place Brand. Public Dipl.* **2017**, *13*, 18–33. [CrossRef]
98. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
99. Wolpert, D.H. Stacked generalization. *Neural Net.* **1992**, *5*, 241–259. [CrossRef]
100. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]

101. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
102. Neshat, M.; Nezhad, M.M.; Sergiienko, N.Y.; Mirjalili, S.; Piras, G.; Garcia, D.A. Wave power forecasting using an effective decomposition-based convolutional Bi-directional model with equilibrium Nelder-Mead optimiser. *Energy* **2022**, *256*, 124623. [CrossRef]
103. Neshat, M.; Nezhad, M.M.; Mirjalili, S.; Piras, G.; Garcia, D.A. Quaternion convolutional long short-term memory neural model with an adaptive decomposition method for wind speed forecasting: North aegean islands case studies. *Energy Convers. Manag.* **2022**, *259*, 115590. [CrossRef]

ARTICLES FOR FACULTY MEMBERS

MULTIMODAL FAKE NEWS DETECTION

Title/Author	**CAF-ODNN: Complementary attention fusion with optimized deep neural network for multimodal fake news detection / Luvembe, A. M., Li, W., Li, S., Liu, F., & Wu, X.**
Source	***Information Processing and Management*** **Volume 61 Issue 3 (2024) 103653 Pages 1-26** **https://doi.org/10.1016/j.ipm.2024.103653** **(Database: ScienceDirect)**

Information Processing and Management 61 (2024) 103653

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

CAF-ODNN: Complementary attention fusion with optimized deep neural network for multimodal fake news detection

Alex Munyole Luvembe, Weimin Li *, Shaohau Li, Fangfang Liu, Xing Wu

School of Computer Engineering and Science, Shanghai University, Shanghai, China

ARTICLE INFO

Keywords:
Multimodal fake news
Image processing
Optimization
Attention fusion
Grid search
Instance normalization

ABSTRACT

Fake news is a real problem; unfortunately, it seems to worsen. Even though some false news detection methods have made significant progress, current multimodal approaches integrate cross-modal features directly without considering uncorrelated semantic representations may introduce noise into the multimodal features. This phenomenon reduces model accuracy by obscuring subtle differences between text and images crucial for identifying fake news. Uncorrelated semantics also reduce the detection accuracy since the identification often relies on these subtle differences. To address these challenges, we propose a unified **C**omplementary **A**ttention **F**usion with an **O**ptimized **D**eep **N**eural **N**etwork **(CAF-ODNN)** that captures subtle cross-modal relationships for multimodal fake news detection. CAF introduces image captioning to represent images semantically, allowing bidirectional complementary attention between modalities based on a scaled dot product to learn fine-grained correlations. A dedicated alignment and normalization component is incorporated to calibrate fused representations based on channel statistics, ensuring the semantics are preserved across modalities during the interaction, thus improving upon the simple concatenation used in existing fusion approaches. To improve feature extraction, an Optimized Deep Neural Network **(ODNN)** is implemented that exploits compositional learning. ODNN is designed with three fully connected layers to learn higher-level representations from CAF-fused features. Model parameters are then systematically tuned beyond standard random search techniques to identify configurations, maximizing feature quality and detection accuracy. Our proposed method outperforms comparable approaches on standard metrics on four real-world datasets, highlighting the importance of complementary attention fusion with optimization in identifying fake news.

1. Introduction

Social media has transformed online communication and interactions, making it easier for individuals to establish relationships and connect online. However, these platforms have also become hubs for fake news and misinformation dissemination (Olan et al., 2022). Since they are user-centric, they rarely focus on the accuracy or reliability of the news they share. As a result, sensational or controversial content, spreads far and wide as people share it with their networks. A misleading claim, for instance, that chloroquine could cure COVID-19 led many people to consume chloroquine phosphate, trusting it would safeguard them from the deadly virus.

* Corresponding author.
E-mail address: wmli@shu.edu.cn (W. Li).

https://doi.org/10.1016/j.ipm.2024.103653
Received 25 June 2023; Received in revised form 3 January 2024; Accepted 4 January 2024
Available online 14 January 2024

Unfortunately, this incident resulted in deaths and hospitalizations.[1] Misinformation akin to these preys on people's trust in biased information, achieving the manipulator's goals. Recent years have seen an increase in fake news, making it more critical than ever to distinguish between true and false information. Thus, there has been an exponential rise of third-party fact-checking services and content moderation to correct fabricated news and prevent the dissemination of misleading and inaccurate information (Kryściński et al., 2020), (Giachanou et al., 2022), (Bagade et al., 2020), (Paschalides et al., 2021), (Munyole et al., 2023), (Li et al., 2022) and (Sengupta et al., 2021). However, despite the progress, it is evident that significant amount of effort is still required to ensure social media remains a safe and reliable source of information and a platform for building relationships and fostering social interactions by effectively preventing misinformation proliferation.

A multimodal fake news detection system analyzes multiple modalities, such as text, images, and videos, to determine whether a piece of content is fake or real. The approach considers that multiple modalities provide deeper understanding of fake news, thus helping to resolve complexities related to misinformation. A key aspect of existing multimodal false news detection algorithms is the direct integration (concatenation) of multimodal representations to enhance performance. Based on this perspective, Singhal et al. (2019) proposed SpotFake model for detecting fake news by leveraging textual and visual information. SpotFake utilized BERT for textual embeddings and VGG19 for images to learn contextual information based on the input data. A universal multimodal framework was also developed by Wang et al. (2018) based on time-sensitive events that learned temporal dynamics and related relationships. In a follow-up study, the authors employed a simulated learning via meta-learning to combine a small number of data instances (Wang et al., 2021). Khattar et al. (2019) proposed a bimodal variational autoencoder (MVAE) that incorporated information from text and image to detect fake news. MVAE learned a shared representation of the input data, which improved classification performance by capturing complementary information in different modalities. Methods based on attention (Sachan et al., 2021) employ cross-modal attention techniques to detect false information online by recognizing interdependencies and relationships between textual and visual elements. However, these approaches result in sub-optimal performance since they fail to capture all relevant information. The above methods have proven effective, but a simple combination of image and text features may not always provide reliable information in every instance. In addition, the veracity of news is not determined by the correlation between the image and text alone since little semantic similarity between the text and image features can result in a noisy representation. The fact that fake news disguises itself as real using subtle differences in the text and image makes detection via direct fusion difficult. The methods that exploit attention, on the other hand, are sensitive to input noise, including variations and perturbations. Even minor disturbances in the input can potentially influence the attention weights and ultimately impact the fusion process.

The concatenation approach utilized by numerous multimodal fake news detection methods may not adequately represent the distinctions between low-level visual features extracted from the image and the high-level semantic concepts conveyed by the text. This phenomenon happens because the approach combines the feature vectors from the image and text modalities into a single vector for classification. However, this method fail to capture the subtle and complex relationships (Yu et al., 2022) between the modalities and introduces noise since it treats them as independent sources of information. As shown in Fig 1(a) and (b), there are subtle differences between the post and the generated image caption. In Fig 1(a), the news post states that "*Biden signs into law Democrats' wide-ranging climate change, health care and tax bill,*" however, the accompanying image shows the photo of the former president, Barack Obama probably signing a landmark health care bill in 2010, different on what the post expresses. On a closer look at the image, Biden can be seen. In the second example, Fig 1(b), the post purports that "*A brave driver tows a wheeler during hurricane Katrina*" while the image generated caption shows that "*A truck and a wheeler parked in a parking lot next to a body of water*". In the first example, Fig 1(a), the post of the text and image captions shows subtle inconsistent and do not match the context, which could be a sign of fake news. In this case, detecting fake news may require careful analysis of the text and the image. In Fig 1(b), while the post indicates that the wheeler is being towed, it is not shown in the picture. Thus, the text and the image caption reveal some subtle inconsistencies. We can infer discrepancies between the generated image caption and the actual content of the image, such as the absence of key individuals or objects mentioned may influence effective multimodal fake news detection.

As social media remains a popular medium for online users to share their views and thoughts, images often accompany text surrounding them. Image captions, thus, can be a helpful feature in fake news detection. Image captioning enables the representation of an image in textual form (Wu & Mebane, 2022) and may generate valuable context and information for understanding and decoding the image. In this study, we refer to image captions as the textual descriptions of the image content, and we utilize them as a substitute rather than a supplement to enhance the interaction between the modalities. This approach is particularly relevant in fake news detection, where images may be unrelated to the textual information or manipulated to support false claims. Generating and analyzing the image captions can help identify patterns of misinformation and disinformation characterizing fake news. Furthermore, although images themselves do not directly authenticate news content, they define significant issues in people's daily lives that can stimulate conversations because they simulate human image interpretation (Al-Malla et al., 2022). The advancement of visual technology has facilitated augmentation of images with text, enabling models to extract semantic relationships between images and text, thus enhancing multimodal analysis and understanding.

This paper addresses the above challenges by proposing a unified **C**omplementary **A**ttention **F**usion with an **O**ptimized **D**eep **N**eural **N**etwork **(CAF-ODNN)** for multimodal fake news detection. First, CAF employs image captioning to represent images semantically, allowing bidirectional complementary attention between modalities based on a scaled dot product to learn fine-grained correlations. Then, a dedicated alignment and normalization component is incorporated to calibrate fused representations based on channel

[1] https://www.nbcnews.com/health/health-news/man-dies-after-ingesting-chloroquine-attempt-prevent-coronavirus-n1167166

Post: Biden signs into law Democrats' wide-ranging climate change, health care and tax bill.

Image generated caption: President Obama signs a landmark health care overhaul in the Oval Office

(a)

Post: A brave driver tows a wheeler during hurricane Katrina

Image generated caption: A truck and a wheeler parked in a parking lot next to a body of water

(b)

Fig. 1. (a) and (b): Fake news relies on subtle differences between the text and image. Harvesting such information can contribute to effective multimodal detection and help eliminate noise in multimodal fusion.

statistics, ensuring the semantics are preserved across modalities during the interaction, thus improving upon the simple concatenation and averaging used in existing fusion approaches. To improve end-to-end feature extraction, an Optimized Deep Neural Network **(ODNN)** that exploits compositional learning is implemented. ODNN is designed with three fully connected layers to learn higher-level representations from CAF-fused features. Model parameters are then systematically tuned beyond standard random search techniques to identify configurations, maximizing feature quality and detection accuracy.

The main contributions of this paper are:

- We propose a novel unified **C**omplementary **A**ttention **F**usion **(CAF)** for multimodal fake news detection that captures subtle cross-modal relationships. CAF introduces image captioning to represent images semantically, allowing bidirectional complementary attention between modalities based on a scaled dot product to learn fine-grained correlations. A dedicated alignment and normalization component is incorporated to calibrate fused representations based on channel statistics, ensuring the semantics are preserved across modalities during the interaction, thus improving upon the simple concatenation used in existing fusion approaches.
- We propose an **O**ptimized **D**eep **N**eural **N**etwork **(ODNN)** for end-to-end refined feature extraction. The ODNN exploits compositional learning designed with three fully connected layers to learn higher-level representations from CAF-fused features. Model parameters are then systematically tuned beyond standard random search techniques to identify configurations, maximizing feature quality and detection accuracy.
- We evaluate the proposed approach for accuracy, recall, F1 score, and precision on four benchmark datasets to validate the detection efficiency.

2. Related work

2.1. Fake news detection using single modality

A single-modality approach to fake news detection relies on information from a single source to identify and capture rich latent features. For example, methods that use texts extract semantic or statistical features from news articles to detect fake news. From this viewpoint, Ma et al. (2018) proposed a recursive neural network to identify fake news on X (formely Twitter) platform. Based on the hierarchical nature of tweets, the authors considered hidden signals within their structure. On the other hand, Mohtarami et al. (2018) introduced a memory network to determine the authenticity of claims. The method combined the strengths of recurrent and convolutional networks to capture temporal and local dependencies in the input data. The model also included a novel similarity and filtering component that minimized irrelevant information in related claims, resulting in improved performance of the memory network. Yu et al. (2017) proposed a supervised learning model incorporating a convolutional neural network (CNN) to extract and classify features from news articles. In addition to the text, it has also been demonstrated that images can be used as a single modality feature to identify fake information (Choras et al., 2018; Cao et al., 2020). Visual methods rely on the assumption that visual cues and patterns offer valuable insights into the information veracity. For example, Qi et al. (2019) introduced a model incorporating pixel–level and frequency-level features to identify fake news. With the complementary information provided by these features, the model

gained understanding of the visual content, strengthening the model's ability to detect fake news. Steinebach et al. (2019) presented a novel technique that involved extracting image features and employing an indexing technique to compare and verify image authenticity using a mesh verifier. While these methods prove effective, they primarily focus on event-level detection that relies on event flags. Event flags involve monitoring the spread of fake news over time; hence, they are resource-intensive and time-consuming, limiting scalability and practicality. The model, may also lack a comprehensive understanding of fake news since it focuses on a specific modality.

Several temporal and network dynamic-based models have also been used to distinguish between fake and credible news. Temporal models (Li et al., 2022; Gong et al., 2023; Li et al., 2023; Alharbi et al., 2023) considers the sequence of events and interactions within a network to gain a deeper understanding and determine how news articles or social media posts propagate and evolve. This phenomenon helps the models identify characteristics distinctive to fake news dissemination. For instance, Li et al. (2023) reconstructed a network based on the infection potential energy and incorporated a source location method that considers rumor propagation centrality and diffusion tendency. On the other hand, Zhang et al. (2022) proposed a heuristic algorithm to enhance the influence of a dynamic GCN (Graph Convolutional Network) network by introducing a leader-fake labeling mechanism. Based on the network topology, the algorithm incorporated adaptive layers to obtain representations of the network nodes and selected seed nodes for training the model. Even though the above methods demonstrate the importance of incorporating network context and relationships between nodes for more effective fake news detection, they may suffer from limited generalizability to different datasets and contexts. Additionally, some of these approaches are computationally expensive and fail to incorporate multimodal information in their detection tasks, limiting their effectiveness in capturing complex relationships between different modalities.

2.2. Multimodal fake news detection

2.2.1. Detection based on multimodal features

Early multimodal approaches to fake news recognition combined different multimodal features to enhance detection accuracy. These methods were primarily based on the complementary nature of media, such as text, images, and metadata to capture more comprehensive and robust features (Wang et al., 2018; Singhal et al., 2019; Khattar et al., 2019; Armin et al., 2021). Wang et al. (2018), for example, proposed an EANN model that combined text-CNN and VGG-19 to improve the correlation between modalities. Khattar et al. (2019) introduced variational encoders to learn probabilistic latent variables and enhance the correlation between different modalities. These approaches effectively demonstrate the benefits of integrating multiple modalities for fake news detection. Besides, Jin et al. (2017) developed a multimodal deep-learning model that combined text, image, and social properties using a combined CNN and RNN networks. The multimodal features extracted were then fused via a multimodal fusion layer to produce a combined representation. This approach, however, did not incorporate the similarity features of images or social engagement into analysis. More recently, Armin et al. (2021) designed a multimodal framework incorporating diverse information with different levels of abstraction. In addition to textual and visual content, the model exploited user comments and metadata. Then, fusing the modalities in phases was adopted to preserve the modality's intrinsic structure. These studies demonstrates that combining different modalities improves detection accuracy. However, simply fusing image and text features may not provide reliable information in all instances because text and image are not often correlated.

2.2.2. Detection based on multimodal consistency

The practice of including unrelated images with textual posts has gained prominence in the creation of multimodal fake news. Hence, in recent years, researchers have sought to identify the level of multimodal consistency between text and visual components in a post (Zhang et al., 2023) (Xiong et al., 2023) (Meel & Vishwakarma, 2021) (Wu et al., 2023) to mitigate the impact of such deceptive content. A SAFE model was proposed by Zhou et al. (2020) that considered similarity measurement when evaluating the consistency of multimodal information. On the other hand, Xue et al. (2021) presented Multimodal Consistency Neural Network (MCNN) that integrated a similarity measurement technique to assess the correlation between text and images. Separate networks were used to extract features from the textual and visual modalities and then mapped into a shared feature space through weight sharing. An essential feature of this model was the incorporation of a similarity measurement module that calculated the cosine similarity between the textual and visual features. While these methods have demonstrated success in multimodal fake news detection, they may encounter challenges in capturing intricate relationships between different modalities or handling missing modalities. They are also susceptible to noise. For instance, if inconsistencies arise in the predictions made by various modalities due to noise or errors in feature extraction, the model's performance is adversely affected.

2.3. Multimodal fusion

Majority of feature fusion algorithms are based on concatenation (Singhal et al., 2019) (Ma et al., 2019), in which features from the textual and visual modes are combined into a single vector and classified using a neural network or a support vector machine. However, this approach has limitations in capturing the complex relationships between the textual and visual modalities since it treats text and visuals as separate sources of information. For example, Nguyen et al. (2019) concatenated document features derived from Word2vec, user events, TF-IDF, and node2vec and fed them into fake news detection algorithm. However, the fusion technique applied did not consider feature interaction. Shu et al. (2019) and Zhang et al. (2019) improved the fusion technique by exploiting the attention mechanism. This technique enabled the model to focus on the interaction of posts and comments, user profiles, and behaviors. However, the method relied primarily on the collaboration of partial and local features rather than global ones.

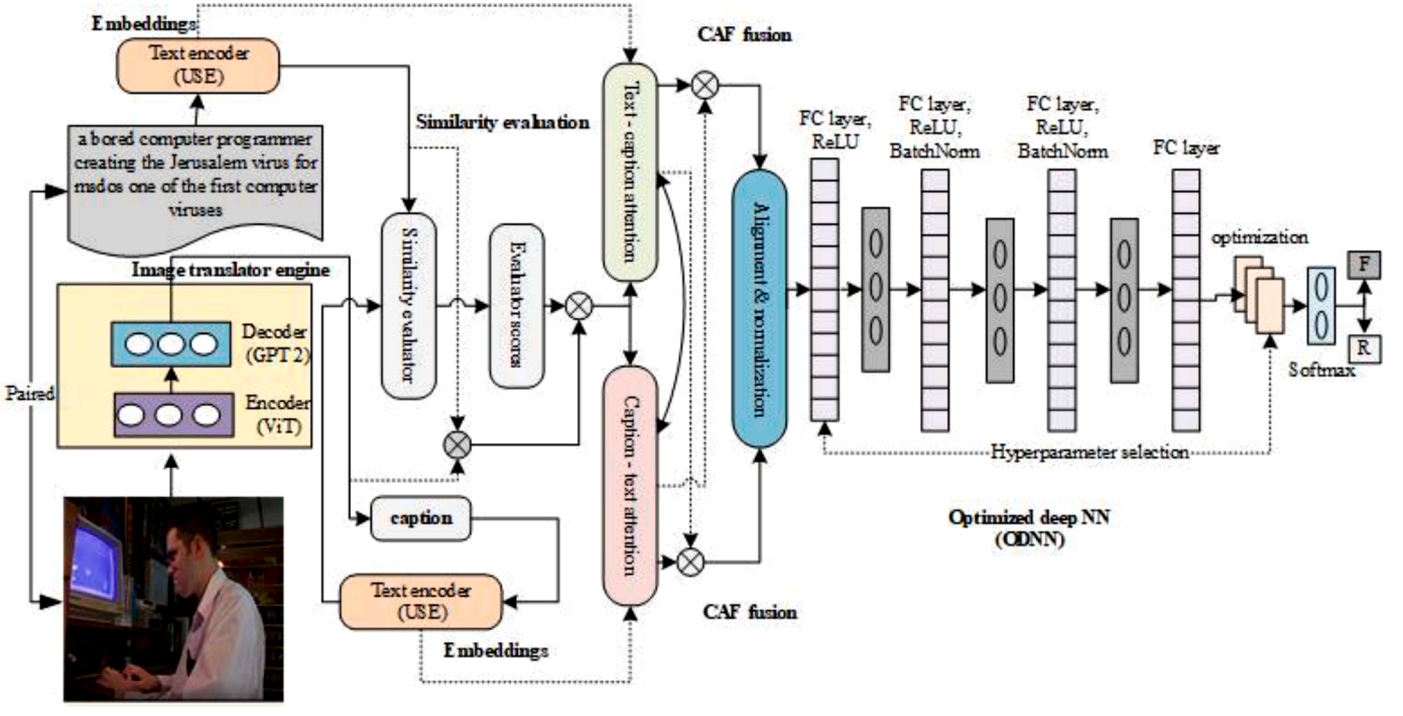


Fig. 2. The proposed CAF-ODNN framework. The Image translator engine comprises a ViT encoder that extracts images as a sequence of patches and a GPT-2 decoder that generates quality image captions. The image captions and the posts are then embedded using a text encoder and passed to the Similarity Evaluator to measure feature similarity. The Complementary Attention Fusion (CAF) applies bidirectional attention to learn fine-grained correlations between the image captions and the text. The fused features are passed to the alignment module for calibration. ODNN improves feature extraction by learning higher-level representations from CAF-fused features and maximizes feature quality and detection accuracy.

There has been a growing interest in deep learning methods that employ different modules to capture multimodal features (Zhou et al., 2020; Song et al., 2021; Kumari & Ekbal, 2021; Xue et al., 2021; Ni et al. 2022). These methods also utilize various fusion approaches to combine multimodal features for evaluation. For example, Zhou et al. (2020) presented a multimodal learning framework to enhance the correlation between text and visual features. The approach entailed extracting features from textual and visual, and integrating them through a fusion layer incorporating a loss function based on the semantic distance between the two. Song et al. (2021) improved the correlation between multimodal features using a Cross-Attention and Recurrent Multimodal Network (CARMN). The model utilized a fusion layer and cross-attention mechanisms to capture the correlation between the textual and visual characteristics. However, the technique failed to account for fine-grained relationships between individual features within each modality. Kumari and Ekbal (2021) leveraged the Attention-based Multimodal Fusion model with Bayesian optimization (AMFB) to maximize the correlation between heterogeneous features in multimodal learning context. An essential characteristics of the model was the assignment mechanism that rationed optimal properties to each modality. For example, the Bayesian optimized each feature based on the correlation with the other modalities. Xue et al. (2021) proposed a framework for multiple modes that considered the consistency of multimodal information and generated general characteristics of social media posts. Though the model was effective, it required feature optimization at the fusion level to standardize different multimodal information. Ni et al. (2022) introduced a collaborative learning approach that combined a relationship representation graph with a two-level attention mechanism. The model facilitated interactions among nodes in a mutual representation network. However, it did not explicitly highlight the heterogeneity of features.

Even though the above methods improve cross-modal correlation, they still have some limitations. The concatenation of features results in a shallow network since the approach fails to choose meaningful features and also introduces noise in the detection algorithm. In addition, employing diverse modules to capture and fuse multimodal elements make the model sensitive to noise in feature extraction, inhibiting the model to capture fine-grained relationships between individual features within each modality. This limitation can result in the model overlooking subtle characteristics that could contribute valuable information for fake news detection. In this work, we introduce image captioning to represent images semantically, allowing bidirectional complementary attention between modalities to learn fine-grained correlations. A dedicated alignment and normalization component is employed to calibrate fused representations based on channel statistics. This strategy ensures the semantics are preserved across modalities during the interaction, improving simple concatenation used in existing fusion approaches. We improve feature extraction by using an optimized deep neural network that exploits compositional learning. The model is designed with three fully connected layers to learn higher-level representations from CAF-fused features. Model parameters are then systematically tuned beyond standard random search techniques to identify configurations, maximizing feature quality and detection accuracy.

3. Methods

3.1. Model framework

Multimodal information presents a challenge for fusion since the modalities typically have different representations and dimensions. The challenge arises because text, images, videos, and other modalities can contain different types of information that may not be directly comparable. This section provides a detailed explanation of our proposed framework, as shown in Fig 2.

3.2. Image translator engine

The image translator engine generates image captions. The concept is that captions help reduce the semantic gap between the image and the text modality. An encoder-decoder architecture is used in the module to convert the image into a latent representation with a textual description. The captions are generated by rescaling images to 224 × 224 resolution and using a beam search with four beams. In this study, ViT-GPT- 2 (Dosovitskiy et al., 2021) is used. It is a publicly available image-to-text captioning pipeline consisting of a Vision Transformer (ViT) neural network as an encoder and a GPT-2 (Radford et al., 2020) neural network as a decoder. The ViT-GPT 2 model is trained on over 330,000 annotated images from the MS COCO captions dataset (Chen et al., 2015). The primary benefit of using ViT-GPT-2 is that it allows end-to-end training of the model, resulting in a better representation of the image features and the caption generated. Additionally, the self-attention mechanisms in ViT captures long-range dependencies between different parts of the image, which is useful in generating more accurate and detailed captions. Given I as an input image with width w and height h, the ViT encoder processes the image to generate a set of K feature vectors denoted as:

$$F = (f_1, \ldots, f_K) \tag{1}$$

where each feature vector $f_i \in \mathbb{R}^d$ has a dimension of d

The GPT-2 encoder takes the set of feature vectors F as the input and generates a caption, which is represented as a sequence of tokens as:

$$C = (c_1, \ldots, c_T) \tag{2}$$

where each token c_i represents a symbol or word in the caption.

The probability of generating the sequence C given the feature vectors F is achieved by:

$$P(C|F) = P(c_1|\ F) * P(c_2\ |\ f_1, \ldots, f_K,\ c_1) * \ldots * P(c_T\ |\ f_1, \ldots,\ f_K,\ c_1, \ldots,\ c_T - 1 \tag{3}$$

where $P\ (c_i\ |f_1,\ \ldots,\ f_K,\ c_1,\ \ldots,\ c_i{-}1)$ denotes the conditional probability of generating the $i-th$ token given the previous tokens and the feature vector.

The generated image caption is represented in a space compatible with the post's textual features.

3.3. Embeddings

The image translator engine is first exploited to generate the text caption describing the image. Then, the text is encoded into 512 high-dimensional embedding vectors to represent the input strings. A Universal Sentence Encoder (USE) (Cer et al., 2018) is employed which comprises two main components: a sentence encoder and a transformer-based encoder. Using this strength, the semantic meaning of text and more complex relationships between words and phrases can be captured since USE considers the context of a text. The text is encoded into a fixed-length vector representation that captures the semantic meaning of the text. Given t as an input text sequence, and h an output representation generated by the USE model, the sentence encoder first learn a fixed-length vector representation $u(t)$for the input text sequence t exploiting convolutional layers as shown below:

$$u(t) = f([c_1,\ \ldots, c_k]) \tag{4}$$

Where c_i denotes the $i-th$ of the convolutional layer and f is a non-linear activation function.

The transformer-based encoder then takes the fixed length vector representation $u(t)$ as input and generates the final output representation $h(t)$ using self- attention mechanism:

$$h(t) = g([\ e_1,\ \ldots,\ e_n]) \tag{5}$$

where e_i denotes the $i-th$ output of the transformer-based encoder, and g is a multi-layer pecerptron that maps concatenated output vectors to the final representation $h(t)$ which is a fixed length vector capturing the semantic meaning of the input text sequence.

3.4. Similarity evaluator

The model generate text (s_{text}) and image caption (s_{image}) embeddings separately and compares the two embeddings to determine similarity. The algorithm can capture the meaning of the sentence based on context, intent, and other semantic nuances by considering the evidence of embedding similarity. Cosine similarity computes text (s_{text}) and image caption (s_{image}) embeddings is used to capture their respective features in a numerical form using Eq. (6). Each dimension (i.e. from 1 to n) of the embedding is iterated. For each dimension, the corresponding elements of s_{text} and s_{image} and sum of their products are multiplied. Furthermore, the length of the embeddings is determined (Eq. (7), Eq. (8)). The idea is that the length of the embeddings serves as useful features for assessing the authenticity of news articles. For instance, if the lengths of both the text and image caption embeddings are large, it indicates that the textual and image captions features are relatively strong and well-represented. This may suggest a higher likelihood of the news article being authentic. On the other hand, if the lengths of the embeddings are relatively small, it could indicate weak or insufficient representation of the textual and visual information. Finally, the cosine similarity with the sigmoid function is computed as shown in Eq. (9). A sigmoid function is included to ensure that the resulting similarity value falls within the range of 0 to 1. It is desirable to utilize this mapping since it provides a more interpretable and normalized measure. In this normalization step, the magnitudes of the embeddings are considered, ensuring similarity value are not biased by the lengths of the vectors. A sigmoid function maps the value to a range between 0 and 1, providing a probabilistic interpretation of the similarity score and making it easier to interpret and compare similarity values across pairs of text and image. The derived similarity value serves as an input to the complementary fusion module.

$$dot_{product} = \sum_{i=1}^{n} s_{text}[i]\ .\ s_{image}\ [i] \tag{6}$$

where $s_{text}[i]$ and $s_{image}\ [i]$ represents the elements in the i th dimension of s_{text} and s_{image} respectively. The index ranges from 1 to n, where n is the dimensionality of the embeddings. Each dimension corresponds to a specific feature of the text or image caption.

$$\|\ s_{text}\| = \sqrt{\sum_{j=1}^{n} s_{text}[j]^2} \tag{7}$$

$$\|\ s_{image}\| = \sqrt{\sum_{j=1}^{n} s_{image}[j]^2} \tag{8}$$

where $s_{text}[j]^2$ and $s_{image}[j]^2$ represent the elements or values in the j-th dimension of the s_{text} and s_{image} respectively. The index j ranges from 1 to n, where n is the dimensionality of the embeddings.

$$cos\ (s_{text},\ s_{image}) = \sigma\left(\frac{s_{text}\ .\ s_{image}}{||s_{text}||.\ ||\ s_{image}||}\right) \tag{9}$$

where σ denotes the sigmoid function that maps its input value to a value between 0 and 1. It is defined as $= \frac{1}{1+e^{-x}}$. $\|\ s_{text}\|$ and $\|\ s_{image}\|$ denotes the lengths or magnitudes of the text and image embeddings, respectively.

3.5. Complementary attention fusion

Instead of concatenating modalities to form a final representation in the fusion process, we use a scaled dot product attention technique to fuse the modalities. This technique involves calculating the output of a set of queries Q to a set of keys K and values V. The dimensions of the keys are denoted as d_k, and the dimensions of the values are denoted as d_v. The output is calculated as follows:

$$Attention\ (Q,\ K,\ V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)\ V \tag{10}$$

In the above calculation, the softmax function is applied to the scaled dot product of the queries Q and the keys K, divided by the key's dimensions. The process produces a set of weights to combine the values V to produce the output linearly.

First, the textual features set $(h_1,\ \ldots,\ h_n)$ is represented as the queries Q, while the image caption features$(I_c,\ \ldots,\ I_w)$ are represented as the keys K and values V. The attention mechanism is used to compute the attention weights for each textual feature for the image caption features:

$$h_i^{'} = \sum_{j=0}^{w} a_j^i I_j \tag{11}$$

For each textual featureh_1, the attention weights a_j^i are computed using the softmax function applied to the dot product h_i and each image caption feature I_j.

$$a_j^i = softmax\ \left(h_i I_j^T\right) \tag{12}$$

The set of textual features is $(h_1^{'},\ \ldots,\ h_n^{'})$. Finally, these integrated features are averaged to obtain a single feature vector f_{text} representing the entire textual input.

$$f_{text} = average\ \left(h_1^{'},\ \ldots,\ h_n^{'}\right) \tag{13}$$

The image caption feature set $I_c,\ \ldots,\ I_w)$ is represented as the queries Q, while the textual features $(h_1,\ \ldots,\ h_n)$ are represented as the keys K and values V. The attention mechanism is used to compute the attention weights for each image caption feature for the textual features. The image caption representations combined with textual representations are thus outputted as below:

$$I_j^{'} = \sum_{i=0}^{n} a_i^j h_i \tag{14}$$

For each image caption feature $I_j^{'}$, the attention weights a_i^j are computed using the softmax function applied to the dot product h_1 and each textual featureh_i. These attention weights represent how much each textual feature contributes to the representation of the image caption feature.

$$a_i^j = softmax\ \left(I_j h_i^T\right) \tag{15}$$

The set of textual features is $(I_1^{'},\ \ldots,\ I_w^{'})$. Finally, these integrated features are averaged to obtain a single feature vector f_{text} representing the entire image caption input.

$$f_{ImageCaption} = average\ \left(I_1^{'},\ \ldots,\ I_w^{'}\right) \tag{16}$$

Overall, this approach takes the advantage of the complementary strengths of text and image modalities and effectively allows the model to integrate information from both modalities.

3.6. Alignment and normalization

The alignment and normalization component aligns and normalizes the text and image captions by calculating the mean along the channel dimension (the dimension of the feature tensor that captures different aspects of the data) for each channel separately (Eqs (17) and 18). The idea is that when text and image caption features are combined, different channels may represent diverse aspects of

the data. Therefore, the model capture the statistical properties specific to each modality or channel by calculating the mean separately. Calculating the mean value for normalization ensures the model aligns and normalizes the features within each modality independently. In this manner, modality-specific biases are removed, and the distributions of the features are aligned, enabling effective fusion. The variance along the channel dimension for each modality is then determined separately (Eqs (19) and 20) to complement the mean calculation and provide insights into the variability of the features. Performing this step allows the model to identify the variances unique to each modality and align the feature distributions more accurately during normalization. The model eliminates the modality-specific biases, scale differences, and feature variations with Eqs (21) and 22 and *f*. This process facilitates effective fusion by ensuring the features have consistent scales and distributions across modalities. Finally, text and image captions features are concatenated into a unified single representation Eq. (23). This concatenation preserves separate modality-specific information while ensuring their alignment. In addition, it ensures consistency in the features' scales and distributions in the subsequent modules. Given the text feature tensor (T) of shape ($batch_s,\ text_d$) and the image caption feature tensor (I) ($batch_s,\ caption_d$) the features are aligned and normalized as below:

$$T = \frac{1}{batch_s} \sum_{i=1}^{batch_s} T_i, : \tag{17}$$

$$I = \frac{1}{batch_s} \sum_{i=1}^{batch_s} I_i, : \tag{18}$$

$$\sigma_T^2 = \frac{1}{batch_s} \sum_{i=1}^{batch_s} (T_i, \ : - \ T)^2 \tag{19}$$

$$\sigma_I^2 = \frac{1}{batch_s} \sum_{i=1}^{batch_s} (I_i, \ : - \ I)^2 \tag{20}$$

$$\widehat{T}_i \ , : = \frac{T_i \ , : \ - T}{\sqrt{\sigma_T^2 + \ \epsilon}} \tag{21}$$

$$\widehat{I}_i \ , : = \frac{I_i \ , : \ - I}{\sqrt{\sigma_I^2 + \ \epsilon}} \tag{22}$$

$$Normalized_f = (\widehat{T}, \ \widehat{I}) \tag{23}$$

where σ_T^2 and σ_I^2 is the text and image caption features calculated along the channel dimension. They are tensor of shape $text_d$ and $caption_d$ representing variance value for each channel in the text and image caption features. i represents the instant index, and (:) denotes all elements along the respective dimension. ε is a small value added for numerical stability to avoid division by zero.

3.7. Optimized deep neural network

To this point, the model has obtained the text's attention-aggregated representation for the image caption, denoted as R_C, and the image caption's aggregated representation for the text denoted as R_T. These two representations capture the most relevant features of each modality. After aligning and normalizing the two features are concatenated to form a multimodal representation of a given post. The fused representation combines the strengths of both modalities and provides a more comprehensive representation of the post. An optimized deep neural network is designed to extract deep features from the fused features. It contains three fully connected layers with different parameters. The first fully connected layer contains 256 neurons in a hidden layer with a ReLU activation function, followed by a 0.5 dropout. The concept is to project high-dimensional fused input into a lower-dimensional hidden space, allowing the model to learn an abstracted nonlinear representation of captions and text modes. The second and third fully connected layers have 128 and 64 neurons in a hidden layer, respectively. These layers continue to project features into lower-dimensional space, allowing the model to discover discovery global, high-level patterns related to real and fake classes. Both layers contain ReLU activation function, a dropout, and a batch normalization layer. The batch normalization layer enables the model to learn independently from each ODNN layer. A ReLU activation function enhances the model's ability to learn nonlinear features and decision boundaries when classifying. It also prevents gradients from vanishing and exploding. The ODNN incorporates three dropout layers, after the ReLU activation function in the first fully connected layer and after batch normalization in the second and third layers. The aim is to prevent overfitting, improve generalization error, and reduce training time. The model incorporates Adam optimizer to generate adaptive estimates, which are computationally efficient. The model uses an initial learning rate of 0.005 to maximize computational efficiency. The fully connected layers enable the model to capture subtle correlations between captions and text semantics in the CAF module by combining simpler features learned in earlier layers to extract higher-level abstracted features. We implement a grid search to optimize the model's hyperparameters (Pedregosa et al., 2011). It is necessary to tune the Hyperparameters of a model before training it to control the learning process. Optimizing the hyperparameters enhances the performance. The grid search involves experimenting with every combination of values of the hyperparameters to find the most appropriate blend for training the model. While it effectively tests

Table 1
Optimization Hyperparameters.

Hyperparameter	Values
Neurons	128, 256, 512
Optimizer	Adam, Nadam, RMSprop
Batch size	10, 30, 60, 90
Epochs	10, 20, 30, 40, 60, 70

various values and quickly finds a near-optimal combination, grid search does not guarantee the best parameters. Consequently, grid search is time-consuming, especially when many hyperparameters are involved. We selected a few parameters to optimize to avoid this situation. These parameters include neurons, epochs, and batch sizes. We chose the parameters based on their impact on the model's performance. It was more efficient to find the best combination by selecting a small number of hyperparameters to optimize. As shown in Table 1, the number of neurons in the hidden layer, the optimized epochs, and batch sizes were considered.

- **Neurons:** A neuron is a unit that receives input from the preceding layer and produces output to the next layer. Neuron in each layer receives input from neurons in previous layer, calculate, and generate an output dispersed to neurons in the subsequent layers.
- **Optimizers:** Optimizers can be tuned to improve the model's efficiency during training. They adjusts the model's weights and biases during training to minimize the loss function. The loss function determines the disparity between the predicted output and the actual output, and the optimizer's goal is to reduce this difference. Different optimization methods, such as Adam, RMSprop, and Nadam, update weights and biases differently. We tuned three optimization techniques: Adams, Root Mean Square Propagation (RMSprop), and Nadam.
- **Batch size:** This parameter can significantly impact the training process and the model's performance. A large batch size can reduce the noise in the gradient estimation and result in a more stable convergence, but it may also require more memory, slowing the training process. In contrast, a small batch expedite the training process, but leads to a less stable convergence due to a noisy gradient approximation. We used batch sizes between 16 and 512. However, a batch size of 32 is a commonly used value.
- **Epochs:** This variable determines the number of successful authorizations in the training dataset. An optimal neural network performance requires the correct number of epochs. A small number of epochs causes under-fitting since the model fails to learn from the training data, whereas a large number of epochs results in overfitting because the model learns noise as well, resulting in poor performance on the test dataset. We used several epochs to update the model's weights and avoid overfitting. We incorporated a grid search to find the best value for epochs.

3.8. Classification

The softmax function converts the output of the fully connected layer into a probability distribution over the two classes of fake and real news. The predicted probability vector is represented by:

$$c = softmax(Wf_m + b) \tag{24}$$

where $c = (c_0,\ c_1)$ denote the predicted vector, and c_0 and c_1 denote the predicted ground truth, 1 being fake news and 0 being real news.

The binary cross-entropy loss function is given by:

$$l_c = [y\log c_0 + (1-y)\log c_1) \tag{25}$$

where $y\ \epsilon\ \{0,\ 1\}$ represents the news label.

4. Experiments

4.1. Datasets

We conducted an empirical analysis on four datasets, GossipCO, PolitiFact, Fakeddit and Pheme to validate the effectiveness of the proposed model. Table 2 displays the statistical information for the datasets:

- FakeNewsNet (Shu et al., 2020) is a popular dataset for fake news detection. It contains two datasets, PolitiFact and GossipCO. A PolitiFact dataset contains tweets related to posts published on the PolitiFact website. The PolitiFact website is a fact-checking website where journalist experts evaluate political claims as fake or real. This dataset consists of tweets mentioning claims and statements made by politicians, and other public figures. To create the FakeNewsNet collection, (Shu et al., 2020) used the headlines of these posts as queries to collect relevant tweets. On the other hand, GossipCO is based on tweets collected using the headlines of articles published and annotated on the GossipCO website.

Table 2
Statistics of our four datasets.

	Veracity	GossipCO			PolitiFact			Fakeddit			Pheme		
		Text	Caption	Image	Text	Caption	Image	Text	Caption	Image	Text	Caption	Image
Category	Real	1003	1003	1003	280	280	280	1399	1399	1399	1086	1086	1086
	Fake	1085	1085	1085	320	320	320	1650	1650	1650	1200	1200	1200
	Total	2088	2088	2088	600	600	600	3049	3049	3049	2286	2286	2286

- Fakeddit (Nakamura et al., 2020) is a dataset specifically designed for multimodal fake news analysis. This dataset contains one million samples, including both false and credible information. The dataset is categorized into six categories with binary ground truths for distinguishing between real and fake news for classification tasks. Furthermore, it offers more fine-grained classification options with three and six-class classifications, providing additional granularity in misinformation analysis.
- The Pheme (Zubiaga et al., 2016) dataset comprises Twitter threads that revolve around nine notable news events, such as the Siege of Sydney, Charlie Hebdo, the Ottawa shooting, and more. Each news thread consists of a tweet that serves as the news source. In our analysis, we leverage the original tweet from the news to perform the classification task.

In the four datasets, we prepare the training, the validation, and test subsets. We adopt Accuracy (A), Precision (P), Recall (R), and F1 score as our evaluation metrics. Table 2 shows the statistics of our datasets.

4.2. Implementation details

We utilize a PyTorch environment with TensorFlow, Keras, and transformers libraries to facilitate the creation, deployment, and utilization of the pre-trained Vit-GPT-2 model. We use a Universal Sentence Encoder, a transformer-based encoder to encode contextualized text and captions generated from the image in 512-dimensional space. Three fully connected layers containing 128 and 64 neurons are incorporated into the model with corresponding activation functions (ReLU) for effective feature extraction. A 0.5 dropout rate is set between the ODNN layers to mitigate overfitting, and we leverage Batch Normalization and Adam optimizer. The model incorporates a sigmoid activation function as a classification layer. The implementation is accelerated using CUDA 11.0 and executed on a GPU with 12 GB of memory. We processed data in batches of size 128 to maximize computational efficiency. The dataset is stratified into training (70 %), testing (20 %), and validation (10 %) sets, ensuring representative distribution across the three sets. We implement a learning rate scheduler to optimize the learning process, with a scheduler reducing the learning rate dynamically if the validation error increases for two consecutive epochs. This reduction continues until the model reaches a minimum learning rate, improving the overall effectiveness of the training performance. We implement an early stopping mechanism to prevent unnecessary training. If the model's performance or validation error does not improve after ten epochs, the training process is terminated early, saving computational resources. We use gradient clipping to stabilize the gradient values and control parameter updates during training. This technique prevents gradients from exploding, which can results in potential issues. We calculate loss using Cross Entropy Loss. This loss function standardizes the impact of different classes by assigning less attention to the input samples. It also aids in achieving balanced performance across class distributions.

4.3. Baselines

To validate the proposed model's effectiveness, we compared the experimental results of the four real datasets as follows:-

4.3.1. GossipCO and PolitiFact datasets

We adopt the following baselines for GossipCO and PolitiFact datasets

- **DEFD** (Obaid et al., 2022) is an approach that employs an ensemble of deep-learning models based on the same feature extractor. Each learner focuses on a distinct aspect of the input news using an attention mechanism and a loss function.
- **Silva et al.** (2021) a multimodal method for fake news detection in cross-domain news datasets that learns from two independent embedding spaces to capture domain-specific and cross-domain information.
- **TRANSFAKE** (Jing et al., 2021) integrates multimodal signals, such as text, images, and comments. A transformer-based architecture is used to extract features from text and image. Further, the model incorporates user comments to enhance the detection process.
- **DEFD-SSL** (Obaid et al., 2023) utilizes an ensemble model to exploit quality pseudo labels. Moreover, it mitigates bias towards the majority category by assessing the class distribution to solve the challenge of imbalanced classes.
- **FR-Detect** (Jarrahi & Safari, 2023) is a multimodal model for efficient multimodal fake news detection. It includes publisher-related features in the detection process.
- **SBERT**(Madhusudhan et al., 2020) utilizes BERT to extract context-rich textual features and ResNet to extract image features. A simple concatenation process combines the extracted features. As a second approach, the model incorporates visual attention, allowing it to focus specifically on the most relevant regions within the image.

- **SAFE** (Zhou et al., 2020) is a multimodal neural network that extracts textual and visual features separately for news representation. It models the relationship between the extracted features on different modalities to predict fake news.

4.3.2. Fakeddit and Pheme datasets

We adopted the following baselines for Fakeddit and Pheme datasets: -

- **MIN** (Zou et al., 2023b) integrates semantic-level image and text representations. The model also incorporates entities and external knowledge. It combines text, images, entities, and external knowledge using a three-level co-attention network.
- **Fakefind** (Sengan et al., 2023) is a method in which convolutional neural networks (CNNs) are combined with recurrent neural networks (RNNs) to effectively fuse multimodal information.
- **CMAC** (Zou et al., 2023a) method uses adversarial learning to align the latent feature distribution of text and image. Moreover, contrastive learning is employed to align the feature distribution in multimodal samples of the same category. The combination of adversarial and contrastive learning allows the model to obtain robust multimodal fusion representations.
- **FDMCE** (Shao et al., 2022) combines two single-modality classifiers and incorporates a similarity classifier to determine feature similarity across modalities. An integrity classifier is also used in the model to leverage integral multimodal information.
- **FakedBits** (Sharma et al., 2023) is a deep neural multi-modal network that utilizes EfficientNet-B0 and distilBERT to process visual and textual information. A feature embedding process is performed for each channel individually, followed by a fusion process at the final classification layer.
- **KMAGCN** (Qian et al., 2021) combines textual information, knowledge concepts, and visual -information within a unified framework to capture semantic representations. The model effectively learns features by representing posts as graphs and utilizing a knowledge-aware multimodal adaptive graph learning approach.
- **CCD** (Chen et al., 2023) is a technique that incorporates causal intervention to mitigate the impact of psycholinguistic bias, which may introduce misleading correlations between text features and new labels for multimodal fake news detection.

4.4. Results and analysis

Tables 3 and 4 present our method's overall performance. According to Table 3, CAF-ODNN outperforms baselines in accuracy, recall, and F1 values on the GossipCO dataset. In particular, it improves accuracy by 0.032 % points, recall by 0.508 % points, and F1 score by 0.399 % points compared to the previous best baselines. However, there is a slight decrease in precision by 0.062 % points. On the other hand, the model exhibits performance improvements of 0.056 % points, 0.215 % points, and 0.147 % points in accuracy, precision, and F1 score, respectively on the PolitiFact dataset. The SAFE method achieves an accuracy rate of 0.838 % and 0.874 % on the GossipCO and PolitiFact datasets, respectively. The approach by Silva et al. (2021), considered the best-performing baseline on the GossipCO dataset, achieves a performance of 0.842 % in accuracy. However, its performance on the PolitiFact dataset decreases by 0.01 % points. This observation suggests that the GossipCO dataset is larger than the PolitiFact dataset, providing the model access to more training instances compared to the PolitiFact dataset. DEFD performs better on the PolitiFact dataset in terms of accuracy. One reason is that DEFD may be more effective because it utilizes multimodal information, which improves the model's performance. Additionally, it employs a focal loss function and diverse learners to enhance accuracy on the minority class (i.e., fake news), enabling more effective detection of fake news. TRANSFAKE consistently performs better, achieving 0.831 % and 0.834 % accuracy on the GossipCO and PolitiFact datasets, respectively, compared to other baselines. However, it shows a slight decrease in accuracy by 0.032 % and 0.055 % points on the GossipCO and PolitiFact datasets, respectively, compared to CAF-ODNN. This performance demonstrates that incorporating users' comments and sentiments enriches multimodal features, and utilizing a transformer for fusion improves the performance of the fake news detection model. DEFD-SSL outperforms FR-Detect on the PolitiFact dataset by 0.039 % points. A quality pseudo-label ensemble model underlies the impressive performance of DEFD-SSL. An advantage of the ensemble method is that it considers the imbalanced nature of the data by estimating the distribution of classes, thereby eliminating bias in favor of the majority category. SBERT also performs better than TRANSFAKE and DEFD-SSL, demonstrating its ability to capture the relationships between text and images for a more comprehensive understanding and analysis.

FDMCE's accuracy decreases by 0.08 % points on the Fakeddit dataset and by 0.023 % points on the Pheme dataset compared to CCD in Table 4. It could be inferred from the model's observation that CCD uses causal intervention and counterfactual reasoning to mitigate image bias and achieve superior performance compared to FDMCE. MIN model outperforms all compared baselines, achieving

Table 3
Comparison performance on GossipCO and PolitiFact datasets.

Dataset	Measure	DEFD	(Silva et al., 2021)	TRANSFAKE	DEFD-SSL	FR-Detect	SBERT	SAFE	CAF-ODNN
GossipCO	Acc. (%)	0.841	0.848	0.831	0.831	0.840	0.837	0.838	0.863
	Prec. (%)	0.912	0.822	0.773	0.592	0.767	0.855	0.877	0.870
	Recall (%)	0.601	0.797	0.826	0.442	0.591	0.945	0.937	0.950
	F1score (%)	0.725	0.808	0.851	0.506	0.668	0.898	0.890	0.905
	Acc. (%)	0.855	0.838	0.834	0.872	0.833	0.855	0.874	0.889
PolitiFact	Prec. (%)	0.705	0.836	0.801	0.730	0.804	0.871	0.889	0.920
	Recall (%)	0.827	0.828	0.862	0.826	0.872	0.926	0.903	0.841
	F1 score (%)	0.761	0.833	0.820	0.775	0.837	0.897	0.896	0.908

Table 4
Comparison performance on Fakeddit and Pheme datasets. The (–) indicates the results in the original paper were not provided, while (*) denotes better performance.

Dataset	Measure	MIN	Fakefind	CMAC	FDMCE	FakedBits	KMAGCN	CCD	CAF-ODNN
Fakeddit	Acc. (%)	0.893	0.848	0.867	0.804	0.888	0.829	0.884	0.900
	Prec. (%)	–	0.841	0.886	0.838	0.852	0.818	0.821	0.961
	Recall (%)	–	0.851	0.896	0.749	0.871	0.808	0.781	0.901
	F1score (%)	0.848	0.846	0.884	0.791	0.860	0.812	0.808	0.930
Pheme	Acc. (%)	0.828	0.900*	0.874	0.836	0.863	0.867	0.859	0.879
	Prec. (%)	–	0.901	0.797	0.846	0.850	0.830	0.764	0.884
	Recall (%)	–	0.901	0.810	0.853	0.921	0.775	0.689	0.948
	F1 score (%)	0.805	0.901	0.780	0.850	0.890	0.800	0.724	0.915

0.893 % accuracy on the Fakeddit dataset. MIN appears to leverage entities and external knowledge to enhance the integration of image representations with their semantics, providing additional semantic information that aids in validating post-rationality. However, the accuracy of MIN decreases by 0.065 % points on the Pheme dataset, indicating potential dataset characteristics. The MIN approach, for example, might have been specifically trained to utilize entities and external knowledge that are more informative in the Fakeddit dataset. Therefore, when applied to the Pheme dataset, which may have distinct characteristics, the model struggles to effectively utilize the same entities and external knowledge, resulting in low performance. On both the Fakeddit and Pheme datasets, FakeBits exhibits consistent performance. This observation suggests that leveraging EfficientNet-B0, distilBERT, and embedding text and visual features in separate channels before final classification improves the performance of models. The CMAC and KMAGCN models also perform comparatively better on the Pheme dataset. This observation suggests that the models exploits counterfactual reasoning and attention to fuse specific features, jointly model textual information, knowledge concepts, and visual information to improve detection performance. Table 4 shows that the Fakefind performs better on the Pheme dataset by 0.021 % points than the proposed model. However, the proposed model outperforms Fakefind in Recall and F1 scores by 0.047 % and 0.014 % points, respectively. A possible explanation could be that the higher scores on Recall and F1 suggest that the proposed model can identify positive instances correctly (correctly identifying fake news). Therefore, it can capture more true positive fake news instances, even though there might be negligible difference in accuracy compared to Fakefind.

In terms of datasets, the proposed model exhibits superior performance on the GossipCO, PolitiFact, Fakeddit, and Pheme datasets, implying its scalability and generalizability. In particular, the model achieves high level of accuracy of 0.900 % on the Fakeddit dataset, 0.889 % on PolitiFact, 0.879 % on Pheme, and 0.863 % on GossipCO. The variation in performance across these datasets can be attributed to the dataset characteristics, model architecture, and the hyperparameters used. During the model's training process, these factors interact, influencing the model's learning ability, as reflected in the changes observed over epochs. In Fig 3, we visualize the convergence points for each dataset. The GossipCO dataset reaches convergence at epoch 5.5, PolitiFact at epoch 9, Fakeddit at epoch 8, and Pheme at epoch 6, indicating the model optimally balances the training and validation losses. A small gap between the training and validation loss reveals that the model fits well with the training data while generalizing well to unseen validation data. This observation demonstrates that the model has effectively learned the underlying patterns in the training data without overfitting or underfitting, demonstrating its capability to capture and generalize from the dataset's features.

CAF-ODNN outperforms the baselines models due to its complementary attention fusion of image captions and text. The performance of CAF-ODNN can be attributed to three key factors: first, the CAF's complementary attention fusion technique captures subtle cross-modal relationships allowing the model to learn fine-grained correlations in a shared fusion space. This approach captures dependencies and fine-grained cross-modal interactions, ensuring accurate and semantically meaningful fusion while eliminating unrelated noisy features. Second, the alignment module calibrates fused features based on channel statistics, ensuring the semantics are preserved across modalities during the interaction mitigating feature variations and distributions. Lastly, through optimization, model hyperparameters are tuned for optimal performance. The optimization technique allows for the efficient feature extraction and enhances utilization of text and image caption information, ultimately resulting in improved accuracy.

4.5. Error analysis

We implemented a Receiver Operating Characteristics Curve (ROC curve) to analyze errors in the model's classification. The ROC curve represents the performance of a model in binary classification. The ROC curve captures the trade-off between the False Positive Rate (FPR) and the True Positive Rate (TPR) at various classification points (Suratkar et al., 2020). The ROC curves are shown in Fig 4 (a - d). In estimating the area under the ROC curve, the Area Under the Curve (AUC) quantifies the general effectiveness of the model. An ideal classification model has an AUC of 1, while random guesses yield an AUC of 0.5. CAF-ODNN performs well on all four datasets, with AUC values of 0.79 %, 0.94 %, 0.96 %, and 0.81 % for the GossipCO, PolitiFact, Fakeddit, and Pheme datasets respectively. This performance suggests a good balance between the TPR and FPR, establishing excellent overall classification performance.

4.6. Ablation analysis

To assess the performance of CAF-ODNN fusion and optimization, we conducted ablation experiments on the GossipCO, PolitiFact, Fakeddit, and Pheme datasets. The ablation sub-experiment results are shown in Fig 5 (a - d) and Fig 6 (a - d). In Fig 5 (a - d), CAF-

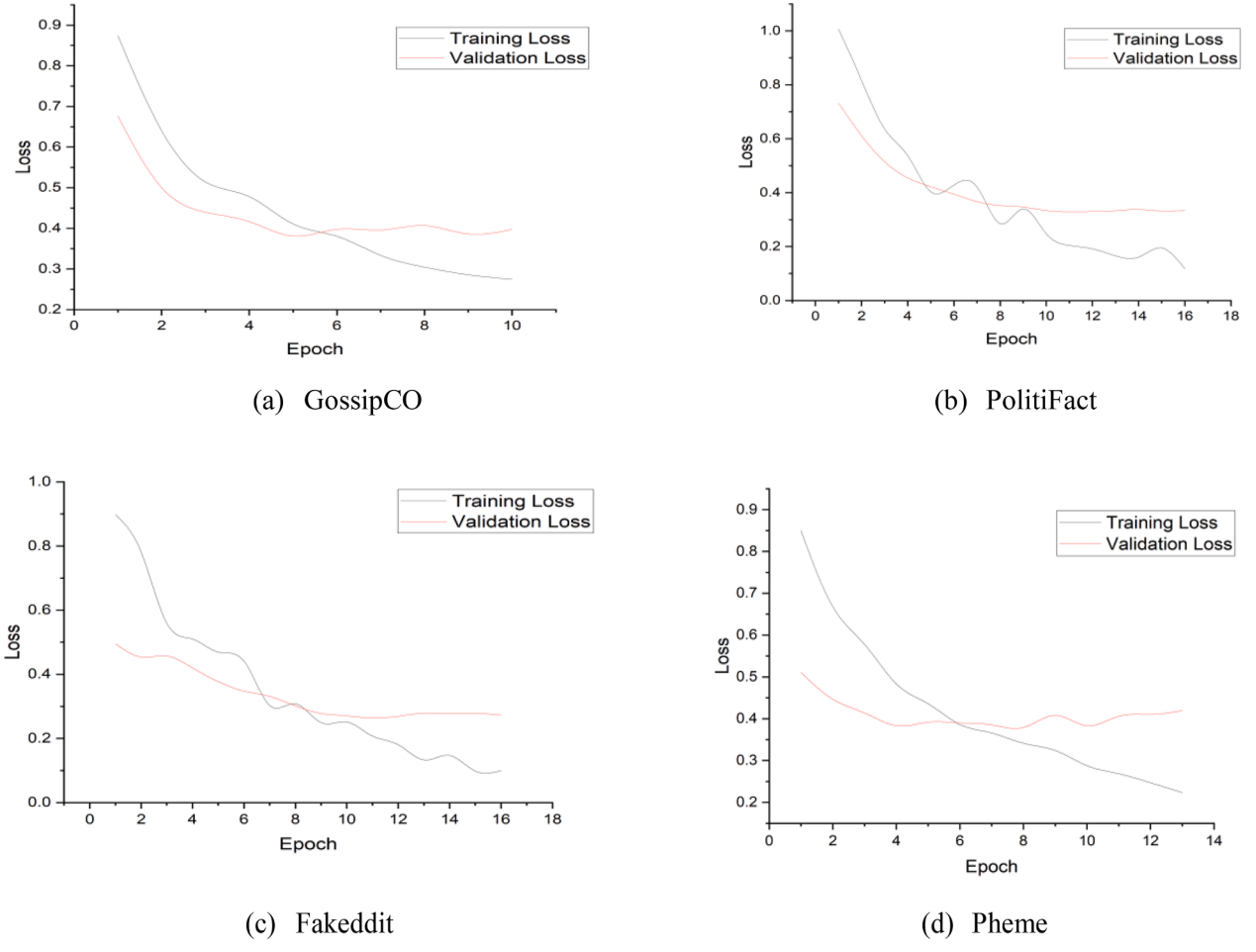


Fig. 3. (a – d). Training and validation loss on GossipCO, PolitiFact, Fakeddit and Pheme dataset.

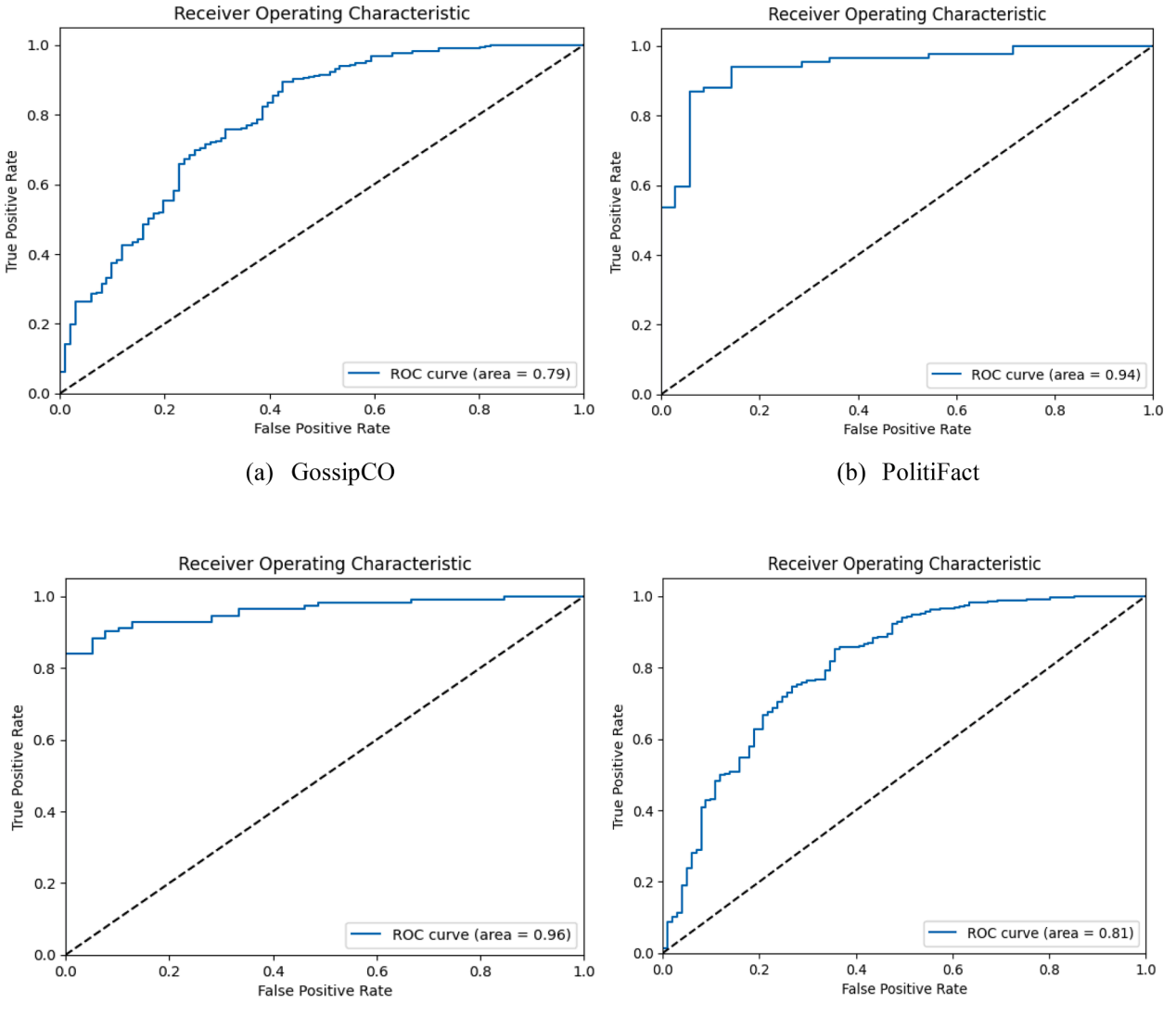


Fig. 4. (a- d). ROC curve on GossipCO, PolitiFact, Fakeddit and Pheme datasets.

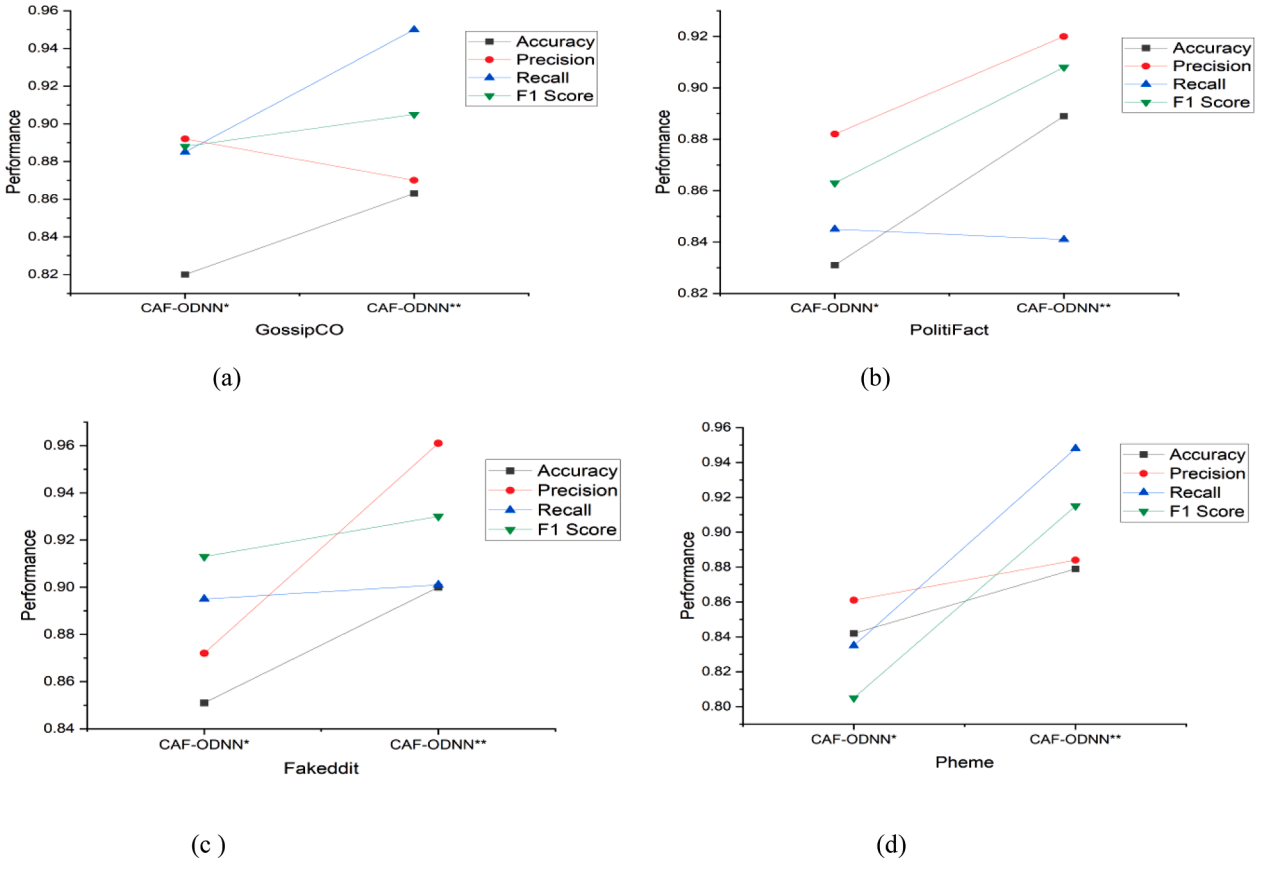


Fig. 5. (a - d). Ablation analysis on GossipCO and PolitiFact datasets. CAF-ODNN* indicates the CAF is not incorporated while CAF-ODNN** incorporates the CAF.

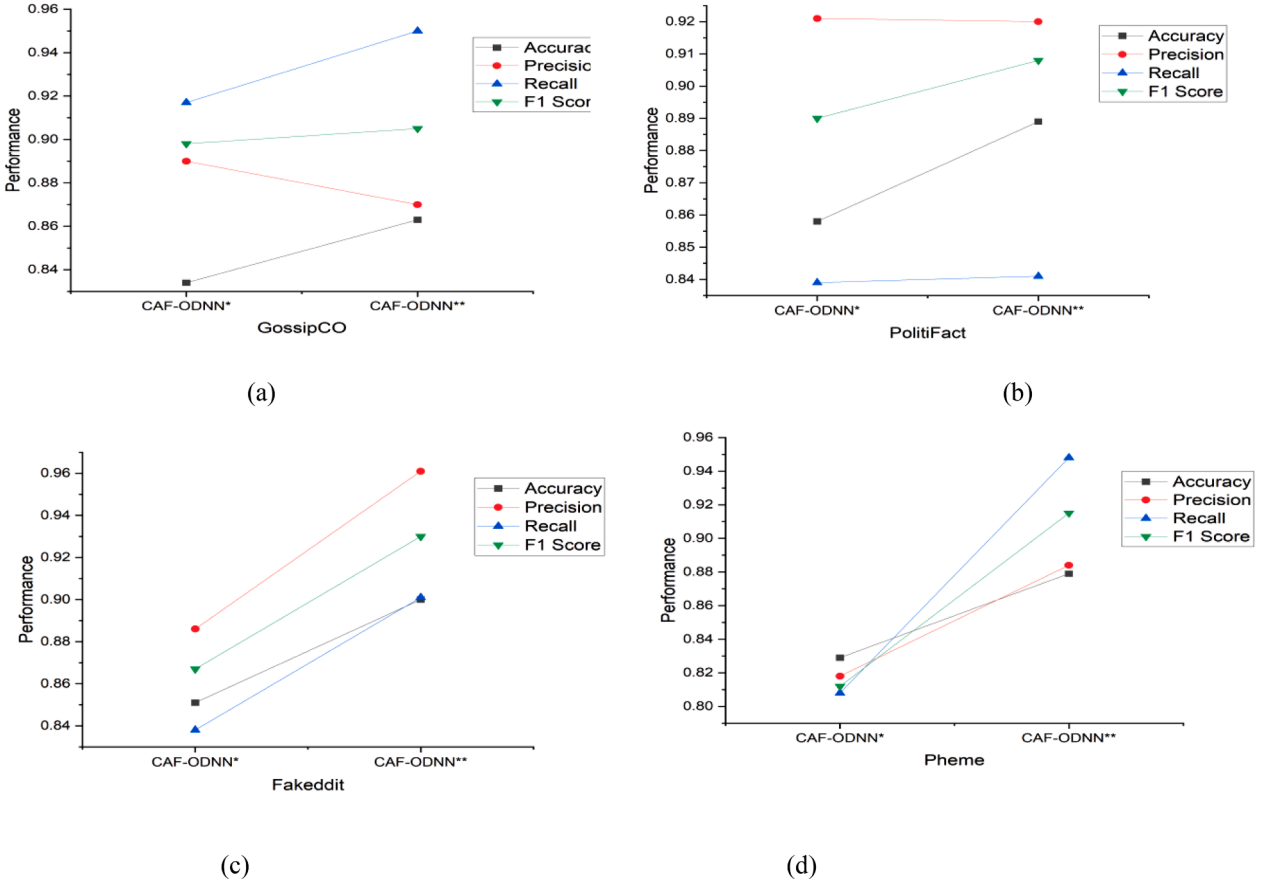
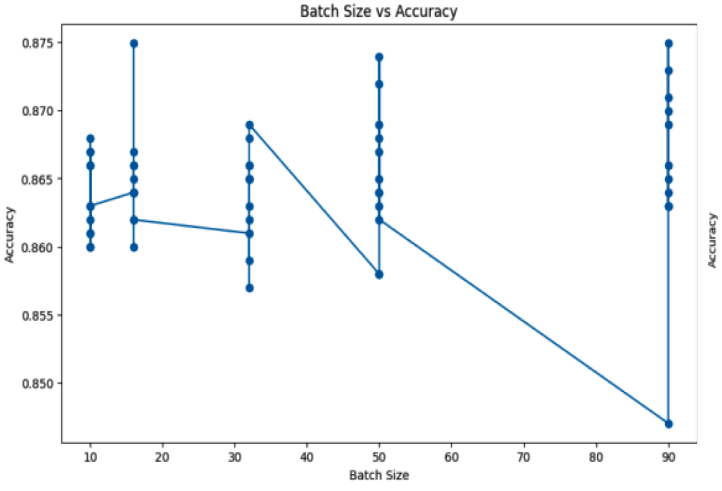
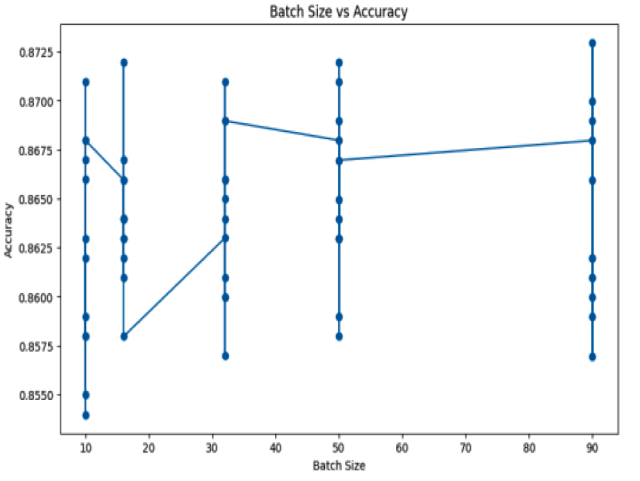


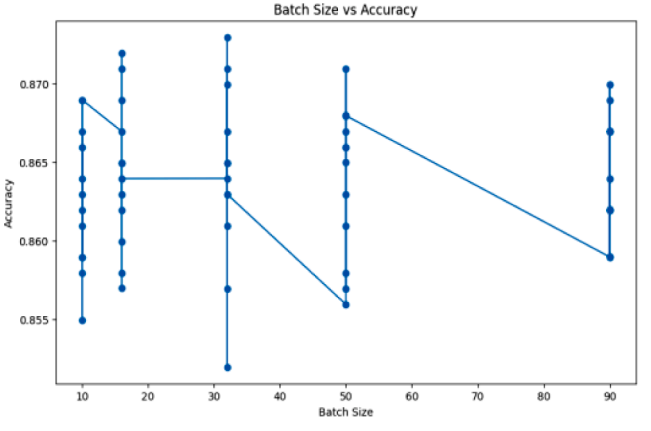
Fig. 6. (a – d) Ablation analysis on GossipCO and PolitiFact datasets. CAF-ODNN* indicates the ODNN is not incorporated while CAF-ODNN** incorporates the ODNN optimization.



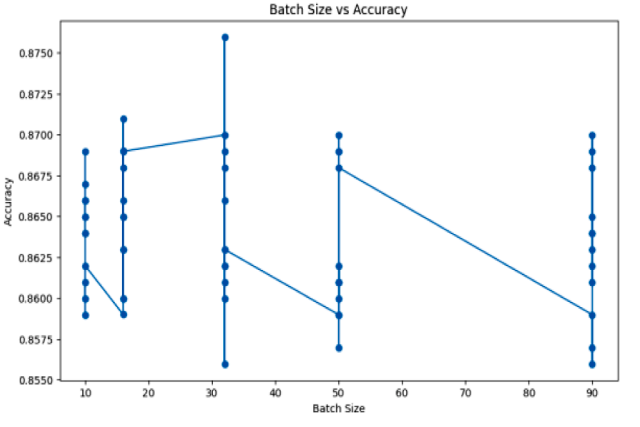
(a) Batch Size and Accuracy on GossipCO



(b) Batch Size and Accuracy on PolitiFact

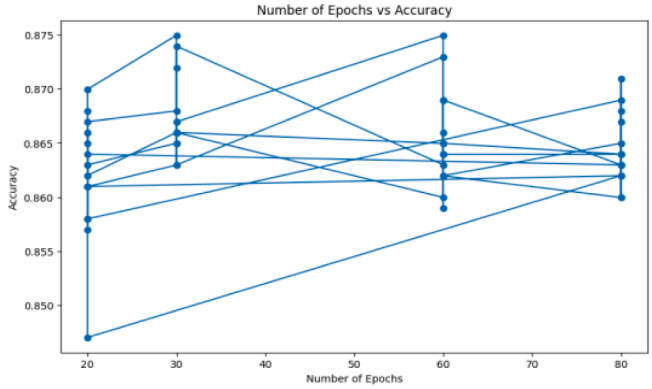


(c) Batch Size and Accuracy on Fakeddit

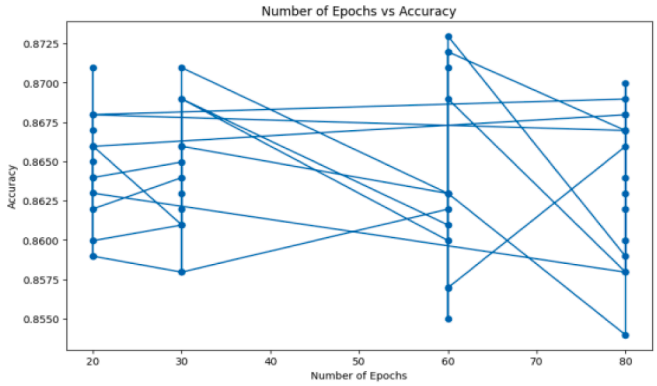


(d) Batch Size and Accuracy on Pheme

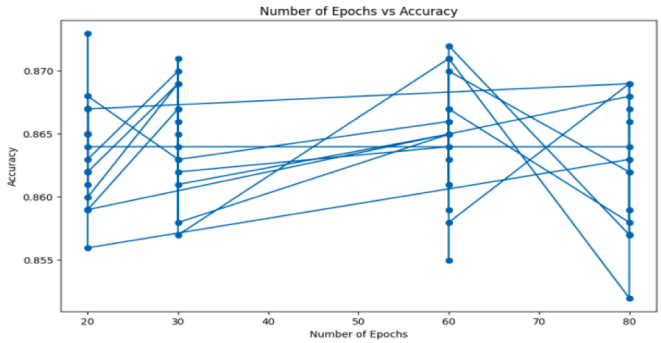
Fig. 7. (a - d) The graph consists of individual data points represented as dots connected by lines. Each dot represents a specific batch size and its corresponding accuracy score. The lines help visualize the trend in the data points, and the circular markers placed at each data point highlight the exact values of batch sizes and accuracy scores.



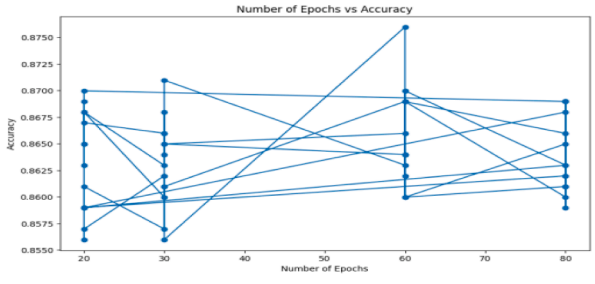
(a) Epochs and Accuracy on GossipCO



(b) Epochs and Accuracy on PolitiFact



(c) Epochs and Accuracy on Fakeddit



(d) Epochs and Accuracy on Pheme

(caption on next page)

Fig. 8. (a - d) The graph consists of individual data points represented as dots connected by lines. Each dot represents a specific batch size and its corresponding accuracy score. The lines help visualize the trend in the data points, and the circular markers placed at each data point highlight the exact values of batch sizes and accuracy scores.

ODNN* indicates the exclusion of CAF fusion, while CAF-ODNN** denotes the inclusion of CAF fusion. In Fig 6 (a - d), CAF-ODNN* excludes the ODNN optimization, while CAF-ODNN** includes ODNN optimization.

In Fig 5, the ablation results show that the performance of CAF-ODNN decreases without incorporating complementary fusion. For instance, the accuracy, recall, and F1 scores for the GossipCO decreased by 0.043 %, 0.065 %, and 0.017 % points respectively in Fig 5 (a). The model's performance falls by 0.058 %, 0.038 %, and 0.045 % points in accuracy, precision, and F1 scores, respectively on PolitiFact (Fig 5(b). Consequently, the model's performance decreases by 0.049 %, 0.089 %, and 0.017% points in accuracy, precision, and F1 scores, respectively for the Fakeddit dataset in Fig 5(c). Fig 5(d) shows the decrease in accuracy, precision, and F1 scores for Pheme dataset by 0.037%, 0.023%, and 0.11% points, respectively. We can attribute this improved performance to three factors. Firstly, when the image and text posts are represented in a joint high-dimensional space, the model compares the features and extract visual and textual features from the data, mapping them to a joint embedding space. Secondly, the image caption attention and text attention mechanisms highlight relevant features in the image caption and text, respectively. This observation implies the model focuses on the most crucial parts of the image captions and text, disregarding irrelevant information. The attention mechanisms detect subtle complexities using image captions as the "post text." Image captions serve as descriptions of the content of an image and represent visual features in a high-dimensional space. The application of image captions as the post text enables the model to capture the most salient visual characteristics of the image in a semantic space that exhibits a significant correlation with textual features, thereby enhancing performance. Thirdly, the alignment module performs separate alignment and normalization of the text and image caption modalities, thereby improving the fusion process and addressing variances and feature distributions. Thus, the benefit of the process is eliminating noisy features, resulting in improved accuracy and semantic meaningfulness of the fusion between image captions and text.

In Fig 6 (a - d), the model's performance improves through optimization. In particular, the model improves accuracy and recall by 0.029 % and 0.033 % points, respectively, on the GossipCO dataset. For the PolitiFact dataset, the accuracy improves by 0.031 % points, recall by 0.002 % points, and F1 score by 0.0018 % points. Similarly, on the Fakeddit dataset, the accuracy improves by 0.049 % points, recall by 0.075 % points, and F1 score by 0.063 % points. The accuracy, recall, and F1 scores of the Pheme dataset are all improved by 0.05 % points, 0.14 % points, and 0.103 % points, respectively. We attribute this improved performance to two factors. Firstly, using an optimized deep neural network enhances the efficiency of feature extraction in image captions and text by learning progressively fine-grained features through compositional learning using fully connected layers. Secondly, the systematic exploration of different combinations of hyperparameters enables the model to find the optimal configuration that maximizes performance. This approach leads to a more accurate and efficient model since it determines the best settings that enhance overall performance.

4.7. Hyperparameter analysis

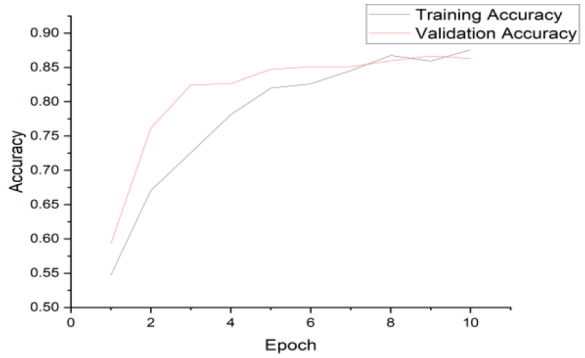
We analyzed the hyperparameters, and due to space constraints, we focused on two specific hyperparameters: batch size and the number of epochs.

4.7.1. Batch size

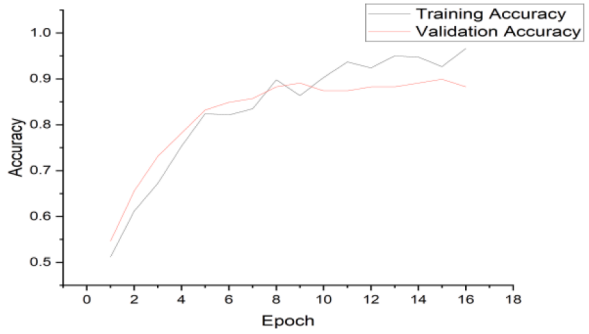
We present the model's performance with different batches in Fig 7 (a–d). In Fig 7(a), we observe that CAF-ODNN convergence occurs after several iterations with a 15-batch size for the GossipCO dataset. The model reaches its optimum point at 90 epochs for the PolitiFact dataset. Fakeddit and Pheme datasets converge with a batch size of 32. We draw some insights based on Fig 7 (a - d). The convergence at a smaller batch size suggests that the model effectively learns from smaller subsets of data. This phenomenon indicates that the dataset is more concentrated with relevant patterns, as seen in the GossipCO dataset. A larger batch size, as observed in the PolitiFact dataset, helps the model generalize better by providing more diverse samples during each update step. Hence, a larger batch size benefits the model by incorporating more significant sampling instances to learn underlying patterns. A moderate batch size, as observed in the Fakeddit and Pheme datasets, implies that the dataset exhibits characteristics that allow the model to learn efficiently. These datasets likely contain a balanced distribution of patterns, enabling the model to capture quality local and global patterns.

4.7.2. Epochs

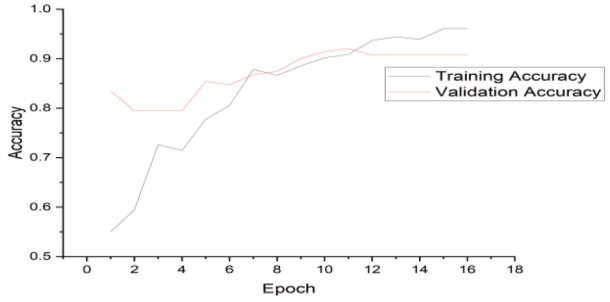
Fig 8 (a - d) shows the relationship between epochs and accuracy on four datasets, demonstrating the CAF-ODNN model's generalizing and learning ability. We conducted experiments using varying numbers of epochs, as shown in Table 1 to determine the optimal value for the CAF-ODNN model. In Fig 8(a), we observe that CAF-ODNN performs significantly better after 30 epochs on GossipCO dataset. We plotted the training and validation accuracy against the number of epochs to visualize the learning curve and identify the optimal number of epochs in Fig 9 (a - d). In Fig 9(a), the validation accuracy reaches a plateau and starts to decline at around 9 epochs, while the training accuracy continues to increase. This phenomenon suggests that the model has already converged, but it may benefit from additional epochs. In Fig 8(b), the model achieves optimal performance and improved accuracy after 60 epochs for the PolitiFact dataset. Correspondingly, in Fig 9(b), the validation and training accuracy curves converge after approximately 10 epochs, exhibiting a zigzag trend in both curves. This observation implies the model is still learning, and additional training may be required to reach a steady state. Similarly, Fig 8(c), shows improvement in CAF-ODNN performance after 20 epochs on Fakeddit.



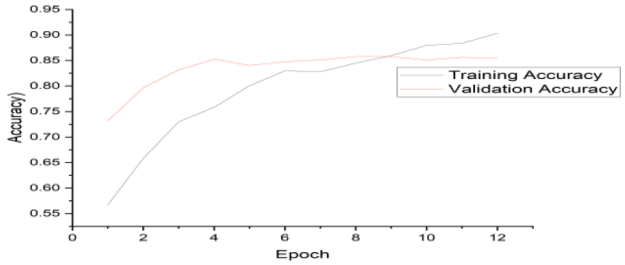
(a) Validation accuracy Vs epochs on GossipCO



(b) Validation accuracy vs epochs on PolitiFact



(c) Validation accuracy Vs epochs on Fakeddit



(d) Validation accuracy Vs epochs on Pheme

(caption on next page)

Fig. 9. (a – d) Validation accuracy Vs epochs on GossipCO, PolitiFact, Fakeddit and Pheme datasets.

Correspondingly, in Fig 9(c), the validation and training accuracy curves converge at around 11 epochs, with validation accuracy plateauing. However, in Fig 8(d), the performance on Pheme's dataset converges after 60 epochs. Correspondingly, in Fig 9(d), the validation accuracy converges after 9 epochs, and then plateaus. These observations suggest that each dataset has varying learning patterns and complexity. For example, initially, the model learns the underlying patterns in the GossipCO dataset, increasing its training and validation accuracy. The training accuracy increases as the model learns to fit the training samples. The behavior observed in both figures for the PolitiFact dataset suggests a complex learning pattern. It appears that the model requires several epochs to achieve optimal performance, indicating that the dataset could have intricate characteristics that necessitate a longer learning process for effective capture. A zigzag trend in the accuracy curves suggests that the optimization landscape of this dataset is not smooth and may have multiple local optima. The model navigates these optima during training, resulting in training and validation accuracy fluctuations. The convergence of the curves after some epoch implies that the model has learned the essential patterns and reached a stable point where further training does not significantly improve performance. The behavior observed in the Fakeddit dataset indicates that the model has learned within a relatively small epoch. The initial improvement in performance suggests that the model quickly captures the relevant patterns in the dataset and adjusts its parameters accordingly. For the Pheme dataset, a longer training process is required for the model to reach its optimal performance. The convergence after 60 epochs suggests that the dataset contains complex patterns that take more time to learn effectively. The plateauing of the validation accuracy indicates that the model has captured the essential information in the dataset, and further training will not yield substantial improvements.

4.8. Qualitative analysis

We provide some instances to illustrate the success and limitations of the proposed model. Specifically, we analyze the fusion aspect of the model. To achieve this goal, we randomly select two pieces of news (one fake and one real) from the PolitiFact test set and qualitatively evaluate the model's decision regarding the categorization into their respective labels. Fig 10 (a**)** and 10 (b) display the results of the quantitative analysis.

The model in Fig 10 (a) correctly classify real news into its respective class label. A closer analysis of the image and corresponding caption suggest that the successful generation of caption words "jetliner … a tarmac next … city" and post-text words "airborne … buildings" may have given the semantics and context complementing each sentence pairs; thus, the correct label prediction. In Fig 10 (b), the caption words "A man in a suit and tie is looking at his cell phone" and the post text "Barack Obama and John Lewis emotional hug" also seems to have given the model the semantic and context to classify the news label. While looking at the image, we can observe two men hugging each other emotionally, and neither is looking at the cell phone. Therefore, the complementary attention fusion captures these subtle differences since it captures the dependencies and interactions between the two modes, and ensures semantically meaningful contexts are incorporated.

Post text: airborne in a field near the buildings

Fig. 10a. Success of fusion. A large jetliner sitting on top of a tarmac next to a city Label: Real [{'score': 0.8758249394330129}] Success: CAFODNN /R.

Post text: Barack Obama and John Lewis emotional hug

Fig. 10b. Success of fusion. A man in a suit and tie is looking at his cell phone Label: Fake [{'score': 0.04367451638718302}] Success: CAF-ODNN /F.

4.9. Discussion

Tables 3 and 4 and Fig 5 (a-d) and Fig 6 (a-d) demonstrates the generalizability of the proposed method on four fake news datasets. This observation highlights the effectiveness of complementary attention fusion technique, which utilizes attention, feature alignment and optimization to learn a shared representation, thus capturing fine-grained correlations and interactions between modalities. Thus, the model is able to eliminate irrelevant and noisy features encouraging semantically meaningful fusion. Additionally, multimodal methods outperform single-modality approaches (La et al., 2022). The reason could be that the intricate relationship between images and text in news pieces enhances classification performance when efficiently combined (Wang et al., 2022). A more comprehensive understanding of the news pieces is obtained when text and visual information are exploited in unison. Therefore, multimodal evidence provides diverse perspectives that complement each other, resulting in superior performance. The advantage offered by the proposed method is its usefulness in multimodal learning tasks, where both image and text information are available. It is crucial to capture the relationships between these modalities in a shared space. The technique eliminates irrelevant and noisy features, addressing a common issue in multimodal datasets containing such information. By removing these features, the model can focus on the most relevant information, avoiding potential misinterpretation. For example, the proposed method improves accuracy on the GossipCO, PolitiFact, and Fakeddit datasets by 0.032 %, 0.056 %, and 0.096 % points, respectively (Tables 3 and 4). Furthermore, the CAF fusion mechanism enhances the model by selectively attending to informative features during the combination, alignment, and learning of sentence pairs (Fig 5a-d) (Kumari & Ekbal, 2021). We also attribute this performance to the proposed model's effective extraction of deep features using deep neural network. Maximizing hyperparameter search via grid algorithm enhances the efficiency and accuracy of feature extraction process. Thus, the model captures relevant information from both modalities with efficiency, resulting in improved performance. Through systematic exploration of diverse hyperparameter combinations, the approach discovers the optimal configuration that maximizes performance, yielding more accurate and efficient results.

4.10. Theoretical and practical implications

Fake news has a detrimental impact that extends far and wide, posing a threat to the safety of individuals and the vitality of democratic societies. Hence, it is essential to identify and combat fake news to maintain the integrity of public discourse and foster an informed citizenry. The proposed approach offers a theoretical framework for effectively incorporating multimodal information in fake news detection systems. Image captions and text data enhances the model's ability to capture broad and relevant information, improving the accuracy of the detection algorithm. An image captioning method is essential in fake news detection since it allows image features to be represented in a shared space, compatible with the text post. This approach enables the model to capture the most subtle yet salient visual features within a semantic space exhibiting a high correlation with textual features. The technique ultimately increases the model's capability to detect fake news containing both text and image content. Moreover, multimodal datasets manifest numerous irrelevant or noisy features that do not contribute to the underlying relationships between image and text data (Jin et al., 2022). Including these irrelevant or noisy features affects the model's performance by introducing noise into the data. The proposed

technique overcomes this problem by employing complementary attention mechanisms that allows the algorithm to concentrate on relevant patterns selectively from both modalities. In addition, the multimodal fused features are aligned and normalized via channel statistics to enhance the fusion process. As a result, irrelevant and noisy features are minimized, preserving semantic meaningfulness. Furthermore, an optimized deep neural network with grid search offers a more effective means of refined feature extraction and parameter optimization enhancing the performance of deep learning models for multimodal data analysis. The optimal alignment of hyperparameters in the proposed approach significantly enhances the accuracy and efficiency of the model.

We have demonstrated the practical implementation of a deep learning multimodal fake news detection framework in this study. A few studies (La et al., 2022), (Meel & Vishwakarma, 2021), (Qi et al., 2021) exist that utilize image captioning information, thus, this study provides a method to mitigate the spread and detection of fake news, particularly multimodal fake news, which has become a global concern. The presented approach can be useful to fact-checkers, organizations, and researchers in improving news authenticity. The advantage of the proposed method is that it can be seamlessly integrated as a plug- in existing single-modal methods and utilized by fact-checkers to enhance detection capabilities.

5. Conclusion and future work

In this study, we proposed a Complementary Attention Fusion with an Optimized Deep Neural Network (CAF-ODNN) for multimodal fake news detection. The approach utilized image captions to represent images in a shared semantic space with text to capture subtle relationships. A complementary attention method based on scaled dot product was applied to fuse captions and text bi-directionally to capture fine-grained cross-modal interactions between the two modalities. To streamline and make the fusion robust, a dedicated alignment and normalization technique was introduced to calibrate fused features based on channel statistics of each channel dimension. This technique enhanced semantic significance and cross-modal interaction, mitigating feature variations and noisy features. ODNN was also implemented for refined feature extraction and model optimization. The ODNN learned higher-level abstracted features progressively from caption and text captured in CAF through compositional learning using three fully connected layers. A grid search was incorporated to optimize the hyperparameters systematically, and identify configurations maximizing feature extraction quality and overall accuracy. The proposed method outperformed comparable methods on standard metrics in four real-world datasets.

The proposed model can be exploited to enhance single-modality models since it uses multimodal features to detect fake news. Single-modality models rely on a single feature type, such as text or image, which can limit detection effectiveness due to individual modality biases and limitations. In the future, we intend to develop a more robust approach for multimodal fusion that effectively integrates image captions with other modalities. Additionally, we will consider incorporating user engagement features into the detection system for improved effectiveness. This approach will enable identify patterns of misinformation and disinformation that would be challenging to detect using any single feature alone.

CRediT authorship contribution statement

Alex Munyole Luvembe: Conceptualization, Methodology, Software. **Weimin Li:** Data curation, Writing – original draft, Writing – review & editing, Supervision. **Shaohau Li:** Visualization, Investigation. **Fangfang Liu:** Supervision. **Xing Wu:** Writing – review & editing.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by the National Key R&D Program of China (No. 2022YFC3302600).

References

Alharbi, R., Jeter, T. R., & B, M. T. T (2023). Detection of fake news through heterogeneous graph interactions (Eds.). In D. Mohaisen, & T. Wies (Eds.), *NETYS 2023, LNCS 14067* (pp. 3–16). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-37765-5, 2023. (pp. 3–16).

Al-Malla, M. A., Jafar, A., & Ghneim, N. (2022). Image captioning model using attention and object features to mimic human image understanding. *Journal of Big Data, 9*(1). https://doi.org/10.1186/s40537-022-00571-w

Al Obaid, A., Khotanlou, H., Mansoorizadeh, M., & Zabihzadeh, D (2022). Multimodal fake-news recognition using ensemble of deep learners. *Entropy, 24*(9), 1–16. https://doi.org/10.3390/e24091242

Armin, K., Djordje, S., & Matthias, Z. (2021). Multimodal detection of information disorder from social media. In *Proceedings of the international workshop on content-based multimedia indexing, 2021-june*. https://doi.org/10.1109/CBMI50038.2021.9461898

Bagade, A., Pale, A., Sheth, S., Agarwal, M., Chakrabarti, S., Chebrolu, K., & Sudarshan, S. (2020). The Kauwa-Kaate fake news detection system: DemO. In *Proceedings of the ACM international conference proceeding series* (pp. 302–306). https://doi.org/10.1145/3371158.3371402

Cao, J., Qi, P., Sheng, Q., Yang, T., Guo, J., & Li, J. (2020). Exploring the role of visual content in fake news detection. In Disinformation, misinformation, and fake news in social media (pp. 141–161). https://doi.org/10.1007/978-3-030-42699-6_8.

St. Cer, D., Yang, Y., Kong, S.yi, Hua, N., Limtiaco, N., John, R., Constant, N., Guajardo-Céspedes, M., Yuan, S., Tar, C., Sung, Y. H., Strope, B., & Kurzweil, R (2018). Universal sentence encoder for English. In *Proceedings of the EMNLP 2018 - conference on empirical methods in natural language processing: system demonstrations, proceedings* (pp. 169–174). https://doi.org/10.18653/v1/d18-2029

Chen, X., Fang, H., Lin, T., Vedantam, R., Zitnick, C.L., Gupta, S., & Doll, P. (2015). Microsoft COCO Captions : Data Collection and Evaluation Server. 1–7.
Choras, M., Gielczyk, A., Demestichas, Konstantinos Puchalski, D., & Kozik, R. (2018). Pattern Recognition Solutions for Fake News Detection. In K. Saeed and W. Homenda (Eds.): CISIM 2018, LNCS 11127, pp. 130–139, 2018. (Vol. 1, pp. 486–498). https://doi.org/10.1007/978-3-319-99954-8.
Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An Image is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. ICLR 2021. http://arxiv.org/abs/2010.11929.
Giachanou, A., Ghanem, B., Ríssola, E. A., Rosso, P., Crestani, F., & Oberski, D. (2022). The impact of psycholinguistic patterns in discriminating between fake news spreaders and fact checkers. *Data and Knowledge Engineering, 138*(March 2021), Article 101960. https://doi.org/10.1016/j.datak.2021.101960
Gong, S., Sinnott, R.O., Qi, J., & Paris, C. (2023). Fake News Detection Through Temporally Evolving User Interactions. In H. Kashima et al. (Eds.): PAKDD 2023, lnai 13938, pp. 137–148, 2023. (Vol. 1, pp. 137–148). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-33383-5_11.
Jarrahi, A., & Safari, L. (2023). Evaluating the effectiveness of publishers' features in fake news detection on social media. *Multimedia Tools and Applications, 82*(2), 2913–2939. https://doi.org/10.1007/s11042-022-12668-8
Jin, Y., Wang, X., Yang, R., Sun, Y., Wang, W., Liao, H., & Xie, X. (2022). Towards fine-grained reasoning for fake news detection. *Proceedings of the AAAI Conference on Artificial Intelligence, 36*(5), 5746–5754. https://doi.org/10.1609/aaai.v36i5.20517
Jin, Z., Cao, J., Guo, H., Zhang, Y., & Luo, J. (2017). Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *MM 2017 - proceedings of the 2017 ACM multimedia conference* (pp. 795–816). https://doi.org/10.1145/3123266.3123454
Jing, Q., Yao, D., Fan, X., Wang, B., Tan, H., Bu, X., & Bi, J. (2021). TRANSFAKE: Multi-task transformer for multimodal enhanced fake news detection. In *Proceedings of the international joint conference on neural networks* (pp. 1–8). https://doi.org/10.1109/IJCNN52387.2021.9533433, 2021-July.
Khattar, D., Gupta, M., Goud, J.S., & Varma, V. (2019). MvaE: Multimodal variational autoencoder for fake news detection. The Web Conference 2019 - Proceedings of the World Wide Web Conference, WWW 2019, 7, 2915–2921. https://doi.org/10.1145/3308558.3313552.
Kryściński, W., McCann, B., Xiong, C., & Socher, R. (2020). Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the EMNLP 2020 - 2020 conference on empirical methods in natural language processing, proceedings of the conference* (pp. 9332–9346). https://doi.org/10.18653/v1/2020.emnlp-main.750
Kumari, R., & Ekbal, A. (2021). AMFB: Attention based multimodal factorized bilinear pooling for multimodal fake news detection. *Expert Systems with Applications, 184*(March), Article 115412. https://doi.org/10.1016/j.eswa.2021.115412
La, T. V., Tran, Q. T., Tran, T. P., Tran, A. D., Dang-Nguyen, D. T., & Dao, M. S. (2022). Multimodal cheapfakes detection by utilizing image captioning for global context. In *ICDAR 2022 - proceedings of the 3rd ACM workshop on intelligent cross-data analysis and retrieval* (pp. 9–16). https://doi.org/10.1145/3512731.3534210
Li, S., Weimin, L., Luvembe, A. M., & Weiqin, T. (2023a). Graph contrastive learning with feature augmentation for rumor detection. *IEEE Transactions on Computational Social Systems*, 590–605. https://doi.org/10.1007/978-3-031-26387-3_36
Li, W., Guo, C., Liu, Y., Zhou, X., Jin, Q., & Xin, M. (2023b). Rumor source localization in social networks based on infection potential energy. *Information Sciences, 634*(March), 172–188. https://doi.org/10.1016/j.ins.2023.03.098
Li, W., Ni, L., Wang, J., & Wang, C. (2022a). Collaborative representation learning for nodes and relations via heterogeneous graph neural network. *Knowledge-Based Systems, 255*, Article 109673. https://doi.org/10.1016/j.knosys.2022.109673
Li, W., Wei, D., Zhou, X., Li, S., & Jin, Q. (2022b). F-SWIR: Rumor Fick-spreading model considering fusion information decay in social networks. *Concurrency and Computation: Practice and Experience*, 1–17. https://doi.org/10.1002/cpe.7166. February.
Ma, J., Gao, W., & Wong, K. (2019). Detect rumors on Twitter by promoting information campaigns with generative adversarial learning. In *Proceedings of the 2019 World Wide Web Conference (WWW'19)*. May 13–17, 2019, San Francisco, CA, USA,.
Ma, J., Gao, W., & Wong, K. F. (2018). Rumor detection on twitter with tree-structured recursive neural networks. In *, 1. Proceedings of the ACL 2018 - 56th annual meeting of the association for computational linguistics, proceedings of the conference (long papers)* (pp. 1980–1989). https://doi.org/10.18653/v1/p18-1184
Madhusudhan, S., Mahurkar, S., & Nagarajan, S. K. (2020). Attributional analysis of multi-modal fake news detection models (grand challenge). In *Proceedings of the 2020 IEEE 6th international conference on multimedia big data, BigMM 2020* (pp. 451–455). https://doi.org/10.1109/BigMM50055.2020.00074
Meel, P., & Vishwakarma, D. K. (2021). HAN, image captioning, and forensics ensemble multimodal fake news detection. *Information Sciences, 567*, 23–41. https://doi.org/10.1016/j.ins.2021.03.037
Mohtarami, M., Baly, R., Glass, J., Nakov, P., Màrquez, L., & Moschitti, A. (2018). Automatic stance detection using end-To-end memory networks. In *, 1. Proceedings of the NAACL HLT 2018 - 2018 conference of the North American chapter of the association for computational linguistics: human language technologies - proceedings of the conference* (pp. 767–776). https://doi.org/10.18653/v1/n18-1070
Munyole, A., Li, W., Li, S., Liu, F., & Xu, G. (2023). Dual emotion based fake news detection : A deep attention-weight update approach. *Information Processing and Management, 60*(4), Article 103354. https://doi.org/10.1016/j.ipm.2023.103354
Nakamura, K., Levy, S., & Wang, W. Y. (2020). r/Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. In *Proceedings of the LREC 2020 - 12th international conference on language resources and evaluation, conference proceedings* (pp. 6149–6157). May.
Nguyen, D. M., Do, T. H., Calderbank, R., Deligiannis, N., & Carolina, N. (2019). Fake news detection using deep Markov random fields. In *Proceedings of the NAACL-HLT 2019* (pp. 1391–1400).
Obaid, A. Al, Khotanlou, H., Mansoorizadeh, M., & Zabihzadeh, D (2023). Robust semi-supervised fake news recognition by effective augmentations and ensemble of diverse deep learners. *IEEE access : practical innovations, open solutions, 11*(May), 54526–54543. https://doi.org/10.1109/ACCESS.2023.3278323
Olan, F., Jayawickrama, U., Arakpogun, E. O., Suklan, J., & Liu, S. (2022). Fake news on social media: The impact on society. *Information Systems Frontiers*. https://doi.org/10.1007/s10796-022-10242-z. January.
Paschalides, D., Christodoulou, C., Orphanou, K., Andreou, R., Kornilakis, A., Pallis, G., Dikaiakos, M. D., & Markatos, E. (2021). Check-It: A plugin for detecting fake news on the web. *Online Social Networks and Media, 25*, 298–302. https://doi.org/10.1016/j.osnem.2021.100156
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., & Grisel, O. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research, 12*, 2825–2830.
Qi, P., Cao, J., Li, X., Liu, H., Sheng, Q., Mi, X., He, Q., Lv, Y., Guo, C., & Yu, Y. (2021). Improving fake news detection by using an entity-enhanced framework to fuse diverse multimodal clues. In *, 1. MM 2021 - proceedings of the 29th ACM international conference on multimedia* (pp. 1212–1220). https://doi.org/10.1145/3474085.3481548
Qi, P., Cao, J., Yang, T., Guo, J., & Li, J. (2019). Exploiting multi-domain visual information for fake news detection. In *Proceedings of the 2019 IEEE International Conference on Data Mining (ICDM) Exploiting, ICDM* (pp. 518–527). https://doi.org/10.1109/ICDM.2019.00062
Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2020). Language Models are Unsupervised Multitask Learners. OpenAI, San Francisco, Californ-Nia, United States. http://arxiv.org/abs/2007.07582.
Sachan, T., Pinnaparaju, N., Gupta, M., & Varma, V. (2021). SCATE: Shared cross attention transformer encoders for multimodal fake news detection. In *Proceedings of the 2021 IEEE/ACM international conference on advances in social networks analysis and mining, ASONAM 2021* (pp. 399–406). https://doi.org/10.1145/3487351.3490965
Sengupta, E., Nagpal, R., Mehrotra, D., & Srivastava, G. (2021). ProBlock: A novel approach for fake news detection. *Cluster Computing, 24*(4), 3779–3795. https://doi.org/10.1007/s10586-021-03361-w
Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu, H. (2020). FakeNewsNet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data, 8*(3), 171–188. https://doi.org/10.1089/big.2020.0062
Shu, K., Zhou, X., Wang, S., Zafarani, R., & Liu, H. (2019). The role of user profiles for fake news detection. In *Proceedings of the IEEE/ACM international conference on advances in social networks analysis and mining* (pp. 436–439).
Singhal, S., Shah, R. R., Chakraborty, T., Kumaraguru, P., & Satoh, S. (2019). SpotFake: A multi-modal framework for fake news detection. In *Proceedings of the 2019 IEEE fifth international conference on multimedia Big Data (BigMM)*. https://ieeexplore.ieee.org/document/8919302.
Song, C., Ning, N., Zhang, Y., & Wu, B. (2021). A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks. *Information Processing and Management, 58*(1), Article 102437. https://doi.org/10.1016/j.ipm.2020.102437

Steinebach, M., Gotkowski, K., & Liu, H. (2019). Fake news detection by image montage recognition. In *Proceedings of the ACM international conference proceeding series*. https://doi.org/10.1145/3339252.3341487

Wang, J., Mao, H., & Li, H. (2022). FMFN: Fine-Grained multimodal fusion networks for fake news detection. *Applied Sciences (Switzerland), 12*(3). https://doi.org/10.3390/app12031093

Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., Su, L., & Gao, J. (2018). EANN: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 849–857). https://doi.org/10.1145/3219819.3219903

Wang, Y., Ma, F., Wang, H., Jha, K., & Gao, J. (2021). Multimodal emergent fake news detection via meta neural process networks. In *, 1. Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining.* Association for Computing Machinery. https://doi.org/10.1145/3447548.3467153.

Wu, L., Long, Y., Gao, C., Wang, Z., & Zhang, Y. (2023). MFIR : Multimodal fusion and inconsistency reasoning for explainable fake news detection. *Information Fusion, 100*(July), Article 101944. https://doi.org/10.1016/j.inffus.2023.101944

Wu, P. Y., & Mebane, W. R. (2022). MARMOT A deep learning framework for constructing multimodal representations for vision-and-language tasks. *Computational Communication Research, 4*(1), 275–322. https://doi.org/10.5117/CCR2022.1.008.WU

Xiong, S., Zhang, G., Batra, V., Xi, L., Shi, L., & Liu, L. (2023). TRIMOON: Two-Round inconsistency-based multi-modal fusion network for fake news detection. *Information Fusion, 93*(December 2022), 150–158. https://doi.org/10.1016/j.inffus.2022.12.016

Xue, J., Wang, Y., Tian, Y., Li, Y., Shi, L., & Wei, L. (2021). Detecting fake news by exploring the consistency of multimodal data. *Information Processing and Management, 58*(5), Article 102610. https://doi.org/10.1016/j.ipm.2021.102610

Yu, C., Ma, Y., An, L., & Li, G. (2022). BCMF: A bidirectional cross-modal fusion model for fake news detection. *Information Processing and Management, 59*(5). https://doi.org/10.1016/j.ipm.2022.103063

Yu, F., Liu, Q., Wu, S., Wang, L., & Tan, T. (2017). A convolutional approach for misinformation identification. *IJCAI International Joint Conference on Artificial Intelligence, 0*, 3901–3907. https://doi.org/10.24963/ijcai.2017/545

Zhang, C., Li, W., Wei, D., Liu, Y., & Li, Z. (2022). Network dynamic GCN influence maximization algorithm with leader fake labeling mechanism. *IEEE Transactions on Computational Social Systems*, 1–9. https://doi.org/10.1109/TCSS.2022.3193583

Zhang, K., Zhang, H., Liu, Q., Zhao, H., Zhu, H., & Chen, E. (2019). Interactive attention transfer network for cross-domain sentiment classification. In *Proceedings of the thirty-third AAAI conference on artificial intelligence (AAAI-19) interactive*.

Zhang, X., Dadkhah, S., Weismann, A. G., Kanaani, M. A., & Ghorbani, A. A. (2023). Multimodal fake news analysis based on image–text similarity. *IEEE Transactions on Computational Social Systems*, 1–14. https://doi.org/10.1109/TCSS.2023.3244068. PP.

Zhou, X., Wu, J., & Zafarani, R. (2020). SAFE: Similarity-Aware multi-modal fake news detection. In *Proceedings of the lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics), 12085 LNAI* (pp. 354–367). https://doi.org/10.1007/978-3-030-47436-2_27

Zubiaga, A., Liakata, M., & Procter, R. (2016). Learning Reporting Dynamics during Breaking News for Rumour Detection in Social Media. http://arxiv.org/abs/1610.07363.

ARTICLES FOR FACULTY MEMBERS

MULTIMODAL FAKE NEWS DETECTION

Title/Author	**Combating multimodal fake news on social media: Methods, datasets, and future perspective / Hangloo, S., & Arora, B.**
Source	***Multimedia Systems*** **Volume 28 Issue 6 (2022) Pages 2391–2422** **https://doi.org/10.1007/S00530-022-00966-Y** **(Database: SpringerLink)**

Multimedia Systems (2022) 28:2391–2422
https://doi.org/10.1007/s00530-022-00966-y

REGULAR PAPER

Combating multimodal fake news on social media: methods, datasets, and future perspective

Sakshini Hangloo[1] · **Bhavna Arora**[1]

Received: 11 January 2022 / Accepted: 3 June 2022 / Published online: 7 July 2022

Abstract
The growth in the use of social media platforms such as Facebook and Twitter over the past decade has significantly facilitated and improved the way people communicate with each other. However, the information that is available and shared online is not always credible. These platforms provide a fertile ground for the rapid propagation of breaking news along with other misleading information. The enormous amounts of fake news present online have the potential to trigger serious problems at an individual level and in society at large. Detecting whether the given information is fake or not is a challenging problem and the traits of social media makes the task even more complicated as it eases the generation and spread of content to the masses leading to an enormous volume of content to analyze. The multimedia nature of fake news on online platforms has not been explored fully. This survey presents a comprehensive overview of the state-of-the-art techniques for combating fake news on online media with the prime focus on deep learning (DL) techniques keeping multimodality under consideration. Apart from this, various DL frameworks, pre-trained models, and transfer learning approaches are also underlined. As till date, there are only limited multimodal datasets that are available for this task, the paper highlights various data collection strategies that can be used along with a comparative analysis of available multimodal fake news datasets. The paper also highlights and discusses various open areas and challenges in this direction.

Keywords Fake news detection · Rumor detection · Transfer learning · Pretrained models · Text embedding · Deep learning · Multimodal

1 Introduction

The propagation of information on social media is a fast-paced process, with millions of individuals participating on these sites. However, unlike traditional news sources, the trustworthiness of content on social media sites is debatable. In the last decade, there is an upsurge in the use of social media and microblogging platforms such as Facebook and Twitter. Billions of users on daily basis use these platforms to convey their opinions through messages, pictures, and videos all over the world. Government agencies also utilize these platforms to disseminate critical information using their official Twitter handles and verified Facebook accounts, since information circulated through these platforms can reach a large population in a short amount of time. Many deceptive practices, including as propaganda and rumor, might however, deceive consumers on a daily basis. Fake news and rumors are quite common in these COVID times, and they are widely circulated, causing havoc in this difficult time. People unknowingly spreading false information is a considerably more serious problem than systematic disinformation tactics. Previously, attempts to influence public opinion were gradual, but now, rumors are targeted at naive users on social media. Once people mistakenly transmit incorrect or fraudulent content, it spreads across trusted peer-to-peer networks in all directions and as a result, in the current situation, the requirement for Fake News Detection (FND) is unavoidable.

Despite ongoing research efforts in the field of FND, ranging from comprehending the problem to building a framework to model evaluation, there is still a need to construct a reliable and efficient model. Various approaches to

Communicated by I. IDE.

✉ Sakshini Hangloo
sakshini.hangloo@gmail.com

[1] Department of Computer Science & Information Technology, Central University of Jammu, Bagla (Rahya Suchani), District-Samba, 181143 Jammu, J&K, India

fake news and rumor detection have been formulated ranging from detection methods based on content [1]–[6], propagation data [7]–[11], user profile [12]–[14], event-specific data [15]–[17], external knowledge [16, 18, 19], temporal data [20]–[24], multimodal data [4, 16, 25]–[28] etc. Source detection [29]–[34], Bot detection [35]–[37], Stance detection [38]–[43], Credibility analysis [44]–[48] are other related areas.

Earlier solutions used ML based approaches [32, 49]–[53] but suffer from the problem of manual feature engineering. With the advancement of Deep Learning (DL) based approaches for computer vision and NLP (Natural Language Processing), recent years have seen a paradigm shift from ML (Machine Learning) to DL-based fake news detection solutions. The DL models are trained using Convolution Neural Network (CNN), Recurrent Neural Network (RNN), Recursive Neural Networks (RvNN), Multi-Layer Perceptron (MLP), Generative Adversarial Networks (GANs) and many more.

It is imperative to spot fake news at the earliest before it reaches the masses. The multimodal aspect of the news article makes the content look much more credible than its counterparts. Most of the existing work focuses on text-only content or the network structure and ignores the most important aspect of the news content i.e., the visual content and consistency between text-image. Currently, due to scarcity of multimodal training datasets, transfer learning and various pre-trained models, like VGGNet [54], ResNet [55], Inception [56], Word2Vec, GloVe, BERT [57], XLNet [58] etc. are utilized for a more efficient DL-based solution. A comparative analysis of available multimodal fake news datasets is provided in Sect. 4. Although various techniques and methods have been developed in the last decade to counter fake news there are still several open research issues and challenges as mentioned in Sect. 5. By evaluating several existing strategies and identifying potential models and approaches that can be used in this area, this paper aims at contributing to the ongoing research in the field of automatic multi-modal fake news detection. Our survey seeks to give an in-depth analysis of current state-of-the-art Multimodal Fake News Detection (MFND) frameworks, with a particular focus on DL-based models. Table 1 compares various

Table 1 A relative comparison of proposed work with various related surveys

Ref.	Discussion	1	2	3	4	5	6	7
[59]	Proposes various visual and statistical features of a visual content	✓	×	×	×	×	×	×
[60]	Presents a comprehensive review of fake news detection techniques on social media from the data mining perspective	✓	×	×	×	✓	×	✓
[61]	Provides an overview of techniques of developing a rumor classification system consisting of detection, tracking, stance classification, and veracity classification modules	✓	✓	×	×	✓	✓	✓
[22]	Examined and compared the relative strength of the user, linguistic, network and temporal features of rumors over time	✓	×	×	×	×	×	×
[62]	provides an extensive study of automatic rumor detection on three paradigms: the hand-crafted feature-based approaches, the propagation structure-based approaches and the neural networks-based approaches	✓	×	×	×	✓	×	✓
[63]	Survey provides a review of techniques for manipulation and detection of face images including DeepFake methods. In particular, facial manipulation are reviewed based on following four types: attribute manipulation, face synthesis, identity swap (DeepFakes), and expression swap	✓	×	×	×	✓	×	✓
[64]	Gives an understanding of fake news creation, source identification, propagation patters, detection and containment strategies	✓	✓	×	×	✓	×	✓
[65]	Presents a detailed review of state-of-the-art FND methods using DL, open issues along with future directions are also suggested	✓	×	✓	×	✓	×	✓
[66]	Reviews the methods for detecting fake news from four verticals: the false information, writing style, propagation patterns, and the source credibility	✓	×	×	×	×	×	✓
[67]	Presents an overview of the state-of-the-art fake news detection methods utilizing users, content, and context features	✓	✓	×	×	✓	×	✓
[68]	Provides an overview of the different forms of fabricated content on social media ranging from text-only to multimedia content and discusses various detection techniques for the same	✓	×	×	×	×	×	✓
[69]	proposed work explores the problem of rumors detection using textual content of social media on collected Twitter data	✓	×	×	×	×	✓	✓
[70]	Compares, reviews and provides insights into twenty-seven popular fake news detection datasets	×	×	×	×	✓	×	✓
Present Study	The prime focus is on various deep learning approaches to fake news detection on social media keeping the multimodal data under consideration	✓	✓	✓	✓	✓	✓	✓

Notes: 1: Overview of ML/DL-based FND; 2: Open tools and initiatives; 3: DL frameworks & tools; 4: Review of MFND frameworks; 5: Datasets; 6: Data collection; 7: Open issues; Notations: ✓:Considered;×: Not considered

existing fake news detection surveys with our survey, demonstrating that our survey not only uses more recent state-of-the-art MFND methods, but also includes widely used DL frameworks and tools, various data collection strategies from online platforms, a comparison of available datasets, and finally open issues and future scope in this direction.

1.1 Scope of the survey

This survey is motivated by the increase in the usage of social networking sites for the spread of fake news where the multimodal nature of post/tweet increases the difficulty level of the detection task. The motivations of this paper can be summed up as follows, (1) Analyzing the text-only content of an article is not sufficient to model a robust and efficient detection system. In this era of social media, it is highly imperative to consider the visual content apart from the textual context and social context to get a complete understanding of overall statistics. (2) Promising DL frameworks and transfer learning approaches are reviewed in this paper, which are advantageous for addressing the challenges and producing an improvement over the existing detection frameworks. (3) The studies performed for the detection of online fake news are diverse but these suffer from a lack of multimodal datasets. So, this study also gives an overview of some existing multimodal datasets and highlights various data collection strategies as well. This survey presents a comprehensive review of the state-of-the-art multimodal fake news detection on online media which was absent in the previous surveys.

An exhaustive comparative analysis of various research surveys is compiled in Table 1 to provide an insight into the dimensions that have not been covered previously. Different from the previous studies, in this work, the prime focus is on various deep learning approaches including the transfer learning and pre-trained models used for fake news detection on social media keeping the multimodal data under consideration. Apart from this, the paper also highlights the data collection methods and the datasets available for this task. Discussion on open areas and future scope is also provided at the end.

1.2 Contribution

The key contributions of this paper are as follows:

- The paper gives a brief introduction to fake news its related terms and provides a clear taxonomy that focuses on different methods for Fake News Detection (FND).
- The paper highlights various DL models and frameworks that are used in the literature and the benefits of using the pre-trained models and the approach of transfer learning are highlighted. Critical analysis of different learning techniques and DL frameworks has also been presented.
- The paper discusses and reviews the various state-of-the-art Multimodal Fake News Detection (MFND) frameworks that are presented in the literature.
- The paper discusses various data collection techniques using APIs and Web crawlers in addition to a comparative analysis of various benchmark multimodal datasets.
- Finally, open issues and future recommendations are provided to combat the issue of fake news.

1.3 Methods and materials

This study is conducted using a suitable methodology to provide a complete analysis of one of the essential pillars in fake news detection, i.e., the multimodal dimension of a given article. To conduct this systematic review, various relevant articles, studies, and publications were examined. Before gathering the essential information for the conducted survey, quality checks are performed on the identified data with a focus on the most cited paper. In this work, the prime focus is on state-of-the-art research on multimodal fake news detection for assessing the authenticity of a news piece using deep learning algorithms. To obtain relevant literature, high-quality, highly cited, and reliable peer-reviewed publications, as well as conferences proceedings, are preferred. Other sources that are referred to for this study include books, technical blogs, and tutorial papers. For the search criteria, keywords like fake news detection, rumor detection, multimodal feature extraction, deep learning, and pretraining have been used. We have analyzed and acknowledged several works related to the reviewed theme of the proposed survey.

1.4 Organization

Figure 1 describes the organization of the presented survey. Section 1 presents the introduction as well as the overall scope of the paper. In Sect. 2 an introduction to fake news along with a taxonomy of fake news detection techniques has been presented. This section also disuses existing techniques and solutions to curb and combat fake news in this era of social media. Section 3 gives an overview of various DL models and transfer learning approaches that are widely in NLP, computer vision, and related fields. This section further presents a review of the DL-based state-of-the-art frameworks for Multimodal Fake news detection (MFND). Various data collection techniques and the details about the related datasets are discussed in Sect. 4. Section 5 and Sect. 6 deal with the challenges, open issues, and future direction, and the conclusion based on the survey.

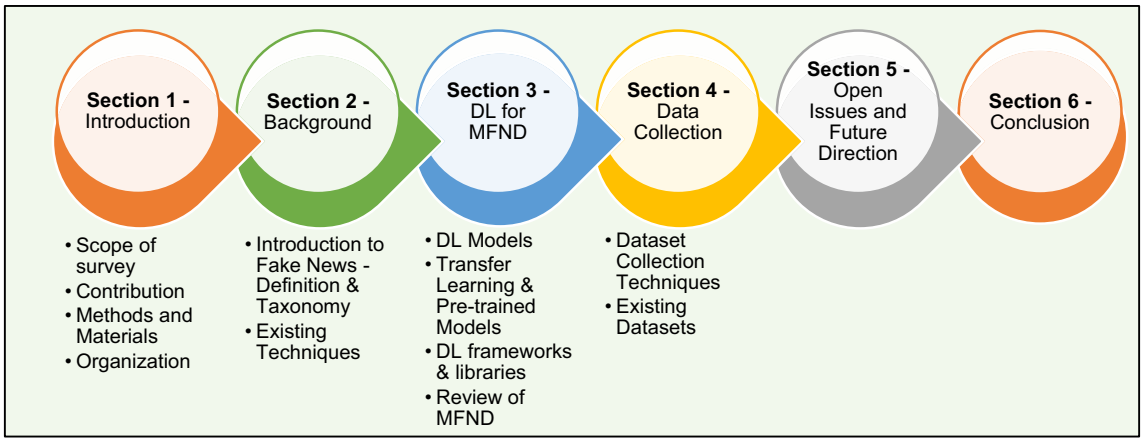


Fig. 1 Roadmap of the proposed survey

2 Background

Social media in the past decade has become the most unavoidable part of our society. With the shift of trend from Web 1.0 to Web 2.0 people have started not only to consume but also to create and spread information online. But, is this information credible? Can these be trusted? Not always. Here is a popular quote by Mark Twain "A lie can travel halfway around the world while the truth is putting on its shoes". This became quite evident in the COVID times when the internet was flooded with all kinds of information related to Government advisories, home remedies, etc.

The graph in Fig. 2 below shows a clear picture of how in the past decade the cases of fake news have increased exponentially. One of the major reasons is the rise in the use of social media and the unchecked circulation of messages on the platforms.

2.1 Introduction to fake news—definition and taxonomy

Fake News is defined as "false stories that are created and spread on the Internet to influence public opinion and appear to be true" The issue of spread of fake news is not new and has been around for centuries but with the use of social media the whole dynamics of the proliferation of information has changed and is quite different from the slow-paced traditional media. These sites provide a platform for intentional propaganda and trolling. Propaganda, fake news, satire, hoax, misinformation, rumors, disinformation, etc., are some of the terms that are used interchangeably. Some of them are discussed below (Fig. 3).

i. *Propaganda:* It is a form of news articles and stories created and disseminated by political parties to shape public opinion.

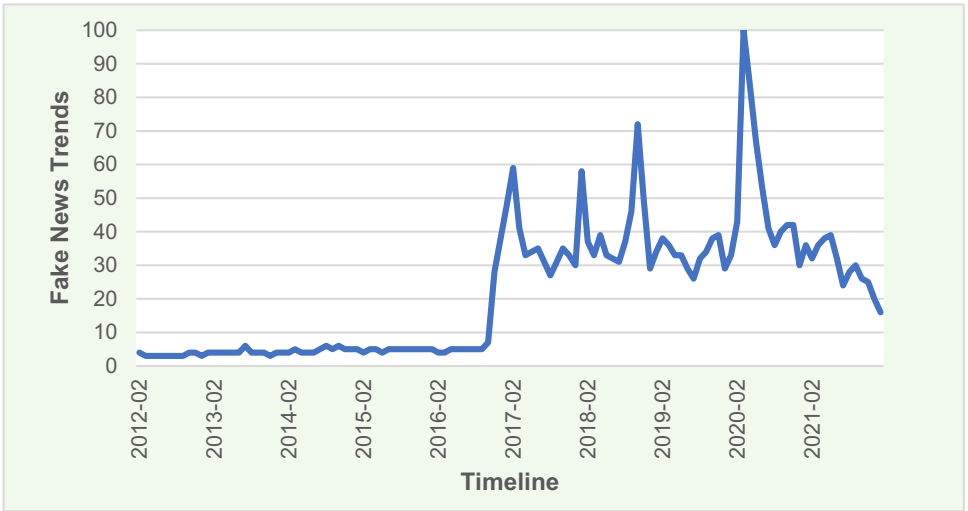


Fig. 2 Fake news trends (2012–2021) [71]

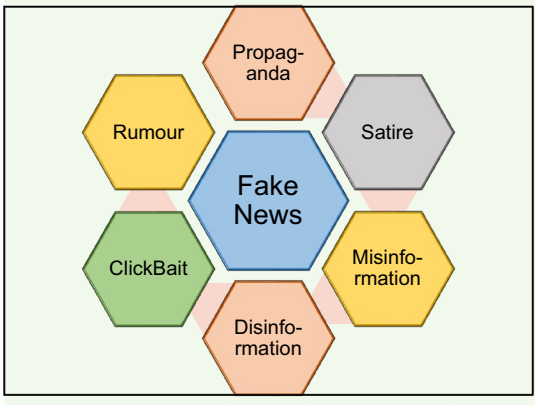


Fig. 3 Key terms related to Fake News

ii. *Misinformation:* It is purposely crafted erroneous information that is broadcast intentionally or accidentally, without regard for the true intent.
iii. *Disinformation:* It refers to a piece of misleading or partial information that is circulated to distort facts and deceive the intended audience.
iv. *Rumors and hoaxes:* These terms are often used interchangeably to refer to the purposeful fabrication of evidence that is intended to appear legitimate. They publish unsubstantiated and false allegations as true claims that are validated by established news outlets.
v. *Parody and Satire:* Humor is frequently used in parody and satire to provide news updates, and they frequently imitate mainstream news sources.
vi. *Clickbait*: Clickbait headlines are frequently used to attract readers' attention and encourage them to click, redirecting the reader to a different site. More advertisement clicks equal more money.

With the increased usage of propaganda, hoaxes, and satire alongside real news and legitimate information, even regular users find it difficult to discriminate between true and fake news. However, there are a number of online tools and IFCN-certified fact-checkers throughout the world such as BSDetector, AltNews, APF Fact Check, Hoaxy, Snopes, and PolitiFact that evaluate, rate, and debunk false news on online platforms[72].

Table 2 provides some of the Fact-checking sites and online tools that are used for debunking false news online. This table also gives an overview of the methodology and set of actions that are taken to detect and combat fake news on online platforms.

2.2 Existing detection techniques

A huge amount of content today is human-generated and most of these get published and people spread that information without even bothering about the credibility of these contents. Many technical giants are now committed to fight against the spread of fake information. Facebook in certain countries has started to work with third-party fact-checker organizations to help identify, review and rate the accuracy of information [73]. These fact-checkers are certified through the non-partisan International Fact-Checking Network (IFCN).

Figure 4 shows some of the online claims that are debunked by fact-checking organizations. Twitter on the other hand took a step forward in May 2020 to curb the misinformation around COVID-19 by introducing new labels and warning messages [74] to provide the users with additional context and information about the Tweets. This has made it easier for the users to find facts and make informed decisions about the tweets. In Jan 2021, Twitter introduced BirdWatch [75], a pilot in the US which is a community-based approach that allows people to identify Tweets that they believe are deceptive and annotate these. The pilot participants can also rate the preciseness of the notes added by other contributors.

The research community is also working tirelessly and many papers have been published to combat rumors and fake news on social media platforms. The earlier approaches [51, 76]–[81] used various ML techniques like SVM, RF, NB, etc. but with the ever-increasing amount of data on social media platforms a shift to DL approaches [11, 23, 82]–[85] can be seen with includes the use of CNN, RNN, LSTM, GAN based approaches.

Figure 5 gives a detailed taxonomy of existing fake news detection methods and techniques. Table 3 below provides a detailed classification of prominent state-of-the art ML/DL FND techniques based on the proposed taxonomy.

In the case of online social media, the rumors proliferate in a short period and hence early detection becomes very important. By exploiting the dissemination structure on social media, Liu et al. [80] offers a model for early identification of misleading news. Each news story's propagation path is treated as a multivariate time series. It employs a hybrid CNN-RNN that gathers global and local fluctuations in user attributes along the propagation path. In just 5 min after it starts spreading, the model detects fake news on Twitter and Sina Weibo with an accuracy of 85 percent and 92 percent respectively. The work proposed by Varol et al. [86] works on the early detection of promoted campaigns on online platforms. It proposed a supervised computational framework that leverages temporal patterns of the message associated with trending hashtags on Twitter to catch how the posts evolve over time and successfully classifies it as either ‘promoted’ or ‘organic’. In addition to this, it also used network structure, sentiment, content features, and user metadata and achieves 75% AUC score for early detection, increasing to above 95% after trending.

Table 2 Fact-checking sites and online tools that are used for debunking false news online

Name	Tool/Extension	Methodology/Action
AltNews	Fact-checking website	Continuously monitors social media and mainstream media platforms for identifying incorrect information related mainly to Indian politics and entertainment, and evaluates the veracity of a claim by Manual Fact-checking
APF Fact Check	Fact-checking website	It uses many simple tools to verify online information. Fact-checking is carried out by editors and a worldwide network of journalists
BS Detector	It is an extension of Google Chrome, Mozilla	Identifies and marks fake and satirical news sites, as well as other suspected news sources. It puts a warning label to the top of potentially dangerous websites, as well as identifies fake links on Facebook and Twitter
Emergent	Fact-checking website	Emergent is a real-time rumor tracker that assesses news credibility and gives a True, False, or Unverified label
Fact-Checker	Fact-checking website	A project of The Washington Post, grades news articles from zero to four "Pinocchios" based on the factual accuracy of their content
FakerFact	A Chrome and Firefox extension	Distinguishes a fake news article from the real one and categorizes it as opinion, satire, agenda-driven, journalism, and sensationalism
InVid Verification Plugin	Use with Chrome, Firefox	A plugin to debunk fake images and videos. The tool uses reverse-image searching to debunk fake videos and also provide the users with metadata to take informed decision
PolitiFact	Fact-checking website	Tests the statements made on the Internet by political analysts and politicians and rate them. Its journalists evaluate original statements and each statement receives a "Truth-O-Meter" rating as "True", "Mostly True", "Half True", "Mostly False", "False", and "Pants on Fire"
Snopes	Fact-checking website	Conducts in-depth fact-checking research on hot issues, which are frequently picked depending on reader interest. "True", "Mostly True", "Mostly False", "False", "Unproven", "Miscaptioned", "Misattributed" are some of the annotations used to classify the content
Reverse Image Search (TinEye)	Browser extension	Can be used to see if the image has been taken from somewhere online. The tool comes with a Compare feature, which can be helpful to see how your image differs from the original one
BOOM	Fact-checking website	Manually checks the posts, debunks fake news, and prevents further spread
SurfSafe	Browser Plugin	Alerts users about misinformation by scanning images and videos on the web pages they're looking at. Performs reverse-image search by looking for the same content that appears on trusted source sites and flagging well-known doctored images
YouTube Data Viewer	A web-based video verification tool	Simple tool for extracting hidden data and metadata from YouTube videos which is particularly valuable for locating original content

Various approaches for early rumor detection have been explored. In case of the content-based approach [18, 50, 59, 90, 103], the content (text + images) within the article is considered, in contrast to the social context-based approaches [7, 13, 118]–[120] where the propagation structures, data from the user profile is considered. The content-based approaches have performed better as compared to the context-based approaches because the propagation structure and user data become available only after the news has traveled masses. Monti et al. [109] proposed a geometric deep learning approach (a non-Euclidean deep learning approach) for fake news detection on Twitter that uses a GRU-based propagation Graph Neural Network to utilize the network structure. In addition to the spreading patterns, it also uses features from the user profile, social network structure, and content. Dou et al. [121] proposes a framework, UPFD, which simultaneously captures various signals from user bias along with the news content to analyze the likelihood of user to forward a post based on his/her existing beliefs. Wu et al. [11] proposes a novel method to construct the network

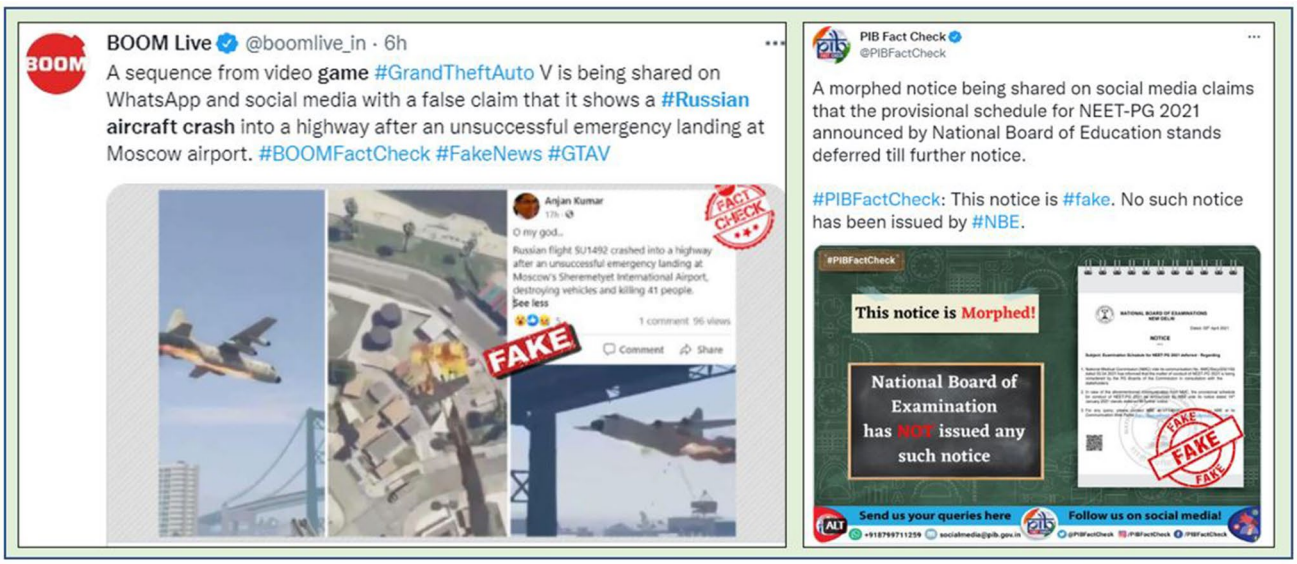


Fig. 4 False claims debunked by fact-checking organizations

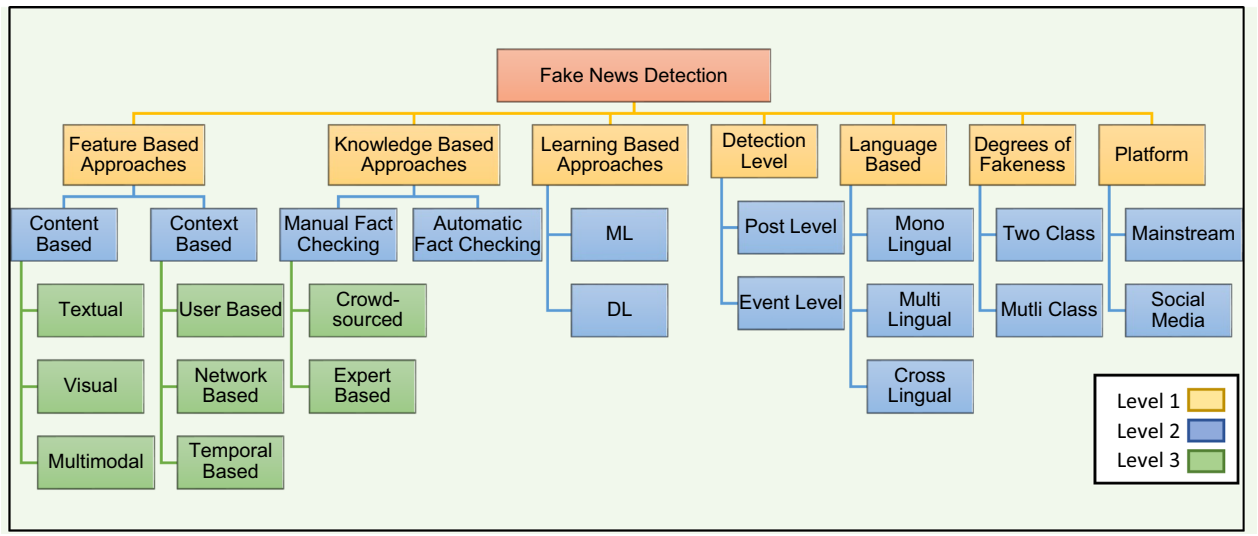


Fig. 5 Taxonomy of Fake News Detection models

graph by taking the "who-replies-to-whom" structure on Twitter.

Most of the existing works are limited to using the datasets that cover domains such as politics, and entertainment. However, in a real-time scenario, a news stream typically covers various domains. Silva et al. [93] proposed a novel fake news detection framework to determine fake news from different domains by exploiting domain-specific and cross-domain knowledge in news records. To maximize domain-coverage, this research merges three datasets (PolitiFact, GossipCop and CoAID).

The content available in a post/ tweet in a microblogging site is very limited and hence detection only based on the content available in that particular post i.e. Post-level becomes difficult in such scenarios but leveraging a complete event is beneficial. [15, 16] are some state-of-the-art Event-level detection models. An event not only includes the particular post but also post-repost structure, comments, likes-dislikes, etc. and this auxiliary information makes the detection process efficient. Guo et al. [101] uses a framework of Hierarchical Networks with Social Attention (HSA-BLSTM) that aims to predict the credibility of a group of

Table 3 Classification of prominent state-of-the art ML/DL FND techniques based on the proposed taxonomy

	Level 1	Level 2	Level 3	Related work
Fake News Detection Methods	Feature Based	Content	Single-Modal	[17, 50, 59, 87]–[90]
			Multi-Modal	[4, 26, 28, 91]–[100]
		Context	Network	[10, 11, 51, 85, 101]
			User	[13, 14, 53, 102]
			Temporal	[21]–[24, 86, 103]
	Knowledge Based	Automatic	–	[16, 19, 99]
		Manual	Expert Based	[18, 40, 104, 105]
			Crowdsourced	[103, 106, 107]
	Learning Based	ML	–	[51, 76, 88, 108]
		DL	–	[11, 26, 85, 92, 96, 101, 109]–[111]
	Detection Based	Post-level	–	[18, 25, 106]
		Event Level	–	[17, 112, 113]
	Language Based	Mono-Lingual	–	[22, 25, 27, 92]
		Multi-Lingual	–	[87]
	Degree of Fakeness	Two-Class	–	[16, 25, 26, 95, 114]
		Multi-Class	–	[106, 107]
	Platform	Main-Stream	–	[18, 106, 115]
		Social Media	–	[60, 82, 86, 116, 117]

posts (reposts and comments) that constitute an event. The model uses user-based features and propagation patterns in addition to the post-based features.

Another popular approach to FND is Evidence-based or Knowledge-based Fact-Checking where the article at hand is verified with external sources. While manual fact checking is done either manually by expert journalists, editors or is sometimes crowdsourced on the other hand Automatic fact checking [18, 115] is done by employing various ML/DL techniques. A novel end-to-end graph neural model, CompareNet [19], compares the news to the knowledge base (KB) through entities for automatic fact checking.

3 Deep learning for multimodal fake news detection

With the rapid development of social media platforms, news content has transformed from traditional text-only pieces to multimedia stories with images and videos that provide more information. Multimodal news items engage more readers than standard text-only news articles because the photos and videos related to these articles make them more credible. The majority of online users are impacted by such material, unwittingly spreading false information, and becoming a part of this entire vicious network. With the growing quantity of articles on the Internet that include visual information and the widespread usage of social media networks, the multimodal aspect is becoming increasingly important in better comprehending the overall structure of the content.

Due to exceptionally promising outcomes in several study areas, including Computer Vision and Natural Language Processing, Deep Learning has become one of the most frequently researched domains by the research community in recent years. Feature extraction, which is a laborious and time-consuming process in traditional machine learning algorithms, is done automatically by deep learning frameworks. Furthermore, these frameworks can learn hidden representations from complex inputs, both in terms of context and content, giving them an advantage in false news detection tasks when identifying relevant features for analysis is difficult.

3.1 Deep learning models

The Deep Learning techniques can be broadly classified as Discriminative and Generative models among these are Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN) are the most widely implemented paradigms. As shown in Fig. 6, the DL models can be broadly classified as Discriminative and Generative models. The models are described under:

A. *Discriminative Models:* These models are supervised learning-based models and are used to solve classification and regression problems. In recent years, several discriminative models (mostly CNN, RNN) have pro-

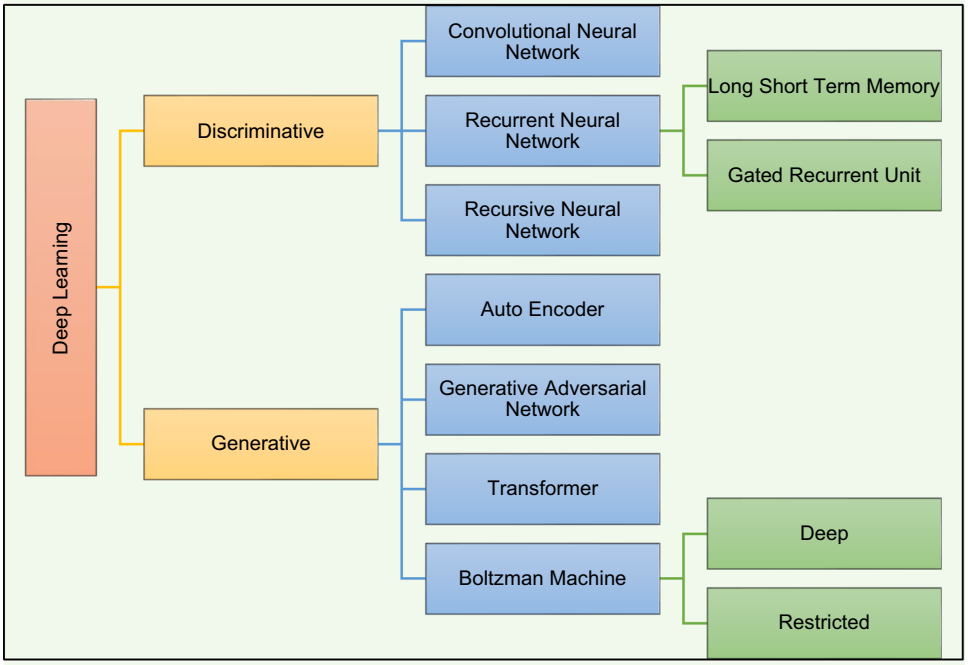


Fig. 6 Classification of Deep Learning Models

vided promising results in detecting fake news on social media platforms.

B. *Convolutional Neural Network (CNN)***:** CNN is a type of neural network that has an input layer, an output layer, and a sequence of hidden layers, in addition to this they include pooling and convolution operations for applying a variety of transformations to the given input. For the Computer Vision tasks, CNNs have been extensively explored and are regarded as state-of-the-art. CNN is also becoming increasingly popular in NLP tasks.

C. *Recurrent Neural Network (RNN):* RNNs are powerful structures that allow modeling of sequential data using loops within the neural network. Deep neural networks (RNN) have shown promising performance for learning representations in recent research. RNN is capable of capturing long-term dependency but fails to hold it as the sequence becomes longer. LSTM and GRU, the two variants of RNN, are designed to have more persistent memory and hence make capturing long-term dependencies easier. Additionally, the two networks also solve the issue of vanishing gradient problem that was encountered in traditional RNNs. LSTM includes memory cells for holding long-term dependencies in the text and includes input, output, and forget gates for memory orchestration. To further capture the contextual information, bidirectional LSTM (Bi-LSTM) and bidirectional GRU (Bi-GRU) are used to model word sequences from both directions.

D. *Recursive Neural Network (RvNN):*A recursive neural network is a deep neural network that applies same set of weights recursively over a structured input, to produce a structured prediction over variable-size input structures, by traversing a given structure in topological order. Ma et al. [122] proposes two recursive neural models based on a bottom-up and a top-down tree-structured neural networks for representing propagation structure of tweet. A tree-structured LSTM with an attention mechanism is proposed in [123] learns the correspondence between image regions and descriptive words.

Fig. 7 provides a descriptive overview of the most widely used descriptive models, namely CNN, RNN and LSTM. In addition to these, several recent works have exploited a combination of RNNs and CNNs in their models for increased efficiency. Nasir et al. [124] proposed a hybrid CNN+RNN model that can generalize across datasets and tasks. This model use CNN and LSTM to extract local features and learn long-term dependencies respectively.

B. *Generative Models:* These models are used in absence of labeled data and belong to the category of unsupervised learning. Among various generative models that have been used widely to solve a vast domain of problems, Generative Adversarial Network (GAN), Auto Encoder (AE), Transformer-based network and Boltzman Machine (BM) are mostly used and have also shown promising results in the field of FND.

i. *Auto Encoder (AE):* An autoencoder is a feed-forward neural network that regenerates the input and

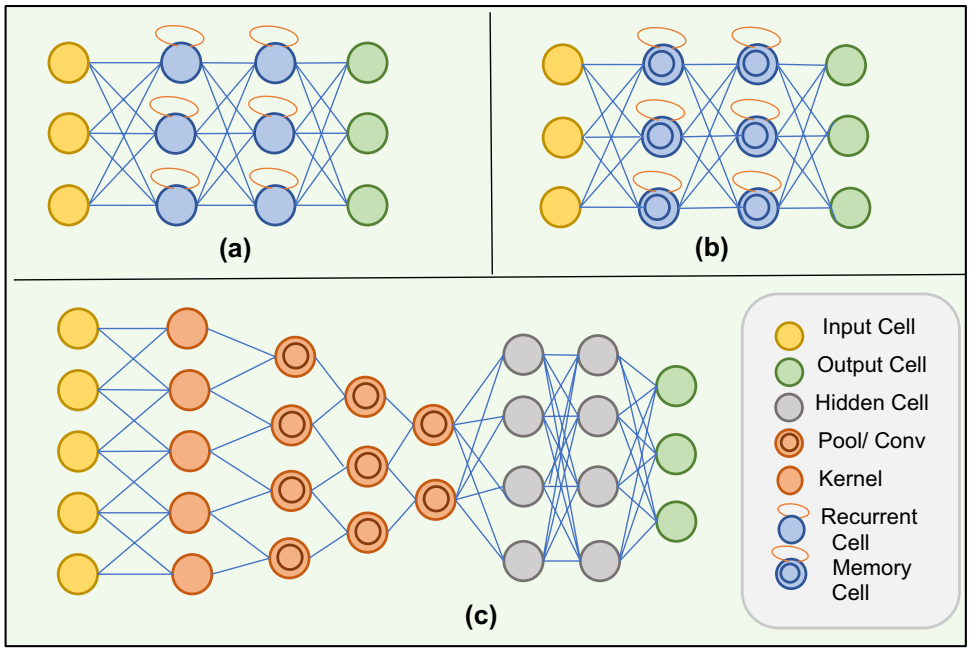


Fig. 7 Discriminative Models (a) RNN (b) LSTM (c) CNN. [126]

creates a compressed latent space representation. An autoencoder consists of i) an encoder that creates an intermediate representation of the input data, and ii) a decoder that regenerates the input as output. It is a self-supervised dimensionality reduction technique that generates its own labels from the training data and creates a lossy compression. A variation of AE is used in [27] that uses Multimodal Variational AE to capture the correlation between text and visual.

ii. *Generative Adversarial Network (GAN):* GAN is a class of unsupervised DL technique that consists of i) a generator that learns the generation of new sample data with the same statistics as training data, and ii) a discriminator that ties to classify the sample as true or fake. The discriminator is updated after every epoch to get better at classifying the samples, and the generator is updated to efficiently generate more believable samples. It is used widely for manipulating images and generating DeepFakes. To detect deepfake [125] detects and extracts a fingerprint that represents the Convolutional Traces (CT) left by GANs during image generation.

iii. *Transformer:* The Transformer-based networks [127] have come into existence in the past few years and have shown tremendous results for various NLP tasks. It aims to solve sequence-to-sequence tasks and handles long-range dependencies. To compute representations of its input and output, it relies on self-attention without using sequence aligned RNNs or CNNs. A transformer network consists of the encoder stack and the decoder stack that have the same number of units. The number of encoder and decoder units act as hyperparameter. In addition to the self-attention and feed-forward layers that are present in both encoder as well as decoder, the decoders also have one more layer of Encoder-Decoder Attention layer to focus on the appropriate parts of the input sequence. BERT[57], RoBERT, ALBERT are some of the widely used transformer-based network that has been applied successfully in FND [26, 128, 129]. These networks are pre-trained and can be fine-tuned for various NLP tasks. Various transformer-based word embeddings are introduced in Sect. 3.2.

iv. *Boltzmann Machine:* It is a type of recurrent neural network where the nodes make binary decisions and are present with certain biases. Several Boltzmann Machines (BM) can be stacked together to make even more sophisticated systems such as a Deep Boltzmann Machine (DBM). These networks have more hidden layers compared to BM and have directionless connections between the nodes. The task of training is to find out how these two sets of variables are actually connected to each other. For the large unlabelled dataset, a DBM incorporates a Markov random field for layer-wise pretraining and then provides feedback to previous layers. Restricted Boltzmann Machine (RBM) shares a similar idea as Encoder-Decoder, but it uses stochastic units with particular distribution instead of deterministic distribution. RBM plays an important role in dimensionality reduction, classification, regression and many more which is used for feature selection and feature extraction. [130] pre-

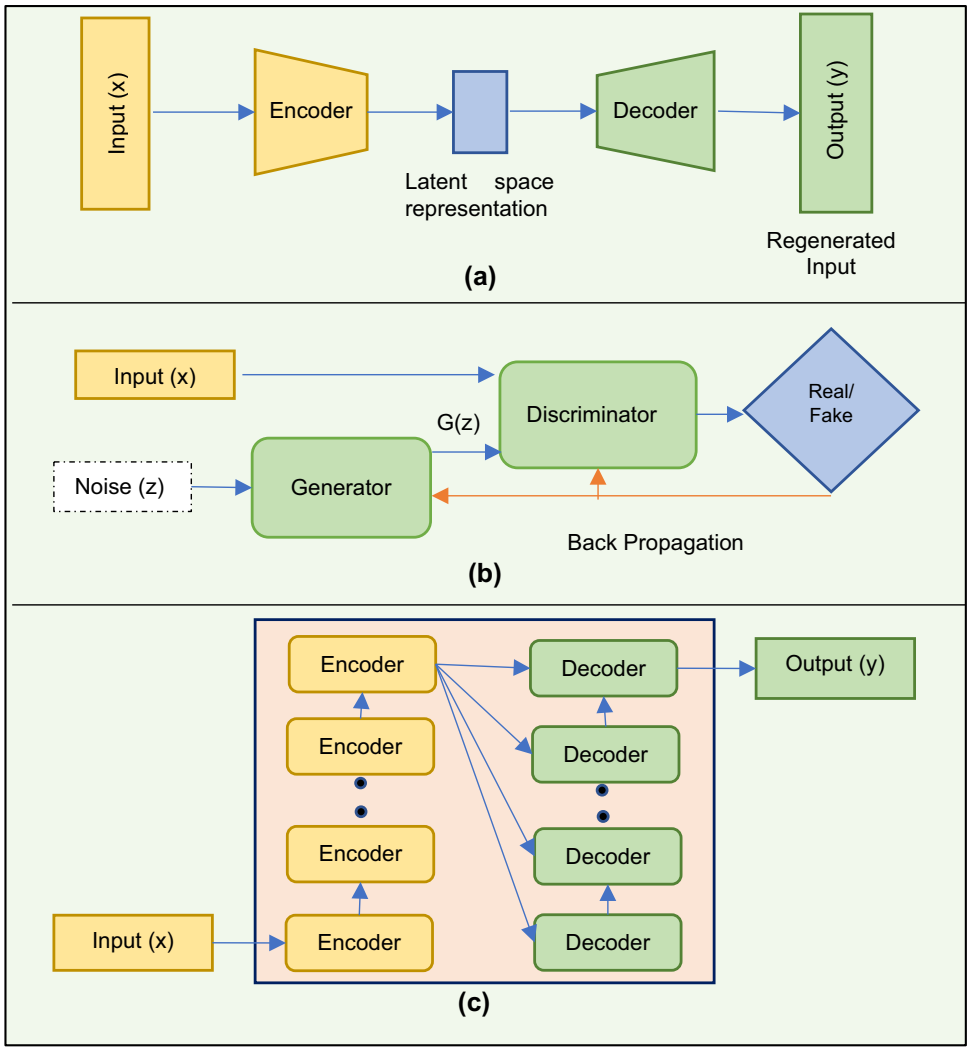


Fig. 8 Generative Models (a) Auto Encoder (b) Generative Adversarial Network (c) Transformer. [131]

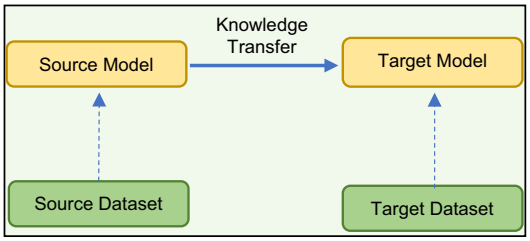


Fig. 9 Transfer learning

sents a Deep Boltzmann Machine based multimodal deep learning model for fake news detection.

Among various generative models (Fig. 8) that have been used widely to solve a vast domain of problems, Generative Adversarial Network (GAN), Auto Encoder (AE), Transformer-based network are widely used and have also shown promising results in the field of FND.

3.2 Transfer learning and pre-trained models

Transfer learning is a machine learning technique that leverages and applies the weights of a model that is trained on one task to some other related tasks with different datasets for improving efficiency as shown in Fig. 9. This approach can be applied in one of the two ways (i) using the pre-trained model as feature extraction, or (ii) fine-tuning a part of the model. The first variant is directly applied without changing the weights of the pre-trained model e.g., using Word2Vec in the NLP task. For the second variant, fine-tuning is done by trial-and-error experiments. For two different tasks around 50% of fine-tuning can be considered and if the tasks are very similar fine-tuning of the last few layers can be done. The nature and amount of fine-tuning needed take time and effort to explore depending on the nature of the task. Using Transfer Learning is beneficial when the target dataset is significantly smaller than the source dataset, as the model can learn features even with less training data without overfitting. Also, such a model exhibits better efficiency and

requires a lesser training time than a custom-made model. As a result, this idea finds vast usage in the field of Computer Vision and Natural Language Processing.

A. *Transfer learning with image data:* Earlier with smaller datasets in the picture, the ML models were effectively used for computer vision tasks. But with the increase in the amount of data available we have seen a shift toward DL-based models that have proven to be very efficient in handling recognition tasks using these huge datasets. Effectively, there are many models available for Computer Vision tasks, this section gives a brief overview of some of these models. Table 4 provides a comparison of pre-trained image models. Many of these models like the VGG, ResNet-50 Inception V3, and Xception models are pre-trained on the ImageNet (contains 1.28 million images divided among 1,000 classes) for object detection tasks.
B. *AlexNET***:** Convolutional Neural Networks (CNNs) have typically been the model of choice for object recognition since they are powerful, easy to train, and control. AlexNET [132], a Deep convolutional network, consists of eight layers (having five convolutional layers and three fully-connected layers), and Rectified Linear Unit (ReLU) function as it does not suffer from the issue of vanishing gradient. It also combats overfitting by employing drop-out layers, in which a link is deleted with a probability of $p = 0.5$ during training. Apart from this, it allows a multi-GPU training environment that helps to train a larger model and even reduces the training time. AlexNet is a sophisticated model that can achieve high accuracies even on huge datasets however, its performance is compromised if any of the convolutional layers are removed.
C. *VGG Model*: The VGGNet (Visual Geometry Group Network) is a CNN model with a multilayered operation and is pre-trained on the ImageNet dataset. VGG [54] is available in two variants with 16 and 19 weight layers namely VGG-16 and VGG-19 respectively. These models are substantially deeper than the previous models and are built by stacking convolution layers but the model's depth is limited because of an issue called diminishing gradient which makes the training process difficult. To reduce the number of parameters in these very deep networks, a 3×3 convolution filter is used in all layers with stride set to 1. The model uses fixed 3×3 sized kernels that can reproduce all of Alexnet's variable-size convolutional kernels (11×11, 5×5, 3×3).
D. *GoogLeNet***:** GoogLeNet (Inception V1) [56] was the first version of the GoogLeNet architecture, which was further developed as Inception V2 and Inception V3. While larger kernels are preferable for global features distributed over a vast area of the image, while smaller kernels detect area-specific features that are scattered across the image frame efficiently, and hence choosing a set kernel size is a challenging task. To resolve the problem of recognition of a variable-sized feature, Inception employs kernels of varied sizes. Instead of increasing the number of layers in the model, it expands it by including several kernels of varied sizes within the same layer. The Xception architecture is a modification of the Inception architecture that uses depth-wise separable convolutions instead of the usual Inception modules.
E. *ResNET:* To address the issue of diminishing gradient in VGG models, the ResNET (Residual Network) [55] model was developed. The primary idea is to use shortcut connections to build residual blocks to bypass blocks of convolutional layers. CNN models can get deeper and deeper using Resnet models. There are many variations for Resnet models but Resnet50 and ResNet101 are used mostly.

B. *Transfer learning with Language data:* Word Embeddings use vector representations of words to encode the relationships between them. The pre-trained word vector is related to the meaning of the word and is one of the most effective ways to represent a text since it intends to learn both the syntactic and semantic meaning of the given word. Figure 10 provides the classification of Word Embeddings that is broadly divided into four categories depending upon (i) whether these can preserve the context or not as Context-independent and Context-dependent word embeddings, (ii) whether the underlying architecture of the model is RNN-based or Transformer based, (iii) the level at which the encoding is produced and (iv) whether the underlying task is supervised or unsupervised. An overview and comparative analysis of

Table 4 Pretrained Image Models

Network	Author(s), Year	Salient Features	Parameters	FLOP	Top 5 Accuracy
AlexNET	Krizhevsky et al. (2012)	Deeper	62 M	1.5B	84.70%
VGGNet	Simonyan et al. (2014)	Fixed-size kernel	138 M	19.6B	92.30%
Inception	Szegedy et al. (2014)	Wider parallel kernel	6.4 M	2B	93.30%
ResNET	He et al. (2015)	Shortcut connections	60.3 M	11B	95.51%

the different types of word embedding techniques categorized under Traditional, Static, and Contextualized word embeddings is provided in [133]. Pretrained Word Embedding is a type of Transfer Learning approach where embeddings learned in one task on a larger dataset are utilized to solve another similar job. These are highly useful in NLP tasks in a scenario where the training data is sparse and number of trainable parameters are quite large. [134] studies the utility of employing pretrained word embeddings in Neural Machine Translation (NMT) from a number of perspectives.

Some of the widely used text embeddings are discussed below (Fig. 10):

i. *One-hot vector representation:* It is one of the first and simplest word embeddings. It represents every word as an $\mathbf{R}^{|V| \times 1}$ vector with all 0 s and only a single 1 at the index of that particular word in the sorted English language, where |V| represents the vocabulary. Though one-hot vectors are easy to construct, but these are not a good choice to represent a large corpus of words as it does not capture the similarity between the words in the corpus.
ii. *Word2Vec:* Word2Vec is developed by Google and is trained on Google News Dataset. It is one of the widely used pre-trained word embeddings, takes text corpus as input, and generates word vectors as output. It is a Shallow Neural Network architecture that uses only one hidden layer in its feed-forward network. It learns vector representations of words after first constructing a vocabulary from the training text input. Distance tools like cosine similarity are used for finding the nearest words for a user-specified term. Depending on how the embeddings are learned, the Word2Vec model can be categorized into one of two approaches: Continuous Bag-of-Words (CBOW) model that learns the target word from the surrounding words, and the Skip-gram model that learns the surrounding words given the target word.
iii. *GloVe* is an unsupervised learning technique that generates word vector representations by leveraging the relationship between the words from Global Statistics. The training produces linear substructures of the word vector space, which are based on aggregated global word-word co-occurrence statistics from a corpus. The basic premise of the model is that ratios of word-word co-occurrence probabilities can contain some form of meaning. A co-occurrence matrix shows the frequency of occurrence of a pair of words together.
iv. *BERT* (Bidirectional Encoder Representations from Transformer) [57] has an advancement over Word-2Vec and generates dynamic word representations based on the context in which the word is being used rather than generating fixed representation like Word-2Vec. A polysemy word e.g., bank, can have multiple embeddings depending on the context in which the word is being used. This has brought context-depend-

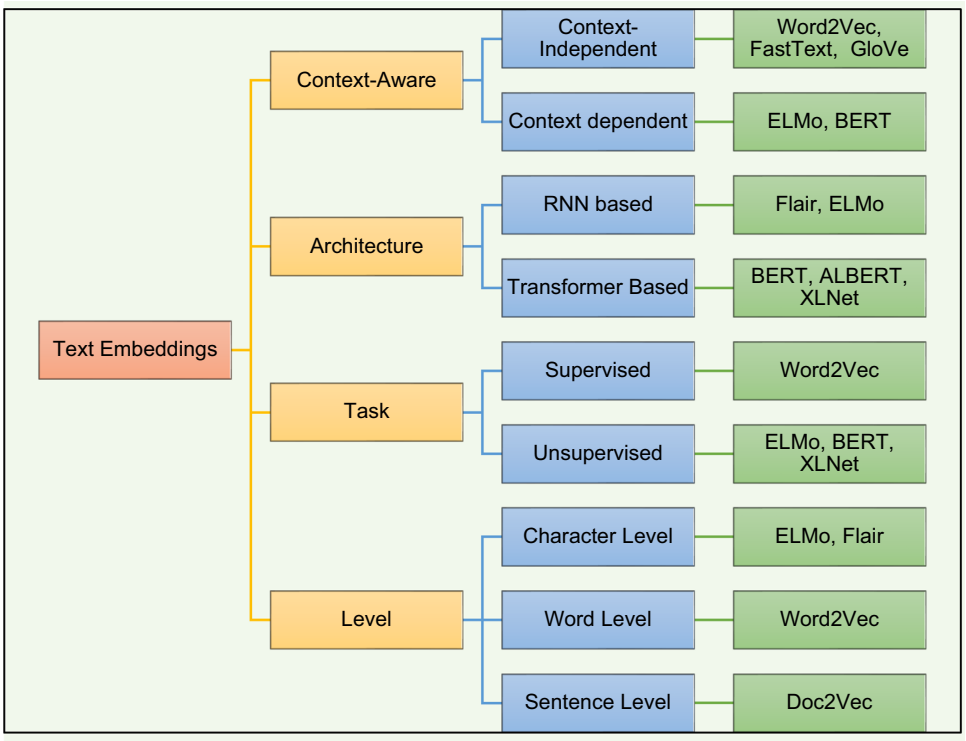


Fig. 10 Taxonomy of Word Embeddings

ent embeddings into the mainstream in present times. BERT is a pre-trained bidirectional transformer-based contextualized word embedding that can be fine-tuned as per the need. Since its introduction, many variants like RoBERTa (Robustly Optimized BERT Approach) [135] and Albert (A Lite BERT) [136] have been introduced to further enhance the state-of-the-art in language representations.

v. *ELMo:* Unlike traditional word-level embeddings like Word2Vec and GLoVe that have the same vector for a given word in the vocabulary, the same word can have distinct word vectors under varied contexts in the case of ELMo representation like the representation of BERT. The ELMo vector assigned to a word is a function of the complete input sentence containing that word.

vi. *XLNet* learns unsupervised language representations based on a novel generalized permutation language modeling aim. It fuses the bidirectional facility of BERT with the autoregressive technology of Transformer-XL.

Table 5 provides a comparison between various embeddings and highlights the advantages and disadvantages of each one of them.

3.3 Deep learning frameworks and libraries

Over the past few years, various detection methods have been proposed for solving the issue of fake news and rumors on online social media. Researchers are constantly working in these domains to find effective solutions and techniques. Deep learning is one of the several techniques that has become increasingly popular in solving problems in various domains. Neural networks such as CNN, RNN, LSTM are becoming increasingly popular. Although, using deep learning techniques complex tasks are performed easily compared to the machine learning counterparts but, successfully building and deploying them is a challenging task. Training a deep learning model takes a little longer when compared with traditional models but testing can be done rapidly. The deep learning frameworks are developed with an intention to accelerate and simplify the processing of the model. These frameworks combine the implementation of contemporary DL algorithms, optimization techniques, with infrastructure support. Figure 11 gives an overview of various DL/ML tools that are widely used to simplify the research problems.

Some of the ML/DL frameworks and libraries are discussed below:

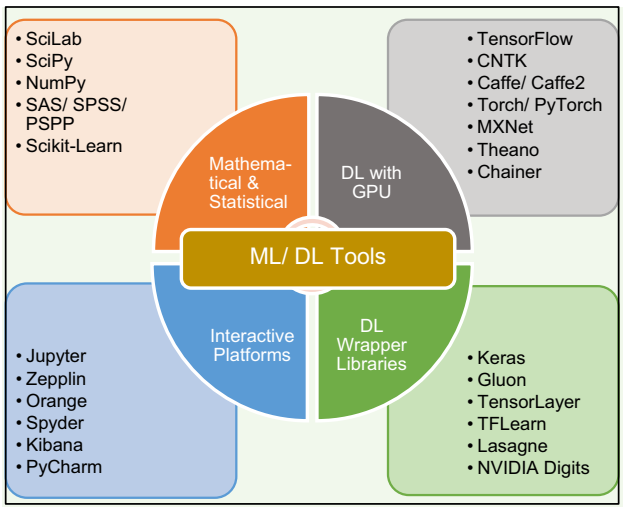


Fig. 11 Overview of ML/DL frameworks and libraries

Table 5 Comparison of various Text Embeddings

Embedding	Advantage	Weakness
Word2Vec	Consume much less space than one-hot encoded vectors Maintain semantic representation of word Capable of capturing multiple degrees of similarity between words using simple vector arithmetic	Can't handle OOV words No shared representation is used at subword level Scaling to new languages requires separate embedding matrices
GloVe	Can handle Out-Of-Vocabulary words	Gives random vectors to OOV words which confuses the model in long run
BERT	Creates contextualized vectors Learns representations at a "subword" (also called WordPieces) level	Computationally intensive Neglects dependency present between the masked positions Suffers from the pretrain-finetune inconsistency
ELMo	Generates contextualized word embeddings Can handle Out-Of-Vocabulary words	Complex Bi-LSTM structure makes train and embedding generation very slow Representing long-term context dependencies becomes difficult
XLNet	Provides autoregressive pretraining Enables bidirectional learning by maximizing the expected likelihood over all permutations of the factorization order	XLNet is pre-trained to capture long-term dependencies but can underperform on short sequences XLNet is generally more resource-intensive and takes longer to train and to infer compared to BERT

i. *TensorFlow:* TensorFlow is an open-source, end-to-end framework for deep learning published under the Apache 2.0 license. It is designed for large-scale distributed training and testing that can run on a single CPU system, GPUs, mobile devices as well as on distributed systems. It provides a large and flexible ecosystem of tools and libraries that allows developers to quickly construct and deploy ML applications. It uses data flow graphs to perform numerical computations. Model building and training with TensorFlow employs high-level Keras API and is quite simple as it offers multiple levels of abstraction, allowing researchers to choose the level of abstraction that best suits their needs.
ii. *Torch and PyTorch:* Torch is one of the oldest frameworks which provides a wide range of algorithms for deep machine learning. It provides a multi-GPU environment. Torch is used for signal processing, parallel processing, computer vision, NLP etc. Pytorch is an open-source Python version of Torch, which was developed by Facebook in 2017 and is released under the Modified BSD license. PyTorch is a library for processing tensors. It is an alternative to NumPy to use the power of GPUs and other accelerators. It contains many pre-trained models and supports data parallelism. It is one of the widely used Machine learning libraries, along with TensorFlow and Keras. It is particularly useful for small projects and prototyping.
iii. *Caffe and Caffe2:* Caffe is a general deep learning framework that is based on C++ launched with speed, and modularity in mind. It was developed by Berkeley AI Research (BAIR). Caffe is designed primarily for speed and includes support for GPU as well as Nvidia's Compute Unified Device Architecture (CUDA). It performs efficiently on image datasets but doesn't produce similar results with sequence modeling. Caffe2, open-sourced by Facebook since April 2017, is lightweight and is aimed towards working on (relatively) computationally constrained platforms like mobile phones. Caffe2 is now a part of PyTorch.
iv. *MXNet:* Apache MXNet is a flexible and efficient deep learning library suited for research prototyping and production. It works on multiple GPUs with fast context switching. MXNet contains various tools and libraries that enable tasks involving computer vision, NLP, time series etc. GluonCV and GlutonNLP are libraries for computer vision and NLP modeling, respectively. The Parameter Server and Horovod support enable scalable distributed training and performance optimization in research and production.
v. *Theano:* Theano is a python-based deep learning library developed by Yoshua Bengio at Université de Montréal in 2007. The latest version of Theano, 1.0.5, is Python 3.9 compatible. Theano is built on top of NumPy and helps to easily define, evaluate, and optimize mathematical operations involving multi-dimensional arrays. It can be run on the CPU or GPU, providing smooth and efficient operation. It offers its users with extensive unit-testing which aids in code debugging. Keras, Lasagne, and Blocks are open-source deep libraries built on top of Theano.
vi. *Chainer:* Chainer is a robust and flexible deep learning framework that supports a wide range of deep networks (RNN, CNN, RvNN etc.). It supports CUDA computation and uses CuPy to leverage a GPU computation. Parallelization with multiple GPUs is also possible. Code debugging with Chainer is quite easy. It provides two DL libraries i) ChainerRL that implements a variety of deep reinforcement algorithms, and ii) ChainerCV which is a Library for Deep Learning in Computer Vision.
vii. *Computational Network Toolkit (CNTK):* The Microsoft Cognitive Toolkit (CNTK) is an open-source toolkit, since April 2015, for providing commercial-grade distributed deep learning services. It is one of the first DL toolkits that supports the Open Neural Network Exchange (ONNX) format for shared optimization and interoperability. The newest release of CNTK, 2.7., supports ONNX v1.0. With CNTK neural networks are represented as a series of computational steps using a directed graph. It allows the user to easily develop and deploy various NN models such as DNN, CNN, RNN etc. CNTK can either be included as a library in Python, and C++ code, or can be used as a standalone ML/DL tool through BrainScript (its own model description language).
viii. *Keras:* Keras is Python wrapper library for DL written in Python, and runs on top of the ML/DL platforms like TensorFlow, CNTK, Theano, MXNet and Deeplearning4j. Given the underlying frameworks, it runs on Python 2.7 to 3.6 on both GPU as well as on CPU. It was launched with a prime focus on facilitating fast experimentation and is available under the MIT license.
ix. *TFLearn:* TFlearn is a modular and transparent deep learning library built on top of Tensorflow and facilitates and speed-up experimentations using multiple CPU/GPU environment. All functions are built over tensors and can be used independently of TFLearn.
x. *TensorLayer*: TensorLayer is a deep learning and reinforcement learning library built on top of TensorFlow framework. Other TensorFlow libraries including Keras and TFLearn hide many powerful features of TensorFlow and provide only limited support for building and training customized models.

Table 6 shows some popular DL frameworks (such as Keras, Caffe, PyTorch, TensorFlow, etc.) along with their

Table 6 Comparison of popular Deep Learning Frameworks

Software	Platform	Written in	Interface	Open MP support	Open CL support	CUDA support	RNN	CNN	Has pre-trained Models
TensorFlow	Windows, Linux, macOS, Android	Python, C++, CUDA	C/C++, R, Python (Keras), Java, JavaScript	×	via SYCL support	✓	✓	✓	✓
PyTorch	Windows, Linux, macOS, Android	C/C++, Python, CUDA	Python, C++	✓	Via separately maintained package	✓	✓	✓	✓
Caffee	Linux, macOS, Windows	C++	Python, C++ MATLAB,	✓	Under development	✓	✓	✓	✓
Theano	Cross-platform	Python	Python (Keras)	✓	Under development	✓	✓	✓	Through Lasagne's model zoo
Chainer	Linux, macOS	Python	Python	×	×	✓	✓	✓	✓
MXNet	Linux, AWS macOS, iOS, Windows, Android, JavaScript	Small C++ core library	C++, Python, MATLAB, JavaScript, Scala, Perl, R	✓	On roadmap	✓	✓	✓	✓
Microsoft Cognitive Toolkit (CNTK)	Linux, Windows, macOS (via Docker on roadmap)	C++	Python (Keras), C++ Command Line	✓	×	✓	✓	✓	✓

comparative analysis. These frameworks and libraries are implemented in Python, are task-specific and allow researchers to develop tools by offering a better level of abstraction

Several machine learning and deep learning frameworks have emerged in the last decade, but their open-source implementations appear to be the most promising for several reasons: (i) openly available source codes, (ii) a large community of developers, and, as a result, a vast number of applications that demonstrate and validate the maturity of these frameworks.

3.4 Review of state-of-the-art multimodal frameworks

With the rapid expansion of social media platforms, news content has evolved from traditional text-only articles to multimedia articles involving images and videos that carry richer information. Multimodal articles have the power to engage more readers as compared to the traditional text-only articles as the images and videos attached to these articles make them more believable. Most of the online users get affected by such information, unknowingly spread the misinformation, and become a part of this whole vicious network.

Traditionally, the great majority of methods for identifying false news have focused solely on textual content analysis and have relied on hand-crafted textual features to do so. However, with the growing quantity of articles on the Internet that include visual information and the widespread usage of social media networks, multimodal aspects are becoming increasingly important in understanding the overall intent of the content in a better way.

Given the contents of a news claim $\mathbb{C}$ with its text set T and image set I, the task of multimodal fake news detector is to determine whether the given claim can be considered as true or fake, i.e., to learn a prediction function $\mathcal{F}(\mathbb{C}) \rightarrow 0, 1$ satisfying:

$$\mathcal{F}(\mathbb{C}) = \begin{cases} 1, & \text{if } \mathbb{C} \text{ is confirmed to be fake} \\ 0, & \text{otherwise} \end{cases}$$

The following figure, Fig. 12, presents a general framework that depicts various channels present in a multimodal fake news detection (MFND) framework. The framework illustrates how the features are extracted individually and then merged to detect the credibility of the claim.

Some multimodal FND frameworks, apart from fusing textual and image data, also evaluate the similarity between

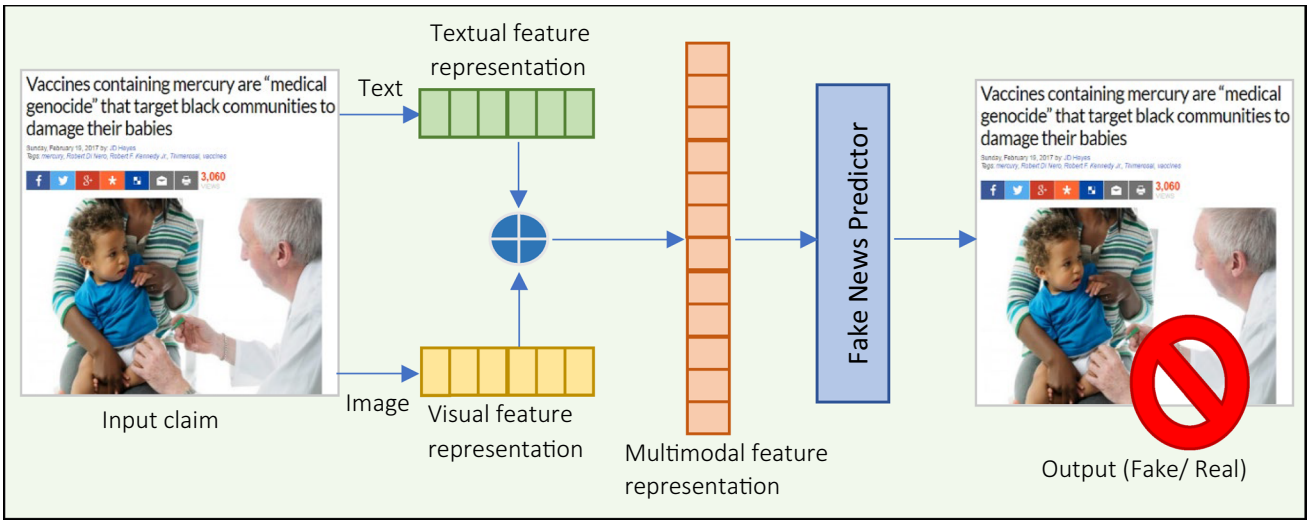


Fig. 12 A general framework for Multimodal Fake News Detection

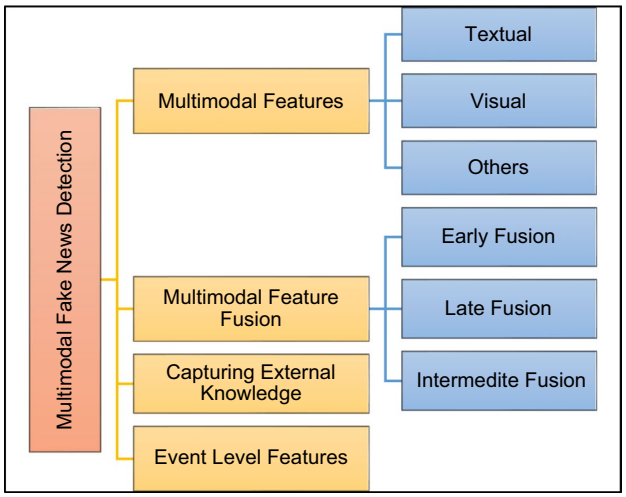


Fig. 13 Various components of Multimodal Fake News Detection

the two [97], or have used external knowledge and event-specific information to check the credibility of a given news article. Others have also used social context including user data[25], propagation structure, sentiments[98], and other auxiliary information to effectively combat and detect fake news.

Figure 13 depicts the taxonomy of a MFND framework by focusing on the various techniques used in processing the individual channels.

A. *Multimodal Features:* With the increasing use of social media, a shift from all text to a multimodal news article can be seen. Now, the news articles comprise images and videos along with the text. The models for MFND have used two different channels for handling the text and image data. The section below summarizes how different models have exploited various techniques for the same.

i. *Text channel:* To capture the context from textual data, the researchers have used various word embeddings and pre-trained model, as discussed in Sect. 3.2. Word2Vec is one of the popular word embedding that is used by [4, 25, 27, 91]. But as word2vec can't handle out-of-vocabulary words, researchers have exploited Glove, BERT, XLNet and other embeddings instead [26, 28, 92, 95, 96, 98, 110, 129]. FND-SCTI [4] considers the hierarchical document structure and uses Bi-LSTM at both word-level (from word to sentence) and sentence-level (from sentence to document) to capture the long-term dependencies in the text. Ying et al. [92] proposed a multi-level encoding network to capture the multi-level semantics in the text. The model KMAGCN [99] proposed by Qian et al. captures the non-consecutive and long-range semantic relations of the post by modeling it as a graph rather than a word sequence and proposes a novel adaptive graph convolutional network handle the variability in the graph data.

ii. *Visual channel:* The Visual sub-network uses the article's visual information as input to generate the post's weighted visual features. The image is initially resized (usually to 224×224 pixel size), after which it is placed into a pre-trained model to extract the image features. The visual channel in the framework captures the manipulation in image data using pre-trained models. VGG-19 is the most widely used model, apart from this VGG-16, ResNet50 are also utilized. [97] uses image2sentence to represent news images by

generating image captions. [110] uses Image forensic techniques—Noise Variance Inconsistency (NVI) and Error Level Analysis (ELA) for the identification of manipulated images. [26, 96] uses the bottom-up attention pre-trained ResNet50 model to extract region features for every image attached with the article. As an article sometimes comes with multiple images, Giachanou et al. in [94] propose visual features that are extracted from multiple images.

iii. Apart from the visual and textual channels some of the researchers have also focused on other aspects of a social media post that might be helpful in detecting fake news. Cui et al. [98] proposed an end-to-end deep framework, named SAME that incorporates user sentiment extracted from users' comments (with VADER -sentiment prediction tool) with the multimodal data. Experiments on PolitiFact and GossipCop shows F1 score of 77% (approx.) and 80%(approx.) which is better than the baseline methods. User profile, network, and propagation features are another set of vectors that are highly exploited [25].

B. *Multimodal Feature Fusion:* Dealing with multimedia data comes with an intrinsic challenge of handling the data of varied modalities while keeping intact the correlation between them. In a multimodal social media post, finding the correlation between text and image is an important step in identifying a fake post. There are three techniques (Fig. 14) that are widely used for multimodal data fusion.

i. *Early fusion/ Data-level Fusion* first fuses the multimodal features and then applies the classifier on the combined representation; The data-level fusion of multimodal features starts with feature extraction of unimodal features and after analysis of the different unimodal feature vectors these features are combined into a single representation. With early fusion as the features are integrated from the start, a true multimedia feature representation is extracted. There are various methods like principal component analysis (PCA), canonical correlation analysis (CCA), independent vector analysis (IVA) and independent component analysis (ICA) which are used to accomplish this task [137]. One of the main disadvantages of this approach is the complexity to fuse the features into a common joint representation. Also, rigorous filtering of data is needed to make a common ground before fusion which poses a challenge if the dataset available is already limited in number.

ii. *Late fusion/ Decision-level Fusion* combines the results obtained from different classifiers trained on different modalities; Late fusion also starts with extracting the unimodal features. But in contrast to early fusion, the late fusion approach learns semantic concepts separately from each unimodal channel and different models are available to determine the optimal approach to combine each of the independently trained models. It is based on the ensemble classifier technique. This method gives the flexibility to concatenate the input data streams that significantly varied in terms of the number of dimensionality and sampling rate. Fusing the features at the decision-level is expensiveness in terms of the learning effort as separate models are employed for each modality. Furthermore, the fused representation requires an additional layer of the learning stage. Another disadvantage is the potential loss of correlation in fused feature space.

iii. *Intermediate Fusion* allows the model to learn a joint representation of modalities by fusing different modalities representations into a single hidden layer. Intermediate fusion changes input data into a higher level of representation (features) through multiple layers and allows data fusion at different stages of model training. Each individual layer uses various linear and non-linear functions to learn specific features and generates a new representation of the original input data.

Several research works in the literature have come up with various techniques, [24] presents a neuron-level attention mechanism for aligning visual features with a joint representation of text and social context, and as a result, greater weights are assigned to visual neurons with semantic meanings related to the word. FND-SCTI proposed in [4] uses a hierarchical attention mechanism to put an emphasis on the important parts of the news article. Giachanou et al. in [94] propose the use of cosine similarity to find the image-text similarity between the title and image tags embeddings. To preserve semantic relevance and representation consistency across different modalities [98] uses an adversarial mechanism. To filter out the noise and highlight the image regions that are strongly related to the target word, [16] uses a word-guided visual attention module. To learn complementary inter-dependencies among textual and visual features [26, 92, 96, 99, 129] uses multiple co-attention layers, hierarchical multi-modal contextual attention network, feature-level attention mechanism, blended attention module and multimodal cross-attention network respectively. The proposed the Crossmodal Attention Residual Network (CARN) in [111] can selectively extract information pertaining to a target modality from another source modality while preserving the target modality's unique information. Another model SAFE [97], jointly learns the text and image features and

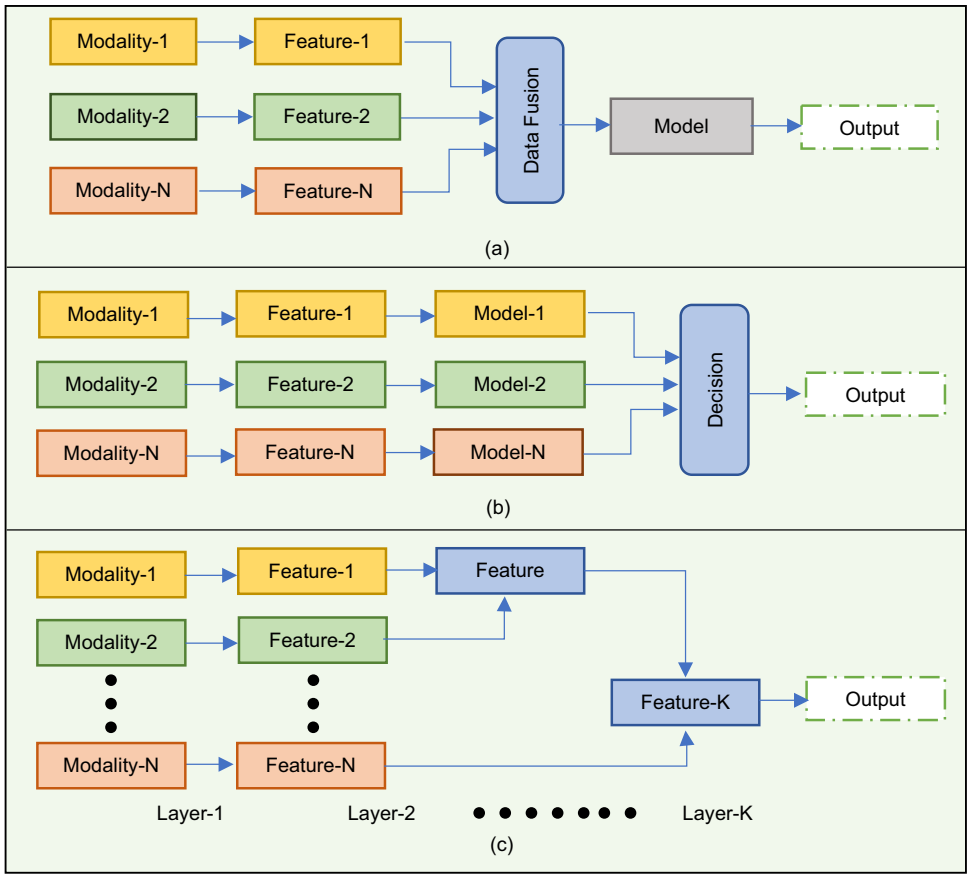


Fig. 14 General schemes for multimodal fusion (a) Early Fusion, (b) Late Fusion, (c) Intermediate Fusion

also learns the similarity between them to evaluate whether the news is credible or not (Fig. 14).

III. *Capturing External Knowledge channel:* The Knowledge module tries to capture background knowledge from a real-world knowledge graph to supplement

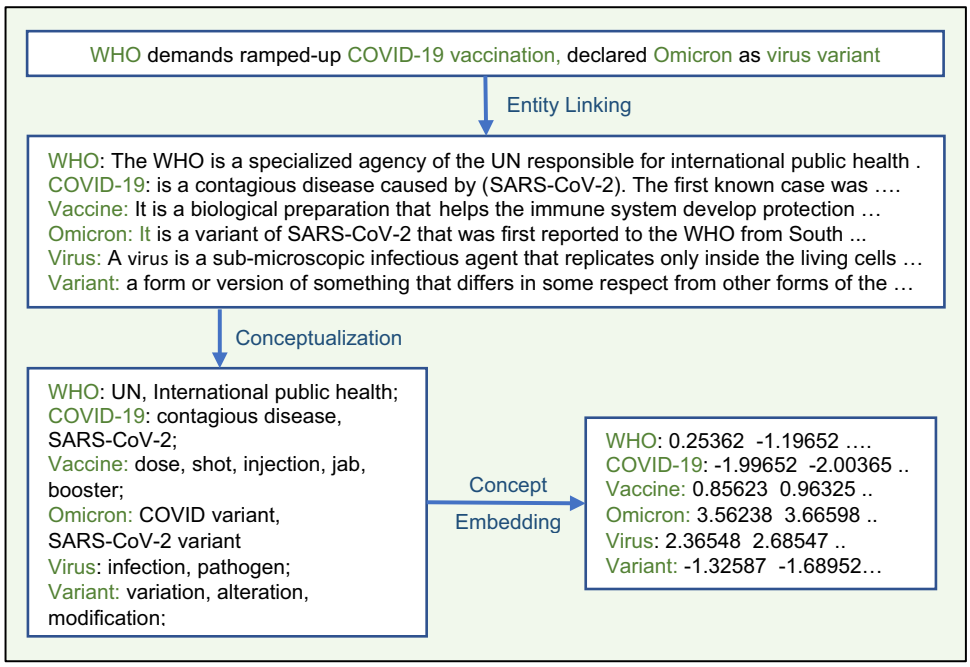


Fig. 15 Illustration of Knowledge Distillation process [16]

the semantic representation of short text posts. An illustration of the knowledge distillation process is presented in Fig. 15. Given a post, [16, 99] utilize the entity linking technique to associate the entity mentions in the text with pre-defined entities in a knowledge graph. Rel-Norm [138], Link Detector [139], and STEL [140] are widely used entity linking techniques. To gain the conceptual data of each distinguished entity from an existing knowledge graph, YAGO [141, 142], and Probase [143] can be exploited.

IV. *Event-level Features:* Individual microblog posts are short and have very limited content. An event generally contains various related posts relevant to the given claim. Detecting fake news at the event level comprises of predicting the veracity of the whole event instead of an individual post. Most of the existing MFND frameworks work at post level [4, 27, 28, 92, 95], and learn event-specific features that are not useful in predicting the unseen future events and suffer from generalization. But [15, 16] intends to accurately categorize the post into one of the K events based on the multimodal feature representations. [96] incorporates topic memory module that captures topic-wise global feature and also learns post representation shared across topics. The Event Adversarial Neural Network (EANN) model proposed in [15] captures the dissimilarities between different events using an event discriminator. The role of the event discriminator is to eliminate the event-specific features and learn shared transferable features across various events. This model is evaluated on two multimedia datasets extracted from Twitter and Weibo and shows an accuracy of 71% and 82% respectively which is better than the baseline models. Ying et al. offers a unique end-to-end Multi-modal Topic Memory Network (MTMN) [115] that captures event-invariant information by merging post representations shared across global latent topics features to address real-world scenarios of fake news in newly emerging posts.

Table 7 gives the comparative analysis of various existing state-of-the-art deep learning-based multimodal fake news detection models focusing on the techniques used for processing individual channels and the methods used for concatenating the features for presenting a combined feature space. The table also discusses about the future perspective. In addition to this, Table 8 provides a detailed analysis of the experimental setup of the DL-based MFND frameworks. These tables provide some ideas about how one can practically approach the problem of fake news.

4 Data collection

Before starting the data collection process, the developers must decide upon the size of the dataset, news domain (e.g., entertainment, politics etc.), media (e.g., texts, images, videos, etc.), type of disinformation (e.g., fake news, propaganda, rumors, hoaxes, etc.) in advance. Various datasets for the task of FND have been developed by various studies and these vary in terms of the news domain, size, type of misinformation, content type, rating scale, language, and media platform.

[44] has laid down the requirements for fake news detection Corpus (Fig. 16). The study in [70] has introduced and divided the requirements of FND dataset into four categories, namely Homogeneity requirements, Availability requirements, Verifiability requirements, Temporal requirements (Fig. 17).

This section describes various data collection and annotation strategies apart from the datasets that are available for the given task.

4.1 Existing datasets

Several datasets are available for FND and related tasks like LIAR, CREDBANK, FEVER but most of these are text-only data. There are only a few datasets that have text along with visual data. The authors in [70] have systematically reviewed and done a comparative analysis of twenty-seven popular FND datasets by providing insights into existing dataset. Table 9 gives an overview of the multimodal datasets that are available and are widely used in the study.

Despite the fact that multimodal fake news datasets are available, but these datasets still have some shortcomings. The above table clearly shows that the multimodal datasets that are available are coarse-grained (have 2 or 3 labels only). These datasets fail to acknowledge fine-grained labeling that may be found in datasets like, LIAR [106], but these datasets are unimodal and can't be used in a multimodal setting.

Even though FakeNewsNet [120] dataset is one of the newest benchmark datasets that contains content, social context as well as socio-temporal data, but is not available as whole and only subsets of dataset can be retrieved using APIs provided. This is due to the fact that the dataset uses Twitter data for capturing user engagement, and so is not entirely publicly accessible according to license regulations.

While some datasets incorporate data from various domains, the existing multimodal datasets focus on limited domains. One of the main reasons that contribute to the compromised performance is due to the fact that the existing datasets only focus on specific topics like politics, entertainment etc. and hence have domain-specific word

Table 7 Review of literature of existing multimodal fake news frameworks

Model, Ref	Contribution	Feature (s) used	Feature Extraction		Multi-modal Fusion	Activation function	Future Scope
			Text	Visual			
Att-RNN [25]	Fuses the textual and visual features along with the social context using attention mechanism	M, SC	Word2Vec	VGG-19	Attention network	Sigmoid, Softmax, ReLU, Tanh	To improve the proposed model's performance
EANN [15]	Uses event discriminator to discover event-specific data to enhance the detection efficiency on new events	M, ES	Text-CNN	VGG-19	Dense layer (Concatenation)	Softmax, ReLU	To improve the fusion network
SAME [98]	Fuses the multimodal information along with user sentiment using adversarial learning	M, SC	GloVe	VGG-19	Adversarial network	Softmax, ReLU	Early detection
MKEMN [16]	The model gathers the event-invariant features shared between different events and captures the external knowledge connections for effective news verification.	M, EK, ES	GloVe	VGG-19	CNN	Softmax, Tanh	Use memory network to exploit the rumor propagation information
MVAE [27]	Model discovers correlations across the modalities leveraging VAE	M	Word2Vec	VGG-19	Concatenation	Tanh, Softmax	Utilizing tweet propagation and user characteristics.
SpotFake [28]	The prime novelty of the proposed model is the use of pre-trained language model BERT	M, SC	BERT	VGG-19	Dense layer (Concatenation)	Sigmoid, ReLU	improvement on longer length articles
SpotFake+ [95]	The prime novelty of the proposed model is the use of the pre-trained language model XLNet	M	XLNet	VGG-19	Dense layer (Concatenation)	Sigmoid, ReLU	Incorporate meta-level feature modalities
MTMN [96]	Address early detection by fusing features shared by various topics with global features of latent topics and modeling intra-modal and inter-modal data in a combined framework	M, ES	BERT	ResNET50	Blended Attention network	Softmax	Explore effective ways to learn background knowledge

Table 7 (continued)

Model, Ref	Contribution	Feature (s) used	Feature Extraction		Multi-modal Fusion	Activation function	Future Scope
			Text	Visual			
SAFE [97]	Produces joint representation of textual and visual features of an article and uses cosine function to measure the similarity between them	M	Text-CNN	Text-CNN (image2sentence)	Cosine Similarity	Softmax, ReLU	Incorporate user and network information
- [94]	proposes visual features that are extracted from multiple images, additionally, cosine similarity of the title and image tags embeddings is calculated to find the image-text similarity	M	BERT	VGG-16	Attention network	Softmax	Improve the performance of the proposed model
MCAN [129]	Proposes multiple co-attention layers that fuse and learn inter-modality relations	M	BERT	VGG-19	Multiple co-attention layers	ReLU	To extend the fusion with the co-attention network to fake news diffusion.
HMCAN [26]	Propose a multi-modal contextual attention network that takes data from different modalities which complement one another	M	BERT	ResNET50	Attention network	Softmax	Explore an effective way to exploit visual data and utilize auxiliary information
KMAGCN [99]	The model represents posts as graphs instead as word sequences to capture long-range non-consecutive semantic relations and leverages knowledge concepts along with multimodal information	M, EK	Adaptive graph convolutional network	VGG-19 (128-D)	Feature-level attention mechanism	Softmax, ReLU	To improve the proposed model's performance
CARMN [91]	The model keeps unique properties intact while reducing the noise induced while fusing different modalities	M	Word2Vec (32-D)	VGG-19	Multichannel CNN	Softmax, ReLU	Event-level multimodal fake news

Table 7 (continued)

Model, Ref	Contribution	Feature (s) used	Feature Extraction		Multi-modal Fusion	Activation function	Future Scope
			Text	Visual			
FND-SCTI [4]	Fuses multi-modal data along with an image-augmented text representation in a multi-task setting	M	Word2Vec	VGG-19	Hierarchical attention network, VAE	tanh, Softmax	Bot detection using user characteristics
- [110]	The proposed system consists of four independent parallel networks with individual predictions that are merged with the max voting ensemble method	M	GloVe	Image caption (CaptionBot)	Ensemble with max voting	Softmax, Tanh, ReLU	Incorporating better image forensic techniques
MMCN [92]	Fuses the text and image embeddings by considering inter-modal relationships using a multi-modal cross-attention network	M	BERT	ResNET50	Cross Attention network	Softmax	To explore effective ways to utilize background knowledge

M Multimodal, *SC* Social Context, *EK* External Knowledge, *ES* Event Specific features

Table 8 Experimental Setup of Multimodal Fake News Detection Models

Model	Ref.	Dataset	Batch size	Learning rate	Dropout	Epochs	Optimizer	Loss function	Performance Evaluation
Att-RNN	[25]	Twitter16, Weibo	128	–	–	100	Stochastic gradient descent	Cross Entropy	Acc- ~78%, ~68%
EANN	[15]	Twitter15, Weibo	100	–	–	100	–	Cross Entropy	Acc- ~71%, ~82%
SAME	[98]	FakeNewsNet	128	0.001	0.5	–	RMSprop	Adversarial, Hybrid similarity, Cross entropy	Acc- ~77%, ~80%
MKEMN	[16]	Twitter15, PHEME	128	–	–	–	–	Cross Entropy	Acc- ~86%, ~81%
MVAE	[27]	Twitter15, Weibo	128	0.00001	–	300	Adam	VAE Loss	Acc- ~74%, ~82%
SpotFake	[28]	Twitter15, Weibo	256	0.0005, 0.001	0.4	–	Adam	–	Acc- ~72%, ~80%
SpotFake+	[95]	FakeNewsNet	–	–	0.4	–	–	–	Acc- ~84%, ~85%
SAFE	[97]	FakeNewsNet	–	–	–	–	–	Cross Entropy	Acc- ~87%, ~83%
–	[94]	FakeNewsNet	32	0.00005	0.2	60	Adam	–	F1 score- ~76%
MCAN	[129]	Twitter16, Weibo	–	–	–	100	Adam	Cross Entropy	Acc- ~80%, ~89%
HMCAN	[26]	Weibo, Twitter15, PHEME	256	0.001	–	150	Adam	Cross Entropy	Acc- ~85%, ~89%, ~88%
KMAGCN	[99]	Weibo, Twitter15, PHEME	128	0.01	–	300	Adam	Cross Entropy	Acc- ~84%, ~78%, ~86%
CARMN	[91]	Twitter16, Weibo	150	–	–	150	Adam	Cross Entropy	Acc- ~74%, ~85%
FND-SCTI	[4]	Twitter15, Weibo	128	0.00001	–	300	Adam	VAE Loss	Acc- ~75%, ~83%
–	[110]	AllData, Kaggle datasets	32	–	–	40	Adam	Cross Entropy	Acc- ~95%, ~95%, ~95%
MMCN	[92]	Weibo, PHEME	64, 256	0.001	–	150	Adam	Cross Entropy	Acc- ~87%, ~87%
MTMN	[96]	Weibo, PHEME	256	0.001	–	200	Adam	Cross Entropy	Acc- ~88%, ~88%

usage, whereas a real news stream typically covers a wide range of domains.

4.2 Dataset collection and annotation

Social media platforms are one of the best ways to reach the masses. Everyday a huge volume of data is circulated, but as most of the data goes unchecked and so the proliferation of fake news on these platforms becomes easy. Many researchers have used these platforms to collect data and create their datasets. But as only a few datasets are available which are suited to our requirement, this section gives a clear understanding of the collection of data from online platforms. Broadly speaking there are two ways of data collection and sampling from online platforms, namely (i) top-down approach, and (ii) bottom-up approach.

The top-down approach involves a collection of information and posts keeping some keywords under consideration. This approach is useful for debunking long-standing rumors that are already known, the data is crawled from fact-checking websites like Snopes, PolitiFact etc. [59] uses this approach for data collection. But, on the contrary, if emerging news articles are to be collected the bottom-up approach comes in handy and it collects all the relevant posts and articles in a given time frame. [12, 145] have used the bottom-up approach.

Collecting reliable datasets for FND is not a trivial task and requires the fact checking of news to label and rate the news items (as binary or multiclass rating scale). As the ground truth is already known if data is collected using the top-down approach, no further annotation is needed. But, in the case of the bottom-up approach annotation is necessary and can be performed in one of the three ways (1) manual expert-oriented fact checking; (2) automated fact checking via knowledge graphs and other web sources; or (3) crowd-sourced fact checking.

Web APIs are one of the simplest ways to access, collect, and store data from social networking networks, and they usually come with documentation that explains how to get the required data. An application can use the APIs to request data using a set of well-defined methods. For example, retrieving all the data posted by some particular user or all the posts containing a specific keyword from a social media platform APIs are used. Twitter and Sina Weibo are two key

Table 9 Multimodal Fake News datasets

Dataset	Year of release	Statistics	Domain	Contents	Labels	Collected from	Used in
Twitter 15 [144]	2015	361 (I) 7032 (F) 5008 (R)	Posts related to 11 events	Text, visual	2	Twitter	[4, 15, 26–28, 99]
Twitter 16 [89]	2016	413 (I) 9596 (F) 6225 (R)	Posts related to 17 events	Text, visual	2	Twitter	[25, 91, 111, 129]
Weibo [25]	2016	9528 (I) 4749 (F) 4779 (R)	Crawl the verfi ed false rumor posts from May, 2012 to Jan, 2016	Text, visual	2	Weibo (Non-rumor tweets are verifi ed by Xinhua News Agency, an authoritative news agency in China)	[4, 15, 25–28, 91, 91, 99]
PHEME [12]	2016	2672 (I) 1972 (F) 3830 (R)	9 different events, which include 5 cases of breaking news	Tweet, conversational threads	3	Twitter	[16, 92, 96, 99]
ALLData [100]	2018	20,015 (I) 11,941 (F) 8074 (R)	2016 US Presidential elections	The title, text, image, author and website	2	Fake and real news scraped from 240 websites and authoritative news websites, i.e., the New York Times, Washington Post, etc. respectively	[100, 110, 111]
FakeNewsNet [120]	2019	19,200 (I) 5367 (F) 17,222 (R)	Politics, Entertainment	Text, image url, conversational threads, location, and timestamp of engagement	2	Content is crawled from PolitiFact, GossipCop, E! online; For user engagements Twitter API is used	[94, 95, 97, 98]

Note: I—Total Number of Images, F—Number of Fake claims, R—Number of Real claims

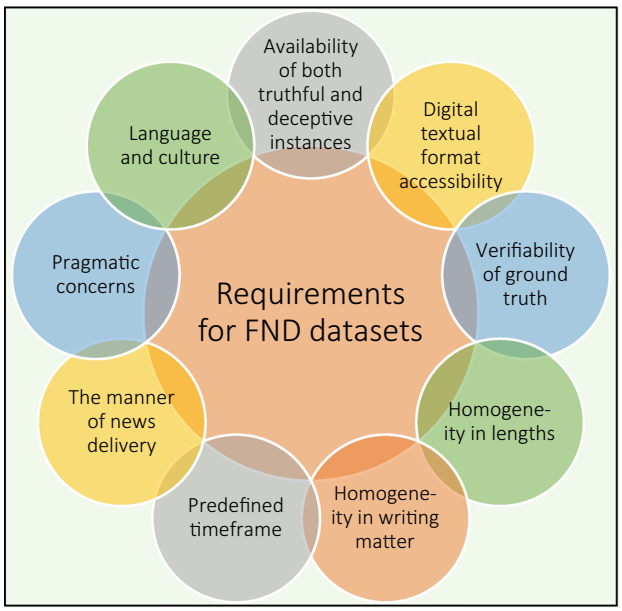


Fig. 16 Requirements for fake news detection datasets defined by [44]

platforms that have been widely explored for studying fake news. [61, 146] have provided and given an overview of several ways of data collection from social media sites. Twitter provides REST APIs to allow users to interact with their service but the user should have a developer account. Tweets based on particular topics can be extracted in real-time from Twitter API using the Tweepy library. Similarly, Sina Weibo provides an API to help users access their data. Apart from this, the data can be also scaped from any website using web crawlers. Beautiful Soup is one of the Python libraries for pulling data out of HTML and XML files.

5 Open issues and future direction—discussion

In this section, we discuss several challenging problems of fake news detection (Fig. 18). To avoid any further spread of fake news on social media, it is particularly important

Homogeneity Requirements
- homogeneity of news length: news artiles in the dataset should be of comparable lengths;
- homogeneity in news domain: the text corpus should be aligned with the news domains;
- homogeneity in type of disinformation: the texts in the dataset must be aligned with the type of misinformation

Availability Requirements
- availability of fake and real news: dataset should be balanced in terms of the composition of True and fake articles
- textual format availability: texts and textual transcriptions of audio and video should be available
- public availability: the dataset should be openly available
- multilingual: the availability of news items in multiple languages;

Verifiability Requirement
- verifiability of ground truth: the ability to identify whether the news is genuine or fabricated;

Temporal Requirements
- belongs to a predefined timeframe: the dataset should be collected within a set time window.

Fig. 17 Requirements for fake news detection datasets as defined by [70]

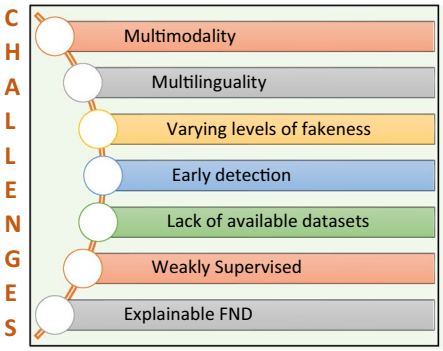


Fig. 18 Fake news detection challenges

to identify fake news at an early stage. Because getting the ground truth and labeling the false news dataset is time-consuming and labor-intensive, investigating the problem in a weakly supervised scenario, i.e. with few or no labels for training, becomes essential. In addition, there has been relatively little research into the multimodal nature of fake news on social media. It's also important to understand why a piece of news is labeled as fake by machine learning models because the resulting explanation can provide fresh data and insights that remain hidden when using content-based models.

There are several challenges associated with the FND problem. After extensive study and evaluation of the literature that is available and has been discussed in Sect. 2 and Sect. 3, the following research gaps have been identified.

i. *Multimodality:* With the rise of social media, there has been a change from all text to multimodal news articles. Images and videos are now included in news pieces, to complement the textual content. As a result, a model that can work with a multi-modal dataset must be built. Although several models have been established to combat such a problem, there has been very little research on the multimodal nature of fake news on social media, as working in a multimodal setting is difficult in and of itself. While some of the offered approaches have been successful in detecting fake news, they still face the challenge of determining the relationships between multiple modalities.

ii. *Multi-linguality:* Most of the existing works in this domain focus on detecting fake news in the English language, very limited work is done on multi-lingual and cross-lingual fake news detection models. Social media platforms are used by a huge population all over the world and hence are not limited to the usage of one language. Also, collecting and annotating fake articles in foreign languages is difficult and time-consuming.

iii. *Varying levels of fakeness:* The majority of existing fake news detection methods tackle the problem from a binary standpoint. However, in practice, a piece of news might be a mixture of factual and false statements. As a result, it's critical to divide fake news into several categories based on the degree of deception. Nevertheless, for multiclass fake news detection, the classifier needs to offer better discriminative power and be more robust as the boundary between classes becomes more intricate as the number of classes increases.

iv. *Early detection:* Fake news or rumor has a very negative impact on the people and society at large. This kind of propaganda can even tarnish the image of a person or organization, so it becomes very crucial to track down a piece of fake news or rumor at an early stage. Fake news early detection aims to curb fake sto-

ries at an early stage by giving prompt signals of fake news during the diffusion process.

v. *Lack of available datasets:* The data source is also a challenge. Unstructured data contains a lot of unnecessary data and junk values that can degrade the algorithm's performance. The number of datasets in the subject of fake news is quite restricted, and only a few are available online. The multi-modal dataset present in this problem domain is scarce. A comparison of major benchmark multimodal fake news datasets is shown in the Table 9 in the above section.

vi. *Weakly supervised fake news detection:* Labels are predicted with low or no supervision labels in a weakly supervised environment. Because getting the ground truth of false news is time-consuming and labor-intensive, it's crucial to investigate fake news detection in a weakly supervised scenario, that is, with few or no labels for training.

vii. *Explainable fake news detection:* The problem of detecting fake news has yielded promising breakthroughs in recent years. However, manually classifying fake news is quite subjective, and there is a vital missing aspect of the study that should explain why a specific item of news is considered to be fake by providing the reader web-based proof. Traditionally, it used manual methods to verify the news content's validity with a variety of sources.

6 Conclusion

In this paper, we have presented an overview of contemporary state-of-the-art techniques and approaches to resolve the issue of detecting fake news on social media platforms with a focus on multimodal context. Our review is primarily concentrated on five key aspects. First, the paper provides a clear definition of Fake News and distinctions between various related terms with an appropriately defined taxonomy of the fake news detection techniques. While surveying we found out that limited work has been done on the multimodal aspect of the news content. Secondly, various DL models, frameworks, libraries, and transfer learning approaches that are widely used in the literature have also been emphasized with TensorFlow being the one that is widely used. Third, we have provided an impression of various state-of-the-art techniques to perform fake news detection on social media platforms using deep learning approaches considering the multimodal data. The review shows that CNN-based models are widely used for handling the image data and the RNN-based models are used for preserving the sequential information present in the text. Additionally, various modifications of the attention network are used to preserve the correlation between text and image. These models take English as their primary language for detection and lack in processing the multi-lingual data which is prevalent with the use of social media. Fourth, the review sheds light on various data collection sources and data extraction techniques. Since this field of research is quite novel, there are only a few multimodal datasets that are available for this particular task. Finally, we have provided some insights to open issues and possible future directions in this area of research and found out that handling the multimodal data while maintaining the correlation becomes a challenge.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. S. Singhania, N. Fernandez, and S. Rao, "3HAN: A Deep Neural Network for Fake News Detection Sneha," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 10635 LNCS, no. October, pp. 118–125, 2017, https://doi.org/10.1007/978-3-319-70096-0_59.
2. H. Ahmed, I. Traore, and S. Saad, "Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques," *First Int. Conf. Intelligent, Secur. Dependable Syst. Distrib. Cloud Environ.*, vol. 10618, pp. 169–181, 2017, https://doi.org/10.1007/978-3-319-69155-8.
3. H. Karimi and J. Tang, "Learning hierarchical discourse-level structure for fake news detection," *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, pp. 3432–3442, 2019, https://doi.org/10.18653/v1/n19-1347.
4. Zeng, J., Zhang, Y., Ma, X.: Fake news detection for epidemic emergencies via deep correlations between text and images. Sustain. Cities Soc. **66**, 102652 (2021). https://doi.org/10.1016/j.scs.2020.102652
5. R. Oshikawa, J. Qian, and W. Y. Wang, "A survey on natural language processing for fake news detection," *Lr. 2020 - 12th Int. Conf. Lang. Resour. Eval. Conf. Proc.*, pp. 6086–6093, 2020.
6. S. Yoon *et al.*, "Detecting incongruity between news headline and body text via a deep hierarchical encoder," *33rd AAAI Conf. Artif. Intell. AAAI 2019, 31st Innov. Appl. Artif. Intell. Conf. IAAI 2019 9th AAAI Symp. Educ. Adv. Artif. Intell. EAAI 2019*, pp. 791–800, 2019, https://doi.org/10.1609/aaai.v33i01.3301791.
7. Bodaghi, A., Oliveira, J.: The characteristics of rumor spreaders on Twitter: a quantitative analysis on real data. Comput. Commun. **160**, 674–687 (2020). https://doi.org/10.1016/j.comcom.2020.07.017
8. S. Kwon, M. Cha, K. Jung, W. Chen, and Y. Wang, "Prominent features of rumor propagation in online social media," *Proc. - IEEE Int. Conf. Data Mining, ICDM*, pp. 1103–1108, 2013, https://doi.org/10.1109/ICDM.2013.61.
9. Rath, B., Gao, W., Ma, J., Srivastava, J.: From Retweet to Believability: Utilizing Trust to Identify Rumor Spreaders on Twitter. Soc. Netw. Anal. Min. **8**(1), 179–186 (2018). https://doi.org/10.1007/s13278-018-0540-z

10. K. Shu, H. R. Bernard, and H. Liu, "Studying Fake News via Network Analysis: Detection and Mitigation," no. January, pp. 43–65, 2019, https://doi.org/10.1007/978-3-319-94105-9_3.
11. Wu, Z., Pi, D., Chen, J., Xie, M., Cao, J.: Rumor detection based on propagation graph neural network with attention mechanism. Expert Syst. Appl. **158**, 113595 (2020). https://doi.org/10.1016/j.eswa.2020.113595
12. Zubiaga, A., et al.: Analysing how people orient to and spread rumours in social media by looking at conversational threads. PLoS ONE **11**(3), 1–29 (2016). https://doi.org/10.1371/journal.pone.0150989
13. K. Shu, S. Wang, and H. Liu, "Understanding User Profiles on Social Media for Fake News Detection," *Proc.—IEEE 1st Conf. Multimed. Inf. Process. Retrieval, MIPR 2018*, pp. 430–435, Jun. 2018, https://doi.org/10.1109/MIPR.2018.00092.
14. K. Shu, X. Zhou, S. Wang, R. Zafarani, and H. Liu, "The role of user profiles for fake news detection," *Proc. 2019 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Mining, ASONAM 2019*, pp. 436–439, 2019, doi: https://doi.org/10.1145/3341161.3342927.
15. Y. Wang *et al.*, "EANN: Event adversarial neural networks for multi-modal fake news detection," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 849–857, Jul. 2018, https://doi.org/10.1145/3219819.3219903.
16. H. Zhang, Q. Fang, S. Qian, and C. Xu, "Multi-modal knowledge-aware event memory network for social media rumor detection," *MM 2019 - Proc. 27th ACM Int. Conf. Multimed.*, pp. 1942–1951, Oct. 2019, https://doi.org/10.1145/3343031.3350850.
17. J. Ma *et al.*, "Detecting rumors from microblogs with recurrent neural networks," *IJCAI Int. Jt. Conf. Artif. Intell.*, vol. 2016-Janua, pp. 3818–3824, 2016.
18. K. Popat, S. Mukherjee, A. Yates, and G. Weikum, "Declare: Debunking fake news and false claims using evidence-aware deep learning," *Proc. 2018 Conf. Empir. Methods Nat. Lang. Process. EMNLP 2018*, pp. 22–32, 2020, https://doi.org/10.18653/v1/d18-1003.
19. L. Hu *et al.*, "Compare to the knowledge: Graph neural fake news detection with external knowledge," *ACL-IJCNLP 2021 - 59th Annu. Meet. Assoc. Comput. Linguist. 11th Int. Jt. Conf. Nat. Lang. Process. Proc. Conf.*, pp. 754–763, 2021, https://doi.org/10.18653/v1/2021.acl-long.62.
20. J. Ma, W. Gao, Z. Wei, Y. Lu, and K. F. Wong, "Detect rumors using time series of social context information on microblogging websites," *Int. Conf. Inf. Knowl. Manag. Proc.*, vol. 19–23-Oct-, no. October, pp. 1751–1754, 2015, https://doi.org/10.1145/2806416.2806607.
21. Jiang, J., Wen, S., Yu, S., Xiang, Y., Zhou, W.: Rumor Source Identification in Social Networks with Time-Varying Topology. IEEE Trans. Dependable Secur. Comput. **15**(1), 166–179 (2018). https://doi.org/10.1109/TDSC.2016.2522436
22. Kwon, S., Cha, M., Jung, K.: Rumor detection over varying time windows. PLoS ONE **12**(1), 1–19 (2017). https://doi.org/10.1371/journal.pone.0168344
23. N. Ruchansky, S. Seo, and Y. Liu, "CSI: A hybrid deep model for fake news detection," *Int. Conf. Inf. Knowl. Manag. Proc.*, vol. Part F1318, pp. 797–806, Nov. 2017, https://doi.org/10.1145/3132847.3132877.
24. Shin, J., Jian, L., Driscoll, K., Bar, F.F.F.: The diffusion of misinformation on social media: Temporal pattern, message, and source. Comput. Human Behav. **83**, 278–287 (2018). https://doi.org/10.1016/j.chb.2018.02.008
25. Z. Jin, J. Cao, H. Guo, Y. Zhang, and J. Luo, "Multimodal fusion with recurrent neural networks for rumor detection on microblogs," *MM 2017 - Proc. 2017 ACM Multimed. Conf.*, pp. 795–816, Oct. 2017, https://doi.org/10.1145/3123266.3123454.
26. S. Qian, J. Wang, J. Hu, Q. Fang, and C. Xu, "Hierarchical Multi-modal Contextual Attention Network for Fake News Detection," *SIGIR 2021 - Proc. 44th Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.*, no. 1, pp. 153–162, 2021, https://doi.org/10.1145/3404835.3462871.
27. D. Khattar, M. Gupta, J. S. Goud, and V. Varma, "MvaE: Multimodal variational autoencoder for fake news detection," *Web Conf. 2019 - Proc. World Wide Web Conf. WWW 2019*, no. May, pp. 2915–2921, May 2019, https://doi.org/10.1145/3308558.3313552.
28. S. Singhal, R. R. Shah, T. Chakraborty, P. Kumaraguru, and S. Satoh, "SpotFake: A multi-modal framework for fake news detection," *Proc.—2019 IEEE 5th Int. Conf. Multimed. Big Data, BigMM 2019*, pp. 39–47, Sep. 2019, https://doi.org/10.1109/BigMM.2019.00-44.
29. Shah, D., Zaman, T.: Finding rumor sources on random trees. Oper. Res. **64**(3), 736–755 (2016). https://doi.org/10.1287/opre.2015.1455
30. Shelke, S., Attar, V.: Source detection of rumor in social network—a review. Online Soc. Netw Media **9**, 30–42 (2019). https://doi.org/10.1016/j.osnem.2018.12.001
31. D. Kr, "On Rumor Source Detection and Its," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 1, no. 840, pp. 110–119, 2017, doi: https://doi.org/10.1007/978-3-319-54472-4.
32. Sam Spencer and R. Srikant, "Maximum likelihood rumor source detection in a star network."
33. Xu, F., Sheng, V.S., Wang, M.: Near real-time topic-driven rumor detection in source microblogs. Knowledge-Based Syst. **207**, 106391 (2020). https://doi.org/10.1016/j.knosys.2020.106391
34. K. Popat, S. Mukherjee, J. Strötgen, and G. Weikum, "Where the truth lies: Explaining the credibility of emerging claims on the web and social media," *26th Int. World Wide Web Conf. 2017, WWW 2017 Companion*, pp. 1003–1012, 2017, https://doi.org/10.1145/3041021.3055133.
35. C. Cai, L. Li, and D. Zeng, "Detecting social bots by jointly modeling deep behavior and content information," *Int. Conf. Inf. Knowl. Manag. Proc.*, vol. Part F1318, pp. 1995–1998, 2017, https://doi.org/10.1145/3132847.3133050.
36. M. Orabi, D. Mouheb, Z. Al Aghbari, and I. Kamel, "Detection of Bots in Social Media: A Systematic Review," *Inf. Process. Manag.*, vol. 57, no. 4, p. 102250, 2020, https://doi.org/10.1016/j.ipm.2020.102250.
37. Kudugunta, S., Ferrara, E.: Deep neural networks for bot detection. Inf. Sci. (Ny) **467**, 312–322 (2018). https://doi.org/10.1016/j.ins.2018.08.019
38. G. Gorrell, K. Bontcheva, L. Derczynski, E. Kochkina, M. Liakata, and A. Zubiaga, "RumourEval 2019: Determining rumour veracity and support for rumours," *arXiv*, pp. 69–76, 2018.
39. G. Gorrell *et al.*, "SemEval-2019 Task 7: RumourEval, Determining Rumour Veracity and Support for Rumours," pp. 845–854, 2019, https://doi.org/10.18653/v1/s19-2147.
40. W. Ferreira and A. Vlachos, "Emergent: A novel data-set for stance classification," *2016 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. NAACL HLT 2016 - Proc. Conf.*, no. 1, pp. 1163–1168, 2016, https://doi.org/10.18653/v1/n16-1138.
41. J. Ma, W. Gao, and K. F. Wong, "Detect Rumor and Stance Jointly by Neural Multi-task Learning," *Web Conf. 2018 - Companion World Wide Web Conf. WWW 2018*, pp. 585–593, 2018, https://doi.org/10.1145/3184558.3188729.
42. S. Dungs, A. Aker, N. Fuhr, and K. Bontcheva, "Can Rumour Stance Alone Predict Veracity?," *Proc. 27th Int. Conf. Comput. Linguist.*, pp. 3360–3370, 2018, [Online]. https://aclweb.org/anthology/papers/C/C18/C18-1284/.

43. S. M. Mohammad, P. Sobhani, and S. Kiritchenko, "Stance and sentiment in Tweets," *ACM Trans. Internet Technol.*, vol. 17, no. 3, 2017, https://doi.org/10.1145/3003433.
44. Rubin, V.L., Chen, Y., Conroy, N.J.: Deception detection for news: Three types of fakes. Proc. Assoc. Inf. Sci. Technol. **52**(1), 1–4 (2015). https://doi.org/10.1002/pra2.2015.145052010083
45. F. Yang, X. Yu, Y. Liu, and M. Yang, "Automatic detection of rumor on Sina Weibo," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2012, https://doi.org/10.1145/2350190.2350203.
46. Boididou, C., Papadopoulos, S., Zampoglou, M., Apostolidis, L., Papadopoulou, O., Kompatsiaris, Y.: Detection and visualization of misleading content on Twitter. Int. J. Multimed. Inf. Retr. **7**(1), 71–86 (2018). https://doi.org/10.1007/s13735-017-0143-x
47. Boididou, C., et al.: Verifying information with multimedia content on twitter: a comparative study of automated approaches. Multimed. Tools Appl. **77**(12), 15545–15571 (2018). https://doi.org/10.1007/s11042-017-5132-9
48. Kakol, M., Nielek, R., Wierzbicki, A.: Understanding and predicting Web content credibility using the Content Credibility Corpus. Inf. Process. Manag. **53**(5), 1043–1061 (2017). https://doi.org/10.1016/j.ipm.2017.04.003
49. M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff, and B. Stein, "A stylometric inquiry into hyperpartisan and fake news," *ACL 2018—56th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf. (Long Pap.*, vol. 1, pp. 231–240, 2018, https://doi.org/10.18653/v1/p18-1022.
50. P. Qi, J. Cao, T. Yang, J. Guo, and J. Li, "Exploiting multi-domain visual information for fake news detection," *Proc. - IEEE Int. Conf. Data Mining, ICDM*, vol. 2019-Novem, pp. 518–527, 2019, https://doi.org/10.1109/ICDM.2019.00062.
51. X. Zhou and R. Zafarani, "Network-based fake news detection: a pattern-driven approach," *arXiv*, 2019, https://doi.org/10.1145/3373464.3373473.
52. Fire, M., Kagan, D., Elyashar, A., Elovici, Y.: Friend or foe? Fake profile identification in online social networks. Soc. Netw. Anal. Min. **4**(1), 1–23 (2014). https://doi.org/10.1007/s13278-014-0194-4
53. M. Del Vicario, W. Quattrociocchi, A. Scala, and F. Zollo, "Polarization and fake news: Early warning of potential misinformation targets," *arXiv*, vol. 13, no. 2, 2018, https://doi.org/10.1145/3316809.
54. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *3rd Int. Conf. Learn. Represent. ICLR 2015—Conf. Track Proc.*, pp. 1–14, 2015.
55. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-Decem, pp. 770–778, 2016, https://doi.org/10.1109/CVPR.2016.90.
56. C. Szegedy *et al.*, "Going deeper with convolutions," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 07–12-June, pp. 1–9, 2015, https://doi.org/10.1109/CVPR.2015.7298594.
57. J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol.—Proc. Conf.*, vol. 1, no. Mlm, pp. 4171–4186, 2019.
58. Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," *Adv. Neural Inf. Process. Syst.*, vol. 32, no. NeurIPS, pp. 1–18, 2019.
59. Jin, Z., et al.: Novel visual and statistical image features for microblogs news verification. IEEE Trans. Multimed. **19**(3), 598–608 (2017). https://doi.org/10.1109/TMM.2016.2617078
60. K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *arXiv*, vol. 19, no. 1, pp. 22–36, 2017, https://doi.org/10.1145/3137597.3137600.
61. A. Zubiaga, A. Aker, K. Bontcheva, M. Liakata, and R. Procter, "Detection and resolution of rumours in social media: A survey," *arXiv*, vol. 51, no. 2, 2017.
62. J. Cao, J. Guo, X. Li, Z. Jin, H. Guo, and J. Li, "Automatic Rumor Detection on Microblogs: A Survey," vol. 1, no. c, pp. 1–14, 2018, [Online]. http://arxiv.org/abs/1807.03505.
63. Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., Ortega-Garcia, J.: Deepfakes and beyond: A Survey of face manipulation and fake detection. Inf. Fusion **64**, 131–148 (2020). https://doi.org/10.1016/j.inffus.2020.06.014
64. P. Meel and D. K. Vishwakarma, "Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities," *Expert Syst. Appl.*, vol. 153, 2020, https://doi.org/10.1016/j.eswa.2019.112986.
65. M. R. Islam, S. Liu, X. Wang, and G. Xu, "Deep learning for misinformation detection on online social networks: a survey and new perspectives," *Soc. Netw. Anal. Min.*, vol. 10, no. 1, 2020, doi: https://doi.org/10.1007/s13278-020-00696-x.
66. X. Zhou and R. Zafarani, "A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities," *ACM Comput. Surv.*, vol. 53, no. 5, 2020, https://doi.org/10.1145/3395046.
67. Zhang, X., Ghorbani, A.A.: An overview of online fake news: characterization, detection, and discussion. Inf. Process. Manag. **57**(2), 1–26 (2020). https://doi.org/10.1016/j.ipm.2019.03.004
68. Shu, K., et al.: "Combating disinformation in a social media age", *Wiley Interdiscip*. Rev. Data Min. Knowl. Discov. **10**(6), 1–23 (2020). https://doi.org/10.1002/widm.1385
69. Kandasamy, N., Murugasamy, K.: Detecting and filtering rumor in social media using news media event. Concurr. Comput. Pract. Exp. **33**(20), 1–10 (2021). https://doi.org/10.1002/cpe.6329
70. D'Ulizia, A., Caschera, M.C., Ferri, F., Grifoni, P.: Fake news detection: A survey of evaluation datasets. PeerJ Comput. Sci. **7**, 1–34 (2021). https://doi.org/10.7717/PEERJ-CS.518
71. "fake news - Explore - Google Trends." https://trends.google.com/trends/explore?date=2011-04-16 2021–04–16&q=fake news (accessed Jan. 04, 2022).
72. S. Hangloo and B. Arora, "Fake News Detection Tools and Methods—A Review," *arXiv*, Nov. 2021, [Online]. http://arxiv.org/abs/2112.11185.
73. "How is Facebook addressing false information through independent fact-checkers? | Facebook Help Centre." https://www.facebook.com/help/1952307158131536 (accessed Jan. 04, 2022).
74. "Updating our approach to misleading information." https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information (accessed Jan. 04, 2022).
75. "Introducing Birdwatch, a community-based approach to misinformation." https://blog.twitter.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation (accessed Jan. 04, 2022).
76. H. Ahmed, I. Traore, and S. Saad, "Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 10618 LNCS, pp. 127–138, 2017, https://doi.org/10.1007/978-3-319-69155-8_9.
77. Zhou, X., Jain, A., Phoha, V.V., Zafarani, R.: Fake news early detection. Digit. Threat. Res. Pract. **1**(2), 1–25 (2020). https://doi.org/10.1145/3377478
78. Tuna, T., et al.: User characterization for online social networks. Soc. Netw. Anal. Min. **6**(1), 1–28 (2016). https://doi.org/10.1007/s13278-016-0412-3

79. P. Rosso and L. C. Cagnina, "Deception Detection and Opinion Spam," pp. 155–171, 2017, https://doi.org/10.1007/978-3-319-55394-8_8.
80. M. Hardalov, I. Koychev, and N. Preslav, "In Search of Credible News," *Int. Conf. Artif. Intell. Methodol. Syst. Appl.*, pp. 172–180, 2016, https://doi.org/10.1007/978-3-319-44748-3_29.
81. P. Galán-García, J. G. de la Puerta, C. L. Gómez, I. Santos, and P. G. Bringas, "Supervised Machine Learning for the Detection of Troll Profiles in Twitter Social Network: Application to a Real Case of Cyberbullying Patxi," *Log. J. IGPL*, pp. 42–53, 2016, https://doi.org/10.1007/978-3-319-01854-6.
82. O. Ajao, D. Bhowmik, and S. Zargari, "Fake news identification on Twitter with hybrid CNN and RNN models," *ACM Int. Conf. Proceeding Ser.*, pp. 226–230, Jul. 2018, doi: https://doi.org/10.1145/3217804.3217917.
83. Umer, M., Imtiaz, Z., Ullah, S., Mehmood, A., Choi, G.S., On, B.W.: Fake news stance detection using deep learning architecture (CNN-LSTM). IEEE Access **8**, 156695–156706 (2020). https://doi.org/10.1109/ACCESS.2020.3019735
84. Kaliyar, R.K., Goswami, A., Narang, P., Sinha, S.: FNDNet—A deep convolutional neural network for fake news detection. Cogn. Syst. Res. **61**, 32–44 (2020). https://doi.org/10.1016/j.cogsys.2019.12.005
85. Y. Liu and Y. B. Wu, "Early Detection of Fake News on Social Media Through Propagation Path Classification with Recurrent and Convolutional Networks," pp. 354–361.
86. O. Varol, E. Ferrara, F. Menczer, and A. Flammini, "Early detection of promoted campaigns on social media," *EPJ Data Sci.*, vol. 6, no. 1, 2017, https://doi.org/10.1140/epjds/s13688-017-0111-y.
87. Faustini, P.H.A., Covões, T.F.: Fake news detection in multiple platforms and languages. Expert Syst. Appl. **158**, 113503 (2020). https://doi.org/10.1016/j.eswa.2020.113503
88. G. B. Guacho, S. Abdali, N. Shah, and E. E. Papalexakis, "Semi-supervised content-based detection of misinformation via tensor embeddings," *Proc. 2018 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Mining, ASONAM 2018*, no. January, pp. 322–325, 2018, https://doi.org/10.1109/ASONAM.2018.8508241.
89. Boididou, C., et al.: Verifying Multimedia Use at MediaEval 2016. CEUR Workshop Proc. **1739**, 4–6 (2016)
90. P. Zhou, H. Xintong, V. I. Morariu, and L. S. Davis, "Learning Rich Features for Image Manipulation Detection," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 1053–1061, 2018, [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2018/papers/Zhou_Learning_Rich_Features_CVPR_2018_paper.pdf%0A, http://dl.acm.org/citation.cfm?doid=3133956.3134027.
91. Song, C., Ning, N., Zhang, Y., Wu, B.: A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks. Inf. Process. Manag. **58**(1), 102437 (2021). https://doi.org/10.1016/j.ipm.2020.102437
92. Ying, L., Yu, H., Wang, J., Ji, Y., Qian, S.: Multi-level Multimodal Cross-attention Network for Fake News Detection. IEEE Access **9**, 132363–132373 (2021). https://doi.org/10.1109/ACCESS.2021.3114093
93. A. Silva, L. Luo, S. Karunasekera, and C. Leckie, "Embracing Domain Differences in Fake News: Cross-domain Fake News Detection using Multi-modal Data," 2021, [Online]. http://arxiv.org/abs/2102.06314.
94. A. Giachanou, G. Zhang, and P. Rosso, "Multimodal multi-image fake news detection," *Proc. - 2020 IEEE 7th Int. Conf. Data Sci. Adv. Anal. DSAA 2020*, pp. 647–654, 2020, https://doi.org/10.1109/DSAA49011.2020.00091.
95. S. Singhal, A. Kabra, M. Sharma, R. R. Shah, T. Chakraborty, and P. Kumaraguru, "SpotFake+: A multimodal framework for fake news detection via transfer learning (student abstract)," *AAAI 2020 - 34th AAAI Conf. Artif. Intell.*, pp. 13915–13916, 2020, https://doi.org/10.1609/aaai.v34i10.7230.
96. Ying, L., Yu, H.U.I., Wang, J., Ji, Y., Qian, S.: Fake news detection via multi-modal topic memory network. IEEE Access **9**, 132818–132829 (2021). https://doi.org/10.1109/ACCESS.2021.3113981
97. X. Z. B, J. Wu, and R. Zafarani, "SAFE : Similarity-Aware Multi-modal Fake," pp. 354–367, 2020, https://doi.org/10.1007/978-3-030-47436-2.
98. L. Cui, S. Wang, and D. Lee, "Same: Sentiment-aware multi-modal embedding for detecting fake news," *Proc. 2019 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Mining, ASONAM 2019*, pp. 41–48, Aug. 2019, https://doi.org/10.1145/3341161.3342894.
99. S. Qian, J. U. N. Hu, Q. Fang, and C. Xu, "Knowledge-aware Multi-modal Adaptive Graph Convolutional Networks for Fake News Detection," *ACM Trans. Multimed. Comput. Commun. Appl.*, vol. 17, no. 3, 2021, https://doi.org/10.1145/3451215.
100. Y. Yang *et al.*, "TI-CNN: Convolutional neural networks for fake news detection," *arXiv*, 2018, [Online]. http://arxiv.org/abs/1806.00749.
101. H. Guo, J. Cao, Y. Zhang, J. Guo, and J. Li, "Rumor detection with hierarchical social attention network," *Int. Conf. Inf. Knowl. Manag. Proc.*, pp. 943–952, Oct. 2018, doi: https://doi.org/10.1145/3269206.3271709.
102. C. Shao, G. L. Ciampaglia, O. Varol, K. C. Yang, A. Flammini, and F. Menczer, "The spread of low-credibility content by social bots," *Nat. Commun.*, vol. 9, no. 1, 2018, https://doi.org/10.1038/s41467-018-06930-7.
103. T. N. Nguyen, C. Li, and C. Niederée, "On early-stage debunking rumors on twitter: Leveraging the wisdom of weak learners," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 10540 LNCS, pp. 141–158, 2017, https://doi.org/10.1007/978-3-319-67256-4_13.
104. S. Ahmed, K. Hinkelmann, and F. Corradini, "Combining machine learning with knowledge engineering to detect fake news in social networks - A survey," *CEUR Workshop Proc.*, vol. 2350, 2019.
105. E. Ortega-fernández, G. Padilla-castillo, S. L. Carcelén-garcía, and M. Arias-oliva, "fact checking agencies and processes to fight against fake news," pp. 219–228, 2020.
106. W. Y. Wang, "'Liar, liar pants on fire': A new benchmark dataset for fake news detection," *ACL 2017 - 55th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf. (Long Pap.*, vol. 2, pp. 422–426, 2017, https://doi.org/10.18653/v1/P17-2067.
107. A. Zubiaga, M. Liakata, R. Procter, G. Wong Sak Hoi, and P. Tolmie, "Analysing how people orient to and spread rumours in social media by looking at conversational threads," *PLoS One*, vol. 11, no. 3, pp. 1–29, 2016, https://doi.org/10.1371/journal.pone.0150989.
108. Hakak, S., Alazab, M., Khan, S., Gadekallu, T.R., Maddikunta, P.K.R., Khan, W.Z.: An ensemble machine learning approach through effective feature extraction to classify fake news. Futur. Gener. Comput. Syst. **117**, 47–58 (2021). https://doi.org/10.1016/j.future.2020.11.022
109. F. Monti, F. Frasca, D. Eynard, D. Mannion, and M. M. Bronstein, "Fake News Detection on Social Media using Geometric Deep Learning," *arXiv*, pp. 1–15, Feb. 2019, [Online]. http://arxiv.org/abs/1902.06673.
110. Meel, P., Vishwakarma, D.K.: HAN, image captioning, and forensics ensemble multimodal fake news detection. Inf. Sci. (Ny) **567**, 23–41 (2021). https://doi.org/10.1016/j.ins.2021.03.037

111. P. Meel and D. K. Vishwakarma, "Multi-modal Fusion using Fine-tuned Self- attention and Transfer Learning for Veracity Analysis of Web Information."
112. V. H. Nguyen, K. Sugiyama, P. Nakov, and M. Y. Kan, "FANG: Leveraging Social Context for Fake News Detection Using Graph Representation," *Int. Conf. Inf. Knowl. Manag. Proc.*, pp. 1165–1174, 2020, https://doi.org/10.1145/3340531.3412046.
113. Wang, Z., Guo, Y.: Rumor events detection enhanced by encoding sentimental information into time series division and word representations. Neurocomputing **397**, 224–243 (2020). https://doi.org/10.1016/j.neucom.2020.01.095
114. S. Singhal, "SpotFake : A Multi-modal Framework for Fake News Detection," 2015.
115. J. Z. Pan, S. Pavlova, C. Li, N. Li, Y. Li, and J. Liu, "Content based fake news detection using knowledge graphs," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11136 LNCS, pp. 669–683, 2018, https://doi.org/10.1007/978-3-030-00671-6_39.
116. C. Boididou, S. Papadopoulos, L. Apostolidis, and Y. Kompatsiaris, "Learning to detect misleading content on Twitter," *ICMR 2017 - Proc. 2017 ACM Int. Conf. Multimed. Retr.*, pp. 278–286, 2017, https://doi.org/10.1145/3078971.3078979.
117. K. Zhou, C. Shu, B. Li, and J. H. Lau, "Evaluating Event Credibility on Twitter," *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, pp. 1614–1623, 2019, https://doi.org/10.18653/v1/n19-1163.
118. K. Shu, S. Wang, and H. Liu, "Beyond news contents: The role of social context for fake news detection," *arXiv*, no. i, pp. 312–320, 2019, https://doi.org/10.1145/3289600.3290994.
119. K. Shu, H. Russell Bernard, and H. Liu, "Studying fake news via network analysis: Detection and mitigation," *arXiv*, pp. 836–837, 2018, https://doi.org/10.1007/978-3-319-94105-9_3.
120. Shu, K., Mahudeswaran, D., Wang, S., Lee, D., Liu, H.: FakeNewsNet: a data repository with news content, social context, and spatiotemporal information for studying fake news on social media. Big Data **8**(3), 171–188 (2020). https://doi.org/10.1089/big.2020.0062
121. Y. Dou, K. Shu, C. Xia, P. S. Yu, and L. Sun, "User Preference-aware Fake News Detection," *SIGIR '21 Proc. 44th Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.*, pp. 2051–2055, 2021, https://doi.org/10.1145/XXXXXX.XXXXXX.
122. J. Ma, W. Gao, and K. F. Wong, "Rumor detection on twitter with tree-structured recursive neural networks," *ACL 2018 - 56th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf. (Long Pap.*, vol. 1, pp. 1980–1989, 2018, https://doi.org/10.18653/v1/p18-1184.
123. Q. You, L. Cao, H. Jin, and J. Luo, "Robust visual-textual sentiment analysis: When attention meets tree-structured recursive neural networks," *MM 2016 - Proc. 2016 ACM Multimed. Conf.*, pp. 1008–1017, 2016, https://doi.org/10.1145/2964284.2964288.
124. Nasir, J.A., Khan, O.S., Varlamis, I.: Fake news detection: A hybrid CNN-RNN based deep learning approach. Int. J. Inf. Manag. Data Insights **1**(1), 100007 (2021). https://doi.org/10.1016/j.jjimei.2020.100007
125. Kim, E., Cho, S.: Exposing fake faces through deep neural networks combining content and trace feature extractors. IEEE Access **9**, 123493–123503 (2021). https://doi.org/10.1109/ACCESS.2021.3110859
126. "The mostly complete chart of Neural Networks, explained | by Andrew Tch | Towards Data Science." https://towardsdatascience.com/the-mostly-complete-chart-of-neural-networks-explained-3fb6f2367464 (accessed Apr. 18, 2022).
127. A. Vaswani *et al.*, "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 2017-Decem, no. Nips, pp. 5999–6009, 2017.
128. Yu, H.U.I., et al.: Multi-level multi-modal cross-attention network for fake news detection. IEEE Access **9**, 132363–132373 (2021). https://doi.org/10.1109/ACCESS.2021.3114093
129. Y. Wu, P. Zhan, Y. Zhang, L. Wang, and Z. Xu, Multimodal fusion with co-attention networks for fake news detection, pp. 2560–2569, 2021, https://doi.org/10.18653/v1/2021.findings-acl.226.
130. Jindal, S., Sood, R., Singh, R., Vatsa, M., Chakraborty, T.: NewsBag: a multimodal benchmark dataset for fake news detection. CEUR Workshop Proc. **2560**, 138–145 (2020)
131. J. Feng *et al.*, "Generative adversarial networks based on collaborative learning and attention mechanism for hyperspectral image classification," *Remote Sens.*, vol. 12, no. 7, 2020, https://doi.org/10.3390/RS12071149.
132. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Adv. Neural Inf. Process. Syst.*, pp. 1097–1105, 2012, https://doi.org/10.1145/3383972.3383975.
133. C. Wang, P. Nulty, and D. Lillis, "A Comparative Study on Word Embeddings in Deep Learning for Text Classification," *PervasiveHealth Pervasive Comput. Technol. Healthc.*, no. December, pp. 37–46, 2020, https://doi.org/10.1145/3443279.3443304.
134. Y. Qi, D. S. Sachan, M. Felix, S. J. Padmanabhan, and G. Neubig, "When and why are pre-trainedword embeddings useful for neural machine translation?," *NAACL HLT 2018 - 2018 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 2, pp. 529–535, 2018, https://doi.org/10.18653/v1/n18-2084.
135. Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," no. 1, 2019, [Online].: http://arxiv.org/abs/1907.11692.
136. Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations," pp. 1–16, 2019, [Online]. http://arxiv.org/abs/1909.11942.
137. Lahat, D., Adalı, T., Jutten, C., Multimodal, C.J.: Multimodal data fusion: an overview of methods, challenges and prospects. Inst. Electr. Electron. Eng. **103**(9), 1449–1477 (2015). https://doi.org/10.1109/JPROC.2015.2460697ï
138. P. Le and I. Titov, "Improving Entity Linking by Modeling Latent Relations between Mentions," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Apr. 2018, vol. 1, pp. 1595–1604, https://doi.org/10.18653/v1/P18-1148.
139. D. Milne and I. H. Witten, "Learning to Link with Wikipedia," 2008.
140. L. Chen, J. Liang, C. Xie, and Y. Xiao, "Short text entity linking with fine-grained topics," *Int. Conf. Inf. Knowl. Manag. Proc.*, pp. 457–466, Oct. 2018, https://doi.org/10.1145/3269206.3271809.
141. F. Suchanek *et al.*, "Yago : A Core of Semantic Knowledge Unifying WordNet and Wikipedia," 2007.
142. Hoffart, J., Suchanek, F.M., Berberich, K., Weikum, G.: YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. Artif. Intell. **194**, 28–61 (2013). https://doi.org/10.1016/j.artint.2012.06.001
143. W. Wu, H. Li, H. Wang, and K. Q. Zhu, "Probase: A Probabilistic Taxonomy for Text Understanding," 2012, Accessed: Dec. 16, 2021. [Online]. http://research.microsoft.com/.
144. C. Boididou *et al.*, "Verifying Multimedia Use at MediaEval 2015," Accessed: Apr. 18, 2022. [Online]. https://github.com/MKLab-ITI/image-verification-corpus/.
145. L. Derczynski *et al.*, "SemEval-2017 Task 8: RumourEval: Determining rumour veracity and support for rumours," Apr. 2017, Accessed: Dec. 30, 2021. [Online]. https://arxiv.org/abs/1704.05972v1.

146. Bondielli, A., Marcelloni, F.: A survey on fake news and rumour detection techniques. Inf. Sci. (Ny) **497**, 38–55 (2019). https://doi.org/10.1016/j.ins.2019.05.035

ARTICLES FOR FACULTY MEMBERS

MULTIMODAL FAKE NEWS DETECTION

Title/Author	**Feature importance in the age of explainable AI: Case study of detecting fake news & misinformation via a multi-modal framework / Kumar, A., & Taylor, J. W.**
Source	***European Journal of Operational Research*** **Volume 317 Issue 2 (2024) Pages 401–413** **https://doi.org/10.1016/J.EJOR.2023.10.003** **(Database: ScienceDirect)**

European Journal of Operational Research 317 (2024) 401–413

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Feature importance in the age of explainable AI: Case study of detecting fake news & misinformation via a multi-modal framework

Ajay Kumar [a,*], James W. Taylor [b]

[a] EMLYON Business School, 23 Av. Guy de Collongue, 69130 Écully, France
[b] Saïd Business School, University of Oxford, Park End Street, Oxford, OX1 1HP, United Kingdom

ARTICLE INFO

Keywords:
Analytics
Feature selection
Machine learning
Explainable data analytics
Optimization algorithm

ABSTRACT

In recent years, fake news has become a global phenomenon due to its explosive growth and ability to leverage multimedia content to manipulate user opinions. Fake news is created by manipulating images, text, audio, and videos, particularly on social media, and the proliferation of such disinformation can trigger detrimental societal effects. False forwarded messages can have a devastating impact on society, spreading propaganda, inciting violence, manipulating public opinion, and even influencing elections. A major shortcoming of existing fake news detection methods is their inability to simultaneously learn and extract features from two modalities and train models with shared representations of multimodal (textual and visual) information. Feature engineering is a critical task in the fake news detection model's machine learning (ML) development process. For ML models to be explainable and trusted, feature engineering should describe how many features used in the ML models contribute to making more accurate predictions. Feature engineering, which plays an important role in the development of an explainable AI system by shaping the features used in the ML models, is an interconnected concept with explainable AI as it affects the model's interpretability. In the research, we develop a fake news detector model in which we (1) identify several textual and visual features that are associated with fake or credible news; specifically, we extract features from article titles, contents, and, top images; (2) investigate the role of all multimodal features (content, emotions and manipulation-based) and combine the cumulative effects using the feature engineering that represent the behavior of fake news propagators; and (3) develop a model to detect disinformation on benchmark multimodal datasets consisting of text and images. We conduct experiments on several real-world multimodal fake news datasets, and our results show that on average, our model outperforms existing single-modality methods by large margins that do not use any feature optimization techniques.

1. Introduction

The recent transformation of social media channels (Facebook, Twitter, YouTube, and WhatsApp, etc.) has changed the whole way we lead our lives and acquire information. Prior to approximately a decade ago, our primary source of news was traditional journalism, established media organizations, and other reliable sources that adhered to specific ethical standards (Hunt & Matthew, 2017). Most of the news we read today, particularly via social media feeds and tweets, might appear true to our eyes; however, that often is not the case. The intent behind misinformation spread is to deceive the readers, and the extensive growth of disinformation spread on media channels has particularly exploded in the last few years via platforms such as YouTube, LinkedIn, Facebook, Instagram, WhatsApp, Twitter, and Sina Weibo, among others. Whether through conspiracy planning or pandemic dealing, social media sites have transformed the dissemination mode of information, and these platforms have always functioned as a platform for 'fake news' (Zafarani et al., 2019; Agarwal et al., 2022; Zhang et al., 2019). The extensive use of social websites to disseminate false and often incendiary information stirred alarm and highlighted the detection of fake news as a critical issue that needed serious attention.

Ironically, the term 'fake news' was actually introduced by Donald Trump to highlight his accusations against the opposing party during the aforementioned US elections; however, he arguably emerged as the greatest beneficiary of the proliferation of misinformation. For example, a widespread claim that Trump fed police officers working protests in Chicago was initiated by a single tweet from a person who was not even present in the city. A study conducted during the election year found

* Corresponding author at: EMLYON Business School, 23 Av. Guy de Collongue, 69130 Écully, France.
E-mail address: akumar@em-lyon.com (A. Kumar).

https://doi.org/10.1016/j.ejor.2023.10.003
Received 12 December 2022; Accepted 2 October 2023
Available online 4 October 2023

that 88 % of US citizens believed that misunderstandings were being fomented by false news; however, 25 % of survey participants accepted the sharing of fake political news online due to ignorance (Barthel, Mitchell, & Holcomb, 2016). Another instance was also related to the 2016 US presidential elections, which was immensely driven by misleading information on social sites. On Twitter, nearly 529 different rumours were spread, and malicious bot accounts which counted for an average of 19 million, tweeted or retweeted tweets in favor of Trump or Clinton. According to research (Silverman, 2016), fake stories and misinformation were more widely used on Twitter, WhatsApp, and Facebook platforms than the real stories in the 2016 election. Silverman and Singer-Vine (2016) proved that most of the people who read false news and misinformation or watch fake videos or stories on Facebook and Twitter, believe them and they also proved that most discussed fake news stories in the 2016 USA presidential election tended to favor Donald Trump over Hillary Clinton (Papanastasiou, 2020; Chang et al., 2022). A recent study conducted in the USA tells that (i) 60 % of people in the United States received the news online from an online media platform (Gottfried & Shearer, 2016) (ii) Most of the time fake news and misinformation was shared on Facebook and Twitter than any other platforms on social media (Silverman, 2016) (iii) Most people report to believe the news to be true which is actually fake news (Silverman and Singer-Vine, 2016). Ever since these cases, various departments including social media companies came under pressure to build something against the expansion of misinformation on media channels.

The widespread disinformation can negatively affect people's behavior and cause the detrimental societal effects. The Impact of fake forwarded messages is well known and how it is plotted to disturb societal peace with the intent to spread violence, cause riots, humiliate emotions, influence elections, or initiate propaganda (Kim & Dennis, 2020; Moravec et al., 2020; Reisach, 2021; Tanınmış et al., 2022). Another example is about Las Vegas massacre in October 2017, which reportedly killed 59 people and injured more than 500 people, began spreading on Facebook and Google with false reports of the deadly tragedy. In 2018, a theme issue about 'Fake News', which reported that fake stories tend to arise human feelings of fear and surprise (Vosoughi et al., 2018) leading to social panic, was published by Science magazine. An example of disinformation impact is a fake video with the name Somalis 'pushed into shallow grave' Ethiopia that lead to a fierce clash between two races in Ethiopia. Another clash between Greek police with travelers or migrants was caused due to rumor spread on social media stating the elevation of onward travel restrictions in Greece. These examples explain the serious threat of the increasing spread of fake news in society (David et al., 2018).

Fake news, misinformation, or disinformation are generally created and shared on social media by manipulating or tempering videos, text, and images. Although technology advancements and enhanced AI tracking have shown great progress in the detection of anomalous information or fake news, but the threat still persists because of the complexity created by disinformation's volume and velocity (George et al., 2018; Han et al., 2020). The issues created by fake news spread through various social media sites and other popular communication apps are worth taking note of for many reasons. There is a lot that goes behind the misinformation spreading such as alteration, distortion, or manipulation of texts, images, and videos with the help of several other contents. There are software and online platforms available with multiple features to manipulate the content and publish it as pieces of real news. Fake news makes the best possible use of videos and images to give an intuitive experience for readers. Facebook, Twitter, and other online media channels have transformed the whole news market evolving from text-only news to news with multimodal news which has images, text, audio, and videos. This transformation provides an interactive storytelling experience and has the magical power to delight and engage readers by leveraging visual context-powered news. Fake news articles can contain misrepresented, irrelevant, and forged images to mislead the readers (Han et al., 2020; Borchert et al., 2023). The recent increase in fake news is affecting the social fabric, dynamics, public affairs, policies, and events negatively. Spreading such news through messages is easily achieved through social media by targeting the vulnerable audience who would forward and share the content. Sean Parker, who is Facebook's first president, is strictly critical of social networks and accuses these platforms of being vulnerable to humans (Moravec et al., 2020).

While most of the existing fake news detectors used only one modality (image or text) to identify the suspicious behavior of fake news propagators but one modality is not enough to handle such a complicated problem. Existing content-based fake news detection methods either solely consider textual information (Zhou et al., 2019), or visual information, or combine textual with visual data to analyze the fakeness of a forwarded message without investigating the behavior of fake news propagators (Wang et al., 2018; Yang et al., 2018). Most of the fake news propagators prefer to write fake news with tempting images whose content does not have any relationship with actual news to engage the users and attract the publish attention. Furthermore, when fake news propagators create a fake message with fictional scenarios, it is very difficult for users or any news agency to find the original and tempered images to verify the realness of these messages. Therefore, a research gap exists in the literature between multimodal (text and images) dimensions of misinformation when fake news propagators use real images (non-manipulated) to spread non-factual news on social media channels (Yang et al., 2018). For example, there was a rumor on the internet in 2013, which claimed that United States President Obama was injured in an explosion in the white house, which crashed the share market and abolished over $150 billion in stock values (Rapoza, 2017). Thus, it has become the greatest threat to economies and freedom of expression. In this scenario, it becomes hard to differentiate between relevant or non-manipulated multimedia and such fictitious news. With non-manipulated images or videos alongside some fictional scenario-based news, there comes a justifying gap between both.

Most of the existing research is still focused on text processing, unimodal features, and using traditional ML techniques to detect misinformation and fake news. With the recent development of advanced ML and deep learning models, there has been less previous evidence for developing multimodal fake news detectors. This is a significant and important gap; given that misinformation, disinformation, and fake news are becoming multimodal in nature in the past couple of years (Begley, 2017).

In recent times, advanced ML and deep learning models have been in trend and are known to achieve excellent ML accuracy in various fields like computer vision, speech recognition, natural language processing, social network filtering, audio recognition, etc. It has shown tremendous results in tasks such as image captioning, visual question answering, and fake news detection. A model was proposed by Jin et al. (2017a) that extracts the visual, textual, and social content features, which are further fused by using the attention mechanism. The Attention mechanism is an effort mechanism by deep neural networks to concentrate on selective or relevant features- while neglecting the others in the network. Another model to learn event-invariant features was proposed, which made use of an adversarial fake news network with a few visual and textual features. However, there remains unclear room for discovering correlations between different modalities and the role of feature engineering to improve the accuracy of the fake news detection models, which is not covered by these models as well. This leads us to propose a novel multi-modal framework for fake news detection. The proposed model takes into account the two modalities present in an article – text, and images. Forging images is one of the most popular ways to tamper with the credibility of news and recent studies have shown that visual features (images) play a very important role in detecting fake news (Wu et al., 2015; Zhang et al., 2019). Rubin et al. (2015) use several ML techniques to analyze the textual data to classify a news item as real or fake but they only used one modality-text to develop the detecting model. Though important, but multimedia content has not been

explored much in the context of fake news detection. Although, basic features of images have been explored by Tian et al. (2013) but these features merely depict the complex distributions of image content and are handcrafted. Therefore, in order to improve the fake news detection systems, the complex relationship between different modes, the role of feature engineering, and varying distributions need to be addressed.

Feature engineering and explainable AI are two interconnected concepts and this connection lies in the interpretability of features. When feature selection is done effectively, it can lead to more interpretable and explainable ML models. By selecting and constructing important features that are capable to capture relevant information from the data, it becomes easier to understand the decision-making process of the ML model, and this process could enable better transparency and interpretability. In summary, feature engineering plays an important role in the development of an explainable AI system by shaping the features used in the ML model, which in turn affects the model's interpretability (Ali et al., 2023). In this paper, in order to effectively identify the disinformation we develop a multimodal fake news detector model, which extracts textual and visual features particularly to detect the disinformation on benchmark multimodal datasets. We extract several novel features e.g., content-based (topics of an image, contrast, brightness, the dominant and fraction representation of color spots in the manipulated area, etc.), emotions-based (sentiment polarity, adult content, violent content or smile, etc.) and manipulation-based (spoofed content, image sharpness/blur, pixel density difference in the manipulated area, etc.). There have been several efforts at developing fake news detection models in recent years (Horne & Adali, 2017). Overall, we found that no published literature that considers specific aspects of feature engineering in developing the multimodal fake news detector has been published to identify misinformation and fake news propagators. To address these research gaps, we aimed to develop a fake news detector model to detect the behavior of fake news spreaders using an innovative modified JAYA feature optimization model.

In addition, to developing the detector model, we also addressed the following research questions to investigate the behavior of fake news propagators:

RQ-1: From the psychology perspective, do fake news propagators use a specific linguistic tone or write the news under a certain rule of the press (e.g., fewer words than real news, more question marks, exclamation, use third-person pronouns more than first-person pronouns and capital letters than real news, etc.)?

RQ-2: From the cognitive perspective in the text features, do fake news writers use exclusive words and negations more frequently than real news writers?

RQ-3: From the emotional perspective in the text features, do fake news propagators use more anger, sexual, and swear words and use more lexical diversity to write the news?

RQ-4: From the emotional perspective in the visual features, do fake news propagators use more artificial text and violence to promote violent extremism?

RQ-5: How do multimodal fake news detection methods perform compared to unimodal methods after combining the visual features with text features?

RQ-6: How do feature engineering methods improve the classification accuracy of the proposed multimodal fake news detector model?

This research work conducts several experiments on three datasets collected from popular social media platforms: Twitter, Weibo, and rFakeEdit, which use textual and visual analytics to extract the important features. The main innovation of our research work is to use the power of textual and visual analytics with an optimization feature engineering algorithm that selects the optimal features to develop the multimodal fake news detector model.

2. Literature review, conceptual background, and theory building

In this literature review and background sections, we discuss the published works related to the use of feature engineering for developing fake news detector models from single-modality and multimodal categories. In the first part, we discuss several theories to extract useful features from the multimodal data and to make theoretical ground for developing a fake news detector model using these features.

Johnson and Kaye, stated that most people specifically use social media for hedonic purposes, connecting with friends and seeking entertainment, rather than for anything that might be of use. When using social media, the user does it with a different mindset than when reading news items found elsewhere on the internet. This difference in the consumption of information affects how the user processes information. As an example, it is known that some product reviews are fake but users do not read them for entertainment, they only read them in order to make an informed choice as to whether to buy an item, knowing there is a monetary incentive when making the best decision. For this reason, users reading fake reviews have a purpose in mind, having a utilitarian mindset, and the goal is to understand the content of the review and whether the information should be considered in making a decision. A paper by Minas et al. investigated the utilitarian mindset in a decision-making context within virtual team interactions. Participants involved in a decision-making team-based chat were found to use confirmation bias. Contrastingly, when reading social media news, the user has a hedonic mindset and the goal is pleasure and enjoyment, not one where they have to decide whether the content is fake or not. The user wants to avoid any activity that feels like work, e.g., thoughtful information processing, and to avoid anything that they do not enjoy, e.g., reading stories where their favorite team has lost. Moravec, Minas and Dennis (2019) state that users want to read feel-good articles, i.e., ones that make them feel happy and that often support what they believe. Moravec, Minas and Dennis (2019) go further to say that on social media, the source of the news is not always clear, whereas when a user is looking at traditional sources for news on the internet, they will have to visit their favorite news network or newspaper online which they would consider trustworthy and would therefore have an understanding of the source's limitations. Facebook differs in that it uses an algorithm to choose the articles it publishes and they are not the user's choice (Yfanti et al., 2023).

Users may subscribe to some information sources by following them on social media platforms, but many sources of information come in the forms of advertisements, shares by friends, and decisions made by the algorithms employed. Therefore, there is a mixture of different sources, reputable and disreputable. For example, a fake news item may appear between a CNN article and an advertisement for Aunt Martha's cookies! Kim and Dennis (2018), and Moravec et al. (2019) both concluded that if the source of the item is not clear, or is even deliberately hidden, users with a hedonic mindset will not bother to make the effort to find and understand the source.

Fake news appears everywhere and in vast amounts resulting in it being hard to separate fact from fiction. Silverman (2016) pointed out that more fake news than real news is shared on social media. Moravec, Minas and Dennis (2019), state that the low cost and massive presence of fake news is a reason why it is so common on social media. They also claim that many fake sites have appeared on Twitter and Facebook solely to destroy a specific individual or to spread propaganda, carefully crafted for a specific reason. The hedonic mindset, not knowing the source and, the sheer volume of fake news, are three reasons, combined to form three contextual factors, why users do not think critically as they otherwise would when presented with news from a known reputable source. Gabielkov et al. (2016) proved further that more than half of the items shared on Twitter have not been completely read in the first place, and the users have not thought about the content in a critical manner.

In the 1980s, two complementary models were proposed for

measuring the cognitive process. These were Elaboration Likelihood Model (ELM), Heuristic-Systematic Model (HSM), proposed by Chaiken (1980), and Chaiken and Eagly (1983), respectively. The basic fundamental concept of HSM and ELM was common and both models discuss two different cognitive processes, which form cognitive attitudes, but differ in cognitive processing, which is used in information evaluation. These are a quantitative difference and a qualitative difference, respectively. HSM and ELM argue that the route that is chosen by an individual is based on the motivation and ability to participate in extensive cognition. Both models have evolved to say that cognition is a continuum of processing and that the routes are not separate. The most popular of the two is ELM and it is widely used today (Cacioppo et al., 2018; Moravec, Minas & Dennis, 2019). That is why we decided to use ELM and framing theory concepts to identify the text and visual features in this research.

Petty & Cacioppo, used the ELM model to show that online users can be influenced whether to accept social media channels' information as true either by using the peripheral (heuristic) or central processing route. The individual needs to take time and cognitive effort to rationally and objectively assess the truth of the item. It relies on people having emotions triggered and requires little cognitive effort. Rational judgment and assessment are often neglected when a news item, for example, uses specific language or tone with images that could trigger certain emotions or feelings: anxiety, violence, and fear. Machine-learning models need to take into account the emotions for developing the fake news detector models and the ELM model is useful in identifying how the peripheral processing route is targeted by text and images for fake news items. However, in order to get the true credibility of an article it is important to integrate images, videos, and text together. In 2005, De Vreese presented the *Framing Theory* which states that an issue or event is defined by the way it is presented, (De Vreese, 2005). In framing theory, a frame consists of a set of ideas which are key, stock phrases, and images that support a specific event interpretation. The interpretation made with text and images is a dominant interpretation created strategically and it is easier to understand, more memorable, and easy to accept. The combination of images and text makes it easy to effectively manipulate ideas and give deniability to the website or the author. It is also hard for automated systems to identify fake news or misinformation if they rely only on text and images (Messaris & Abraham, 2001).In the work presented here, many theories are used in developing the fake news detector model. Lang used the LC4MP model that enables the understanding of how information-rich, multi-modal messages are processed. Apparently, there are several mechanisms that could determine that a person will register cognitively his profile on social media websites and share fake news consciously or unconsciously. It is, therefore, important that the deep learning and advanced machine-learning models that are used in the current work identifies fake reviewers and captures several media channels of information so that the salient features within a modality are processed as far as possible.

In data science, feature engineering or feature selection (*selecting the best features for building the ML models*) is the most important data pre-processing challenge that is used to reduce the dimensions of the datasets by removing unimportant, noisy, and irrelevant features or variables (Kumar et al., 2022). Generally, we classify the feature selection methods into two categories: wrapper, and filter-based techniques. Filter-based methods (LDA, ANOVA, Chi-square, etc.) are used to rank the variables based on the scores of several statistical tests. All Filter-based methods use the similarity methods (fisher or Laplacian score), and statistical methods (F-score, Gini index, or Chi-square statistic score) to select the best features for building ML models based on specific criteria. One of the major issues with the Filter-based methods is that method does not remove multicollinearity (Kumar et al., 2022). In wrapper-based methods, we use the inferences that we have drawn from the previous ML model and evaluate the quality of variables based on this score. For adding or removing the variables in the model, this method first searches the best subset of variables and after that, the model evaluates the fitness of the subset of variables before developing the ML model building block Kumar et al., 2022). The Wrapper-based methods are still unexplored and researchers are developing new methods using feature optimization techniques these days. Several wrapper-based feature selection methods have been developed by the researchers with their own challenges. Too et al. developed the feature optimization algorithms to find the best subset of suitable features for developing the ML models. All these feature optimization algorithms are developed by specific controlling parameters such as weights, number of iterations, crossover probability, etc. that need to be tuned to get the best subset of optimized features. To achieve good accuracy in the ML building process, parameter tuning is essential and a good choice of parameters can make the ML model succeed after getting the best subset of important features (Peng & Xintong, 2022).

Recently, deep learning models have been used to develop fake news multimodal detector models. Most of the published literature (Singhal et al., 2019; Shah & Kobti, 2020; Peng & Xintong, 2022; Uppada & Patel, 2022) used convolutional neural network (CNN), and Visual Geometry Group (VGG)−19 deep learning architectures that use the attention-based mechanism and combines it with other ML model - long short-term memory (LSTM) to develop a microblogging detector model for identifying the characteristics of fake news propagators. The published literature used the CNN model to extract several emotions and manipulation-based fine-grained features and used the LSTM for coarse-grained feature extraction and after that, these extracted features are classified into fake or credible news stories to achieve better accuracy of the misinformation detector models.

However, all of the above detection methods (unimodal and multimodal) have delivered promising results but we observed several limitations in the published literature for identifying the behavioral characteristics of fake news propagators using feature engineering. The multimodal nature of a dataset is always a challenge for researchers for extracting useful features for building ML models. Feature engineering (selecting the best features/variables) or data pre-processing is the most important and critical step in developing fake news ML detector models. Although a few researchers have used feature engineering on the unimodal dataset and developed JAYA and modified JAYA algorithms (Rao, 2016; Das et al., 2020) for finding an optimal feature set of best variables to detect fake news propagators, feature engineering remains uncommon in multimodal dataset modeling. In addition, to the best of our knowledge, a few studies have used unimodal text datasets with preexisting feature sets to detect the behavior of fake news propagators, accounting for feature engineering in multimodal data remains insufficiently explored. Our proposed fake news detector model also differs from other published literature by focusing on investigating the behavior of fake news propagators on psychological, emotional, and cognitive perspectives in a hierarchical manner using the optimized set of best features. Singh et al. (2017) research could be considered the first step towards understanding the behavior of fake news propagators on multimodal datasets; however, a number of questions regarding feature engineering to improve the accuracy of fake news detector models remain to be addressed. Similarly, Kumari and Ekbal (2021), Zhang et al. (2022) contributed to the multimodal fake news detection literature mainly focused on developing the model using deep learning but they devoted less attention to identifying the behavior of fake news propagators and improving the accuracy of the detector with feature engineering tasks. Thus, the use of feature engineering in predicting the behavior of fake news propagators and developing the multimodal fake news detector model is a promising area of research. The main contribution of our research study is summarized below.

1 To the best knowledge, we develop an approach that investigates the behavior of fake news propagators and establishes the relationship between textual, and visual features in predicting misinformation & fake news;
2 We propose a novel multimodal (with text and image features) fake news detector model to examine the behavior of fake news propagators through psychological, cognitive, and emotional perspectives; and
3 After developing the fake news multimodal detector model, we conduct several experiments on real-world datasets to prove the effectiveness of our proposed model.

3. Feature design, data description, and fake news detector model development

The theoretical concepts, Framing Theory, LC4MP, and ELM, described in the last section, helped us to identify the features for developing the fake news detector model. Five broad sub-categories (user-based, content-based, style-based, emotions-based, and manipulation-based) were identified into two categories (text and visual) from this review:

3.1. Text-based features

Linguistic approaches that are used to detect news, which is fake generally, rely on the use of language and its analysis (Singh, Ghosh & Sonagara, 2020). There are two user-based features: user profiling features and user credibility features. User profiling features (account name, geolocation information, registration information, verified, not verified (Zhang & Ghorbani, 2019), has a description and does not have a description) and user credibility features (user's credibility score (Chu et al., 2012), number of users' friends and followers, the ratio between friends and followers, number of tweets post/tweets (Castillo et al., 2011; Zhang & Ghorbani, 2019) have been used in the investigation of suspicious user and non-human accounts by capturing their unique characteristics. Content-based features correspond to the Framing Theory's Issue Selection and central persuasion route for ELM. LIWC-based categories of textual content were used in accordance with the work done by Horne and Adali (2017). Tausczik and Pennebaker give an extensive list of LIWC-based features. LIWC contains five main categories and a few sub-categories, e.g., social, cognitive, affective, and perceptual (Pennebaker, Mehl & Niederhoffer, 2003).

Style-based features correspond to the peripheral persuasion route for ELM. In order to analyze the overall complexity of fake messages or stories, punctuation (e.g., question marks, exclamation points), quotes, negations (e.g., no, never, not), and grammar are used (Singh et al., 2020). The framing theory and previous research is done to investigate fake news to accentuate the importance of the relatability and impression of the overall article, Horne and Adali (2017). Ahmed found that fake writers take longer to write articles and make more mistakes because of the timespan features, the namely the average interval between words, average time span of a word, timespan of document, and editing patter features, such as a number of deletions, arrow keystrokes and "Mouse-Up"s. In the Emotions-based features, Messaris and Abraham (2001), point out that a major determinant of the quality of information is considered by the ELM to be sentiment polarity, Osatuyi and Hughes (2018). LIWC dictionaries are used to measure the effect of words, sentiment score, and SentiStrength which are important measures to identify the intensity of positive and negative emotions (Dickerson et al., 2014; Horne & Adali, 2017; Zhang & Ghorbani, 2019).

3.2. Visual features

Images require a smaller cognitive load than text and are not as intrusive; therefore considered powerful framing tools. This means that the peripheral route in the brain is activated and the users are much more interested to see the visual frames in their mind without questioning it (Singh et al., 2020). Identification of a fake news item using content-based features can be made by examining the image content for clues. It corresponds to Framing Theory's Issue Selection and the central persuasion route for ELM. Features identified can be objects in an image, and the presence and number of faces in an image. Other indicators of fake news are an analysis of color components and properties of images. Style-based features discuss that Fake news can also be identified from features such as size, width, etc., of the image (Blei, Ng & Jordan, 2003; Papadopoulou et al., 2017), and it corresponds to the issue salience component for framing theory and peripheral persuasion route for ELM. In the Emotions-based features, emotions and expressions can also be portrayed by the faces in the image and the presence of violence and so these are also included as features. Pantti and Sirén (2015) showed that when fake news is compared to credible news, it is more visually striking and eye-catching. The images also tend to show accidents, abuse, injury, conflict, and other disturbing material (Jin et al., 2017b).

Manipulation-based features in visual features correspond to the peripheral persuasion route for ELM. Lin et al. (2009) showed that any tampering or manipulation of the image in an article is very important for the determination of fake news. Jin et al. (2017b), Zhang and Ghorbani (2019) investigated visual-based features, namely: similarity distribution histogram, image ratio, clarity score, diversity score, long & hot image ratios, clarity score, and coherence score for identifying the fake news propagators from the data sets. Castillo et al. (2011) shows that the statistical features of an image can also be useful in detecting fake news. The *imagetweet ratio,* from basic image statistics, is already used as a feature for distinguishing fake news, Jin et al. (2017b). One of the most important inspections of images is that most images have a standard resolution, while others do not. Therefore, the different types of tweets can be described in two ways- popularity (hot image) and long image. A hot image is the news event's most popular tweet and popularity is the number of re-tweets of the hot image. A long image is an image with a length-to-width ratio that is greater than 1.9. Generally, it is composed of several different images. The summary of textual and visual features used in the research can be seen in Table 1.

3.3. Data description and model development

One of the most important challenges of building fake news multimodal detector models is obtaining news items that are clearly classified as fake or real. We used several real-world multimodal datasets (Twitter, Weibo, and rFakeEdit) to test our proposed detector model, which has several textual and visual behavioral characteristics of fake news propagators.

We used the Twitter dataset for developing, and testing the fake news detector model that was released by Boididou et al. (2015) as part of a data science challenge (MediaEval, 2015), and the aim of the challenge was to detect the misinformation & fake news content on Twitter website. The Twitter dataset is publically available and has text (in the form of tweets), visual (in the form of images), and additional social contextual dimensions of users. The Twitter dataset encompasses around 18,000 multimodal tweets, of which 10,000 are fake news and 8000 belong to the real news category. In the data pre-processing stage, we deleted a few tweets or messages from the original dataset that do not have any text or image because we are focusing on developing a multimodal detector tool. The other dataset, we used to test the accuracy of our proposed fake news detector model was the Weibo dataset, which is also publically available and used in this research work (Wang et al., 2018). In the Weibo dataset, authors collected the multimodal dataset from the official Chinese news agency (Xinhua agency) and released this dataset to track the misinformation and behaviors of fake news

Table 1
Summary of features (textual and visual) used in developing the fake news detector model.

Feature's Classification	Theoretical Research Support	Sample Text Features	Sample Visual (Image) Features	Empirical Literature Support
User-based Features	Central route processing with ELM framing issue	Username length, number of tweets, has personal URL, is account verified, friends count, followers count, follower-friend ratio, the time listed (number of times the user has been listed/tagged)	Number of hot, and long images in a tweet, popularity score of users (number of shares, re-tweets, and comments obtained by the multimodal tweet)	Singh et al. (2020), Castillo et al. (2011), Jin et al. (2017b)
Content-based Features	Framing: Issue selection ELM: central route processing	Sentistrength, topics for personal concern like-work, home, money, religion, number of external links, tags, and URLs in a tweet, social words like- friends, male referents, family, female referents	Topics of image labels, fraction representation of colours, contrast, dominant color in the focused area, celebrity presence, brightness, gender, number of faces in image	Hong (2013), Shu et al. (2017b), Pérez-Rosas et al., (2017), Singh et al. (2020)
Style-based Features	Peripheral route processing with ELM framing issue	Word or sentence level of features, Average length of tweet, average sentence length, sentence complexity, use of punctuation, use of these types of negations (e.g., no, nope, not, never, fake, etc.), slang terms (e.g., lol, brb, etc.)	Multi-image and hot image ratio, number of images, long image ratio	Castillo et al. (2011), Hong (2013), Conroy et al. (2015), Rubin et al. (2016), Jin et al. (2017b), Horne and Adali (2017), Singh et al. (2020)
Emotions-based Features	ELM: peripheral route processing using LC4MP	Average polarity of sentence, swear words, affect words, anger words, assent, emotional tone, sadness, anxiety related words	Adult content in image, violent content in image, blood content or any other medical content, emotions portrayed in image	Gupta et al. (2013), Jin et al. (2017b), Singh et al. (2020)
Manipulation-based Features	Issue salience Framing	Discrepancy, certainty, number of impersonal and personal pronouns, tentativeness,	Spoofed content, blurred ratio, multi-image and hot image ratio, image ratio II, image sharpness score, long image ratio	Farid (2006), Lin et al. (2009), Pantti and Sirén (2015), Jin et al. (2017b), Horne and Adali (2017), Singh et al. (2020)

propagators in the China region. The authors scrapped this Weibo dataset from 2012 to 2016 and verified it with the Weibo official rumor debunking system.

Fig. 1 presents the four different phases of fake news detector model development: (i) visual data processing using Google vision deep learning architecture, and textual data processing using LC4MP; (ii) textual and visual feature extraction and adding new variables in the original dataset; (iii) modified JAYA feature engineering algorithm development for identifying the important features; and (iv) fake news detector model development. This detector model receives multimodal textual and visual data and the relevant information of Twitter users or news propagators as input.

For the first task, we used the LC4MP framework to extract several textual features (see Table 1) from the original dataset and after that, we used the Google vision deep learning architecture to extract a wide variety of visual features from the images that are associated with the text data. In the next stage, we perform feature engineering and develop a novel M-JAYA algorithm to select the important features for ML model building and remove the noisy and irrelevant features from the dataset. After that, we develop several ML models for identifying the fake news propagators and use the majority voting concept to select the best three ML models based on their rank of classifiers' performances. We use a simple rule in the majority voting section, and output label Y could be predicted based on three top ML hyperplanes (output = mode{M1(Z), M2(Z), M3(Z)}). We can understand it with this example, if the proposed fake news detector model selects the best three ML models based on accuracy and if model_1 classifies the instance as fake, model_2 classifies the same instance in the fake category, and model_3 classifies it in the real news propagator category, then the detector model uses the majority voting concept and provide the final output as a fake class Y= mode {1, 1, 0} =1. This is the final output of our proposed detector model.

4. Feature engineering, M-JAYA algorithm development, experimentation, results, and discussion

Based on the data set, we had 11,498 news with 6643 of them identified as fake news and 4864 of them identified as real news. We accessed them in different features including language features, word features, linguistic features, grammar features, affective processes features, text search, manipulation techniques, objectives, and visual features. The general null hypothesis was set that for each feature the fake news was not statistically different from real news, i.e. fake news propagators did not use any specific features or techniques to produce the fake news. To check the details of the difference between fake news and real news based on the effect of the feature, the list of null hypotheses from research questions with corresponding features were developed in Table 2.

To give the answer to the RQ1, RQ2, RQ3, and RQ4, and to test the significance of the hypotheses, we used several t-tests to test the difference in the language features between fake news and real news in the data set. Tables 2–7 presented the mean and standard deviation for each variable as well as the p-values.

From Table 2, we found that the P value was less than 0.05 for the language features (analytical thinking words, clout words, authentic words, and emotional tone). Therefore, the null hypothesis was rejected, and the conclusion was presented that the fake news propagators write the news using a specific linguistic tone, i.e., using less analytical thinking words, fewer clout words, less authentic words, and less emotional tone. From Table 3, we found that fake news propagators write the same count of total words and dictionary words as real news writers. Also, the fake news propagators write fewer sentences, but more big words (words larger than 6 letters) than the real news. From Table 4 results, we found that fake news propagators write the news using negations (e.g., 'no', 'not') more frequently than real news writers, but use conjunctions less than real news writers. From Table 5 results, we concluded that both fake news and real news use more first-person pronouns but there was no difference in the use between fake news and real news. From Table 6, the conclusion was presented that fake news propagators use more verbs, but fewer comparisons, interrogatives, numbers, and quantifiers than the real news. From Table 7 results, we found that the P value was less than 0.05 for the use of anger, sexual, sadness, and swear words. Therefore, the null hypothesis was rejected, and the conclusion was presented that the fake news propagators use more anger words and swear words, but fewer sexual words and sadness words than the real news propagators. From Tables 8 and 9

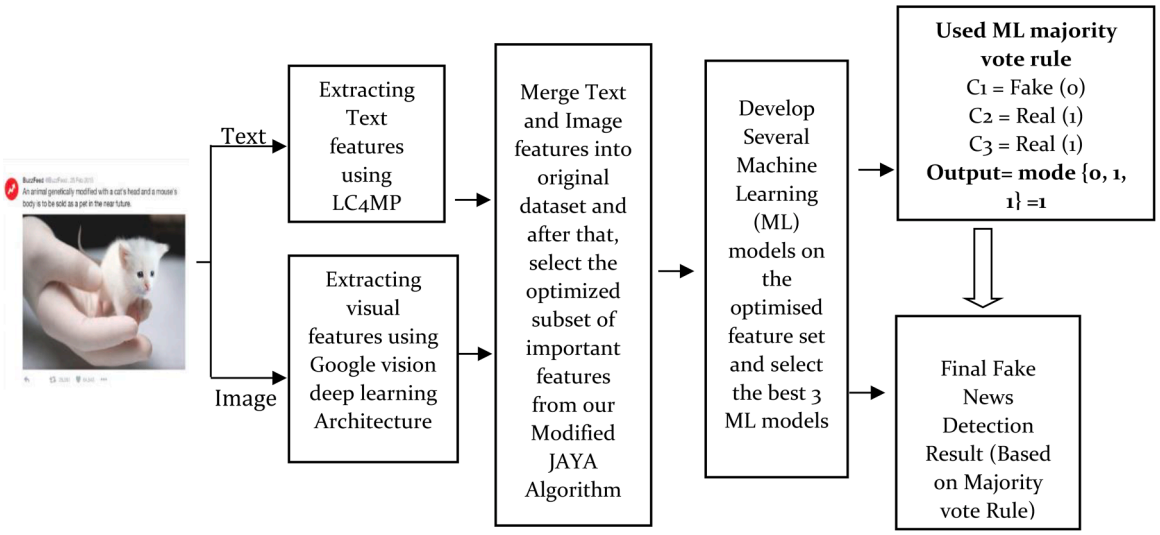


Fig. 1. Proposed Multimodal fake News detector Model.

Table 2
.

Hypothesis	Features	Test
Fake news propagators use the same linguistic tone as real news writers do	Language Features	*t* Test
Fake news is produced under the same rule of the press as the real news	Word Features	*t* Test
Frequency of exclusive words and negations used in fake news is the same as in real news	Linguistic Features	*t* Test
No difference for use of third- and first-person pronouns between fake news and real news	Linguistic Features	*t* Test
Fake news propagators use the same lexical diversity as real news writers do	Grammar Features	*t* Test
The use of anger, sexual, sadness, and swear words in fake news are the same as in real news	Processes Features	*t* Test
No difference in the use of artificial text and violence between fake news and real news	Text Search	Chi-Square Test
Fake news propagators do not use specific manipulation techniques	Manipulation Technique	*t* Test
No specific pattern of difference from objective between fake news and real news	Objectives	Chi-Square Test
Fake news propagators use the same visual features as real news writers do	Visual Features	Chi-Square Test

Table 3
Summary Statistics of Language Features by Fake or Real News.

	Total	Fake News	Real News	*P Value*
Language Features, MEAN (STD)				
Analytical thinking	86.35 (19.19)	85.7 (20.24)	87.24 (17.62)	***<0.0001***
Clout	54.3 (21.34)	52.21 (20.91)	57.15 (21.59)	***<0.0001***
Authentic	15.38 (23.49)	14.71 (23.03)	16.29 (24.08)	***0.0004***
Emotional tone	42.1 (35.57)	40.2 (34.76)	44.7 (36.48)	***<0.0001***

results, the conclusion was presented that the fake news propagators use more artificial text and violence than real news to promote violent extremism and we can identify patterns in the behavior of fake news propagators by checking the "Object" column, i.e., Animal, Building, Person, Sculpture, and Plants and "object" variable could be an important factor for understanding the behavior of fake news propagators.

In the next phase, we used the features in the univariate and multivariate logistic regression (LR) model to understand the behavior of fake news propagators and to evaluate the effect of each feature to develop the fake news detector model. We used the GEE model (Generalized Estimating Equations) to process the analysis. The importance of the GEE model is that it produces efficient estimates of all coefficients by taking the over-time correlations into account when producing the estimates and it will typically be a block-diagonal matrix. The models were fitted to check the effect on the outcome, and the fake news detection rate, from each dependent variable in different categorical features. We also identified whether the effect varies by each variable throughout the feature. For this, we used multivariate logistic regression to control for the specific effect of each variable in the feature, which is allowed by the panel dimension of our dataset, and estimate the following equation for the detection rate of fake news.

Table 4
Summary Statistics of Word Count and Language Features by Fake or Real News.

	Total	Fake News	Real News	*P Value*
Word Features, MEAN (STD)				
Word Count	16.65 (20.65)	16.4 (20.44)	17 (20.93)	*0.1247*
Words / Sentence	10.68 (5.58)	10.58 (5.69)	10.81 (5.43)	***0.0304***
Words > 6 letters	22.9 (9.79)	23.37 (10.06)	22.27 (9.37)	***<0.0001***
Dictionary words	49.36 (14.5)	49.5 (15.05)	49.19 (13.71)	*0.2553*

Table 5
Summary Statistics of Linguistic Features by Fake or Real News (from cognitive perspectives).

	Total	Fake News	Real News	*P Value*
Linguistic Features, MEAN (STD)				
Conjunctions	1.02 (2.50)	0.96 (2.51)	1.10 (2.47)	***0.0030***
Negations	0.62 (2.02)	0.68 (2.14)	0.54 (1.83)	***0.0004***

Table 6
Summary Statistics of Linguistic Features by Fake or Real News (from psychology perspectives).

	Total	Fake News	Real News	*P Value*
Linguistic Features, MEAN (STD)				
Difference of singular use	−0.45 (2.14)	−0.48 (2.31)	−0.40 (1.89)	*0.0502*
Difference of plural use	−0.04 (1.56)	−0.02 (1.61)	−0.07 (1.50)	*0.0601*

$$P(FakeNewsDetectioni) = \alpha + \beta 1 Languagei + \beta 2 Wordi + \beta 3 Linguistici + \beta 4 Grammari + \beta 5 Processi + \beta 6 Texti + \beta 7 Manipulationi + \beta 8 Objectivei + \beta 9 Visuali + \mu i + \varepsilon i$$

The outcome variables were identified as the probability of news being detected as fake news. The independent variables included language features, word features, linguistic features, grammar features, process features, text search, manipulation techniques (copy-move techniques and splicing techniques), objectives, and visual features. Error term, μi included an idiosyncratic error term, εi and fixed effect (μi). The idea is that if given news was with some specific features, i.e., specific linguistic tone, certain rules of the press, more lexical diversity, or particular objectives, etc., we would expect to see the rate pattern of detection as fake news. Therefore, β1 through β9 capture the average effect of the different features on this pattern. Table 10 presented the output of the model below.

Table 7
Summary Statistics of Grammar Features by Fake or Real News.

	Total	Fake News	Real News	*P Value*
Grammar Features, MEAN (STD)				
Common verbs	5.45 (6.06)	5.67 (6.33)	5.15 (5.66)	***<0.0001***
Common adjectives	8.89 (6.08)	8.89 (6.44)	8.90 (5.54)	*0.8930*
Comparisons	0.72 (2.15)	0.63 (2.02)	0.85 (2.31)	***<0.0001***
Interrogatives	0.46 (1.65)	0.43 (1.67)	0.50 (1.62)	***0.0262***
Numbers	0.90 (2.95)	0.81 (2.66)	1.04 (3.30)	***<0.0001***
Quantifiers	0.43 (1.59)	0.36 (1.48)	0.52 (1.73)	***<0.0001***

Table 8
Summary Statistics of Affective Processes Features by Fake/Real News (emotional perspectives).

	Total	Fake News	Real News	*P Value*
Affective Processes Features, MEAN (STD)				
Anger words	0.57 (2.18)	0.61 (2.28)	0.50 (2.04)	***0.0113***
Sexual words	0.14 (1.04)	0.12 (0.94)	0.17 (1.17)	***0.0026***
Sadness words	0.22 (1.20)	0.15 (0.95)	0.32 (1.46)	***<0.0001***
Swear words	0.64 (2.41)	0.71 (2.53)	0.54 (2.23)	***0.0003***

Based on the model output, the predicted probability plots of fake news detection are shown in Fig. 2, we can conclude that

- Fake news detection was significantly affected by language features. The decrease of analytical thinking words, clout words, authentic words, and emotional tone would increase the possibility of the detection of fake news. In addition, the clout words had more effect than other words.
- From the word features, sentences had a positive effect and big words (words larger than 6 letters) had a negative effect on the fake news detection rate. The strengths of the effect were almost the same. But dictionary works did not have any effect on the detection rate.
- For the linguistic features, if negations would be found 20 % more in a news than others, the probability of this news being detected as fake news would be increased by 15 %−20 %. In addition, if news found 20 % more conjunctions, the probability of fake news detection would be decreased by 10 %.
- By increasing the detection rate of 10 % for the grammar features (including verbs, comparisons, interrogatives, numbers, and quantifiers), fake news detected possibility varied from less than 5 % to around 20 %. Among these variables, quantifiers had the largest effect. In other words, more quantifiers in a news would make less possibility of fake news.
- Anger words and swear words had a positive effect, and sexual words and sadness words had a negative effect on fake news detection. All the effects were significant.
- Fake news detection rate was significantly affected by image ratio. If this ratio was increased by 60 %, the probability of fake news detection would be increased by 20 %, which means that the copy-move technique played an important role in the process of fake news producing.
- For the splicing technique used in producing fake news, the effect from the visual score (visual clarity score, visual clustering score, visual coherence score, and visual diversity score) reflected this situation. A lower score indicated a higher rate of fake news detection.

Table 9
Summary Statistics of Text Search by Fake/Real News (emotional perspective in visual features).

	Total	Fake News	Real News	*P Value*
Text Search: Adult, N (%)				***<0.0001***
Very unlikely	11,188	6412 (96.7 %)	4776 (98.2 %)	
Unlikely	307	222 (3.3 %)	85 (1.7 %)	
Possible	3	–	3 (0.1 %)	
Text Search: Spoof, N (%)				***<0.0001***
Very unlikely	9498	5227 (78.8 %)	4271 (87.8 %)	
Unlikely	1142	963 (14.5 %)	179 (3.7 %)	
Possible	382	310 (4.7 %)	72 (1.5 %)	
Likely	101	76 (1.2 %)	25 (0.5 %)	
Very likely	375	58 (0.9 %)	317 (6.5 %)	
Text Search: Medical, N (%)				***<0.0001***
Very unlikely	10,965	6123 (92.3 %)	4842 (99.5 %)	
Unlikely	533	511 (7.7 %)	22 (0.5 %)	
Text Search: Violence, N (%)				***<0.0001***
Very unlikely	8887	4735 (71.4 %)	4152 (85.4 %)	
Unlikely	2607	1898 (28.5 %)	709 (14.5 %)	
Possible	4	1 (0.1 %)	3 (0.1 %)	
Text Search: Racy, N (%)				***<0.0001***
Very unlikely	9601	5426 (81.8 %)	4175 (85.8 %)	
Unlikely	1677	1066 (16.1 %)	611 (12.6 %)	
Possible	55	27 (0.4 %)	28 (0.6 %)	
Likely	139	89 (1.3 %)	50 (1.0 %)	
Very likely	26	26 (0.4 %)	–	

Table 10
Multivariate Generalized Linear Mixed Model for Fake New Detection by each Feature through Logistic Regression.

	aOR (95 %CI)	*P*-Value
Total Number of News	11,498	–
Language Features		
Analytical thinking	0.996 (0.994,0.998)	0.0003
Clout	0.989 (0.988,0.991)	<0.0001
Authentic	0.996 (0.994,0.997)	<0.0001
Emotional tone	0.997 (0.996,0.998)	<0.0001
Word Features		
Words / Sentence	0.991 (0.984,0.998)	0.0078
Words > 6 letters	1.014 (1.010,1.018)	<0.0001
Dictionary words	1.005 (1.002,1.007)	0.0006
Linguistic Features		
Conjunctions	0.976 (0.961,0.990)	0.0011
Negations	1.038 (1.018,1.058)	0.0002
Grammar Features		
Common verbs	1.020 (1.013,1.027)	<0.0001
Common adjectives	1.005 (0.999,1.012)	0.1248
Comparisons	0.954 (0.937,0.971)	<0.0001
Interrogatives	0.960 (0.938,0.982)	0.0005
Numbers	0.975 (0.963,0.988)	0.0001
Quantifiers	0.943 (0.921,0.966)	<0.0001
Affective Processes Features		
Anger words	1.002 (0.974,1.031)	0.8927
Sexual words	0.893 (0.856,0.930)	<0.0001
Sadness words	0.888 (0.859,0.917)	<0.0001
Swear words	1.053 (1.025,1.081)	0.0002
Copy-move technique		
Image Ratio	1.012 (1.008,1.016)	<0.0001
Long Image Ratio	1.000 (0.996,1.003)	0.8277
Hot Image Ratio	0.964 (0.953,0.974)	<0.0001
Image Sharpness	1.030 (1.015,1.044)	<0.0001
Blurred Ratio	1.017 (1.013,1.021)	<0.0001
Splicing technique		
Visual Diversity Score	0.993 (0.991,0.995)	<0.0001
Visual Clarity Score	0.988 (0.987,0.989)	<0.0001
Visual Coherence Score	0.958 (0.955,0.961)	<0.0001
Visual Clustering Score	1.034 (1.032,1.036)	<0.0001
Test Search		
Adult	0.387 (0.269,0.557)	<0.0001
Spoof	0.559 (0.523,0.597)	<0.0001
Medical	2.255 (1.420,3.582)	<0.0001
Violence	5.061 (4.347,5.892)	<0.0001
Racy	0.907 (0.820,1.003)	0.0571
Objective		
Animal	4.293 (3.321,5.551)	<0.0001
Building	0.861 (0.755,0.982)	0.0257
Person	1.617 (1.466,1.784)	<0.0001
Sculpture	3.223 (1.820,5.708)	<0.0001
Plant	0.942 (0.680,1.304)	0.7175
Visual Feature		
Face of Joy	2.447 (2.038,2.937)	<0.0001
Face of Exposed	1.210 (1.101,1.329)	<0.0001
Face of Blurred	0.415 (0.379,0.455)	<0.0001
Face of Headwear	0.606 (0.514,0.714)	<0.0001

- By checking text search of “Adult”, “Medicine”, “Racy”, “Violence”, and “Spoof”, except “Spoof”, all the other texts had a positive effect on the detection of fake news, which means that fake news propagators use more artificial text and violence to promote violent extremism.
- On the behavior of information identification, if the objectives came to “Sculpture”, “Plant” and “Animal”, the probability of fake news increased by 35 %−45 %; if the objectives were chosen from “Person” or “building” the rate decreased by 15 %−25 %.
- From the last section of the plot, all the visual feature (face of joy, face of exposed, face of blurred, and face of headwear) had a negative effect on the fake news detection possibility, which presented opposite results to our research questions. That was because the sample power was too small. Based on the summary statistics, news without faces shown took place about 90–95 %. Therefore, only around 5 %−10 % of the sample was used to check the effect of visual features on fake news detection, which was a lack of power. Fake news multimodal messages use a specific manipulation technique, i.e., resampling/copy-move/ splicing. For instance, fake news was detected using a larger image ratio, but less hot image ratio and less image sharpness. In addition, fake news was found with a significantly lower score of visual diversity, visual clarity, visual coherence, and visual clustering.

After understanding the characteristics and behavior of fake news propagators, we started to work on the first phase of any ML project-data processing and feature engineering. One of the main contributions of this research is to build a novel fake news detector framework based on an optimized set of important features. In order to identify the best features and understand the importance of individual variables of the fake news detector model, we tried to control the parameters of the JAYA algorithm and we modified it to design the optimal subset of important features for improving the ML classifier’s performance. The detailed description of the modified JAYA algorithm is described in detail below. In the first step, we initialize JAYA algorithm parameters and four parameters need to be initialized: Population Size **(P)**, Number of Executions **(E)**, Maximum Number of Iteration **(MaxIter)**, and the minimum number of features **(D)**. Once parameters are initialized, it is essential to define the training and testing datasets. After that, the algorithm needs the objective function to start the process. This objective function aims to select the optimal subset of the features that minimize the fitness function of the selected classification models. The error is computed as the difference between the actual output, and the selected model estimation is presented by Eq. (1), where $k = 1, 2, \ldots m$ and m is the number of testing observations $m = M_{ma}^{Te}$ for a dataset.

The fitness function values of the model are computed by dividing the summation of errors by the number of observations presented by Eq. (2).

$$\boldsymbol{Error}(\boldsymbol{k}) = (\hat{\boldsymbol{y}}(\boldsymbol{k}) \neq \boldsymbol{y}(\boldsymbol{k})) \tag{1}$$

For each population p where, $p = 1, 2, \ldots, P$, the objective function (Eq. (2)) is calculated using Eq. (1).

$$Fitness(p) = \frac{\sum_{k=1}^{m} Error(k)}{m} \tag{2}$$

After that, the binary JAYA population is reconstructed using the S-shape transfer function and a selected subset of features and this S-shape function is used to convert the input to binary values (Kennedy & Eberhart, 1997). After that, the updated solution vector $X(d)$ will be assigned into Eq. (3) to calculate its binary vector.

$$\boldsymbol{B}(\boldsymbol{X}_{\boldsymbol{new}}(\boldsymbol{p}, \boldsymbol{d})) = \frac{1}{1 + \boldsymbol{exp}^{\left(-5 \times X_{new(p,d)} - 0.25\right)}} \tag{3}$$

The binary value from Eq. (3) further needs to label X_{new} into either `1′ or `0′. Two methods are proposed here. For *Random* r

$$\boldsymbol{X}_{\boldsymbol{new}} = \begin{cases} 1, & B(\boldsymbol{X}_{\boldsymbol{new}}(\boldsymbol{p}, \boldsymbol{d}))\rangle rand \\ 0 & Otherwise \end{cases} \tag{4}$$

Here, rand () is the random number distributed between the range of 0 and 1.

Default r

$$X_{new} = \begin{cases} 1, & B(X_{new}(p, d))\rangle 0.5 \\ 0 & Otherwise \end{cases} \tag{5}$$

The next step would be to execute the binary JAYA algorithm and construct the initial binary populations. In this step, each binary population is generated using a random selection of features. Later on, the fitness function cost of each generated population is calculated. In the next step, the best and the worse solutions are selected based on the objective function’s value. If it selects (objective function = error function) option, the minimum value will generate the best solution, and

the maximum value will generate the worse. If it selects (objective function = AUC score) option, the maximum value will generate the best solution, and the minimum value will generate the worse.

In the next improvement step, each and every decision variable of all iterations are modified using the JAYA operator formulated in Eq. (6) (Awadallah, Al-Betar, Hammouri & Alomari, 2020).

$$X_{new}(p,d) = x(p,d) + r1 \times (bestX - |x(p,d)|) - r2(worseX - |x(p,d)|) \quad (6)$$

In the next step of the selection process, when a new solution is generated, it compares to the current solution Xp that is stored in the population's p^{th} position. Suppose the fitness value of a new solution Xnew is better than the current solution Xp and in this case, the new solution will replace the current solution. This new fitness is compared with the old fitness.

- Objective function = error function: If the new fitness value is less than the old fitness value, its corresponding population is selected for the next iteration and otherwise dropped.
- Objective function = AUC score: If the new fitness value is more than the old fitness value, its corresponding population is selected for the next iteration and otherwise dropped.

The last step is the stop criterion. The Second to fourth steps are repeated until either one of the following conditions is satisfied:

- The maximum number of iterations (i.e., Max_itr) are satisfied.
- Only one population is left.

In the next phase, the effectiveness of the modified JAYA algorithm is tested and compared with several other methods on a few benchmark datasets (PIMA, Musk, Sonar, Madelon, Colon, Leukemia, and Vehicle) that are available freely on the UCI machine learning repository (Table 11–13).

5. Method validation and performance comparison with other studies

For methodological validation, we tested our fake news detector model on the other two datasets – Weibo and, r/Fakeddit. We developed and repeated the data pre-processing and feature optimization steps on two new datasets and our proposed model provided better classification accuracy as compared to other studies. Table 14 shows the accuracy and performance comparison of the proposed fake news detector model with other several published works of literature EANN (Wang et al., 2018), MVAE (Khattar et al., 2019), and cultural algorithm-based multimodal (Shah & Kobti, 2020), and observed that all the key learning performance metrics (Recall and Accuracy) of our model are better than other models. In comparison with other fake news detector models, our modified-JAYA model achieves better Recall (95.5 %), and AUC (96.9 %) scores on the Weibo dataset. We have achieved high accuracy and Recall scores on the third rFakeEdit dataset. Over proposed fake news detector model on the Weibo and rFakeEdit datasets gains 4 % and 8 % improvement in accuracy over the cultural algorithm-based model,

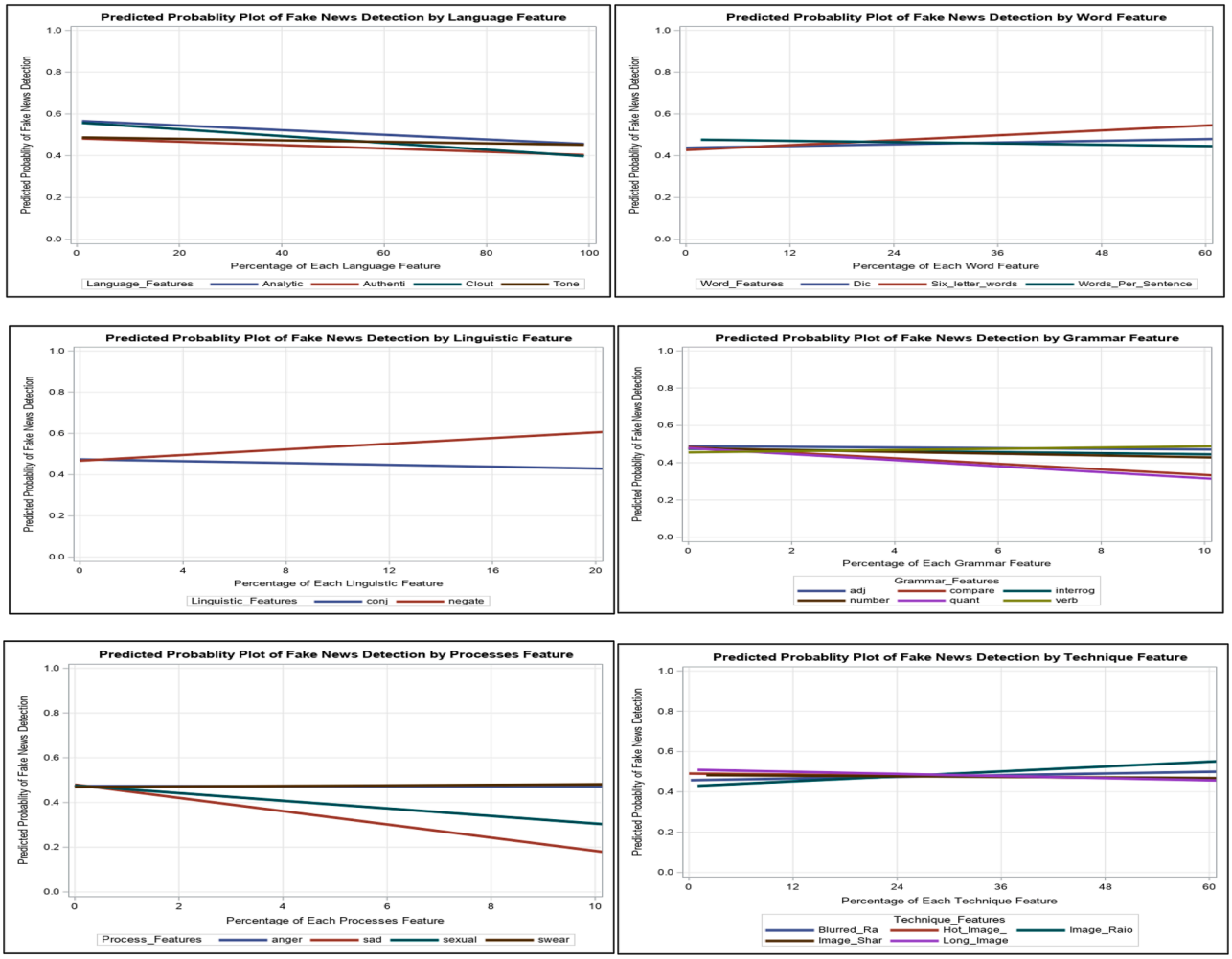


Fig. 2. Predicted Probability Plots of Fake News Detection by each Features.

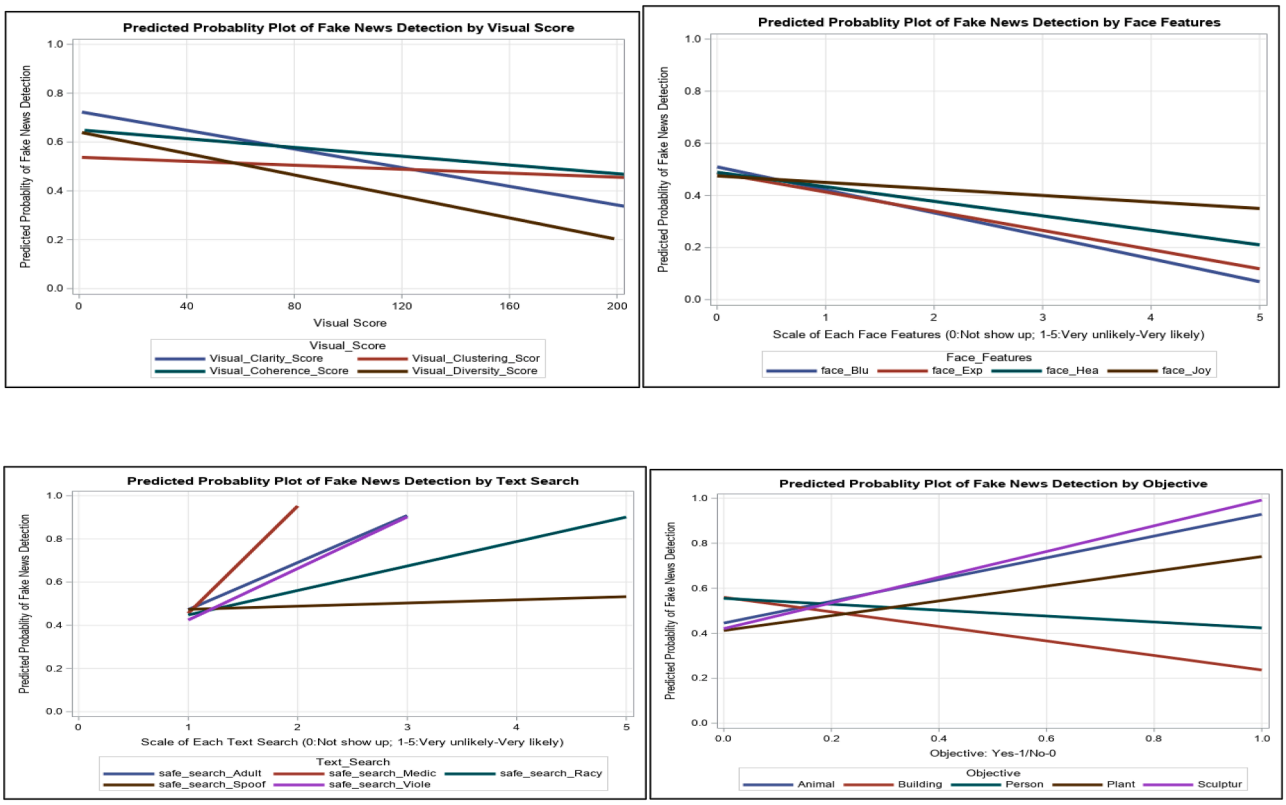


Fig. 2. (*continued*).

Table 11
Performance comparison of all ML models with and without feature engineering.

	Without feature engineering				After using feature engineering by the M-JAYA algorithm			
Algorithms	**Precision**	**Recall**	**F-1**	**AUC**	**Precision**	**Recall**	**F-1**	**AUC**
XGBoost Method	0.716	0.713	0.732	0.713	0.873	0.842	0.882	0.880
LR (Logistic Regression)	0.712	0.696	0.705	0.759	0.804	0.811	0.843	0.856
LSTM	0.693	0.682	0.663	0.747	0.826	0.816	0.811	0.831
ANN	0.670	0.641	0.681	0.731	0.690	0.713	0.742	0.775
RNN	0.641	0.632	0.649	0.736	0.659	0.717	0.729	0.784
RBF- SVM	0.603	0.581	0.592	0.641	0.677	0.688	0.729	0.797
Logistic Regression	0.567	0.734	0.657	0.731	0.739	0.723	0.745	0.781
k-NN ($k = 10$)	0.556	0.712	0.656	0.723	0.652	0.723	0.744	0.784
Naïve Bayes	0.642	0.591	0.614	0.671	0.689	0.614	0.723	0.741
Random Forest	0.644	0.581	0.589	0.647	0.666	0.681	0.682	0.728

Table 12
Final Model Results.

	After using feature engineering by the M-JAYA algorithm (Top 3 ML models)				After applying majority vote rule on feature engineering results Y= mode {0, 1, 1} =1			
Algorithms	**Precision**	**Recall**	**F-1**	**AUC**	**Precision**	**Recall**	**F-1**	**AUC**
XGBoost	0.873	0.842	0.882	0.880	0.932	0.955	0.966	0.969
LR	0.804	0.811	0.843	0.856				
LSTM	0.826	0.816	0.811	0.831				

Table 13
Performance improvement in ML after incorporating visual features.

Algorithms	**Precision (%)**	**Recall (%)**	**F-1 (%)**	**AUC (%)**
XGBoost method	+8.6***	+3.9***	+6.8 %***	+6.2***
LR	+2.6	+4.6	−2.6	+2.4
LSTM	3.9	+4.8	+6.2	+6.8

*** denote the significance at the 0.001 level.

MVAE, and EANN multimodal models.

6. Conclusion, managerial implications, and future research directions

In this research work, we developed a novel M-JAYA-based fake news detector model that can help social media managers, society, organizations, and online users to detect misinformation and identify the

Table 14
Performance comparison of the proposed fake news detector model with other models on Weibo dataset.

Algorithms	Precision	Recall	F-1	AUC
EANN Model	0.795	0.806	0.795	0.800
MVAE Model	0.824	0.854	0.769	0.809
Cultural algorithm-based multimodal	0.891	0.873	0.822	0.932
Our proposed Model	0.932	0.955	0.966	0.969

suspicious behavior of fake news propagators through social media posts. However several fake news detector models have already been developed, but none of these fully exploits data pre-processing challenges and optimized feature selection, and the characteristics of fake propagators' behavior to understand the fake news phenomenon in new dimensions. Our analysis revealed several new deceptions dimensions in investigating the behavior of fake news propagators. We proved that the decrease of analytical thinking words, clout words, authentic words, and emotional tone can increase the possibility of the detection of fake news. Fake news propagators use more anger and swear words and use the copy-move technique to temper the images in writing the fake news. The fake news detection rate was significantly affected by the image ratio. If this ratio was increased by 60 %, the probability of fake news detection would be increased by 20 %. The effect from the visual score reflected this situation and a lower score indicated a higher rate of fake news detection. We also proved that multimodal analysis can improve the accuracy of the fake news detector model if feature engineering is performed on the dataset before building the model. For the feature engineering task, a new modified JAYA algorithm has been proposed in this research work to find an optimal subset of important features and to improve the accuracy of the fake news detector model. The proposed model selects the best variables/features from the raw dataset and removes the unimportant, noisy, redundant, and irrelevant features that do not contribute to improving the accuracy of the detector model. The experiments and results show that our proposed model is removing the unimportant features effectively and giving better classification results as compared to other traditional feature selection models. The limitation of the proposed feature engineering approach is the high computational cost which is due to searching for the optimal subset of important variables and removing the noisy and irrelevant variables which lead sometimes to the loss of information or important variables.

Developing a fake news & disinformation detection model carries significant managerial implications in businesses, as it involves a combination of technical, ethical, and strategic considerations. A disinformation detection model can improve the decision-making power of managers and it can help to make better decisions by providing them with accurate information about the news they are reading on social media channels; especially important in situations where the news is related to business decisions, such as investment decisions or new product launches. A fake news detection model can help to reduce the risk of businesses being harmed by fake news and increase trust in organizations. For example, a fake news about a company's financial performance can lead to investors selling their shares, which could damage the company's stock price. When investors and customers know that companies are taking steps to combat fake news, they are more likely to believe the information that they are being told. This can help businesses to protect their reputation and avoid negative publicity. So, businesses can use the detector model and can enhance their brand reputation to identify and remove fake news stories that are damaging their brand reputation. This can help businesses to improve the public's perception and make it more attractive to customers and investors. Overall, a misinformation or disinformation detection model development can have a number of positive implications for organizations. By helping to improve decision-making, reduce risk, increase trust, improve public relations, and enhance brand reputation, a fake news detection model can be a valuable tool for businesses in today's digital age. Feature engineering, which plays an important role in the development of an explainable fake news detection system by shaping the features used in the ML models, is an interconnected concept with explainable AI as it affects the model's interpretability.

In future research work, multimodal visual and text variables identified in this research could be used to develop a single web application or dashboard that could help social media managers and online users to understand the consumption of their own misinformation and disinformation on daily/hourly basis. Our proposed model can also be applied to other broad and diversified applications (e.g. big NLP datasets, biomedical and clinical datasets, etc.) in which the datasets have a large number of features or variables.

References

Agarwal, P., Al Aziz, R., & Zhuang, J (2022). Interplay of rumor propagation and clarification on social media during crisis events-A game-theoretic approach. *European Journal of Operational Research, 298*(2), 714–733.

... & Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., & Herrera, F. (2023). Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial Intelligence. *Information Fusion*, Article 101805.

Awadallah, M. A., Al-Betar, M. A., Hammouri, A. I., & Alomari, O. A. (2020). Binary JAYA algorithm with adaptive mutation for feature selection. *Arabian Journal for Science and Engineering*.

Barthel, M., Mitchell, A., & Holcomb, J. (2016). Many Americans believe fake news is sowing confusion. https://www.pewresearch.org/journalism/2016/12/15/many-americans-believe-fake-news-is-sowing-confusion/.

Begley, J. (2017). The rise of the image: Every NY times front page since 1852 in under a minute. Colossal. Retrieved from https://www.thisiscolossal.com/2017/02/the-rise-of-the-image-everyny-times-front-page-since-1852-in-under-a-minute/.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Boididou, C., Andreadou, K., Papadopoulos, S., Dang-Nguyen, D. T., Boato, G., Riegler, M., & Kompatsiaris, Y. (2015). Verifying multimedia use at mediaeval 2015. *MediaEval, 3*(3), 7.

Borchert, P., Coussement, K., De Caigny, A., & De Weerdt, J. (2023). Extending business failure prediction models with textual website content using deep learning. *European Journal of Operational Research, 306*(1), 348–357.

Cacioppo, J. T., Cacioppo, S., & Petty, R. E. (2018). The neuroscience of persuasion: A review with an emphasis on issues and opportunities. *Social Neuroscience, 13*(2), 129–172.

Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on Twitter. In *Proc. 20th Int. Conf. World Wide Web* (pp. 675–684).

Chaiken, S., & Eagly, A. H. (1983). Communication modality as a determinant of persuasion: The role of communicator salience. *Journal of Personality and Social Psychology, 45*(2), 241–256.

Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology, 39*(2), 752–766.

Chang, Y., Keblis, M. F., Li, R., Iakovou, E., & White, C. C., III (2022). Misinformation and disinformation in modern warfare. *Operations Research, 70*(3), 1577–1597.

Chu, Z., Gianvecchio, S., Wang, H., & Jajodia, S. (2012). Detecting automation of twitter accounts: Are you a human, bot, or cyborg? *IEEE Transactions on Dependable and Secure Computing, 9*(6), 811–824.

Conroy, N. J., Rubin, V. L., & Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. In *Proceedings of the 78th ASIS&T annual meeting*.

Das, H., Naik, B., & Behera, H. (2020). A Jaya algorithm based wrapper method for optimal feature selection in supervised classification. *Journal of King Saud University - Computer and Information Sciences*.

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). Science of fake news. *Science (New York, N.Y.), 359*(6380), 1094–1096.

De Vreese, C. H. (2005). News framing: Theory and typology. *Information design journal+ document design, 13*(1), 51–62.

Dickerson, J. P., Kagan, V., & Subrahmanian, V. (2014). Using sentiment to detect bots on Twitter: Are humans more opinionated than bots?. In *2014 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM)* (pp. 620–627).

Farid, H. (2006). Digital doctoring: How to tell the real from the fake. *Significance, 3*(4), 162–166.

Gabielkov, M., Ramachandran, A., Chaintreau, A., & Legout, A. (2016). Social clicks: What and who gets read on Twitter?. In *ACM Sigmetrics /IFIP Performance 2016*. France: Antibes Juan-les-Pins.

George, J. F., Gupta, M., Giordano, G., Mills, A. M., Tennant, V. M., & Lewis, C. C (2018). The effects of communication media and culture on deception detection accuracy. *MIS Quarterly, 42*(2), 551–575.

Gottfried, J., & Shearer, E. (2016). News use across social media platforms 2016.

Gupta, A., Lamba, H., Kumaraguru, P., & Joshi, A. (2013). Faking sandy: Characterizing and identifying fake images on twitter during hurricane sandy. In *Proceedings of the 22nd international conference on world wide web* (pp. 729–736). ACM.

Han, Y., Lappas, T., & Sabnis, G. (2020). The importance of interactions between content characteristics and creator characteristics for studying virality in social media. *Information Systems Research*.
Hong, S. C. (2013). Scare sells? A framing analysis of news coverage of recalled Chinese products. *Asian Journal of Communication, 23*(1), 86–106.
Horne, B. D., & Adali, S. (2017). This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. *Association for the Advancement of Artificial Intelligence*.
Hunt, A., & Gentzkow, M. (2017). Social Media and fake news in the 2016 election. *Journal of Economic Perspectives, 31*(2), 211–236.
Jin, Z., Cao, J., Zhang, Y., Zhou, J., & Tian, Q. (2017a). Novel visual and statistical image features for microblogs news verification. *IEEE Transactions on Multimedia, 19*(3), 598–608.
Jin, Z., Cao, J., Guo, H., Zhang, Y., & Luo, J. (2017b). Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 2017 ACM on multimedia conference* (pp. 795–816). ACM.
Kennedy, J., & Eberhart, R. C. (1997). A discrete binary version of the particle swarm algorithm. In *, 5. IEEE International conference on systems, man and cybernetics. Computational and cybernetics and simulations* (pp. 4104–4108).
Khattar, D., Goud, J. S., Gupta, M., & Varma, V. (2019). Mvae: Multimodal variational autoencoder for fake news detection. In *The web conference* (pp. 2915–2921).
Kim, A., & Dennis, A. R. (2020). Says who? The effects of presentation format and source rating on fake news in social media. *MIS Quarterly, 43*, 3.
Kumar, A., Gopal, R. D., Shankar, R., & Tan, K. H. (2022). Fraudulent review detection model focusing on emotional expressions and explicit aspects: investigating the potential of feature engineering. *Decision Support Systems, 155*, 113728.
Kumari, R., & Ekbal, A. (2021). Amfb: Attention based multimodal factorized bilinear pooling for multimodal fake news detection. *Expert Systems with Applications, 184*, Article 115412.
Lin, Z., He, J., Tang, X., & Tang, C. K. (2009). Fast, automatic and fine-grained tampered JPEG image detection via DCT coefficient analysis. *Pattern Recognition, 42*(11), 2492–2501.
MediaEval (2015). http://www.multimediaeval.org/mediaeval2015/.
Messaris, P., & Abraham, L. (2001). The role of images in framing news stories. *Framing public life* (pp. 231–242). England, UK: Routledge: Abingdonon- Thames.
Moravec, P., Minas, R. A., & Dennis, A. R. (2019). Fake news on social media: People believe what they want to believe when it makes no sense at all. *MIS Quarterly, 43*(4), 1343–1360.
Moravec, P., Minas, R. A., & Dennis, A. R. (2020). Fake news on social media: People believe what they want to believe when it makes no sense at all. *MIS Quarterly* (in press).
Osatuyi, B., & Hughes, J. (2018). A tale of two internet news platforms-real vs. fake: An elaboration likelihood model perspective. In *Proceedings of the 51st Hawaii International Conference on System Sciences*.
Pérez-Rosas, V., Kleinberg, B., Lefevre, A., & Mihalcea, R. (2017). Automatic detection of fake news. arXiv preprint arXiv: 1708.07104.
Pantti, M., & Sirén, S. (2015). The fragility of photo-truth: Verification of amateur images in Finnish newsrooms. *Digital Journalism, 3*(4), 495–512.
Papadopoulou, O., Zampoglou, M., Papadopoulos, S., & Kompatsiaris, Y. (2017). Web video verification using contextual cues. In *Proceedings of the 2nd international workshop on multimedia forensics and security, June* (pp. 6–10). ACM.
Papanastasiou (2020). Fake news propagation and detection: A sequential model, Management Science.
Peng, X., & Xintong, B. (2022). An effective strategy for multi-modal fake news detection. *Multimedia Tools and Applications, 81*(10), 13799–13822.
Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology, 54*(1), 547–577.
Rao, R. (2016). Jaya: A simple and new optimization algorithm for solving constrained and unconstrained optimization problems. *International Journal of Industrial Engineering Computations, 7*(1), 19–34.
Rapoza K. 2017. "Can 'fake news' impact the stock market?" www.forbes.com/sites/kenrapoza/2017/02/26/can-fake-news-impact-the-stock-market/).
Reisach, U. (2021). The responsibility of social media in times of societal and political manipulation. *European Journal of Operational Research, 291*(3), 906–917.
Rubin, V. L., Chen, Y., & Conroy, N. J. (2015). Deception detection for news: Three types of fakes. In *Proceedings of the 78th ASIS&T annual meeting: information science with impact: Research in and for the community* (p. 83). American Society for Information Science.
Rubin, V., Conroy, N., Chen, Y., & Cornwell, S. (2016). Fake news or truth? Using satirical cues to detect potentially misleading news. In *Proceedings of the second workshop on computational approaches to deception detection* (pp. 7–17).
Shah, P., & Kobti, Z. (2020). Multimodal fake news detection using a cultural algorithm with situational and normative knowledge. In *2020 IEEE Congress on evolutionary computation (CEC), July* (pp. 1–7).
Shu, K., Wang, S., & Liu, H. (2017). Exploiting tri-relationship for fake news detection. ArXiv e-prints.
Silverman, C. 2016. This analysis shows how fake election news stories outperformed real news on Facebook. https://www.buzzfeed.com/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook.
Silverman, C. and J. Singer-Vine. 2016. Most Americans who see fake news believe it, new survey says. https://www.buzzfeed.com/craigsilverman/fake-news-survey.
Singh, V. K., Ghosh, I., & Sonagara, D. (2020). Detecting fake news stories via multimodal analysis. *Journal of the Association for Information Science and Technology*, 1–15.
Singh, V., Dasgupta, R., Sonagra, D., Raman, K., & Ghosh, I. (2017, July). Automated fake news detection using linguistic analysis and machine learning. In *International conference on social computing, behavioral-cultural modeling, & prediction and behavior representation in modeling and simulation (SBP-BRiMS)* (pp. 1–3).
Singhal, S., Shah, R. R., Chakraborty, T., Kumaraguru, P., & Satoh, S. (2019). SpotFake: A multi-modal framework for fake news detection. In *Proceedings of the 2019 IEEE fifth international conference on multimedia big data (BigMM)*.
Tanınmış, K., Aras, N., & Altınel, İ. K. (2022). Improved x-space algorithm for min-max bilevel problems with an application to misinformation spread in social networks. *European Journal of Operational Research, 297*(1), 40–52.
Tian, D., et al. (2013). A review on image feature extraction and representation techniques. *International Journal of Multimedia and Ubiquitous Engineering, 8*(4), 385–396.
Uppada, S. K., & Patel, P. (2022). An image and text-based multimodal model for detecting fake news in OSN's. *Journal of Intelligent Information Systems*, 1–27.
Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science (New York, N.Y.), 359*(6380), 1146–1151.
Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., & Gao, J. (2018). EANN: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 849–857).
Wu, K., Yang, S., & Zhu, K. Q (2015). False rumors detection on sina weibo by propagation structures. In *Data engineering (ICDE), 2015 IEEE 31st international conference on* (pp. 651–662). IEEE.
Yang, Y., Zheng, L., Zhang, J., Cui, Q., Li, Z., Yu, P.S.: Ti-CNN: Convolutional neural networks for fake news detection. arXiv preprint arXiv:1806.00749 (2018).
Yfanti, S., Karanasos, M., Zopounidis, C., & Christopoulos, A. (2023). Corporate credit risk counter-cyclical interdependence: A systematic analysis of cross-border and cross-sector correlation dynamics. *European Journal of Operational Research, 304*(2), 813–831.
Zafarani, R., Zhou, X., Shu, K., & Liu, H. (2019). Fake news research: Theories, detection strategies, and open problems. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 3207–3208).
Zhang, X., & Ghorbani, A. A. (2019). An overview of online fake news: Characterization, detection, and discussion. *Information Processing and Management*, 1–26.
Zhang, C., Gupta, A., Kauten, C., Deokar, A. V., & Qin, X. (2019). Detecting fake news for reducing misinformation risks using analytics approaches. *European Journal of Operational Research, 279*(3), 1036–1052.
Zhang, G., Giachanou, A., & Rosso, P. (2022). SceneFND: Multimodal fake news detection by modelling scene context information. *Journal of Information Science*, Article 01655515221087683.
Zhou, X., Jain, A., Phoha, V.V., Zafarani, R. 2019. Fake news early detection: A theory-driven model. arXiv preprint arXiv:1904.11679.

ARTICLES FOR FACULTY MEMBERS

MULTIMODAL FAKE NEWS DETECTION

Title/Author	**Multimodal Fake News Detection / Segura-Bedmar, I., & Alonso-Bartolome, S.**
Source	***Information*** **Volume 13 Issue 6 (2022) Pages 1-16** **https://doi.org/10.3390/info13060284** **(Database: MDPI)**

information

Article

Multimodal Fake News Detection

Isabel Segura-Bedmar *,†, Santiago Alonso-Bartolome †

Computer Science Department, University Carlos III of Madrid, Avenida de la Universidad, 30, 28911 Madrid, Spain; santiagoalonsobartolome@gmail.com
* Correspondence: isegura@inf.uc3m.es
† These authors contributed equally to this work.

Abstract: Over the last few years, there has been an unprecedented proliferation of fake news. As a consequence, we are more susceptible to the pernicious impact that misinformation and disinformation spreading can have on different segments of our society. Thus, the development of tools for the automatic detection of fake news plays an important role in the prevention of its negative effects. Most attempts to detect and classify false content focus only on using textual information. Multimodal approaches are less frequent and they typically classify news either as true or fake. In this work, we perform a fine-grained classification of fake news on the Fakeddit dataset, using both unimodal and multimodal approaches. Our experiments show that the multimodal approach based on a Convolutional Neural Network (CNN) architecture combining text and image data achieves the best results, with an accuracy of 87%. Some fake news categories, such as Manipulated content, Satire, or False connection, strongly benefit from the use of images. Using images also improves the results of the other categories but with less impact. Regarding the unimodal approaches using only text, Bidirectional Encoder Representations from Transformers (BERT) is the best model, with an accuracy of 78%. Exploiting both text and image data significantly improves the performance of fake news detection.

Keywords: Multimodal Fake News Detection; Natural Language processing; deep learning learning; BERT

Citation: Segura-Bedmar, I.; Alonso-Bartolome, S. Multimodal Fake News Detection. *Information* **2022**, *13*, 284. https://doi.org/10.3390/info13060284

Academic Editor: Diego Reforgiato Recupero

Received: 18 April 2022
Accepted: 31 May 2022
Published: 2 June 2022

1. Introduction

Digital media has provided a lot of benefits to our modern society, such as facilitating social interactions, boosting productivity, and improving sharing information. However, it has also led to the proliferation of fake news [1]; that is, news articles containing false information that has been deliberately created [2]. The effects of this kind of misinformation and disinformation spreading can be seen in different segments of our society. The Pizzagate incident [3], as well as the mob lynchings that occurred in India [4], are some of the most tragic examples of the consequences of fake news dissemination. Changes in health behavior intentions [5], an increase in vaccine hesitancy [6], and significant economic losses [7] are also some of the negative effects that the spread of fake news may have.

Every day, a huge quantity of digital information is produced, making the detection of fake news by manual fact-checking impossible. Due to this, it becomes essential to use techniques that help us to automate the identification of fake news so that more immediate action can be taken.

During the last few years, several studies have already been carried out to perform the automatic detection of fake news [8–13]. Most previous works only exploit textual information for identifying fake news. These approaches can be considered unimodal methods because they only use a type of input data to deal with the task. The last few years have shown great advances in the field of machine learning by combining multiple types of data, such as audio, video, images, and text [14], for different tasks such as text classification [15] or image recognition [16]. These systems are known as multimodal

approaches [14]. The use of multimodal data (combining texts and images) for detecting fake news has been explored little [10,11,17]. These approaches have shown promising results, obtaining better results than the unimodal approaches. However, these studies typically address the problem of fake news detection as a binary classification task (that is, consisting of classifying news as either true or fake).

The main goal of this paper is to study both unimodal and multimodal approaches to deal with a finer-grained classification of fake news. To do this, we use the Fakeddit dataset [18], made up of posts from Reddit. The posts are classified into the following six different classes: true, misleading content, manipulated content, false connection, imposter content, and satire. We explore several deep learning architectures for text classification, such as Convolutional Neural Network (CNN) [19], Bidirectional Long Short-Term Memory (BiLSTM) [20], and Bidirectional Encoder Representations from Transformers (BERT) [21]. As a multimodal approach, we propose a CNN architecture that combines both texts and images to classify fake news.

2. Related Work

Since the revival of neural networks in the second decade of the current century, many different applications of deep learning techniques have emerged. Many Natural Language Processing (NLP) advances are due to the incorporation of deep neural network approaches [22,23] .

Text classification tasks such as sentiment analysis or fake news detection are also one of the tasks for which deep neural networks are being extensively used [24]. Most of these works have been based on unimodal approaches that only exploit texts. More ambitious architectures that combine several modalities of data (such as text and image) have also been tried [25–29]. The main intuition behind these multimodal approaches is that many texts are often accompanied by images, and these images may provide useful information to improve the results of the classification task [30].

We review the most recent studies for the detection of fake news using only the textual content of the news. Wani et al. [31] use the Constraint@AAAI COVID-19 fake news dataset [32], which contains tweets classified as true or fake. Several methods were evaluated: CNN, LSTM, Bi-LSTM + Attention, Hierarchical Attention Network (HAN) [33], BERT, and DistilBERT [34], a smaller version of BERT. The best accuracy obtained was 98.41% by the DistilBERT model when it was pre-trained on a corpus of COVID-19 tweets.

Goldani et al. [35] use a capsule network model [36] based on CNN and pre-trained word embeddings for fake news classification of the ISOT [37] and LIAR [38] datasets. The ISOT dataset is made up of fake and true news articles collected from Reuters and Kaggle, while the LIAR dataset contains short statements classified into the following six classes: pants-fire, false, barely-true, half-true, mostly-true, and true. Thus, the authors perform both binary and multi-class fake news classification. The best accuracies obtained with the proposed model were 99.8% for the ISOT dataset (binary classification) and 39.5% for the LIAR dataset (multi-class classification).

Girgis et al. [39] perform fake news classification using the above-mentioned LIAR dataset. More concretely, they use three different models: vanilla Recurrent Neural Network [40], Gated Recurrent Unit (GRU) [41], and LSTM. The GRU model obtains an accuracy of 21.7%, slightly outperforming the LSTM (21.66%) and the vanilla RNN (21.5%) models.

From this review on approaches using only texts, we can conclude that deep learning architectures provide very high accuracy for the binary classification of fake news; however, the performance is much lower when these methods address a fine-grained classification of fake news. Curiously enough, although BERT is reaching state-of-the-art results in many text classification tasks, it has hardly ever been used for the multiclassification of fake news.

Recently, some efforts have been devoted to the development of multimodal approaches for fake news detection. Singh et al. [10] study the improvement in performance on the binary classification of fake news when textual and visual features are combined as opposed to using only text or image. They explored several traditional machine learning methods: logistic

regression (LR) [42], classification and regression tree (CART) [43], linear discriminant analysis (LDA) [44], quadratic discriminant analysis (QDA) [44], k-nearest neighbors (KNN) [44], naïve Bayes (NB) [45], support vector machine (SVM) [46], and random forest (RF) [47]. The authors used a Kagle dataset of fake news [48]. Random forest was the best model, with an accuracy of 95.18%.

Giachanou et al. [11] propose a model to perform multimodal classification of news articles as either true or fake. In order to obtain textual representations, the BERT model [21] was applied. For the visual features, the authors used the VGG (Visual Geometry Group) network [49] with 16 layers, followed by an LSTM layer and a mean pooling layer. The dataset used by the authors was retrieved from the FakeNewsNet collection [50]. More concretely, the authors used 2745 fake news and 2714 real news collected from the GossipCop posts of the collection. The proposed model achieved an F1 score of 79.55%.

Finally, another recent architecture proposed for multimodal fake news classification can be found in the work carried out by [17]. The authors proposed a model that is made up of four modules: (i) ABS-BiLSTM (attention-based stacked BiLSTM) for extracting the textual features), (ii) ABM-CNN-RNN (attention based CNN-RNN) to obtain the visual representations, (iii) MFB (multimodal factorized bilinear pooling), where the feature representations obtained from the previous two modules are fused, and (iv) MLP (multilayer perceptron), which takes the fused feature representations provided by the MFB module as input, and then generates the probabilities for each class (true of fake). In order to evaluate the model, two datasets were used: Twitter [51] and Weibo [52]. The Twitter dataset contains tweets along with images and contextual information. The Weibo dataset is made up of tweets, images, and social context information. The model obtains an accuracy of 88.3% on the Twitter dataset and an accuracy of 83.2% on the Weibo dataset.

Apart from the previous studies, several authors have proposed fake news classification models and have evaluated them using the Fakeddit dataset. Kaliyar et al. [53] propose the DeepNet model for the binary classification of fake news. This model is made up of one embedding layer, three convolutional layers, one LSTM layer, seven dense layers, ReLU for activation, and, finally, the softmax function for the binary classification. The model was evaluated on the Fakeddit and BuzzFeed [54] datasets. The BuzzFeed dataset contains news articles collected within a week before the U.S. election, and they are classified as either true or fake. The models provided an accuracy of 86.4% on the Fakeddit dataset (binary classification) and 95.2% on the BuzzFeed dataset.

Kirchknopf et al. [55] use four different modalities of data to perform binary classification of fake news over the Fakeddit dataset. More concretely, the authors used the textual content of the news, the associated comments, the images, and the remaining metadata belonging to other modalities. The best accuracy obtained was 95.5%. Li et al. [56] proposed the Entity-Oriented Multimodal Alignment and Fusion Network (EMAF) for binary fake news detection. The model is made up of an encapsulating module, a cross-modal alignment module, a cross-model fusion module, and a classifier. The authors evaluated the model on the Fakeddit, Weibo, and Twitter datasets, obtaining accuracies of 92.3%, 97.4%, and 80.5%, respectively.

Xie et al. [57] propose the Stance Extraction and Reasoning Network (SERN) to obtain stance representations from a post and its associated reply. They combined these stance representations with a multimodal representation of the text and image of a post in order to perform binary fake news classification. The authors use the PHEME dataset [58] and a reduced version of the Fakeddit dataset created by them. The PHEME dataset contains 5802 tweets, of which 3830 are real, and 1972 are false. The accuracies obtained are 96.63% (Fakeddit) and 76.53% (PHEME).

Kang et al. [59] use a heterogeneous graph named News Detection Graph (NDG) that contains domain nodes, news nodes, source nodes, and review nodes. Moreover, they proposed a Heterogeneous Deep Convolutional Network (HDGCN) in order to obtain the embeddings of the news nodes in NDG. The authors evaluated this model using reduced versions of the Weibo and Fakeddit datasets. For the Weibo dataset, they obtained an

F1 score of 96%, while for the Fakeddit dataset they obtained F1 scores of 88.5% (binary classification), 85.8% (three classes), and 83.2% (six classes).

As we can see from this review, most multimodal approaches evaluated on the Fakkeddit dataset have only addressed the binary classification of fake news. Thus far, only work [59] has addressed the multi-classification of fake news using a reduced version of this dataset. To the best of our knowledge, our work is the first attempt to perform a fine-grained classification of fake news using the whole Fakeddit dataset. Furthermore, contrary to the work proposed in [59], which exploits a deep convolutional network, we propose a multimodal approach that simply uses a CNN, obtaining a very similar performance.

3. Materials and Methods

In this section, we describe our approaches to dealing with the task of fake news detection. First, we present the unimodal approaches that only use texts. Then, we describe our multimodal approach, exploiting texts and images.

3.1. Dataset

In our experiments, we train and test our models using the Fakeddit dataset [18], which consists of a collection of posts from Reddit users. It includes texts, images, comments, and metadata. The texts are the titles of the posts submitted by users, while the comments are made by other users as an answer to a specific post. Thus, the dataset contains over 1 million instances.

One of the main advantages of this dataset is that it can be used to implement systems capable of performing a finer-grained classification of fake news than the usual binary classification, which only distinguishes between true and fake news. In the Fakeddit dataset, each instance has a label that distinguishes five categories of fake news, besides the unique category of true news. We briefly describe each category:

- True: this category indicates that the news is true.
- Manipulated Content: in this case, the content has been manipulated by different means (such as photo editing, for example).
- False Connection: this category corresponds to those samples in which the text and the images are not in accordance.
- Satire/Parody: this category refers to the news in which the meaning of the content is twisted or misinterpreted in a satirical or humorous way.
- Misleading Content: this category corresponds to the news in which the information has been deliberately manipulated or altered in order to mislead the public.
- Imposter Content: in the context of this project, all the news that belongs to this category include content generated by bots.

The Fakeddit dataset is divided into training, validation, and test partitions. Moreover, there are two different versions of the dataset: the unimodal dataset, whose instances only contains texts, and the multimodal dataset, whose instances have both text and image. The full dataset contains a total of 682,661 news with images. There are almost 290,000 additional texts without images. Therefore, 70% of the instances include both texts and images, while 30% only contain texts. Actually, all texts of the multimodal dataset are also included in the unimodal dataset.

Table 1 shows the distribution of the classes in the unimodal dataset. Table 2 provides the same information for the multimodal dataset. As we can see, all classes follow a similar distribution in both versions of the dataset (unimodal and multimodal) as well as in the training, validation, and test splits. Moreover, both datasets, unimodal and multimodal, are clearly imbalanced (the classes true, manipulated content, and false connection have more instances than the other classes satire, misleading content, and imposter content, which are much more underrepresented in both datasets). This imbalance may cause the classification task to be more difficult for those classes with fewer instances.

Table 1. Class distribution for the unimodal scenario.

Class	Training	Validation	Test
True	400,274 (49.86%)	42,121 (0.5%)	42,326 (0.5%)
Satire/Parody	42,310 (5.27%)	4450 (0.05%)	4446 (0.05%)
Misleading Content	141,965 (17.68%)	14,964 (0.18%)	14,928 (0.18%)
Imposter Content	23,812 (2.97%)	2514 (0.03%)	2471 (0.03%)
False Connection	167,857 (20.91%)	17,810 (0.21%)	17,472 (0.21%)
Manipulated Content	26,571 (3.31%)	2677 (0.03%)	2838 (0.03%)
Total	802,789	84,536	84,481

Table 2. Class distribution for the multimodal scenario.

Class	Training	Validation	Test
True	222,081 (39.38%)	23,320 (0.39%)	23,507 (0.4%)
Satire/Parody	33,481 (5.94%)	3521 (0.06%)	3514 (0.06%)
Misleading Content	107,221 (19.01%)	11,277 (0.19%)	11,297 (0.19%)
Imposter Content	11,784 (2.09%)	1238 (0.02%)	1224 (0.02%)
False Connection	167,857 (29.76%)	17,810 (0.3%)	17,472 (0.29%)
Manipulated Content	21,576 (3.83%)	2176 (0.04%)	2305 (0.04%)
Total	564,000	59,342	59,319

3.2. Methods

We now describe our approaches to deal with the task of fake news detection. First, we present the unimodal approaches that only use texts. Three models only using the texts are proposed: CNN, BiLSTM, and BERT. Then, we describe our multimodal approach, exploiting texts and images.

All texts were cleaned by removing stopwords, punctuations, numbers, and multiple spaces. Then, we split each text into tokens and we apply lemmatization. After lemmatization, we transform the texts into sequences of integers. This is performed first, by learning the vocabulary of the corpus and building a dictionary where each word is mapped to a different integer number. This dictionary is used to transform each text into a sequence of integers. Every non-zero entry in such a sequence corresponds to a word in the original text. The original order of the words in the text is respected.

As we need to feed the deep learning models with vectors of the same length, we pad and truncate the sequences of integers so that they have the same number of entries. This has the disadvantage that those vectors that are too long will be truncated, and some information will be lost. In order to select the length of the padded/truncated vectors, we computed the percentage of texts that are shorter than 10, 15, 20, and 25 tokens. We saw that 98% of the texts have less than 15 tokens.

Since the number of texts that will have to be truncated is very small (less than 2%), very little information is lost. Therefore, we selected 15 as the length of the vectors after padding and truncating.

Then, an embedding layer transforms each integer value from the input sequence into a vector of word embeddings. Thus, each text is represented as a sequence of word embeddings, which is the input of each deep learning model. In particular, every text is transformed into a matrix of 15 rows and 300 columns (300 being the dimension of the word embeddings).

3.2.1. CNN

We now explain the CNN architecture for the text classification of fake news. As was mentioned above, the first layer is an embedding layer. We initialize the embedding matrix using both random initialization and the pre-trained GloVe word embeddings of dimension 300. We chose this size for the word embeddings over other options (50, 100 or 200) because word embeddings of a larger dimension have been proven to give better results [60].

After the embedding layer, we apply four different filters in a convolutional layer. A convolutional operation is essentially the multiplication of the embedding matrix with a filter to extract the most representative features from the matrix. Each of these filters slides across the (15×300) matrix with the embeddings of the input sequence and generates 50 output channels. The 4 filters have sizes (2×300), (3×300), (4×300), and (5×300), respectively, since these are the typical filter sizes of a CNN for text classification [61]. As a consequence, the outputs of the previous filters have shapes (14×1), (13×1), (12×1), and (11×1), respectively.

The next step is to pass the outputs obtained from the previous layer through the ReLU activation function. This function is applied element-wise, and, therefore, it does not alter the size of the outputs obtained after the previous step. The effect of this function is to set all the negative values to 0 and leave the positive values unchanged.

To reduce the size of the model, after going through the ReLU activation, we will apply a maxpooling layer that selects the biggest element out of each of the 200 feature maps (50 feature maps per each of the 4 filters). Thus, 200 single numbers are generated.

These 200 numbers are concatenated, and the result is passed through 2 dense layers with 1 ReLU activation in between [24]. The resulting output is a vector of six entries (each entry corresponding to a different class of the Fakeddit dataset) that, after passing through the logsoftmax function, can be used to obtain the predicted class for the corresponding input text.

Early stopping [62] with the train and validation partitions is used in order to select the appropriate number of epochs. We use the Adam optimization algorithm [63] for training the model and the negative log-likelihood as the loss function.

3.2.2. BiLSTM + CNN

We now present a hybrid model that uses a bidirectional LSTM followed by a CNN layer. First, texts are processed as was described above, and these inputs are passed through the same embedding layer that was used for the CNN model. Therefore, each input vector of length 15 is transformed into a matrix of shape 15×300.

Then, the matrix with the word embeddings goes through a bidirectional LSTM layer with hidden states of length 70. The output of this layer is a matrix of size 15×140 that contains two hidden states (corresponding to the two directions of the BiLSTM) for each word embedding. The output of the BiLSTM layer is the input of a convolutional layer, which applies 240 filters of size (3×140). Therefore, it generates 240 output arrays of size (13×1). Then, the ReLU activation is applied, followed by a maxpooling layer that selects the largest element within each of the 240 feature maps. Thus, this layer outputs a sequence of 240 numbers.

Similar to what was done for the CNN model, the output of the maxpooling layer is concatenated and passed through two dense layers with ReLU activation in between. The resulting vector goes through the logsoftmax function, and the predicted class is obtained. Figure 1 shows the architecture of the BiLSTM for text classification.

Early stopping is again used for selecting the optimal number of epochs. We use Adam as the optimization algorithm and the negative log-likelihood as the loss function.

3.2.3. BERT

In this case, instead of using random initialization of the pre-trained Glove embeddings, we now use the vectors provided by BERT to represent the input tokens. As opposed to the GloVe model [64], BERT takes into account the context of each word (that is, the words that surround it).

For the preprocessing of the texts, the steps are similar to those described above. The main differences are that we tokenize the texts by using the BertTokenizer class from the transformers library [65]. This class has its own vocabulary with the mappings between words and ID, so it was not necessary to train a tokenizer with the corpus of texts. We also add the [CLS] and [SEP] tokens at the beginning and at the end of each tokenized sequence.

It was also necessary to create an attention mask in order to distinguish what entries in each sequence correspond to real words in the input text and what entries are just 0 s resulting from padding the sequences. Thus, the attention mask is composed of 1 s (indicating non-padding entries) and 0 s (indicating padding entries). We use the BERT base model in its uncased version (12 layers, 768 hidden size, 12 heads, and 110 million parameters).

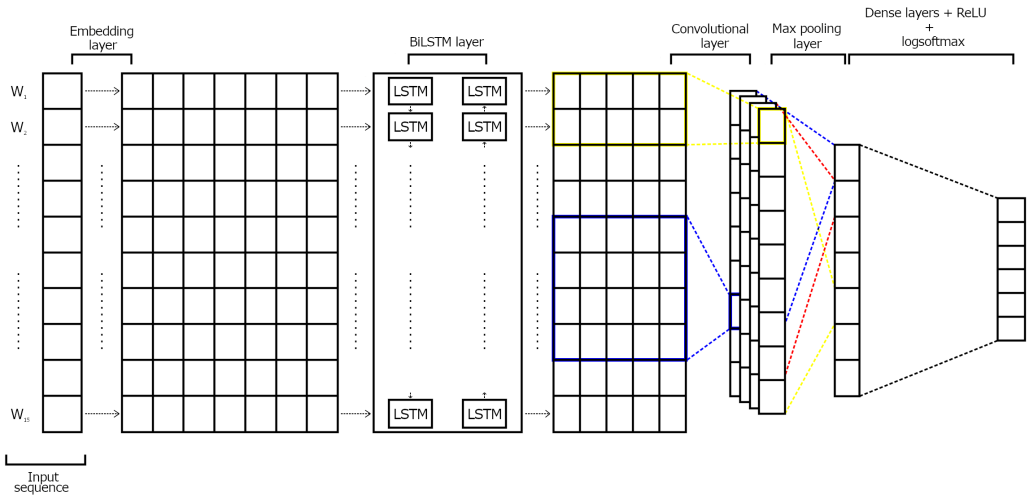


Figure 1. BiLSTM + CNN for text classification.

Then, we fine-tune it on our particular problem; that is, the multi-classification of fake news. To do this, we add a softmax layer on top of the output of BERT. The softmax layer receives a vector of length 768 and outputs a vector of length 6 (see Figure 2), which contains the probabilities for each class.

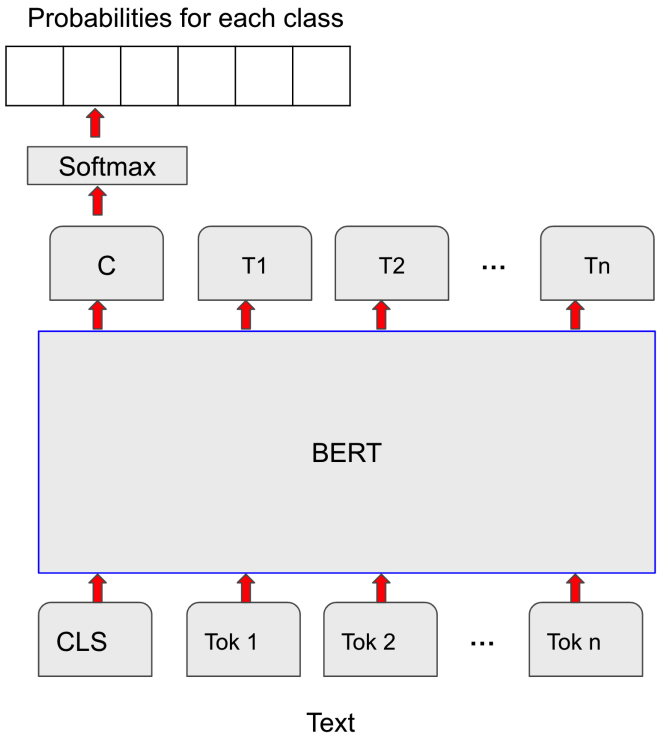


Figure 2. BERT for text classification.

For the training process, we used the Adam algorithm for optimization with a learning rate of 2×10^{-5}. We trained the model for two epochs since the authors of BERT recommended using between two and four epochs for fine-tuning on a specific NLP task [21].

3.2.4. Multimodal Approach

Our multimodal approach uses a CNN that takes both the text and the image corresponding to the same news as inputs. The model outputs a vector of six numbers, out of which the predicted class is obtained. In the following lines, we describe the preprocessing steps applied before feeding the data into the network, as well as the architecture of the network (see Figure 3).

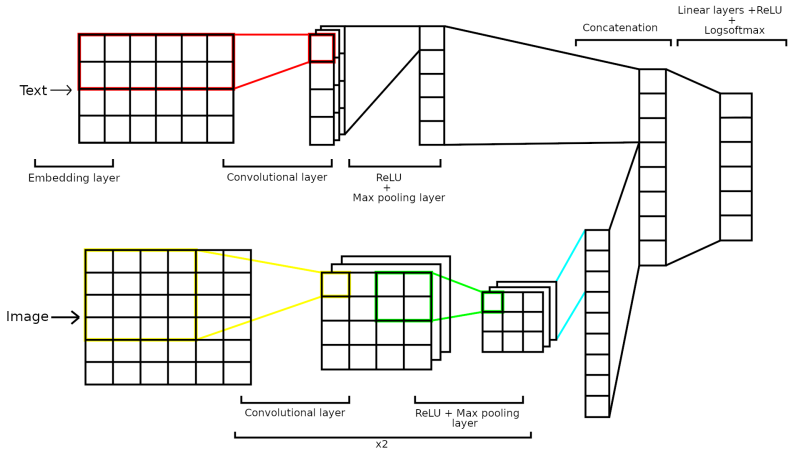


Figure 3. Architecture of the multimodal approach for fake news detection.

Regarding the preprocessing of the images, we only reshaped them so that they all have the same shape (560×560). Once the preprocessed data are fed into the network, different operations are applied to the texts and images. We use the same CNN architecture that we have used for the unimodal scenario, except for the fact that we eliminate the last two dense layers with ReLU activation in between.

We now describe the CNN model to classify the images. The data first goes through a convolutional layer. Since each image is made up of three channels, the number of input channels of this layer is also three. Moreover, it has six output channels. Filters of size (5×5) are used with a stride equal to 1 and no padding. The output for each input image is, therefore, a collection of 6 matrices of shape (556×556). The output of the convolutional layer passes through a non-linear activation function (ReLU), and then maxpooling is applied with a filter of size (2×2) and a stride equal to 2. The resulting output is a set of six matrices of shape (278×278). The output from the maxpooling layer again passes through another convolutional layer that has 6 input channels and 3 output channels. The filter size, stride length, and padding are the same as those used in the previous convolutional layer. Then, the ReLU non-linear activation function and the maxpooling layer are applied again over the feature maps resulting from the convolutional layer. Thus, for a given input (image), we obtain a set of 3 feature maps of shape (137×137). Finally, these feature maps are flattened into a vector of length 56,307.

The texts are also processed by using the same CNN model for texts, which was described previously. However, instead of feeding the output of the dense layer to the softmax layer in the CNN model, this output vector representing the text is concatenated to the vector obtained by the CNN model for images. Then, this vector is passed through two dense layers with a ReLU non-linear activation in between. Finally, the logsoftmax function is applied, and the logarithm of the probabilities is used in order to compute the predicted class of the given input.

4. Results

In this section, we present the results obtained for each model. We report the recall, precision, and F1 scores obtained by all the models for each class. The accuracy is computed over all the classes. It helps us to compare models and find the best approach. Moreover, we are also interested in knowing which model is better at detecting only those news containing false content. For this reason, we also compute the micro and macro averages of the recall, precision, and F1 metrics only over five classes of fake news without the true news. Macro averaging computes the metrics for each class and then calculates the average. Micro averaging calculated the sum of all true positives and false positives for all the classes, and then computes the metrics. We use the $F1_{micro}$ score and the accuracy to compare the performance of the models.

4.1. CNN Results

Our first experiment with CNN uses random initialization to initialize the weights of the embedding layer, which are updated during the training process. This model obtains an accuracy of 72%, a micro F1 of 57%, and a macro F1 of 49% (see Table 3). We can also see that True and Manipulated content are the classes with the highest F1 (79%). A possible reason for this could be that they are the majority classes. On the other hand, the model obtains the lowest F1 (13%) for Imposter content, which is the minority class in the dataset (see Table 1). Therefore, the results for the different classes appear to be related to the number of instances per class. However, the model achieves an F1 of 61% for the second minority class, Misleading content. As was explained before, the content of this news has been deliberately manipulated. Identifying these manipulations appears to be easier than detecting humor or sarcasm in the news (Satire) or fake news generated by bots (Imposter content).

Table 3. Results of CNN with random initialization.

Class	P	R	F1
True	0.71	0.87	0.79
Manipulated content	0.75	0.84	0.79
False connection	0.70	0.48	0.57
Satire	0.63	0.26	0.37
Misleading content	0.71	0.54	0.61
Imposter content	0.72	0.07	0.13
micro-average	0.73	0.62	0.57
macro-average	0.70	0.44	0.49

Interestingly, although the model only exploits the textual content of the news, it achieves an F1 of 57% for classifying the instances of False connections. In these instances, the text and the image are not in accordance.

We also explore CNN with static (see Table 4) and dynamic (see Table 5) GloVe embeddings [64]. In both models, the embedding layer is initialized with the pre-trained Glove vectors. When dynamic training is chosen, these vectors are updated during the training process. On the other hand, if static training is chosen, the vectors are fixed during the training process. The model with dynamic vectors overcomes the one with static vectors, with a slight improvement in accuracy (74%) (roughly one percentage point). However, in terms of micro F1, the static model is better than the dynamic one. Both models provide the same macro F1 (69%). Regarding the classes, there are no significant differences, except for Imposer content. For this class, updating the pre-trained Glove vectors results in a decrease of seven percentage points in F1.

Table 4. Results of CNN with static Glove vectors.

Class	P	R	F1
True	0.76	0.81	0.79
Manipulated content	0.75	0.82	0.79
False connection	0.65	0.59	0.62
Satire	0.60	0.40	0.48
Misleading content	0.71	0.59	0.64
Imposter content	0.35	0.21	0.26
micro-average	0.70	0.67	0.69
macro-average	0.61	0.52	0.56

Table 5. Results of CNN with dynamic Glove vectors.

Class	P	R	F1
True	0.74	0.87	0.80
Manipulated content	0.76	0.83	0.80
False connection	0.71	0.54	0.61
Satire	0.67	0.35	0.46
Misleading content	0.74	0.58	0.65
Imposter content	0.70	0.11	0.19
micro-average	0.74	0.65	0.69
macro-average	0.71	0.48	0.54

We also compared the effect of the pre-trained Glove vectors with random initialization (see Table 3). In both dynamic and static approaches, initializing the model with the pre-trained GloVe word embeddings gets better results than random initialization. The reason for this is that the GloVe vectors contain information about the relationship between different words that random vectors can not capture.

As the dataset is highly unbalanced, we use the micro F1 to assess and compare the overall performances of the three models. Thus, the best model is a CNN with dynamic Glove vectors. However, dynamic training takes much more time than static training (around 6000 to 8000 s more). This is due to the fact that, in a dynamic approach, word embeddings are also learned, and this significantly increases the training time.

4.2. BiLSTM + CNN Results

As a second deep learning model, we explore a hybrid model based on a BiLSTM followed by a CNN. We replicate the same experiments as described for CNN; that is, using random initialization and pre-trained Glove vectors.

The BiLSTM initialized with random vectors (see Table 6) very similar results to those achieved by CNN with random initialization (see Table 3). In fact, both models provide the same accuracy of 0.72. However, in terms of micro F1, the BiLSTM model obtains up to nine points more than the CNN model with random initialization. This improvement may be because the BiLSTM improved its scores for Imposter content.

Table 6. Results of BiLSTM with random initialization.

Class	P	R	F1
True	0.70	0.88	0.78
Manipulated content	0.73	0.85	0.79
False connection	0.73	0.44	0.55
Satire	0.58	0.25	0.35
Misleading content	0.71	0.54	0.61
Imposter content	0.86	0.08	0.14
micro-average	0.73	0.61	0.66
macro-average	0.74	0.41	0.48

The use of static Glove vectors (see Table 7) appears to have a positive effect on the performance of the BiLSTM model. The model shows significant improvements for False connection, Satire, Misleading content, and Imposter content, with increases of 6, 12, 3, and 10 points, respectively. The model obtains an accuracy of 73%. Therefore, the pre-trained Glove vectors achieve better results than random initialization.

Table 7. Results of BiLSTM with static Glove vectors.

Class	P	R	F1
True	0.74	0.85	0.79
Manipulated content	0.77	0.82	0.79
False connection	0.68	0.55	0.61
Satire	0.55	0.41	0.47
Misleading content	0.77	0.55	0.64
Imposter content	0.45	0.17	0.24
micro-average	0.72	0.65	0.69
macro-average	0.65	0.50	0.55

Table 8 shows the results obtained by BiLSTM with dynamic Glove vectors. If these vectors are updated during the training of the BiLSTM model, an accuracy of 75% is achieved; that is, two points more than BiLSTM with static Glove vectors. Moreover, this model with dynamic Glove vectors improves the results for all classes, with increases ranging from one to four points. In terms of micro F1, using dynamic Glove vectors is the best approach for BiLSTM. Moreover, this model slightly overcomes the CNN model with dynamic Glove vectors by roughly one percentage point. However, as mentioned above, dynamic training takes much more time than static training.

Table 8. Results of BiLSTM with dynamic Glove vectors.

Class	P	R	F1
True	0.75	0.86	0.80
Manipulated content	0.77	0.84	0.80
False connection	0.72	0.55	0.63
Satire	0.63	0.41	0.50
Misleading content	0.78	0.57	0.66
Imposter content	0.57	0.18	0.28
micro-average	0.74	0.67	0.70
macro-average	0.69	0.51	0.57

4.3. Bert Results

Table 9 shows the results obtained by BERT. This model achieves an accuracy of 78% and a micro F1 of 74%. Therefore, it outperforms all the previous unimodal deep learning approaches. This proves the advantage of the pre-trained contextual text representations provided by BERT, as opposed to the context-free GloVe vectors or random initialization for neural networks.

Table 9. BERT results.

Class	P	R	F1
True	0.81	0.86	0.83
Manipulated content	0.80	0.86	0.83
False connection	0.72	0.64	0.68
Satire	0.70	0.53	0.61
Misleading content	0.77	0.70	0.73
Imposter content	0.61	0.28	0.38
micro-average	0.76	0.73	0.74
macro-average	0.72	0.60	0.65

Moreover, BERT is better in all classes. Comparing the classes, the behavior of BERT is very similar to the previous deep learning models; that is, the more training instances for a class, the better predictions for it. In this way, True and Manipulated content both get the highest F1 (83%), while the worst-performing class is Imposter content (F1 = 38%). As in previous models, Misleading content gets better scores than Satire, despite the fact that this class is more represented than the first one, Misleading content (see Table 2).

4.4. Multimodal Approach Results

The multimodal approach obtains an accuracy of 87% and a micro F1 of 72% (see Table 10), which are the highest scores out of all the unimodal models.

Table 10. Multimodal approach results.

Class	P	R	F1
True	0.85	0.88	0.86
Manipulated content	1	1	1
False connection	0.77	0.76	0.76
Satire	0.82	0.72	0.77
Misleading content	0.75	0.79	0.77
Imposter content	0.46	0.25	0.32
micro-average	0.88	0.86	0.87
macro-average	0.76	0.70	0.72

As expected, the training set size for each class strongly affects the model scores. While True and Manipulated content, the majority classes, get the highest scores, Imposter content, the minority class, shows the lowest F1 (32%), even six points lower than that provided by BERT for the same class (F1 = 38%). Thus, we can say that the image content provides little information for identifying instances of Imposter content. Manipulated content shows an F1 of 100%. This is probably due to the fact that the images in this category have been manipulated. These manipulations may be easily detected by CNN.

As expected, the use of images significantly improves the results for False connection. The multimodal model shows an F1 of 76%, 8 points higher than that obtained by BERT, the best unimodal approach, and 15 points higher than the unimodal CNN model using only texts. The improvement is even greater for detecting instances of Satire, with an increase of 16 points higher than those obtained by BERT and by the unimodal CNN model.

5. Discussion

In addition to the deep learning algorithms, we also propose a Support Vector Machine (SVM) as a baseline for the unimodal approaches. SVM is one of the most successful algorithms for text classification. For this algorithm, the texts were represented using the tf-idf model. Table 11 shows a comparison of the best models (traditional algorithms, CNN, BiLSTM, BERT, and multimodal CNN) according to their accuracy and micro average scores.

Table 11. Comparison of the best models (micro_averages).

Model	P	R	F1	Acc.
SVM	0.71	0.64	0.67	0.72
CNN (Dynamic + GloVe)	0.74	0.65	0.69	0.74
BiLSTM + CNN (Dynamic + GloVe)	0.74	0.67	0.70	0.75
BERT	0.76	0.73	0.74	0.78
Multimodal CNN	0.88	0.86	0.87	0.87

We can see that the multimodal CNN outperforms all the unimodal approaches. In fact, the multimodal approach achieves higher accuracy than that provided by the best model of the unimodal approaches, BERT, with a difference of 9% in overall accuracy.

In terms of micro-F1, the improvement is even greater, 13 points over the micro F1 of BERT. This proves the usefulness of combining texts and images for a fine-grained fake news classification.

Focusing on the unimodal approaches, the BERT model is the best both in terms of accuracy and micro F1 score, which shows the advantage of using contextual word embeddings. In terms of accuracy, BERT achieves a significant improvement over the other deep learning models. The third best approach is BiLSTM + CNN with dynamic Glove vectors, with an accuracy of 0.75 (three points lower than the accuracy achieved by BERT). The fourth approach is the CNN model, with an accuracy of 0.74 (four points lower than the accuracy provided by BERT). In terms of micro F1, BERT also outperforms the other deep learning models, with improvements of around 4–5%. Finally, all the deep learning approaches outperform our baseline SVM, with an accuracy of 0.72. This also shows that when a large dataset is available, as in the case of the Fakeddit dataset, the deep learning models provide better performance than traditional machine learning algorithms.

6. Conclusions

Fake news could have a significant negative effect on politics, health, and economies. Therefore, it becomes necessary to develop tools that allow for the rapid and reliable detection of misinformation.

Apart from the work carried out by the creators of the Fakeddit dataset [18], this is, to the best of our knowledge, the only study that addresses a fine-grained classification of fake news by performing a comprehensive comparison of unimodal and multimodal approaches based on the most advanced deep learning techniques.

The multimodal approach overcomes the approaches that only exploit texts. BERT is the best model for the task of text classification. Moreover, using dynamic GloVe word embeddings outperforms random initialization for the CNN and BiLSTM architectures.

In future work, we plan to use pre-trained networks to generate the visual representations. In particular, we will use the network VGG, which was pre-trained on a large dataset of images, such as ImageNet. We also plan to explore different deep learning techniques, such as LSTM, BiLSTM, GRU, or BERT, as well as different methods of combining the visual and textual representations. In our current study, we have built our multimodal CNN using an early fusion approach, which consists of creating textual and visual representations, combining them, and then applying a classifier over the resulting combined representation to get the probabilities for each class. Instead of this, we plan to study a late fusion approach, which would require two separate classifiers (one for the textual inputs and the other for the image inputs). The predictions from both classifiers are then combined, and the final prediction is obtained.

Author Contributions: Conceptualization, I.S.-B. and S.A.-B.; methodology, I.S.-B.; software, S.A.-B.; validation, I.S.-B. and S.A.-B.; formal analysis, I.S.-B.; investigation, I.S.-B.; writing—original draft preparation, I.S.-B. and S.A.-B.; writing—review and editing, I.S.-B.; supervision, I.S.-B.; project administration, I.S.-B.; funding acquisition, I.S.-B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Madrid Government (Comunidad de Madrid) under the Multiannual Agreement with UC3M in the line of "Fostering Young Doctors Research" (NLP4RARE-CM-UC3M) and in the context of the V PRICIT (Regional Programme of Research and Technological Innovation) and under the Multiannual Agreement with UC3M in the line of Excellence of University Professors (EPUC3M17).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The source code is available from https://github.com/isegura/MultimodalFakeNewsDetection (accessed on 1 June 2022).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

BERT	Bidirectional Encoder Representations from Transformer
BiLSTM	Bidirectional Long Short-Term Memory
CNN	Convolutional Neural Network
NLP	Natural Language Processing
SVM	Support Vector Machine

References

1. Finneman, T.; Thomas, R.J. A family of falsehoods: Deception, media hoaxes and fake news. *Newsp. Res. J.* **2018**, *39*, 350–361. [CrossRef]
2. Hunt, A.; Gentzkow, M. Social Media and Fake News in the 2016 Election. *J. Econ. Perspect.* **2017**, *31*, 211–236. [CrossRef]
3. Hauck, G. Pizzagate Shooter Sentenced to 4 Years in Prison. CNN. 2017. Available online: https://edition.cnn.com/2017/06/22/politics/pizzagate-sentencing/index.html (accessed on 1 June 2022).
4. Mishra, V. India's Fake News Problem is Killing Real People. Asia Times. 2019. Available online: https://asiatimes.com/2019/10/indias-fake-news-problem-is-killing-real-people/ (accessed on 1 June 2022).
5. Greene, C.M.; Murphy, G. Quantifying the effects of fake news on Behavior: Evidence from a study of COVID-19 misinformation. *J. Exp. Psychol. Appl.* **2021**, 27, 773–784. http://dx.doi.org/10.1037/xap0000371. [CrossRef] [PubMed]
6. Islam, M.; Kamal, A.H.; Kabir, A.; Southern, D.; Khan, S.; Hasan, S.; Sarkar, T.; Sharmin, S.; Das, S.; Roy, T.; et al. COVID-19 vaccine rumors and conspiracy theories: The need for cognitive inoculation against misinformation to improve vaccine adherence. *PLoS ONE* **2021**, *16*, e0251605. [CrossRef] [PubMed]
7. Brown, E. Online Fake News is Costing us $78 Billion Globally Each Year. ZDNet. 2019. Available online: https://www.zdnet.com/article/online-fake-news-costing-us-78-billion-globally-each-year/ (accessed on 1 June 2022).
8. Thota, A.; Tilak, P.; Ahluwalia, S.; Lohia, N. Fake news detection: A deep learning approach. *SMU Data Sci. Rev.* **2018**, *1*, 10.
9. Choudhary, M.; Chouhan, S.S.; Pilli, E.S.; Vipparthi, S.K. BerConvoNet: A deep learning framework for fake news classification. *Appl. Soft Comput.* **2021**, *110*, 107614. [CrossRef]
10. Singh, V.K.; Ghosh, I.; Sonagara, D. Detecting fake news stories via multimodal analysis. *J. Assoc. Inf. Sci. Technol.* **2021**, *72*, 3–17. [CrossRef]
11. Giachanou, A.; Zhang, G.; Rosso, P. Multimodal multi-image fake news detection. In Proceedings of the 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), Sydney, Australia, 6–9 October 2020; pp. 647–654.
12. Singhal, S.; Shah, R.R.; Chakraborty, T.; Kumaraguru, P.; Satoh, S. SpotFake: A multi-modal framework for fake news detection. In Proceedings of the 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM), Singapore, 11–13 September 2019; pp. 39–47.
13. Wang, Y.; Ma, F.; Jin, Z.; Yuan, Y.; Xun, G.; Jha, K.; Su, L.; Gao, J. EANN: Event adversarial neural networks for multi-modal fake news detection. In Proceedings of the Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK, 19–23 August 2018; pp. 849–857.
14. Parcalabescu, L.; Trost, N.; Frank, A. What is Multimodality? In Proceedings of the 1st Workshop on Multimodal Semantic Representations (MMSR), Online, 14–18 June 2021; pp. 1–10.
15. Kiela, D.; Firooz, H.; Mohan, A.; Goswami, V.; Singh, A.; Ringshia, P.; Testuggine, D. The hateful memes challenge: Detecting hate speech in multimodal memes. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 2611–2624.
16. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning. PMLR, Virtual Event, 13–14 August 2021; pp. 8748–8763.
17. Kumari, R.; Ekbal, A. AMFB: Attention based multimodal Factorized Bilinear Pooling for multimodal Fake News Detection. *Expert Syst. Appl.* **2021**, *184*, 115412. [CrossRef]
18. Nakamura, K.; Levy, S.; Wang, W.Y. Fakeddit: A New Multimodal Benchmark Dataset for Fine-grained Fake News Detection. In Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, 21–23 June 2020; European Language Resources Association: Marseille, France, 2020; pp. 6149–6157.
19. Goodfellow, I.; Bengio, Y.; Courville, A. Convolutional Networks. In *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016; pp. 321–362.
20. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
21. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the NAACL-HLT 2019, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
22. O'Mahony, N.; Campbell, S.; Carvalho, A.; Harapanahalli, S.; Hernandez, G.V.; Krpalkova, L.; Riordan, D.; Walsh, J. Deep learning vs. traditional computer vision. In Proceedings of the Science and Information Conference, Leipzig, Germany, 1–4 September 2019; pp. 128–144.
23. Deng, L.; Liu, Y. *Deep Learning in Natural Language Processing*, 1st ed.; Springer: Berlin/Heidelberg, Germany, 2018.

24. Minaee, S.; Kalchbrenner, N.; Cambria, E.; Nikzad, N.; Chenaghlu, M.; Gao, J. Deep Learning–based Text Classification: A Comprehensive Review. *ACM Comput. Surv. (CSUR)* **2021**, *54*, 1–40. [CrossRef]
25. Abavisani, M.; Wu, L.; Hu, S.; Tetreault, J.; Jaimes, A. Multimodal Categorization of Crisis Events in Social Media. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; IEEE Computer Society: Los Alamitos, CA, USA, 2020; pp. 14667–14677.
26. Bae, K.I.; Park, J.; Lee, J.; Lee, Y.; Lim, C. Flower classification with modified multimodal convolutional neural networks. *Expert Syst. Appl.* **2020**, *159*, 113455. [CrossRef]
27. Yu, J.; Jiang, J. Adapting BERT for Target-Oriented Multimodal Sentiment Classification. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19, Macao, 10–16 August 2019; pp. 5408–5414.
28. Viana, M.; Nguyen, Q.B.; Smith, J.; Gabrani, M. Multimodal Classification of Document Embedded Images. In Proceedings of the International Workshop on Graphics Recognition, Nancy, France, 22–23 August 2017; pp. 45–53.
29. Gaspar, A.; Alexandre, L.A. A multimodal approach to image sentiment analysis. In Proceedings of the International Conference on Intelligent Data Engineering and Automated Learning, Manchester, UK, 14–16 November 2019; Springer: Cham, Switzerland, 2019; pp. 302–309.
30. Baheti, P. Introduction to Multimodal Deep Learning. 2020. Available online: https://heartbeat.comet.ml/introduction-to-multimodal-deep-learning-630b259f9291 (accessed on 13 November 2021).
31. Wani, A.; Joshi, I.; Khandve, S.; Wagh, V.; Joshi, R. Evaluating deep learning approaches for COVID-19 fake news detection. In Proceedings of the Combating Online Hostile Posts in Regional Languages during Emergency Situation: First International Workshop, CONSTRAINT 2021, Collocated with AAAI 2021, Virtual Event, 8 February 2021; p. 153.
32. Patwa, P.; Sharma, S.; Pykl, S.; Guptha, V.; Kumari, G.; Akhtar, M.S.; Ekbal, A.; Das, A.; Chakraborty, T. Fighting an infodemic: COVID-19 fake news dataset. *arXiv* **2020**, arXiv:2011.03327.
33. Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.; Hovy, E. Hierarchical attention networks for document classification. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, CA, USA, 12–17 June 2016; pp. 1480–1489.
34. Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv* **2019**, arXiv:1910.01108.
35. Goldani, M.H.; Momtazi, S.; Safabakhsh, R. Detecting fake news with capsule neural networks. *Appl. Soft Comput.* **2021**, *101*, 106991. [CrossRef]
36. Sabour, S.; Frosst, N.; Hinton, G.E. Dynamic routing between capsules. *arXiv* **2017**, arXiv:1710.09829.
37. Ahmed, H.; Traore, I.; Saad, S. Detection of online fake news using n-gram analysis and machine learning techniques. In Proceedings of the International Conference on Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments, Vancouver, BC, Canada, 26–28 October 2017; pp. 127–138.
38. Wang, W.Y. "Liar, Liar Pants on Fire": A new benchmark dataset for fake news detection. *arXiv* **2017**, arXiv:1705.00648.
39. Girgis, S.; Amer, E.; Gadallah, M. Deep Learning Algorithms for Detecting Fake News in Online Text. In Proceedings of the 2018 13th International Conference on Computer Engineering and Systems (ICCES), Cairo, Egypt, 18–19 December 2018; pp. 93–97. [CrossRef]
40. Aggarwal, C.C. Recurrent Neural Networks. In *Neural Networks and Deep Learning: A Textbook*; Springer International Publishing: Cham, Switzerland, 2018; pp. 271–313.
41. Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv* **2014**, arXiv:1412.3555.
42. Kleinbaum, D.G.; Klein, M. *Logistic Regression*, 3rd ed.; Springer: New York, NY, USA, 2010.
43. Hastie, T.; Tibshirani, R.; Friedman, J. Additive Models, Trees, and Related Methods. In *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer: New York, NY, USA, 2009; pp. 295–336.
44. Murphy, K. Kernels. In *Machine Learning: A Probabilistic Perspective*; The MIT Press: Cambridge, MA, USA, 2012; pp. 479–512.
45. Barber, D. Naive Bayes. In *Bayesian Reasoning and Machine Learning*; Cambridge University Press: Cambridge, UK, 2012; pp. 243–255. [CrossRef]
46. Hastie, T.; Tibshirani, R.; Friedman, J., Support Vector Machines and Flexible Discriminants. In *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer: New York, NY, USA, 2009; pp. 417–458. [CrossRef]
47. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
48. Kaggle. Getting Real about Fake News. Available online: https://www.kaggle.com/mrisdal/fake-news (accessed on 13 October 2021).
49. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
50. Shu, K.; Mahudeswaran, D.; Wang, S.; Lee, D.; Liu, H. FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media. *Big Data* **2020**, *8*, 171–188. [CrossRef] [PubMed]
51. Boididou, C.; Andreadou, K.; Papadopoulos, S.; Dang-Nguyen, D.T.; Boato, G.; Riegler, M.; Kompatsiaris, Y. Verifying Multimedia Use at MediaEval 2015. *MediaEval* **2015**, *3*, 7.
52. Jin, Z.; Cao, J.; Guo, H.; Zhang, Y.; Luo, J. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In Proceedings of the 25th ACM international conference on Multimedia, Mountain View, CA, USA, 23–27 October 2017; pp. 795–816.

53. Kaliyar, R.K.; Kumar, P.; Kumar, M.; Narkhede, M.; Namboodiri, S.; Mishra, S. DeepNet: An Efficient Neural Network for Fake News Detection using News-User Engagements. In Proceedings of the 2020 5th International Conference on Computing, Communication and Security (ICCCS), Patna, India, 14–16 October 2020; pp. 1–6. [CrossRef]
54. Kaggle. FakeNewsNet. Available online: https://www.kaggle.com/mdepak/fakenewsnet (accessed on 14 October 2021).
55. Kirchknopf, A.; Slijepcevic, D.; Zeppelzauer, M. Multimodal Detection of Information Disorder from Social Media. *arXiv* **2021**, arXiv:2105.15165.
56. Li, P.; Sun, X.; Yu, H.; Tian, Y.; Yao, F.; Xu, G. Entity-Oriented Multi-Modal Alignment and Fusion Network for Fake News Detection. *IEEE Trans. Multimed.* **2021**, *99*, 1. [CrossRef]
57. Xie, J.; Liu, S.; Liu, R.; Zhang, Y.; Zhu, Y. SERN: Stance Extraction and Reasoning Network for Fake News Detection. In Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual Conference, 6–12 June 2021; pp. 2520–2524. [CrossRef]
58. Zubiaga, A.; Liakata, M.; Procter, R. Exploiting context for rumour detection in social media. In Proceedings of the International Conference on Social Informatics, Oxford, UK, 13–15 September 2017; pp. 109–123.
59. Kang, Z.; Cao, Y.; Shang, Y.; Liang, T.; Tang, H.; Tong, L. Fake News Detection with Heterogenous Deep Graph Convolutional Network. In Proceedings of the Advances in Knowledge Discovery and Data Mining, Virtual Event, 11–14 May 2021; Karlapalem, K.; Cheng, H.; Ramakrishnan, N.; Agrawal, R.K.; Reddy, P.K.; Srivastava, J.; Chakraborty, T., Eds.; Springer International Publishing: Cham, Switzerland, 2021; pp. 408–420.
60. Patel, K.; Bhattacharyya, P. Towards lower bounds on number of dimensions for word embeddings. In Proceedings of the Eighth International Joint Conference on Natural Language Processing, Taipei, Taiwan, 27 November–1 December 2017; pp. 31–36.
61. Voita, E. Convolutional Neural Networks for Text. 2021. Available online: https://lena-voita.github.io/nlp_course/models/convolutional.html (accessed on 5 October 2021).
62. Brownlee, J. A Gentle Introduction to Early Stopping to Avoid Overtraining Neural Networks. Available online: https://machinelearningmastery.com/early-stopping-to-avoid-overtraining-neural-network-models/ (accessed on 5 October 2021).
63. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
64. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
65. Face, H. Transformers. Available online: https://huggingface.co/transformers/ (accessed on 5 October 2021).

ARTICLES FOR FACULTY MEMBERS

MULTIMODAL FAKE NEWS DETECTION

Title/Author	**Positive unlabeled fake news detection via multi-modal masked transformer network / Wang, J., Qian, S., Hu, J., & Hong, R.**
Source	***IEEE Transactions on Multimedia*** **Volume 26 (2024) Pages 234–244** **https://doi.org/10.1109/TMM.2023.3263552** **(Database: IEEE Xplore)**

Positive Unlabeled Fake News Detection via Multi-Modal Masked Transformer Network

Jinguang Wang, Shengsheng Qian, *Member, IEEE*, Jun Hu, and Richang Hong, *Senior Member, IEEE*

***Abstract*—Fake news detection has gotten continuous attention during these years with more and more people have been posting and reading news online. To enable fake news detection, existing researchers usually assume labeled posts are provided for two classes (true or false) so that the model can learn a discriminative classifier from the labeled data. However, this supposition may not hold true in reality, as most users may only label a small number of posts in a single category that they are interested in. Furthermore, most existing methods fail to mask the noise or irrelevant context (i.e., regions or words) between different modalities to assist in strengthening the correlations between relevant contexts. To tackle these issues, we present a curriculum-based multi-modal masked transformer network (CMMTN) for positive unlabeled multi-modal fake news detection by jointly modeling the inter-modality and intra-modality relationships of multi-modal information and masking the irrelevant context between modalities. In particular, we adopt BERT and ResNet to obtain better representations for texts and images, separately. Then, the extracted features of images and texts are fed into a multi-modal masked transformer network to fuse the multi-modal content and mask the irrelevant context between modalities by calculating the similarity between inter-modal contexts. Finally, we design a curriculum-based PU learning method to handle the positive and unlabeled data. Massive experiments on three public real datasets prove the effectiveness of the CMMTN.**

***Index Terms*—Fake news detection, multi-modal learning, social media.**

I. Introduction

OVER the past decades, more and more people have been posting and reading news online because of the increasing ease of social media. With the rise in the number of internet users, multifarious information data has appeared on social media platforms. However, since users do not evaluate the dependability of the given information, the authenticity of the information data is difficult to guarantee, resulting in the widespread propagation of significant fake news. Besides, the widespread dissemination of misinformation has a significant detrimental influence on individuals and society due to its malicious distortion and fabrication of facts [1]. For example, a conspiracy theory that claims that 5 G internet is behind the coronavirus outbreak has led to arson attacks on more than 70 cell phone towers in the U.K. [2]. Hence, discovering fake news is desirable and beneficial to society.

In recent years, many approaches are proposed to identify fake news. They can basically be separated into two groups: (1) One is the traditional hand-crafted feature based methods [3], [4], which generally obtain features from post content and train a classifier to debunk fake news. However, the content of fake news is highly complicated and difficult to be fully captured with hand-crafted features. (2) The other is deep learning based approaches [5], [6], [7], which are good at capturing deep features by using neural networks. For example, Ma et al. [5] extract hidden features of posts through Recurrent Neural Networks. Yu et al. [6] utilize Convolutional Neural Networks to learn latent representations and capture the high-level relationships of fake news. Lately, with the multimedia technology developing by leaps and bounds, the news content has changed from mere text to multi-modal auxiliary descriptions such as images or videos, which has largely deepened readers' understanding. Different modalities (e.g., images and text) have imbalanced and complementary relations that include imbalanced information when representing the same semantic meaning. For instance, as shown in Fig. 1, images usually provide some details that textual descriptions cannot convey and vice versa. Unfortunately, most of the approaches listed above only consider text content and disregard the posts containing multi-modal content (e.g., text, images.) that is an important constituent of social media sites. Recently, many approaches [8], [9] are proposed to deal with multi-modal news content to accelerate the detection of fake news. For example, the multimodal Variational Autoencoder (MVAE) was proposed by Khattar et al. [9] for learning and extracting the latent complicated multi-modal features and these multimedia posts were categorized through the binary classifier. Although this methods [8], [9] prove competitive performance in detecting fake news, most of them simply concatenate different modal features together or build a collective space for different modalities to explore the potential alignment between them, they still have some shortcomings in using multi-modal data.

Manuscript received 23 June 2022; revised 18 November 2022 and 2 March 2023; accepted 16 March 2023. Date of publication 31 March 2023; date of current version 8 January 2024. This work was supported in part by the National Key Research and Development Program of China under Grant 2022YFB3105400, and in part by the National Natural Science Foundation of China under Grants 61932009, 62276257, and 62106262. The Associate Editor coordinating the review of this manuscript and approving it for publication was Dr Ramanathan Subramanian. *(Corresponding author: Richang Hong.)*

Jinguang Wang and Richang Hong are with the School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230009, China, and also with the Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei 230009, China (e-mail: wangjinguang502@gmail.com; hongrc.hfut@gmail.com).

Shengsheng Qian is with the State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: shengsheng.qian@nlpr.ia.ac.cn).

Jun Hu is with the School of Computing, National University of Singapore, Singapore 117417 (e-mail: jun.hu@nus.edu.sg).

Digital Object Identifier 10.1109/TMM.2023.3263552

Fig. 1. A multi-modal post composed of text and picture.

As mentioned above, the core challenge of multi-modal feature fusion is to tackle the heterogeneous problem between modalities and explore the essential correlation between them. In general, different modalities have imbalanced and complementary relations that supply varying amounts of information in representing the same semantics, because some modality-specific features within one modality cannot be completely matched with other modalities. However, there is also irrelevant context and mutual interference between different modalities. Therefore, it is not enough to only consider the complementary relationship between modalities of multi-modal data, but also consider the noise or irrelevant context (i.e., regions or words) between modalities. Recently, a hierarchical multi-modal contextual attention network [10] was proposed for detecting fake news by applying two transformer units to jointly model the multi-modal data. However, it only captures the intra-modality and inter-modality relationship of multi-modal data, while ignoring the noise or irrelevant context between different modalities.

Furthermore, the general depth learning model usually needs a big quantity of manually labeled data during training. However, it is costly and time-consuming to get a big quantity of labeled data. Recently, positive unlabeled (PU) learning [11], [12], [13] has been proposed and widely employed in many tasks, which train a classifier through smidgen positive data and abundant unlabeled data. Many works have confirmed the superior performance of PU learning in semi-supervised learning [14], [15]. For example, Wang et al. [14] apply PU learning for high-quality content recognition. Wu et al. [15] first apply positive unlabeled learning to the graph for node classification to solve the problem of only a small number of labeled samples. However, these approaches ignore the learning capability of the model itself, which may have given dependable supervision.

Therefore, the following challenges need to be addressed in order to establish a productive framework for detecting fake news:

- *Challenge 1:* How to propose a more effective method to better fuse multi-modal data to capture complementary information between different modalities? How to mask the noise or irrelevant context between modalities to strengthen the detection of fake news?
- *Challenge 2:* How to perform better fake news detection with little labeled data? And how to consider the learning capability of the model itself, so as to provide reliable supervision?

To address the aforementioned challenges, we present a Curriculum-Based Multi-modal Masked Transformer Network (CMMTN) for positive unlabeled multi-modal fake news detection by modeling the inter-modality and intra-modality relationships of multi-modal information and masking the noise or irrelevant context between modalities. (1) For *Challenge 1*, we present a multi-modal masked transformer network to make full use of the similarity and difference between modalities, which can obtain both inter-modality and intra-modality relationships and mask the noise or irrelevant context between modalities to assist in strengthening the correlations between relevant contexts. (2) For *Challenge 2*, we introduce a curriculum-based PU learning, which can adaptively discover and augment confident positive examples and negative examples as training progress to investigate the model's ability to learn on its own, so as to provide reliable supervision.

To sum up, the following are the contributions of our article:

- We explore positive unlabeled multi-modal fake news detection and present a curriculum-based multi-modal masked transformer network (CMMTN) by fusing the multi-modal information in a unified network as a solution.
- A multi-modal masked transformer network is proposed to consider the similarity and differences between modalities, which can obtain both inter-modality and intra-modality relationships, and mask the noise or irrelevant context between modalities to assist in strengthening the correlations between relevant contexts.
- We introduce a curriculum-based PU learning method to adaptively discover and augment confident positive/negative examples as the training proceeds to investigate the model's ability to learn on its own.
- A large number of experiments have proved the superior robustness and effectiveness of our presented CMMTN compared with state-of-the-art methods on three public real datasets.

II. Related Work

A. Fake News Detection

Recently, detecting fake news intelligently and effectively has become a popular research issue on social media platforms. To date, there are many methods [8], [16], [17], [18], [19] which have been proposed, and existing methods are mainly separated into two groups: single-modal approaches and multi-modal approaches.

In the single-modal approaches, existing approaches [16], [17], [18], [20] usually simply concatenate the extracted text features and visual features together (i.e. the concatenate operation). For instance, Yu et al. [6] capture high-level interactions and important information about the post by utilizing CNNs. Ma et al. [5] extract hidden features of posts through RNNs. In addition, some methods take into account emotional

signals to detect fake news [21], [22]. However, social media platforms often contain multiple modal data (e.g. images, text, and video), which can supplement each other semantically and be useful to the understanding of social media [23], [24], [25], [26].

Researchers have realized that multi-modal representation plays a critical role in fake news detection, thanks to the enormous effectiveness of deep neural networks in learning image and word representations. In recent years, multi-modal fake news detection has attracted large quantities of interest. Some methods [8], [27], [28], [29], [30] use multi-modal content to detect fake news and achieve competitive performance. Khattar et al. [9] design a multimodal variational autoencoder that is capable of learning a common representation for texts and images. Shivangi et al. [31] introduce a multi-modal framework that takes into consideration information from different modalities (text and image) and then concatenated them together to classify the post. In [32], the author proposed a model, which can simultaneously learn the characteristics of news text information and visual information and capture the relationship between them according to their similarity. Zhang et al. [33] utilize the post information replied to by users to increase the ability to detect fake news. Moreover, a hierarchical multi-modal contextual attention network [10] was proposed for detecting fake news by applying two transformer units to jointly model the multi-modal data.

Despite the fact that the approaches listed above perform well, they still have shortcomings in making full use of multi-modal data, and they ignore the noise between different modalities. In this article, we develop a multi-modal masked transformer network by fusing the multi-modal information in a unified network for fake news detection.

B. Positive Unlabeled Learning

Positive unlabeled (PU) learning is a research direction of semi-supervised learning, and it utilizes smidgen positive (P) data and abundant unlabeled (U) data to learn a classifier. Existing positive unlabeled learning approaches can be classified into two groups according to how they handle unlabeled data U. The first group is known as the two-step technique, which first determines potentially negative (N) data in U, and then conducts regular supervised learning (PN learning) from both trustworthy positive and negative instances [34], [35]. The second group is known as a direct learning approach, and it views U data as N data with lesser weights. Direct learning methods (e.g. One-class SVM [36], Biased-SVM [37]) learn a classification model from the P data and U data directly. However, the first is largely dependent on heuristic methods that recognize N data, while the second group is largely dependent on various selections of U data weights, which is computationally costly to tune.

In order to address these limitations, some unbiased positive unlabeled learning approaches [11], [12], [13], [38] are proposed. The main strategy of these methods is to use an unbiased risk estimator to eliminate the PU classification bias. Niu et al. [12] provide an unbiased risk estimator to prevent the intrinsic bias for unbiased PU learning. And then, a non-negative risk estimator [13] is presented, which is more resistant to overfitting when the loss function is minimized, allowing some models to be utilized with a limited quantity of P data. Recently, some researchers have applied PU learning to the fake news detection task [39], [40]. Liu et al. [39] propose a novel deep learning framework for fake news early detection and utilize PU-Learning to improve fake news early detection given unlabeled and imbalanced data. They first conduct undersampling over the unlabeled news samples. Then, they use PN learning to train an instance on the combination of the pseudo-true news samples and the positive news samples. M. C. de Souza et al. [41] propose a new approach for detecting fake news based on PU-LP to minimize the news labeling effort, where the authors use Label propagation to handle unlabeled documents. Then, M. C. de Souza et al. [40] add a representative terms selection module based on the previous work [41] to further improve the model. Although these methods have made a fairly good performance, most of the existing methods can not make full use of unlabeled data, and ignore the existence of some confident data in unlabeled data that can be used for supervised learning. In contrast to other works, we try to improve the learning capability of the model itself by using the confident data in the unlabeled data. In our article, we introduce a curriculum-based PU learning method to adaptively discover and augment confident positive/negative examples as the training proceeds to investigate the self-learning ability of the model.

III. Problem Statement

A. General Fake News Detection

The issue of general fake news detection can be described as a binary classification task in which the goal is to determine whether or not the posts on social media are fake. Suppose a multi-modal news post composed of text and image $P = \{W, I\}$ (W represents the text and I represents the image), the model will output $Y = \{0, 1\}$ to signify the post's label, with $Y = 0$ and $Y = 1$ denoting true real news and fake news, respectively.

1) Positive Unlabeled Fake News Detection: Assume that a set of data $\mathcal{D} = P \bigcup U$, with P representing labeled posts ($\forall d_i \in P$, $y_i = 1$) and U representing unlabeled posts. Positive Unlabeled Fake news detection intends to develop a binary classification model, $f : (D; P) \mapsto Y$, to predict the unlabeled posts' class labels. We are the first to present a complete deep learning model for positive unlabeled fake news detection.

IV. Method

A. Overall Framework

We present a curriculum-based multi-modal masked transformer network (CMMTN) to fuse multi-modal information to increase the ability to detect fake news. By trying to exploit a multi-modal masked transformer network, the model CMMTN can obtain the intra-modality and inter-modality relationship and mask the noise or irrelevant context between modalities to strengthen the correlations between relevant contexts. As shown in Fig. 2, the components of our proposed model are as follows:

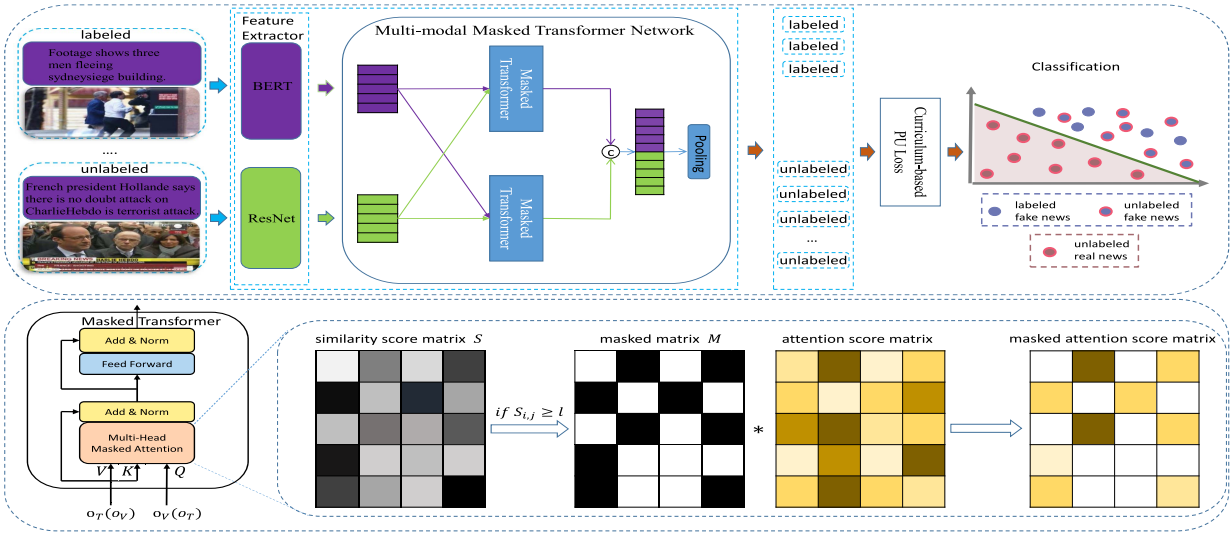


Fig. 2. Overview of our Curriculum-Based Multi-modal Masked Transformer Network (CMMTN) architectures. For both labeled and unlabeled posts, we adopt BERT and ResNet to obtain better representations for texts and images, separately. And then a multi-modal masked transformer network is employed to fuse the multi-modal content. Finally, a curriculum-based PU loss is used to optimize the model.

- *Text and Image Encoding Network:* For the given input text and picture, we use BERT [42] and ResNet50 [43] to extract the text content embedding and visual content embedding respectively.
- *Multi-modal Masked Transformer Network:* Because different modalities contain unequal amount of information and they can supplement each other, we introduce a multi-modal masked transformer network to fuse the multi-modal data, which can obtain both intra-modality and inter-modality relationships and mask the noise or irrelevant context between modalities.
- *Curriculum-based Positive Unlabeled Learning for Fake News Detection:* To address the issue of a scarcity of labeled samples, we propose a curriculum-based positive unlabeled (PU) learning for fake news detection, which can adaptively discover confident instances from the unlabeled data, which will then be labeled into trusted positive (or negative) classes.

B. Text and Image Encoding Network

As stated in the problem statement, the input of the CMMTN is a multi-modal news $P = \{W, I\}$, where W indicates text information and I indicates visual information. Qian et al. [10] and Ying et al. [44] proved that BERT and ResNet50 can obtain better text and image representation. Based on this, we use BERT and ResNet50 to extract the text content embedding and visual content embedding respectively.

1) Text Encoding Network: For the text content, we utilize BERT [42] to obtain the features that contain the semantic of the word and the linguistic contexts.

Given a sequence of words $W = \{w_1, w_2, \ldots, w_n\}$ (n denotes the number of words in the text) of a text content, we utilize the BERT [42] to generate a set of embedding $o_T = \{t_1, \ldots, t_n\}$, where t_i denotes the embedding feature of w_i. Each word representation t_i is obtained using pre-trained BERT:

$$\mathbf{o}_T = \{t_1, \ldots, t_n\} = \text{BERT}(\mathbf{W}) \tag{1}$$

where $\mathbf{t}_i \in \mathbb{R}^{d_t}$ denotes the output feature of the i-th token by BERT, and d_t denotes the dimensionality of the word embedding.

2) Image Encoding Network: We extract region features of a given visual content I using the pre-trained ResNet50 [43], which is pre-trained on ImageNet [45]. Before applying the pre-trained ResNet50 model, we resize the image to 112×112 pixels. A $4 \times 4 \times 2048$-dimensional feature tensor is generated from the last convolution layer as the high-level semantic representation, then it is flattened to a 16×2048-dimensional matrix. So we obtain a feature set of image regions $o_V = \{v_1, \ldots, v_m\}$ (m means the total number of image's regions), where we consider the penultimate pooling layer as the output, and each v_j represents the mean-pooled convolutional feature of the j-th region. During training, the pre-trained model is fixed. Given the input visual content I, the output of the visual feature extractor in the penultimate pooling layer can be described as follows:

$$\mathbf{o_V} = \{v_1, \ldots, v_m\} = \text{ResNet50}(I) \tag{2}$$

where $\mathbf{v}_j \in \mathbb{R}^{d_v}$ and d_v denotes the dimensionality of the region embedding of image.

In addition, we also add a 2D-convolutional layer to transform the region embedding dimension d_v to the word embedding dimension d_t to fit our task.

C. Multi-Modal Masked Transformer Network

In order to effectively fuse multi-modal features and mask the noise or irrelevant context between modalities, we design a multi-modal masked transformer network to establish the multi-modal information and capture complementary information between different modalities. As depicted in Fig. 2, the

multi-modal masked transformer network is made up of two masked transformer units. The above one takes image features as Q and text features as K, V. On the contrary, the next one takes the text feature as Q and the image feature as K, V.

In the transformer [46] model, the key component is the self-attention module, and it can build the long-range dependency between inputs and outputs. Given a set of input $X \in \mathbb{R}^{L\times D}$ (L represents the number of tokens and D represents the embedded dimension), the following is a definition of single-headed self-attention:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{3}$$

where QK^T represents the attention score between queries Q and keys K, and $\sqrt{d}$ denotes the scaling factor.

In order to capture the relationship between different modalities and mask the noise or irrelevant context between modalities, we have improved the self attention in Transformer [46]. Take the o_V as Q and o_T as K, V as an example. For a multi-modal input $o_T = \{t_1, \ldots, t_n\} \in \mathbb{R}^{n\times d_t}$ and $o_V = \{v_1, \ldots, v_m\} \in \mathbb{R}^{m\times d_v}$, the similarity score matrix S and masked matrix M are calculated as follows:

$$\text{S} = \text{softmax}\left(\frac{o_V o_T^T}{\sqrt{d}}\right) \tag{4}$$

$$\text{M}_{i,j} = \begin{cases} 1, & if\ \text{S}_{i,j} \geq l \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

$$l = FFN([o_V||o_T]) \tag{6}$$

where $\text{S}_{i,j}$ represents the similarity score of the i-th region in o_V and the j-th word in o_T, l is the context similarity threshold, FFN is a two-layer fully-connected network, and $||$ denotes the concatenate operation. Note that, we do not set the threshold l as a fixed constant, but use a two-layer full connection according to the original characteristics of the text and the image and map it to a one-dimensional space to get l, so that the network can adaptively set different threshold l for different samples and the model will be more robust. Then, the modified single-headed multi-modal self-attention can be defined as:

$$\begin{aligned} &\text{Masked Attention}(Q, K, V, M) \\ &= \text{softmax}\left(\frac{M * (QK^T)}{\sqrt{d}}\right)V \end{aligned} \tag{7}$$

where $Q = W_Q^1 o_V, K = W_K^1 o_T, V = W_V^1 o_T$ and W_Q^1, W_K^1, W_V^1 are different linear transformations that project the input into queries, keys and values respectively. $*$ means the hadamard product.

D. Curriculum-Based Positive Unlabeled Learning for Fake News Detection

After integrating the textual and visual features via a multi-modal masked transformer network module, we will gain a new features $O = \{\mathbf{o}_1, \ldots, \mathbf{o}_N\}, \mathbf{o}_k \in \mathbb{R}^{(n+m)\times d_t}$ in the final layer, where N indicates the number of posts. According to the problem statement (Section III), one crucial question is *how can we use this new representation to achieve positive unlabelled learning for fake news detection*?

1) Traditional Fake News Detection: In the traditional fake news detection methods, researchers generally regard this as a binary classification task and learn a model $f : \mathcal{O} \to \mathcal{Y}$ to categorize news posts into the predefined classes $\mathcal{Y} = \{+1, -1\}$, where $+1$ refers to positive samples and -1 refers to negative samples.

Assume that $\mathcal{L}(y,' y)$ is the predicting loss of the model, where $\mathcal{L}$ indicates a loss function, y' indicates the output and y indicates the ground truth. Let f be a mapping function that maps the input o between 0 and 1. The general binary classification issue is recast as a risk minimization problem:

$$R(f) = \mathbb{E}\left[\mathcal{L}(f(O), Y)\right] = \pi_p R_p^+(f) + \pi_n R_n^-(f) \tag{8}$$

which $R_p^+(f) = \mathbb{E}_p[\mathcal{L}(f(O), +1)]$ and $R_n^-(f) = \mathbb{E}_p[\mathcal{L}(f(O), -1)]$ is the expectation loss of positive samples and negative samples respectively. Here, the *class-prior probability* is denoted by $\pi_p = p(Y = +1)$ and $\pi_n = p(Y = -1) = 1 - \pi_p$. π_p is assumed to be known throughout the work and can be derived from positive data [47].

As a result, we may minimize an approximated $R(f)$ for a traditional binary classification problem (such as PN Learning) by,

$$\hat{R}_{pn}(f) = \pi_p \hat{R}_p^+(f) + \pi_n \hat{R}_n^-(f) \tag{9}$$

where $\hat{R}_p^+(f) = (1/n_p)\sum_{i=1}^{n_p} \mathcal{L}(f(\mathbf{o}_i^p), +1)$ and $\hat{R}_n^-(f) = (1/n_n)\sum_{i=1}^{n_n} \mathcal{L}(f(o_i^n), -1)$. Besides, n_p indicates the number of positive data, and n_n indicates the number of negative data.

2) Positive Unlabeled Learning: In order to address the issue of less labeled posts in fake news detection, we first recast the general binary classification task as a risk reduction problem, and then offer two effective positive unlabeled learning approaches (*unbiased risk estimator* [12] and *non-negative risk estimator* [13]) to approximate the risk for fake news detection.

Unbiased Risk Estimator for Positive Unlabeled Learning: Negative training data, on the other hand, is not available for PU learning. As a consequence, we must use (9) to estimate $\hat{R}_n^-(f)$. In order to approximate $\hat{R}_n^-(f)$, we employ an unique unbiased risk estimator [11]. Specifically, through the positive data loss $\hat{R}_p^-(f)$ and unlabeled data loss $\hat{R}_u^-(f)$, we can acquire the negative loss $\hat{R}_n^-(f)$, and its calculation formula is as follows:

$$\pi_p \hat{R}_n^-(f) = -\pi_p \hat{R}_p^-(f) + \hat{R}_u^-(f) \tag{10}$$

where $\hat{R}_p^-(f) = (1/n_p)\sum_{i=1}^{n_p} \mathcal{L}(f(o_i^p), -1)$, and $\hat{R}_u^-(f) = (1/n_u)\sum_{i=1}^{n_u} \mathcal{L}(f(o_i^u), -1)$. In addition, n_p indicate the number of positive data and n_u indicate the number of unlabeled data.

As a result, the risk $R(f)$ can be obtained by the following formula:

$$\hat{R}_{pu}(f) = \pi_p \hat{R}_p^+(f) - \pi_p \hat{R}_p^-(f) + \hat{R}_u^-(f) \tag{11}$$

Non-negative Risk Estimator for Positive Unlabeled Learning: Even though the unbiased risk estimator can handle the positive unlabeled learning issue effectively. Nevertheless, since

$\hat{R}_p^-(f)$ is preceded by a minus sign, (11) may lead the risk value to be negative. The $\hat{R}_p^-(f)$ refers to the sample with negative label expected risk value obtained through model projection in the positive data set, i.e., if the prediction is a negative sample, the loss is 0; otherwise, the loss is positive. For positive unlabeled learning, this will result in an overfitting issue. We apply a non-negative risk estimator $\hat{R}_{pu}(f)$, which is motivated by [13], as follows:

$$\hat{R}_{pu}(f) = \pi_p \hat{R}_p^+(f) + \max\left\{0, \hat{R}_u^-(f) - \pi_p \hat{R}_p^-(f)\right\}. \quad (12)$$

However, the above two PU learning methods ignore the learning capability of the model itself, which may have given dependable supervision.

3) Curriculum-Based Positive Unlabeled Learning: Following existing studies [48], [49] that the model should be trained from easy samples to hard ones. Inspired by the above work, we rank the prediction results of all posts to discover simple examples and confidently label them, and then label it positive and add it to the labeled pool for the next training step. As the training process continues, the data in the trusted (confidence) set will become larger and larger to provide confident full supervision.

We can calculate the output $f(x)$ and then the probability of o being positive as $p(x) = P(Y = +1|x) = f(x)$ using the model f, an input sample o, and corresponding label y. A higher $p(x)$ indicates a greater likelihood that o belongs to the positive class as indicated by f, and vice versa. We can choose n most confident positive and n most confident negative examples from the current unlabeled data set U by sorting $p(x)$ in ascending order at each epoch. They will then be deleted from U and put into our trustworthy subset D_t, which will be referred to as labeled training examples during the later training process. Finally, our hybrid loss for curriculum-based positive unlabeled learning becomes:

$$\mathcal{L}(o, y) = \sum_{(o,y)\in D_t} \mathcal{L}_{ce}(o, y) + \sum_{o\in(U-D_t)} \mathcal{L}_{nnPU}(o) \quad (13)$$

where $\mathcal{L}_{ce}$ is the cross entropy loss. It's worth noting that prior studies have chosen only confidently positive data [50] or negative data [34], while our curriculum-based positive unlabeled learning selects both positive data and negative data. One benefit of using cross entropy as supervised loss is that the size of the trusted set of positive / negative samples is balanced at each sampling step, thus avoiding the possible risk of excessive class imbalance that can occur when only sampling one class incrementally.

V. Experiments and Results

A. Experimental Setup

1) Datasets: We compare our model CMMTN with state-of-the-art baseline methods on three public datasets: WEIBO [19], TWITTER [9], [19], and PHEME [51]. WEIBO dataset is used in [19], where the real news is gathered from reliable Chinese media platforms like the Xinhua News Agency,[1] and the fake news was collected from the Weibo[2] data verified by Weibo's official rumor debunking system. The TWITTER dataset comes from the MediaEval Verifying Multimedia Use benchmark [52], which is designed to identify fraudulent Twitter posts. The PHEME dataset [51] is built on five breaking news stories, each of which comprises several posts. Here, all three datasets contain a large quantity of labeled textual content (text) and labeled visual content (images). Following [9], we divide WEIBO and PHEME into a train set and test set by a 4:1 ratio, while TWITTER is employed with the development set for training and the test set for testing to maintain the same data segmentation scheme as the baseline. Table I shows the records of the three benchmark datasets.

TABLE I
The Statistics of Three Real-World Datasets

News	WEIBO	TWITTER	PHEME
# of Fake News	4749	7898	1972
# of Real News	4779	6026	3830
# of Images	9528	514	3670

2) Evaluation Metrics: In the binary categorization task (e.g., fake news detection), people usually adopt Accuracy as the evaluation metric. Nonetheless, if a dataset has the problem of class imbalance, its trustworthiness will be significantly degraded. Thus, in our experiments, in addition to the Accuracy metric, macro F_1 and weighted F_1 were also considered as supplementary performance indicators for the task.

3) Implementation Details: We utilize the BERT [42] to obtain word embeddings and the ResNet50 [43] to obtain image features from multi-modal data of posts. The post embedding dimension $d = 768$, the word embedding dimension is 768, and the image features dimension is 2048. To match our task, we employ a convolution layer to convert the image region feature dimension from 2048 to 768. The Adam [53] optimizer is used to train our algorithm, which is built on the Pytorch deep learning framework [54]. We train the model for 200 epochs and set the learning rate of the model to 0.01. The Minibatch size during the training process is set to 256. We divided each PN dataset into positive sets and unlabeled sets at random for fairness in comparison. Following [13], we sample $r * N_P$ (N_P denotes the overall amount of positive posts in the training set) posts from P set as the positive data, and the remainder positive posts and negative posts are regarded as the unlabeled posts (r is the sampling ratio of positive posts in the training set). Each experiment in our article is repeated 10 times to obtain a stable performance, and both the mean and standard deviation are reported.

B. Baselines

Before introducing the baselines, let's introduce two PU learning approaches as follows:

- *uPU [12]:* uPU is a positive unlabeled learning method based on an unbiased estimator. For loss functions that meet specific linear-odd requirements, it is convex.

[1][Online]. Available: http://www.xinhuanet.com/

[2][Online]. Available: https://weibo.com/

TABLE II
COMPARISON RESULTS OF DIFFERENT MODELS ON THREE DATASETS

Dataset	Methods	1%(r)			2%(r)		
		Accuracy	macro F1	weighted F1	Accuracy	macro F1	weighted F1
WEIBO	SAFE_uPU	0.701±0.038	0.699±0.029	0.699±0.029	0.724±0.012	0.722±0.009	0.723±0.011
	SAFE_nnPU	0.715±0.068	0.717±0.045	0.718±0.069	0.746±0.009	0.744±0.017	0.742±0.027
	SpotFake_uPU	0.780±0.048	0.779±0.049	0.779±0.049	0.806±0.007	0.804±0.007	0.804±0.007
	SpotFake_nnPU	0.802±0.028	0.801±0.028	0.801±0.028	0.815±0.003	0.814±0.003	0.814±0.003
	HMCAN_uPU	0.808±0.048	0.805±0.050	0.805±0.050	0.813±0.045	0.811±0.056	0.810±0.058
	HMCAN_nnPU	**0.811±0.025**	**0.810±0.025**	**0.810±0.025**	0.818±0.016	0.816±0.016	0.816±0.016
	$CMMTN(Ours)$	0.806± 0.026	0.804±0.027	0.804±0.027	**0.825±0.014**	**0.824±0.013**	**0.824±0.013**
TWITTER	SAFE_uPU	0.566±0.029	0.439±0.069	0.460±0.064	0.571±0.046	0.420±0.091	0.443±0.085
	SAFE_nnPU	0.621±0.038	0.545±0.068	0.559±0.064	0.694±0.100	0.628±0.164	0.640±0.155
	SpotFake_uPU	0.750±0.092	0.723±0.119	0.730±0.114	0.779±0.158	0.761±0.188	0.761±0.192
	SpotFake_nnPU	0.774±0.128	0.762±0.147	0.761±0.152	0.780±0.083	0.760±0.109	0.765±0.104
	HMCAN_uPU	0.781±0.120	0.748±0.152	0.753±0.158	0.801±0.092	0.788±0.097	0.790±0.088
	HMCAN_nnPU	0.789±0.131	0.758±0.171	0.764±0.164	0.810±0.110	0.796±0.130	0.799±0.126
	$CMMTN(Ours)$	**0.819±0.097**	**0.804±0.118**	**0.808±0.113**	**0.858±0.091**	**0.848±0.117**	**0.850±0.112**
PHEME	SAFE_uPU	0.682±0.014	0.544±0.050	0.623±0.036	0.704±0.009	0.623±0.031	0.678±0.016
	SAFE_nnPU	0.694±0.013	0.553±0.038	0.633±0.026	0.714±0.012	0.616±0.027	0.677±0.019
	SpotFake_uPU	0.742±0.013	0.683±0.025	0.726±0.019	0.759±0.010	0.710±0.026	0.747±0.017
	SpotFake_nnPU	0.746±0.014	0.695±0.023	0.734±0.019	0.767±0.011	0.724±0.028	0.759±0.018
	HMCAN_uPU	0.740±0.015	0.700±0.020	0.734±0.016	0.761±0.014	0.723±0.032	0.754±0.024
	HMCAN_nnPU	0.746±0.015	0.709±0.022	0.741±0.017	0.766±0.011	0.733±0.020	0.762±0.015
	$CMMTN(Ours)$	**0.772±0.024**	**0.740±0.037**	**0.768±0.030**	**0.784±0.017**	**0.757±0.022**	**0.782±0.018**

- *nnPU [13]:* nnPU is a positive unlabeled learning method based on a non-negative risk estimator that is more resistant to overfitting when minimized. As a result, given a limited set of P (positive) data, some flexible models can be utilized.

We chose the following baselines with appropriate adaptations to develop an impartial comparison and evaluate the availability of our model.

- *SpotFake [31]:* SpotFake applies the BERT to obtain better textual information and uses the VGG-19 that pre-train on ImageNet [55] to get better visual information to recognize whether a post is true or fake.
- *SAFE [32]:* SAFE is a multi-modal fake news detection approach, which uses TextCNN [56] to extract textual and image features. Then a cross-modal similarity module is applied to explore the correlation between modalities and generate the final representation.
- *HMCAN [10]:* HMCAN is a model by applying two transformer units to jointly model the multi-modal data. In addition, it utilizes BERT [42] to extract the hierarchical semantic information for textual content.

C. Results and Analysis

The results about CMMTN as well as all baseline approaches are listed in Table II. We have got the following conclusions: (1) *SpotFake_** (* means uPU or nnPU model) has superior performance than *SAFE_**, which shows that the BERT [42] and ResNet50 [43] can obtain better representations to enhance the model's performance. (2) *HMCAN_** can achieve better results than *SpotFake_** and *SAFE_**, indicating that better fusion of multi-modal features by capturing the inter-modality and intra-modality relationships of multi-modal data can contribute to improving the recognition of fake news. (3) The proposed CMMTN outperforms all the baselines on the TWITTER and PHEME datasets. Meanwhile, when $r = 0.02$, CMMTN outperforms all the baselines on the WEIBO dataset. The findings show that our suggested curriculum-based PU Learning method can obtain better performance on the positive and unlabeled data by adaptively discovering and augmenting confident positive/negative examples as the training proceeds to investigate the model's ability to learn on its own.

D. Analysis of CMMTN Components

Because our CMMTN contains more than one component, we also compared various variants of CMMTN from the following viewpoint to prove the availability of CMMTN—(1) influence of the visual content, (2) impact of the curriculum-based PU Learning, and (3) impact of masked strategy (V, C, M) for the multi-modal masked transformer network module. The CMMTN variants listed below are offered for comparison.

- *CMMTN$\neg V$:* A variant of CMMTN in which the visual information is removed and just textual data is used.
- *CMMTN$\neg C$:* A variant of CMMTN in which the curriculum-based PU loss is removed and only applies the general nnPU loss.
- *CMMTN$\neg M$:* A variant of CMMTN in which the masked strategy is omitted, and only uses the general transformer module.

Table III shows the results of the ablation study.

1) Influence of the Visual Information: We make a comparison in performance between CMMTN and CMMTN$\neg V$ on three datasets to show that the visual information is effective. As a result of the findings, we can conclude that CMMTN performs better than CMMTN$\neg V$, demonstrating that the visual information can be adopted to enhance our model.

TABLE III
COMPARISON RESULTS OF DIFFERENT VARIANTS IN CMMTN ON THREE DATASETS

Dataset	Methods	1%(r)			2%(r)		
		Accuracy	macro F1	weighted F1	Accuracy	macro F1	weighted F1
WEIBO	CMMTN$\neg V$	0.650±0.024	0.648±0.025	0.648±0.025	0.685±0.020	0.683±0.020	0.683±0.020
	CMMTN$\neg C$	0.792±0.019	0.791±0.020	0.791±0.020	0.795±0.023	0.793±0.024	0.793±0.024
	CMMTN$\neg M$	0.778±0.039	0.777±0.039	0.777±0.039	0.811±0.019	0.810±0.019	0.810±0.019
	$CMMTN(Ours)$	**0.806±0.026**	**0.804±0.027**	**0.804±0.027**	**0.825±0.014**	**0.824±0.013**	**0.824±0.013**
TWITTER	CMMTN$\neg V$	0.767±0.097	0.747±0.130	0.752±0.124	0.825±0.032	0.822±0.035	0.823±0.034
	CMMTN$\neg C$	0.771±0.111	0.740±0.152	0.747±0.145	0.845±0.065	0.841±0.074	0.842±0.071
	CMMTN$\neg M$	0.803±0.076	0.794±0.088	0.797±0.085	0.854±0.043	**0.851±0.047**	**0.852±0.046**
	$CMMTN(Ours)$	**0.819±0.097**	**0.804±0.118**	**0.808±113**	**0.858±0.091**	0.848±0.117	0.850±0.112
PHEME	CMMTN$\neg V$	0.727±0.017	0.663±0.041	0.708±0.030	0.758±0.014	0.724±0.018	0.754±0.015
	CMMTN$\neg C$	0.747±0.011	0.708±0.019	0.741±0.014	0.768±0.013	0.730±0.018	0.761±0.015
	CMMTN$\neg M$	0.738±0.029	0.690±0.067	0.728±0.048	0.772±0.019	0.735±0.029	0.768±0.022
	$CMMTN(Ours)$	**0.772± 0.024**	**0.740±0.037**	**0.768±0.030**	**0.784±0.017**	**0.757±0.022**	**0.782±0.018**

TABLE IV
COMPARISON RESULTS OF DIFFERENT MODELS ON THREE DATASETS IN THE LABEL-COMPLETE SCENARIO. (THE BASELINE RESULTS ARE FROM THE ARTICLE OF *HMCAN* [10])

Dataset	Methods	Accuracy	Fake news			Real news		
			Precision	Recall	F1	Precision	Recall	F1
WEIBO	SAFE	0.763	0.833	0.659	0.736	0.717	0.868	0.785
	EANN	0.782	0.827	0.697	0.756	0.752	0.863	0.804
	MVAE	0.824	0.854	0.769	0.809	0.802	0.875	0.837
	SpotFake	**0.892**	0.902	0.964	**0.932**	0.847	0.656	0.739
	SpotFake+	0.870	0.887	0.849	0.868	0.855	0.892	0.873
	HMCAN	0.885	0.920	0.845	0.881	0.856	0.926	0.890
	$MMTN(Ours)$	0.889±0.005	0.886±0.011	0.893±0.013	0.889±0.005	0.892±0.010	0.885±0.014	**0.893±0.005**
TWITTER	SAFE	0.766	0.777	0.795	0.786	0.752	0.731	0.742
	EANN	0.648	0.810	0.498	0.617	0.584	0.759	0.660
	MVAE	0.745	0.801	0.719	0.758	0.689	0.777	0.730
	SpotFake	0.777	0.751	0.900	0.820	0.832	0.606	0.701
	SpotFake+	0.790	0.793	0.827	0.810	0.786	0.747	0.766
	HMCAN	0.897	0.971	0.801	0.878	0.853	0.979	**0.912**
	$MMTN(Ours)$	**0.903±0.029**	0.870±0.047	0.917±0.033	**0.892±0.029**	0.927±0.028	0.881±0.049	0.903±0.029
PHEME	SAFE	0.811	0.827	0.559	0.667	0.806	0.940	0.866
	EANN	0.681	0.685	0.664	0.694	0.701	0.750	0.747
	MVAE	0.852	0.806	0.719	0.760	0.871	0.917	0.893
	SpotFake	0.823	0.743	0.745	0.744	0.864	0.863	0.863
	SpotFake+	0.800	0.730	0.668	0.697	0.832	0.869	0.850
	HMCAN	0.881	0.830	0.838	**0.834**	0.910	0.905	0.908
	$MMTN(Ours)$	**0.887±0.008**	0.870±0.023	0.789±0.022	0.827±0.011	0.895±0.010	0.938±0.014	**0.915±0.007**

2) Impact of the the Curriculum-Based PU Learning: We have conducted comparative experiments CMMTN and CMMTN$\neg C$ on both three datasets, and checked out the advantages of the curriculum-based PU Learning component. From the experiment results, we find our CMMTN outperforms CMMTN$\neg C$, confirming the advantage of curriculum-based PU Learning for positive and unlabelled data.

3) Impact of the Masked Strategy: We have conducted comparative experiments CMMTN and CMMTN$\neg M$ on both three datasets and checked out the advantages of the masked strategy. From the experiment results, we can find that our CMMTN outperforms CMMTN$\neg M$, which demonstrates the efficacy of the masked strategy in our model. In addition, we find that the performance difference between CMMTN$\neg M$ and CMMTN is the smallest on the TWITTER dataset and the larger on the WEIBO and the PHEME dataset, which shows that our mask strategy has the most obvious superiority when visual information is not missing, and its superiority decreases when visual information is missing.

E. Statistical Tests

To verify whether our method (CMMTN) is significantly better than other methods, we adopt the Friedman test and Nemenyitest [57] to further compare the performance of CMMTN with that of its rivals.

We first perform the Friedman test at the 0.05 significance level under the null-hypothesis which states that the performance of all algorithms is the same on all datasets and all r values (i.e., the average ranks of all algorithms are equivalent). The average ranks of CMMTN and its rivals when using different evaluation metrics are summarized in Table V. From Table V, we can see that the null hypothesis is rejected on these two evaluation metrics. We also note that CMMTN performs better than its rivals (the lower rank value is better).

To further analyze the difference between CMMTN and its rivals, we perform the Nemenyi test, which states that the performance levels of two algorithms are significantly different if the corresponding average ranks differ by at least one critical

TABLE V
THE AVERAGE RANKS OF CMMTN AND ITS RIVALS FOR ACCURACY AND MACRO F1

Methods		SAFE_uPU	SAFE_nnPU	SpotFake_uPU	SpotFake_nnPU	HMCAN_uPU	HMCAN_nnPU	CMMTN
Avg rank	Accuracy	7	6	4.83	3.5	3.25	2.08	**1.33**
	macro F1	6.83	6.17	4.83	3.5	3.33	2	**1.33**

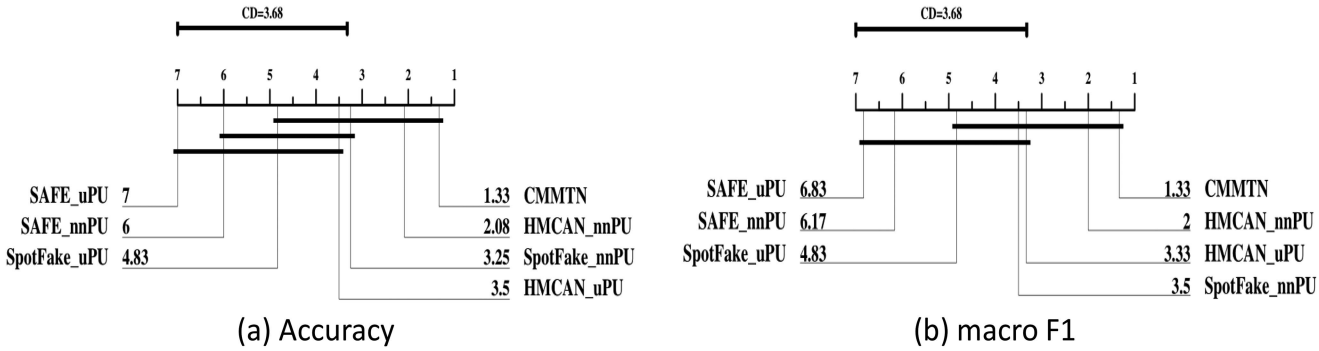


Fig. 3. Crucial difference diagram of the Nemenyi test for Accuracy and macro F1 on three datasets and two r values.

difference (CD). Fig. 3 provides the CD diagrams, where the average rank of each algorithm is marked along the axis (lower ranks to the right). From Fig. 3, we observe that CMMTN achieves a comparable performance against *HMCAN_nnPU*, *HMCAN_uPU*, *SpotFake_nnPU* and *SpotFake_uPU*, and CMMTN significantly outperforms the *SAFE_nnPU* and *SAFE_uPU*. CMMTN is the only algorithm that achieves the lowest rank value for both accuracy and macro F1.

F. Results and Analysis in Label-Complete Scenario

To further verify the availability of the multi-modal masked transformer network (MMTN), we compare it with some of the latest methods in the label complete scenario (ie. $p = 100\%$). In addition, we added three comparison models including *EANN* [8], *MVAE* [9], *SpotFake+* [58]. Table IV shows the results, and the following conclusions can be drawn: 1) *Accuracy:* The proposed *MMTN* is higher than all baselines on TWITTER and PHEME datasets. 2) *F1 of fake news:* The *MMTN* is higher than all baselines on TWITTER datasets. 3) *F1 of real news:* On WEIBO and PHEME datasets, the *MMTN* is higher than all baselines. On the whole, the proposed *MMTN* outperforms *HMCAN* and other baselines. This shows that capturing the intra-modal relationship and inter-modal relationship of multi-modal data and masking the irrelevant context between modalities can assist in fake news detection.

G. Impact of Selecting Confident Samples

In the curriculum learning, we compared different selection strategies in the case of $r = 1\%$, and the results are depicted in Fig. 4. Here, 'B,' 'P,' and 'N' respectively mean that both positive posts and negative posts are selected, only positive posts are selected and only negative posts are selected. It can be found from Fig. 4 that 'B' shows a better performance than 'P' and 'N' on the three datasets, which indicates that the strategy of selecting both positive samples and negative samples is better than the other two.

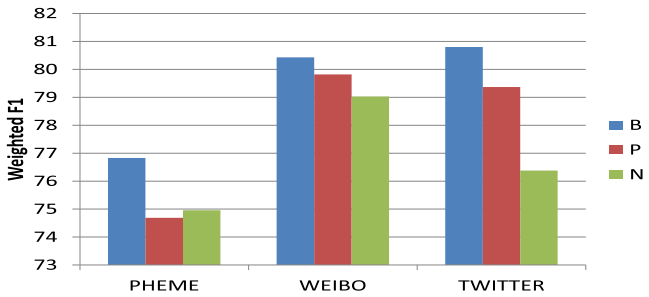


Fig. 4. Impact of selecting both confident positive samples and negative samples ($r = 1\%$).

VI. CONCLUSION

This article presents a curriculum-based multi-modal masked transformer network for positive unlabeled fake news detection. The majority of existing approaches are tough to make the best of utilizing the intra-modality relationship and inter-modality relationship of multi-modal data. Simultaneously, they ignore the noise or irrelevant context between modalities. Additionally, existing PU learning approaches mainly ignore the learning capability of the model itself, which may have given dependable supervision. To address the aforementioned issues, CMMTN is presented to model the inter-modality and intra-modality relationships of multi-modal data and mask the noise or irrelevant context between modalities. Our strategy is based on three technical breakthroughs: (1) We apply BERT and ResNet to obtain better features for texts and images, separately. (2) A multi-modal masked transformer network is used to better fuse the multi-modal feature information, which is capable of capturing the intra-modality and inter-modality relationship and masking the noise or irrelevant context between modalities. (3) We introduce a curriculum-based PU learning method to deal

with positive and unlabeled data. Experiments and comparisons show the superiority of our model CMMTN for detecting fake news. In the future, we will try to use additional knowledge or user comments to discover explainable information for detecting fake news. Besides, a more efficient method of extracting visual content information also will be explored, which might provide helpful complementary information.

REFERENCES

[1] G. Ruffo, A. Semeraro, A. Giachanou, and P. Rosso, "Studying fake news spreading, polarisation dynamics, and manipulation by bots: A tale of networks and language," *Comput. Sci. Rev.*, vol. 47, 2023, Art. no. 100531.

[2] L. Cui and D. Lee, "COAID: COVID-19 healthcare misinformation dataset," 2020, *arXiv:2006.00885*.

[3] X. Liu, A. Nourbakhsh, Q. Li, R. Fang, and S. Shah, "Real-time rumor debunking on twitter," in *Proc. 24th ACM Int. Conf. Inf. Knowl. Manage.*, 2015, pp. 1867–1870.

[4] J. Ma, W. Gao, Z. Wei, Y. Lu, and K.-F. Wong, "Detect rumors using time series of social context information on microblogging websites," in *Proc. 24th ACM Int. Conf. Inf. Knowl. Manage.*, 2015, pp. 1751–1754.

[5] J. Ma et al., "Detecting rumors from microblogs with recurrent neural networks," in *Proc. Int. Joint Conf. Artif. Intell.*, 2016, pp. 3818–3824.

[6] F. Yu et al., "A convolutional approach for misinformation identification," in *Proc. Int. Joint Conf. Artif. Intell.*, 2017, pp. 3901–3907.

[7] J. Ma, W. Gao, and K.-F. Wong, "Detect rumors on twitter by promoting information campaigns with generative adversarial learning," in *Proc. World Wide Web Conf.*, 2019, pp. 3049–3055.

[8] Y. Wang et al., "EANN: Event adversarial neural networks for multi-modal fake news detection," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2018, pp. 849–857.

[9] D. Khattar, J. S. Goud, M. Gupta, and V. Varma, "MVAE: Multimodal variational autoencoder for fake news detection," in *Proc. World Wide Web Conf.*, 2019, pp. 2915–2921.

[10] S. Qian, J. Wang, J. Hu, Q. Fang, and C. Xu, "Hierarchical multi-modal contextual attention network for fake news detection," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2021, pp. 153–162.

[11] M. C. D. Plessis, G. Niu, and M. Sugiyama, "Analysis of learning from positive and unlabeled data," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2014, pp. 703–711.

[12] G. Niu and M. Sugiyama, "Convex formulation for learning from positive and unlabeled data," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1386–1394.

[13] R. Kiryo, G. Niu, M. C. du Plessis, and M. Sugiyama, "Positive-unlabeled learning with non-negative risk estimator," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1675–1685.

[14] J. Wang, J. Hu, S. Qian, Q. Fang, and C. Xu, "Multimodal graph convolutional networks for high quality content recognition," *Neurocomputing*, vol. 412, pp. 42–51, 2020.

[15] M. Wu, S. Pan, L. Du, and X. Zhu, "Learning graph neural networks with positive and unlabeled nodes," *ACM Trans. Knowl. Discov. From Data*, vol. 15, no. 6, pp. 1–25, 2021.

[16] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD Explorations Newslett.*, vol. 19, no. 1, pp. 22–36, 2017.

[17] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," in *Proc. 20th Int. Conf. World Wide Web*, 2011, pp. 675–684.

[18] S. Kwon, M. Cha, K. Jung, W. Chen, and Y. Wang, "Prominent features of rumor propagation in online social media," in *Proc. IEEE 13th Int. Conf. Data Mining*, 2013, pp. 1103–1108.

[19] Z. Jin, J. Cao, H. Guo, Y. Zhang, and J. Luo, "Multimodal fusion with recurrent neural networks for rumor detection on microblogs," in *Proc. 25th ACM Int. Conf. Multimedia*, 2017, pp. 795–816.

[20] A. Gupta, P. Kumaraguru, C. Castillo, and P. Meier, "Tweetcred: Real-time credibility assessment of content on twitter," in *Proc. Int. Conf. Social Informat.*, 2014, pp. 228–243.

[21] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.

[22] A. Giachanou, P. Rosso, and F. Crestani, "The impact of emotional signals on credibility assessment," *J. Assoc. Inf. Sci. Technol.*, vol. 72, no. 9, pp. 1117–1132, 2021.

[23] X. Wu, C.-W. Ngo, and A. G. Hauptmann, "Multimodal news story clustering with pairwise visual near-duplicate constraint," *IEEE Trans. Multimedia*, vol. 10, no. 2, pp. 188–199, Feb. 2008.

[24] I. Kalamaras, A. Drosou, and D. Tzovaras, "Multi-objective optimization for multimodal visualization," *IEEE Trans. Multimedia*, vol. 16, no. 5, pp. 1460–1472, Aug. 2014.

[25] X. Yang, T. Zhang, and C. Xu, "Cross-domain feature learning in multimedia," *IEEE Trans. Multimedia*, vol. 17, no. 1, pp. 64–78, Jan. 2015.

[26] R. Ji, F. Chen, L. Cao, and Y. Gao, "Cross-modality microblog sentiment prediction via Bi-layer multimodal hypergraph learning," *IEEE Trans. Multimedia*, vol. 21, no. 4, pp. 1062–1075, Apr. 2019.

[27] D. Khattar, J. S. Goud, M. Gupta, and V. Varma, "MVAE: Multimodal variational autoencoder for fake news detection," in *Proc. World Wide Web Conf.*, 2019, pp. 2915–2921.

[28] S. Singhal, R. R. Shah, T. Chakraborty, P. Kumaraguru, and S. Satoh, "SpotFake: A multi-modal framework for fake news detection," in *Proc. 5th IEEE Int. Conf. Multimedia Big Data*, 2019, pp. 39–47.

[29] A. Giachanou, G. Zhang, and P. Rosso, "Multimodal multi-image fake news detection," in *Proc. IEEE 7th Int. Conf. Data Sci. Adv. Anal.*, 2020, pp. 647–654.

[30] G. Zhang, A. Giachanou, and P. Rosso, "SceneFND: Multimodal fake news detection by modelling scene context information," *J. Inf. Sci.*, 2022, Art. no. 01655515221087683.

[31] S. Singhal, R. R. Shah, T. Chakraborty, P. Kumaraguru, and S. Satoh, "SpotFake: A multi-modal framework for fake news detection," in *Proc. IEEE 5th Int. Conf. Multimedia Big Data*, 2019, pp. 39–47.

[32] X. Zhou, J. Wu, and R. Zafarani, "Safe: Similarity-aware multi-modal fake news detection," in *Proc. Pacific-Asia Conf. Knowl. Discov. Data Mining*, 2020, pp. 354–367.

[33] H. Zhang, S. Qian, Q. Fang, and C. Xu, "Multi-modal meta multi-task learning for social media rumor detection," *IEEE Trans. Multimedia*, vol. 24, pp. 1449–1459, 2021.

[34] X. Li and B. Liu, "Learning to classify texts using positive and unlabeled data," in *Proc. Int. Joint Conf. Artif. Intell.*, 2003, vol. 3, pp. 587–592.

[35] D. H. Fusilier, M. Montes-y Gómez, P. Rosso, and R. G. Cabrera, "Detecting positive and negative deceptive opinions using pu-learning," *Inf. Process. Manage.*, vol. 51, no. 4, pp. 433–443, 2015.

[36] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Comput.*, vol. 13, no. 7, pp. 1443–1471, 2014.

[37] B. Liu, Y. Dai, X. Li, W. S. Lee, and P. S. Yu, "Building text classifiers using positive and unlabeled examples," in *Proc. IEEE Int. Conf. Data Mining*, 2003, pp. 179–186.

[38] J. Zhang, Z. Wang, J. Meng, Y.-P. Tan, and J. Yuan, "Boosting positive and unlabeled learning for anomaly detection with multi-features," *IEEE Trans. Multimedia*, vol. 21, no. 5, pp. 1332–1344, May 2019.

[39] Y. Liu and Y.-F. B. Wu, "FNED: A deep network for fake news early detection on social media," *ACM Trans. Inf. Syst.*, vol. 38, no. 3, pp. 1–33, 2020.

[40] M. C. de Souza et al., "A network-based positive and unlabeled learning approach for fake news detection," *Mach. Learn.*, vol. 111, no. 10, pp. 3549–3592, 2022.

[41] M. C. d. Souza, B. M. Nogueira, R. G. Rossi, R. M. Marcacini, and S. O. Rezende, "A heterogeneous network-based positive and unlabeled learning approach to detect fake news," in *Proc. Braz. Conf. Intell. Syst.*, 2021, pp. 3–18.

[42] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, J. Burstein, C. Doran, and T. Solorio, Eds., 2019, vol. 1, pp. 4171–4186, doi: 10.18653/v1/n19-1423.

[43] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[44] L. Ying, H. Yu, J. Wang, Y. Ji, and S. Qian, "Multi-level multi-modal cross-attention network for fake news detection," *IEEE Access*, vol. 9, pp. 132363–132373, 2021.

[45] O. Russakovsky et al., "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[46] A. Vaswani et al., "Attention is all you need," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, I. Guyon et al., Eds. 2017, vol. 30, pp. 5998–6008. [Online]. Available: http://papers.nips.cc/paper/7181-attention-is-all-you-need

[47] S. Jain, M. White, and P. Radivojac, "Estimating the class prior and posterior from noisy positives and unlabeled data," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2693–2701.
[48] E. A. Platanios, O. Stretcu, G. Neubig, B. Poczos, and T. M. Mitchell, "Competence-based curriculum learning for neural machine translation," in *Proc. Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2019, pp. 1162–1172.
[49] T. Kocmi and O. Bojar, "Curriculum learning and minibatch bucketing in neural machine translation," in *Proc. Recent Adv. Natural Lang. Process.*, 2017, pp. 379–386.
[50] M. Xu, B. Li, G. Niu, B. Han, and M. Sugiyama, "Revisiting sample selection approach to positive-unlabeled learning: Turning unlabeled data into positive rather than negative," 2019, *arXiv:1901.10155*.
[51] A. Zubiaga, M. Liakata, and R. Procter, "Exploiting context for rumour detection in social media," in *Proc. Int. Conf. Social Informat.*, 2017, pp. 109–123.
[52] C. Boididou et al., "Verifying multimedia use at mediaeval 2015," *MediaEval*, vol. 3, no. 3, 2015, Art. no. 7.
[53] D. P. Kingma and J. Ba, "ADAM: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015.
[54] A. Paszke et al., "Automatic differentiation in pytorch," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst. Workshop*, 2017.
[55] J. Deng, W. Dong, R. Socher, L. J. Li, and F. F. Li, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.
[56] Y. Chen, "Convolutional neural network for sentence classification," M.S. thesis, Univ. Waterloo, Waterloo, ON, Canada, 2015.
[57] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, 2006.
[58] S. Singhal, A. Kabra, M. Sharma, R. R. Shah, and P. Kumaraguru, "Spotfake+: A multimodal framework for fake news detection via transfer learning (student abstract)," in *Proc. 34th AAAI Conf. Artif. Intell.*, 2020, pp. 13915–13916.

Jinguang Wang received the B.E. and master's degrees from the School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, China, where he is currently working toward the Ph.D. degree. His research interests include fake mews detection and multimedia computing.

Shengsheng Qian (Member, IEEE) received the B.E. degree from Jilin University, Changchun, China, in 2012, and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2017. He is currently an Associate Professor with the Institute of Automation, Chinese Academy of Sciences. His research interests include social media data mining and social event content analysis.

Jun Hu received the Ph.D. degree from the School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, China, in 2020. He is currently a Research Fellow with the School of Computing, National University of Singapore, Singapore. His research interests include graph deep learning and social multimedia.

Richang Hong (Senior Member, IEEE) received the Ph.D. degree from the University of Science and Technology of China, Hefei, China, in 2008. From 2008 to 2010, he was a Research Fellow with the School of Computing, National University of Singapore, Singapore. He is currently a Professor with the Hefei University of Technology, Hefei, China. He has coauthored more than 70 publications in his research areas, which include multimedia content analysis and social media. He is a Member of the ACM and the Executive Committee Member of the ACM SIGMM China Chapter. He was the recipient of the Best Paper Award from the ACM Multimedia 2010, Best Paper Award from the ACM ICMR 2015, and Honorable Mention of the IEEE T RANSACTIONS ON M ULTIMEDIA Best Paper Award. He was the Technical Program Chair of the MMM 2016 and an Associate Editor for the *Information Sciences* (Elsevier) and *Signal Processing* (Elsevier).

ARTICLES FOR FACULTY MEMBERS

MULTIMODAL FAKE NEWS DETECTION

Title/Author	**QMFND: A quantum multimodal fusion-based fake news detection model for social media / Qu, Z., Meng, Y., Muhammad, G., & Tiwari, P.**
Source	***Information Fusion*** **Volume 104 (2024) 102172 Pages 1-11** **https://doi.org/10.1016/J.INFFUS.2023.102172** **(Database: ScienceDirect)**

Information Fusion 104 (2024) 102172

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

QMFND: A quantum multimodal fusion-based fake news detection model for social media

Zhiguo Qu [a,b,1], Yunyi Meng [b], Ghulam Muhammad [c,*], Prayag Tiwari [d,1]

[a] Engineering Research Center of Digital Forensics, Ministry of Education, Nanjing University of Information Science and Technology, Nanjing, 210044, Jiangsu, China
[b] School of Computer Science, Nanjing University of Information Science and Technology, Nanjing, 210044, Jiangsu, China
[c] Department of Computer Engineering, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia
[d] School of Information Technology, Halmstad University, Halmstad, Sweden

ARTICLE INFO

Dataset link: https://github.com/olivia-2333/QMFND_quantum_fake_news_detection

Keywords:
Fake news detection
Multimodal fusion
Social network
Quantum convolutional neural network

ABSTRACT

Fake news is frequently disseminated through social media, which significantly impacts public perception and individual decision-making. Accurate identification of fake news on social media is usually time-consuming, laborious, and difficult. Although the leveraging of machine learning technologies can facilitate automated authenticity checks, the time-sensitive and voluminous nature of the data brings considerable challenge for fake news detection. To address this issue, this paper proposes a quantum multimodal fusion-based model for fake news detection (QMFND). QMFND integrates the extracted images and textual features, and passes them through a proposed quantum convolutional neural network (QCNN) to obtain discriminative results. By testing QMFND on two social media datasets, Gossip and Politifact, it is proved that its detection performance is equal to or even surpasses that of classical models. The effects of various parameters are further investigated. The QCNN not only has good expressibility and entangling capability but also has good robustness against quantum noise. The code is available at

1. Introduction

With the prevalence of social media, individuals prefer to access news from social media networks rather than from traditional news outlets. However, the rise of social media has also been a double-edged sword, offering easily accessible and convenient short news at the risk of propagating fake news. Fake news can deceive the public potentially through fabricated text, images, audio, video, or mismatched headlines and graphics intended to attract attention. Consequently, it can drastically distort the truth and influence individuals' decisions, leading to illegal profits and the manipulation of public events [1]. For example, during the Covid-19 epidemic, online fake news spread lead to incorrect treatments and false vaccine effects. This could lead the public to make wrong medical decisions and harm their health. Due to the fast-spreading nature of social media and people's need for instant news, it becomes essential to quickly detect the authenticity of the news. In these circumstances, rapid and accurate detection of fake news has become a critical issue that urgently needs to be addressed in social media.

Due to billions of users, the number of news released by them on social media exhibits massive and explosive growth, which causes fake news detection on social media to become more and more challenging. News on current social networks is often presented in a multimodal manner, with text and image combinations being the most common form of multimodal information. Compared to unimodal data, multimodal data requires information fusion techniques to process the data, which is more complicated. Information fusion techniques encompass data fusion [2], feature fusion [3], decision fusion [4], and more [5]. They involve integrating and processing data, knowledge, and features from multiple data sources. They can help to provide a more comprehensive, accurate, and reliable representation of information. For instance, in medical diagnosis and treatment, combining different medical images (such as MRI and CT scans of a patient) can result in a more comprehensive and accurate diagnosis.

In view of this, researchers have explored the application of machine learning (ML) techniques, including graph neural networks (GNN) [6], natural language processing (NLP) [7,8], and generative adversarial networks [9], to detect fake news automatically. However,

* Corresponding author.
E-mail addresses: 002359@nuist.edu.cn (Z. Qu), 20211221061@nuist.edu.cn (Y. Meng), ghulam@ksu.edu.sa (G. Muhammad), prayag.tiwari@ieee.org (P. Tiwari).

[1] Zhiguo Qu and Prayag Tiwari contribute equally and share co-first authorship.

https://doi.org/10.1016/j.inffus.2023.102172
Received 30 June 2023; Received in revised form 28 November 2023; Accepted 29 November 2023
Available online 30 November 2023

these methods still suffers to process ambigous features and fail to quantify level of uncertainity. Many studies have shown that quantum machine learning (QML) based on quantum computing principles [10] can be used to perform ML tasks [11–13] such as pattern matching [11], binary classification [12], support vector machines [13], and others. Quantum convolutional neural network (QCNN), as one of the emerging research branches of QML, has been widely used in multiple application scenarios, such as smart healthcare [14] intelligent transportation [15], etc.

In summary, the main contributions can be presented, as follows.

- A quantum multimodal fusion-based fake news detection (QM FND) model is proposed for social media networks. The model integrates quantum encoding, and quantum convolutional neural networks (QCNNs) to process high-dimensional data processing.
- To achieve lower complexity and better accuracy, multimodal features are encoded into a VQC (variational quantum circuit) through amplitude encoding.
- QCNN is proposed for efficient training of quantum multimodal data. The QCNN circuit not only has excellent expressibility and entangling capability, but also good robustness against quantum noise.
- QMFND achieved detection accuracies of up to 87.9% and 84.6% on the Gossip and Politifact datasets, respectively. Additionally, QMFND can alleviate the barren plateau phenomenon.

The remaining part of this paper is organized as follows. Section 2 reviews the related work on fake news detection and quantum neural networks (QNNs). Section 3 presents the proposed QMFND model which comprises data pre-processing, data encoding and QCNN training. Section 4 provides the experimental results and analysis, datasets, runtime environment, baselines, and performance comparison. Finally, the conclusions are given in Section 5.

2. Related work

2.1. Traditional fake news detection models

Traditional fake news detection primarily includes based on article information, social context, and their combinations with external knowledge.

The detection based on article information includes text-based, image-based, and multimodal analysis. In 2019, Ma et al. [16] improved classification performance by using adversarial training to generate more data. Singhal et al. [17,18] combined visual and textual information using the visual geometry group 19 (VGG-19), bidirectional encoder representations from transformers (BERT), and the generalized autoregressive pretraining method of XLNet. In 2022, Segura [19] used unimodal and multimodal approaches for a more detailed classification of fake news. In the same year, Jayakody and Mohammad [20] proposed a system for fake news detection, employing federated learning and blockchain methods to address resource allocation and privacy concerns. In 2023, Luvembe et al. [21] proposed a method that achieved high accuracy, using dual sentiment features.

In fake news detection based on social context, user trustworthiness and news dissemination are commonly used criteria. For instance, in 2020, Nguyen et al. [22] proposed a novel graphical social context representation and learning framework called FANG to detect fake news. FANG captures social context well and is robust despite limited training data. In 2022, based on Transformer, Raza et al. [23] proposed an effective labeling technique in a fake news detection framework to address the lack of labeled data in training models.

GNNs and attention mechanisms are also used to detect fake news by integrating external information. In 2021, Li et al. [24] constructed a star-shaped knowledge graph for factual evidence and news content, using GNNs to identify fake news. In the same year, Hu et al. [25] presented a novel end-to-end GNN model called CompareNet to compare news with a knowledge base of entities for fake news detection.

Deep learning models, particularly CNNs, GNNs, recurrent neural networks (RNNs), and Transformers, have been widely used for fake news detection. These models can automatically learn text and image data patterns and help identify fake news. Moreover, the emergence of pre-trained language models (such as BERT, GPT, etc.) has changed the field of natural language processing, providing new opportunities for fake news detection.

Despite the increasing diversity and sophistication of fake news detection methods, multimodal fake news detection is still the focus of research. Fake news exists not only in text but also in the dissemination of images, videos, and audio, simultaneously. Therefore, the need for multimodal fake news detection, which involves the simultaneous analysis of various types of media content, is growing. In addition, the enormous computational power of quantum computing has not yet been fully investigated in fake news detection.

2.2. Quantum neural networks

QNNs leverage the properties of quantum computing to enhance the performance of neural networks in some tasks. For example, in 2021, Narottama et al. [26] proposed a reinforcement learning-inspired quantum neural network (RL-QNN) to enhance resource allocation efficiency in wireless communication. In 2023, to capture the complexity and uncertainty of sarcasm and sentiment elements in human language, Tiwari et al. [27] proposed a quantum fuzzy neural network (QFNN) with a multi-task learning capability for multimodal sarcasm and sentiment detection. The algorithm combines the fuzzy system and QNN to enhance the expressiveness of sentiment and sarcastic features and exhibits superior performance compared to various existing algorithms.

Quantum circuits are used as convolutional kernels in QCNNs with promising feature extraction capability that may match or even surpass classical convolutional kernels [28]. As shown in Fig. 1, the quantum circuit of the convolutional kernel consists of many unitary operators. Similar to the classical convolution kernel, the quantum convolution kernel passes through the entire image according to the size of a given receptive field and step size. However, the quantum convolution kernel first encodes the input into a quantum state consisting of a number of qubits. Gate operations are then performed in a VQC containing entangled modules with trainable weights. Finally, measurement is performed on these qubits to obtain output. Also, the pooling layer of the QCNN reduces the circuit dimensionality, so as to lower the number of qubits while preserving as much information as possible from previously learned values.

At present, QCNNs are widely used in various application fields such as image classification [29,30], traffic prediction [15], speech recognition [31] and medical diagnosis [14,32,33]. In 2020, Li et al. [29] proposed a quantum deep convolutional neural network (QDCNN) model for image recognition based on a parameterized quantum circuit. Network complexity analysis indicates that the proposed model offers exponential acceleration compared to classical models. In 2021, Yang et al. [31] addressed privacy preservation issues in speech recognition by using a decentralized feature extraction approach that employs a QCNN. In 2022, Qu et al. [15] proposed a novel algorithm using a quantum graph convolutional network to simultaneously capture the temporal and spatial features of traffic data for traffic congestion prediction. In 2022, Ovalle-Magallanes et al. [14] used quantum computing in coronary artery X-ray angiography to construct a hybrid neural network. This network employs a mixed-transfer learning approach, in which a quantum network enhances the performance of a pre-trained classical network. In 2023, Qu et al. [32] introduced a QNN-based multimodal fusion (QNMF) system for intelligent medical diagnosis. QNMF can process multimodal medical data transmitted by Internet of Things (IoT) devices, fuse data from different modalities,

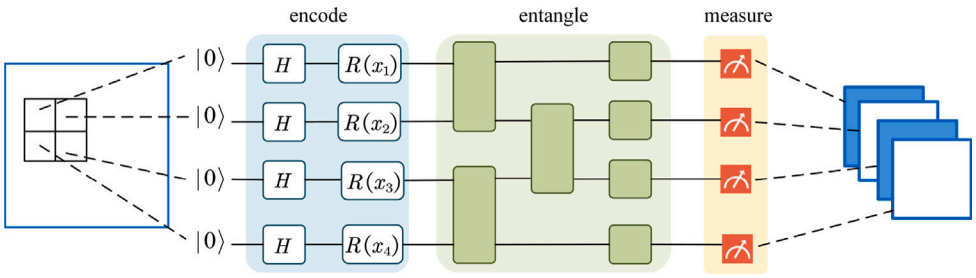


Fig. 1. The quantum circuit of QCNN convolution kernel. $(x_1, x_2, x_3, x_4$ are the parameters. $R(x_i)\,(1 \le i \le 4)$ are the rotation gates. H is the Hadamard gate, and $H = \frac{1}{\sqrt{2}}(|0\rangle\langle 0| + |0\rangle\langle 1| + |1\rangle\langle 0| - |1\rangle\langle 1|)$.

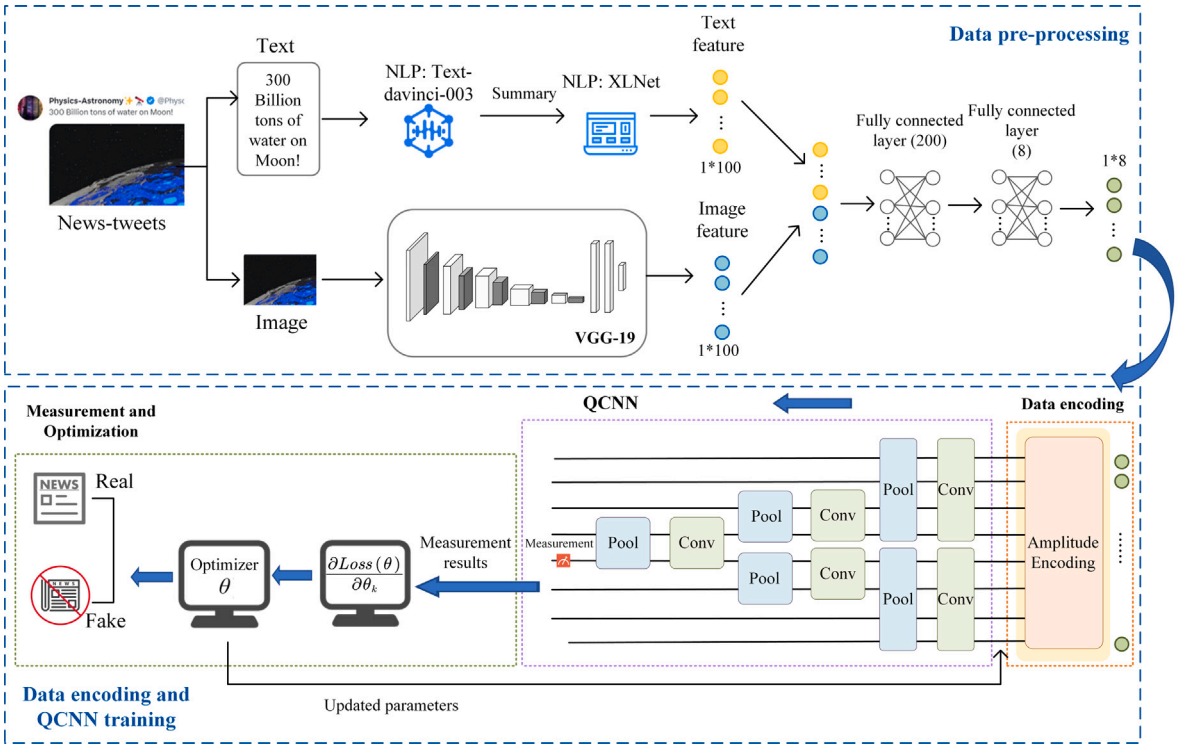


Fig. 2. The flowchart of QMFND.

and enhance the performance of intelligent diagnosis. In 2023, Chen et al. [30] proposed two scale-inspired local feature extraction methods based on QCNN for binary image classification.

The advantages of using QCNNs can be summarized as follows: (1) QCNNs can leverage entanglement, superposition, and interference to process complex tasks. (2) For datasets with specific complex features, QNNs have a classification advantage over classical neural networks [34]. (3) Furthermore, models trained using quantum circuits can exhibit higher performance and superior generalization capabilities because of quantum entanglement [35].

3. The proposed QMFND model

The QMFND model is divided into two main phases. The first phase is the data pre-processing, which separates raw classical data into text and image, and then performs feature fusion. The second phase involves data encoding and QCNN training. The data are fed into the quantum circuit, and the measurement results are used to generate the model prediction results. Then, the output is used to compute the loss function in a classical computer environment, followed by the parameter optimization. The flowchart of QMFND is shown as Fig. 2 and the details are presented as the follow.

3.1. Data pre-processing

3.1.1. Text pre-processing

For news, usually, not all the textual content is strictly relevant to the expressed topic. For example, a news piece may be about the birth of a celebrity's daughter but the text may contain substantial passages describing the celebrity's history.

Quantum natural language processing (QNLP) [36] aims to design and implement NLP models that run on quantum hardware. The release of the Lambeq toolkit by Cambridge Quantum in 2019 marked the first high-level Python library for QNLP. However, our attempts of utilizing Lambeq library revealed that it could only process brief expressions and was time-consuming. QNLP's capability to identify news segments is limited, and the circuit cost is excessively high. Therefore, the classical NLP technique is used to extract text features. The text-davinci-003 model, introduced by OpenAI, is an advanced language model trained to understand and generate human-like text and has been widely recognized for its text-processing capability. Therefore, we uses the API of the OpenAI model to extract the textual content summaries. By calling the API, a non-local model can accurately summarize a relatively lengthy news article by condensing it into a few sentences or phrases. This enhances the training efficiency of the model. Then, news

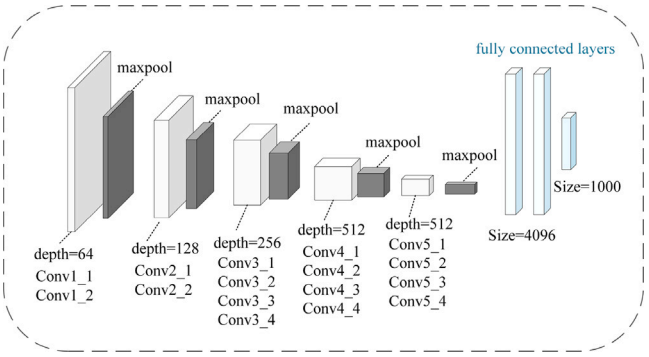


Fig. 3. The structure diagram of VGG-19.

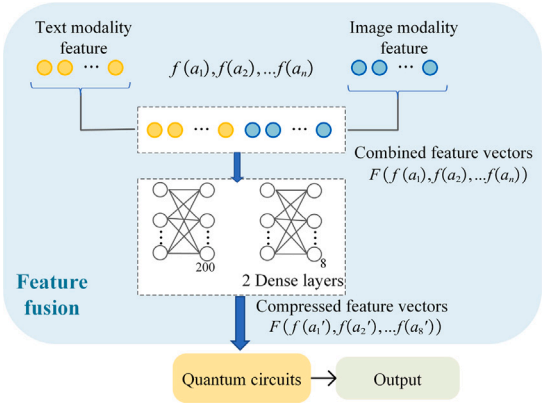


Fig. 4. The flowchart of multimodal feature fusion.

summaries are embedded using a pre-trained XLNet model from the Hugging Face Transformers library to obtain a matrix of eigenvalues. Finally, fully connected layer is used.

3.1.2. Image pre-processing

For news, image information usually plays a vital role in conveying its theme. A pre-trained VGG-19 model is used to extract the image features. VGG-19 is a deep CNN architecture initially proposed by a research team at the University of Oxford in 2014 [37]. As depicted in Fig. 3, VGG-19's architecture comprises 16 convolutional layers and three fully connected layers. The number of convolutional layers in the five VGG blocks is (2, 2, 4, 4, 4), along with three fully connected layers, resulting in 19 parameter layers. VGG-19 is used in the image pre-processing phase because it is typically pre-trained using large-scale image datasets. Thus, its weights already contain a vast amount of image knowledge. This makes VGG-19 highly versatile for various computer vision tasks. Additionally, the VGG-19 network is suitable for extracting rich feature representations. These factors make VGG-19 one of the most popular deep CNN models currently in use.

Considering the limitations of quantum resources, the dimension of image feature vectors is compressed and flattened to 100.

3.1.3. Feature fusion

Information fusion techniques are adept at managing multimodal data. Based on different fusion times, information fusion can be categorized into early, mid, and late fusion. An early feature fusion technique is used, where multiple layers of features are fused at the input layer. Compared with mid fusion and late fusion, early fusion can better utilize complementary information among various modalities, and the extracted features can be effectively used for training the QCNN.

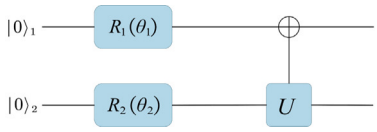


Fig. 5. The quantum circuit that encodes a four-dimensional vector on two qubits.

Let x_i represents the i^{th} data source, and $f(x_i)$ denotes the extracted features from image or text modalities. QMFND concatenates the extracted text and image features, denoted as $F\left(f\left(x_1\right), f\left(x_2\right), \ldots, f\left(x_n\right)\right)$. Through two fully connected layers with the sizes of 200 and 8, the compressed fused multimodal features, denoted as $F\left(f\left(x_1'\right), f\left(x_2'\right), \ldots, f\left(x_8'\right)\right)$, were obtained. A flowchart of multimodal feature fusion is illustrated in Fig. 4. The extracted fused multimodal information will be used at the next encoding step.

3.2. Data encoding and QCNN training

3.2.1. Data encoding

The processed classical data cannot be used directly for training the quantum circuit. Consequently, classical input data must be encoded into a quantum state for further processing in VQC. Given the constraints of quantum resources, circuit simplification costs must be considered in experiments. Reducing the number of qubits is one of the widely accepted approaches. First, amplitude encoding is used to encode classical data into quantum states. It is an effective method for reducing the number of qubits. Specifically, $x = \left(x_1, x_2 \ldots x_N\right)$ represents a normalized vector. Then, the vector can be represented by using the amplitude encoding as follows:

$$|x\rangle = \sum_{i}^{N} x_i |i\rangle, i = 1, 2, \ldots, N. \tag{1}$$

For example, a 4-dimensional classical vector $x = (\alpha, \beta, \gamma, \eta)$ needs to be encoded into the VQC and the initial quantum state is $|00\rangle_{12}$. Applying the rotation gate $R_1\left(\theta_1\right)$ on the first qubit, we get

$$|0\rangle_1 \rightarrow \left(\sqrt{|\alpha|^2 + |\beta|^2}\right)|0\rangle + \left(\sqrt{|\gamma|^2 + |\eta|^2}\right)|1\rangle. \tag{2}$$

Then, the second qubit is also rotated through a rotation gate $R_2\left(\theta_2\right)$:

$$|0\rangle_2 \rightarrow \frac{\alpha|0\rangle + \beta|1\rangle}{\sqrt{|\alpha|^2 + |\beta|^2}}. \tag{3}$$

It can be obtained that:

$$\begin{aligned} |00\rangle_{12} \rightarrow & \left(\sqrt{|\alpha|^2 + |\beta|^2}\right)|0\rangle \frac{\alpha|0\rangle + \beta|1\rangle}{\sqrt{|\alpha|^2 + |\beta|^2}} \\ & + \left(\sqrt{|\gamma|^2 + |\eta|^2}\right)|1\rangle \frac{\alpha|0\rangle + \beta|1\rangle}{\sqrt{|\alpha|^2 + |\beta|^2}}. \end{aligned} \tag{4}$$

Let us assume that there exists a unitary gate U satisfying with:

$$U\left(\frac{\alpha|0\rangle + \beta|1\rangle}{\sqrt{|\alpha|^2 + |\beta|^2}}\right) = \frac{\gamma|0\rangle + \eta|1\rangle}{\sqrt{|\gamma|^2 + |\eta|^2}}. \tag{5}$$

Subsequently, the circuit passes through one more controlled U-gate with the control bit as the first bit. The circuit for amplitude encoding is illustrated in Fig. 5.

After amplitude encoding, it can be obtained that x is encoded into a quantum circuit of the form:

$$\left(\sqrt{|\alpha|^2 + |\beta|^2}\right)|0\rangle \frac{\alpha|0\rangle + \beta|1\rangle}{\sqrt{|\alpha|^2 + |\beta|^2}} + \left(\sqrt{|\gamma|^2 + |\eta|^2}\right)|1\rangle \frac{\gamma|0\rangle + \eta|1\rangle}{\sqrt{|\gamma|^2 + |\eta|^2}}$$

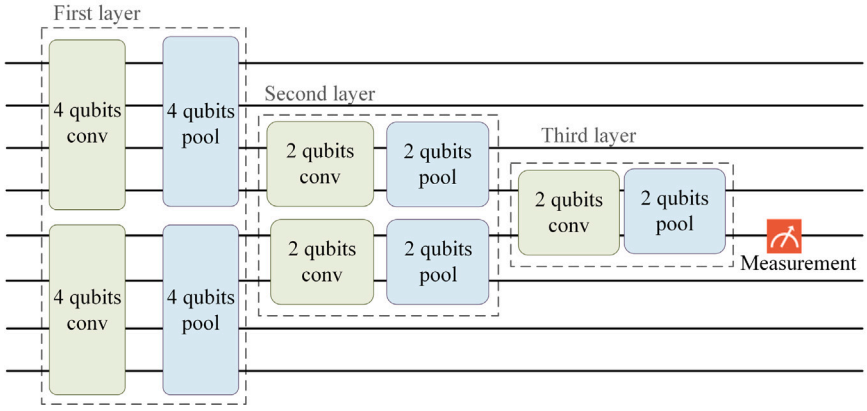


Fig. 6. The QCNN circuit in QMFND (conv represents the convolutional layer and pool represents the pooling layer).

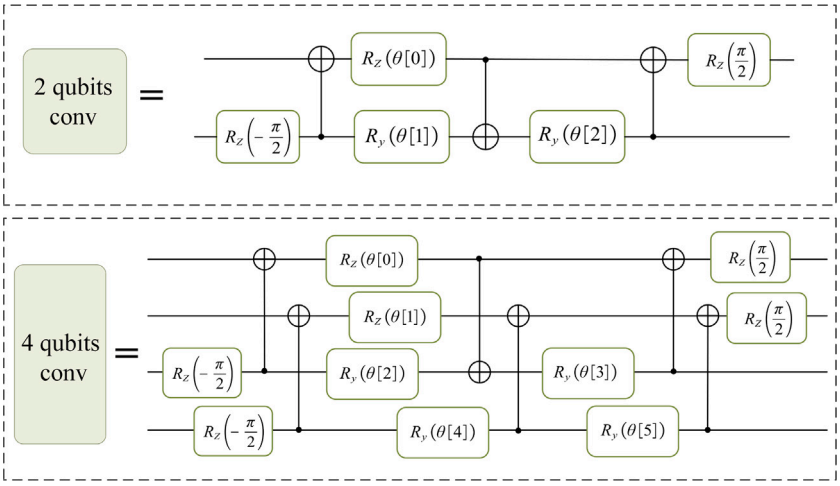


Fig. 7. The quantum circuit of the convolutional layer ($\theta[i]$ is the i^{th} parameter of θ).

$$= \alpha|00\rangle + \beta|01\rangle + \gamma|10\rangle + \eta|11\rangle. \quad (6)$$

It can be seen that only $\lceil \log_2 n \rceil$ qubits are required to represent an n-dimensional vector. In QMFND, eight qubits are used for feature representation.

3.2.2. Quantum circuit of the QCNN

The next step is to construct the QCNN circuit. In this paper, a QCNN is proposed for efficient training of quantum data. The QCNN circuit constructed in this paper is divided into three layers, as shown in Fig. 6. The first layer performs quantum convolution and pooling operations on the first four and last four qubits, respectively. The second layer performs convolution and pooling operations on the third and fourth, and fifth and sixth qubits, respectively. The third layer performs convolution and pooling operations on the fourth and fifth qubits. The quantum measurements are performed on the fifth qubit to obtain the measurement results. The circuits of convolutional and pooling layers are shown in Figs. 7 and 8, respectively. $R_y(\theta)$ and $R_z(\theta)$ are the rotation gates: $R_y(\theta) = \begin{pmatrix} \cos\frac{\theta}{2} & -\sin\frac{\theta}{2} \\ -\sin\frac{\theta}{2} & \cos\frac{\theta}{2} \end{pmatrix}$, $R_z(\theta) = \begin{pmatrix} e^{-i\theta/2} & 0 \\ 0 & e^{i\theta/2} \end{pmatrix}$. By running quantum convolution and pooling operations in the three layers, the network can extract discriminative features from the input qubits that help detect fake news. The QCNN makes full use of entanglement operations and rotation gates to make the circuits have good expressibility and entangling capability. This helps to represent the correlation between fused features more efficiently so that QMFND can obtain better detection performance.

QCNN discards certain qubits in the circuit, and the discarded qubits are no longer involved in operations and measurements. This successfully compresses the dimensionality of the data and reduces the cost of subsequent circuits.

3.2.3. Measurement and optimization

Multiple measurements are required to improve the performance. The predicted results are the expectation of qubits obtained through several measurements. Cross entropy loss function is used to assess model loss. In addition, we implemented constrained optimization by linear approximation (COBYLA), which is gradient-free. COBYLA ensures efficient optimization of the model parameters, to enhance the optimal performance and reliability. A flowchart of the measurement and optimization is shown as Fig. 9, and the specific steps are presented as below.

Let assume classical inputs are $\{x_i, y_i\}$. x_i is the i^{th} classical data and y_i is the i^{th} label. The real news is labeled as 1 and the fake news is labeled as 0. Set the initial states of the quantum circuit as $|0\rangle^{\otimes n}$ where n is the number of qubits. After the encoding layer, $|\psi\rangle_i$ is obtained.

$$|\psi\rangle_i = U_e|x_i\rangle, \quad (7)$$

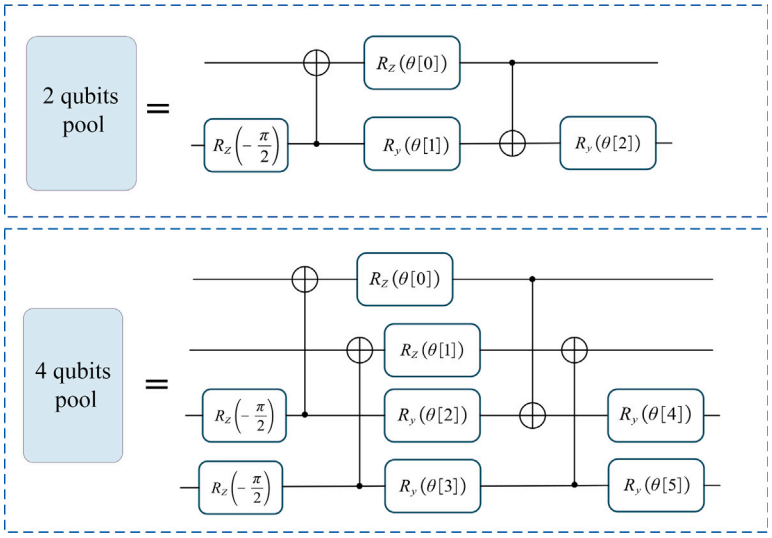


Fig. 8. The quantum circuit of the pooling layer.

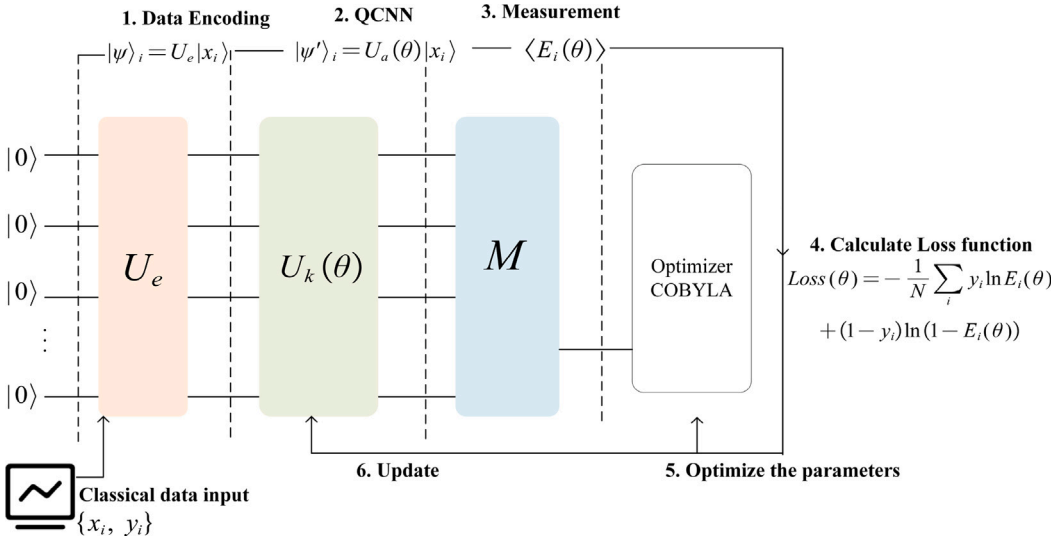


Fig. 9. The flowchart of measurement and optimization.

where U_e denotes the unitary operator of the encoding layer (without parameters). The unitary operator of the QCNN circuit is set as $U_a(\theta)$:

$$U_a(\theta) = \prod_k U_k\left(\theta_k\right). \tag{8}$$

where θ is the parameter to be optimized, θ_k is the k^{th} θ, and U_k is the unitary operators associated with θ_k. After the QCNN layer, the state becomes

$$\left|\psi'\right\rangle_i = U_a(\theta)|\psi\rangle_i, \tag{9}$$

which is measured by using the measurement operator M. $E_i(\theta)$ is the expectation of the predicted values.

$$M = I^{\otimes n-1} \otimes Z, E_i(\theta) = \left\langle\psi'|M|\psi'\right\rangle_i. \tag{10}$$

Then, using cross entropy as the loss function, the equation is given as follows:

$$Loss(\theta) = -\frac{1}{N}\sum_i y_i \ln E_i(\theta) + \left(1-y_i\right)\ln\left(1-E_i(\theta)\right). \tag{11}$$

Here, N is the number of samples in a batch. For the parameter k to be optimized, the gradient of θ_k is

$$\frac{\partial Loss(\theta)}{\partial\theta_k} = -\frac{1}{N}\sum_i \frac{y_i}{E_i(\theta)}\frac{\partial E_i(\theta)}{\partial\theta_k} - \frac{1-y_i}{1-E_i(\theta)}\frac{\partial E_i(\theta)}{\partial\theta_k} \tag{12}$$

It is not difficult to determine whether only the partial derivative of E_i with respect to θ_k is required. Here, E_i is denoted by E. Taking the partial derivative of E with respect to θ_k, we can obtain

$$\begin{aligned}\frac{\partial E(\theta)}{\partial\theta_k} &= \frac{\partial\langle\psi'|M|\psi'\rangle}{\partial\theta_k} = \frac{\partial\left\langle\psi\left|U_a^\dagger M U_a\right|\psi\right\rangle}{\partial\theta_k} \\ &= \left\langle\psi\left|U_a^\dagger M\frac{\partial U_a}{\partial\theta_k}\right|\psi\right\rangle + H.c.\end{aligned} \tag{13}$$

Here, $H.c$ is the complex conjugate of $\left\langle\psi\left|U_a^\dagger M\frac{\partial U_a}{\partial\theta_k}\right|\psi\right\rangle$. As the quantum gates in the circuit used in this paper are all Pauli rotation gates,

Table 1
Dataset descriptions.

	Gossip	Politifact
Description	Genuine news data and some instances of false or misleading gossip.	Legitimate and fabricated news data.
Training Set	10 010	381
Real: Fake	2036:7974	135:246
Test Set	1000	100

Table 2
The hyperparameter setting in experiments.

Qubits in QCNN	Learning rate	Batch size	Epoch	Measurement
8	0.001	32	100	Z Gate

according to the conclusions drawn by Schuld et al. [38], there is

$$\begin{aligned}&\left\langle \psi \left| U_a^{\dagger} M \frac{\partial U_a}{\partial \theta_k} \right| \psi \right\rangle + H.c. \\ &= \frac{1}{2}\left(\left\langle \psi \left| U_a^{\dagger}\left(\theta_k + \frac{\pi}{2}\right) M U_a\left(\theta_k + \frac{\pi}{2}\right) \right| \psi \right\rangle \right. \\ &\left. - \left\langle \psi \left| U_a^{\dagger}\left(\theta_k - \frac{\pi}{2}\right) M U_a\left(\theta_k - \frac{\pi}{2}\right) \right| \psi \right\rangle\right)\end{aligned} \tag{14}$$

The method for computing the gradient is called the parameter-shift rule. In this way, the partial derivative of each parameter is obtained and fed into the corresponding optimizer for optimization.

4. Experimental results

4.1. Datasets

This study uses multimodal fake news datasets, namely, Gossip [39] and Politifact [39,40], both collected from the social media platform (Twitter). Each dataset incorporates both text and image. Initially, in each multimodal news, we select its first image (cover image) to participate in the subsequent training. This is because the first image of news usually can represent all the images in this news and carries its most important information. The data in datasets are labeled as real or fake news and divided into training and testing sets, with the validation set extracted directly from the training set. For clarity, a description of two datasets is provided in Table 1.

4.2. Experimental setting

The experimental environment is established on an x86 platform. Data embedding is conducted on a graphics processing unit (NVIDIA RTX3080 10G), and the QCNN training is performed on a central processing unit (Intel Core 12400F).

The hyperparameters used in the experiments are listed in Table 2. Due to the limitation of quantum resources, and considering the effects of the representation capability and the barren plateau phenomenon on the model performance, we use 8 qubits for QCNN representation. Also, the learning rate is set to 0.001, batch size to 32, and epoch to 100. To measure the probability of a quantum state result being 1, the Z measurement operator is employed on the quantum circuit measurements.

When extracting text summaries, the maximum length of each sentence in the news is limited to 30 tokens. A quantum circuit is constructed and trained by using the Pennylane library, with the initial state of the quantum circuit set to $|0\rangle$. Quantum circuit is trained in the environment with simulated quantum noise.

Table 3
The performance comparison of QMFND with other fake news detection models.

Dataset	Model	Accuracy	Recall	Precision	TNR	F1
Gossip	BERT	0.871	0.914	0.924	0.897	0.919
	XLNet	0.884	0.943	0.917	0.897	0.929
	MCAN	0.869	0.806	0.890	0.877	0.846
	MCAN-A	0.851	0.859	0.877	0.859	0.824
	SpotFake	0.789	0.723	0.815	0.788	0.766
	SpotFake+	0.839	0.799	0.853	0.842	0.825
	CNN	0.806	0.643	0.882	0.812	0.744
	QMFND	**0.879**	**0.958**	**0.899**	**0.882**	**0.928**
Politifact	BERT	0.843	0.696	0.904	0.864	0.786
	XLNet	0.847	0.704	0.905	0.867	0.792
	MCAN	0.846	0.829	0.851	0.847	0.84
	MCAN-A	0.809	0.758	0.827	0.81	0.792
	SpotFake	0.779	0.693	0.815	0.775	0.749
	SpotFake+	0.789	0.753	0.803	0.789	0.777
	CNN	0.776	0.614	0.857	0.763	0.715
	QMFND	**0.846**	**0.853**	**0.927**	**0.904**	**0.888**

4.3. Evaluation metrics

Evaluation metrics are adoped as follows: Accuracy, Recall, Precision, True Negative Rate (TNR), and F1, as expressed in Eq. (15). Among them, TP (True Positive) indicates that it correctly predicts the positive class when the sample is actually positive. TN (True Negative) indicates that it correctly predicts the negative class when the sample is actually negative. FP (False Positive) indicates that it incorrectly predicts the positive class when the sample is actually negative. FN (False Negative) indicates that it incorrectly predicts the negative class when the sample is actually positive.

$$\begin{aligned}&Accuracy = \frac{TN + TP}{TP + TN + FP + FN}, Recall = \frac{TP}{TP + FN}, \\ &Precision = \frac{TP}{TP + FP}, \\ &F1 = \frac{2 * Precision * Recall}{Precision + Recall}, TNR = \frac{TN}{TN + FN}.\end{aligned} \tag{15}$$

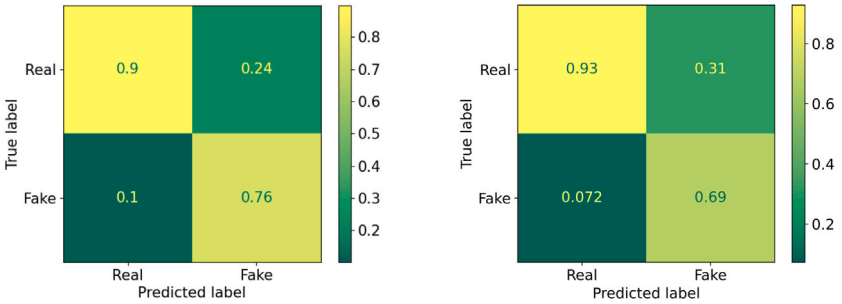
4.4. Baselines

To evaluate the effectiveness of the QMFND model, seven baselines are considered for comparison. These baselines include:

- XLNet: XLNet model [41] is used to verify the authenticity of unimodal (text) data.
- BERT: BERT model [42] to verify the authenticity of unimodal (text) data.
- MCAN: It is a multimodal fake news detection model using a co-attention network, enabling better fusion of textual and visual features for fake news detection [43].
- MCAN-A: It is a model similar to MCAN but without the part of fusing multimodal features. Spatial-domain features, frequency-domain features, and text features are simply connected.
- SpotFake: Multimodal model for detecting fake news without considering additional subtasks such as event discrimination [17].
- SpotFake+: It leverages transfer learning to capture semantic and contextual information from news articles and related images, so as to improve the accuracy of fake news detection [18].
- CNN: It uses the same pre-processing methods as QMFND (XLNet, VGG-19, etc.) and uses a classical CNN for classification tasks.

4.5. Performance analysis

The confusion matrices of the QMFND performance on two datasets are shown in Fig. 10. The comparisons between QMFND and other baselines are presented in Table 3. As shown in the Table 3, QMFND achieves a detection accuracy of 87.9% and 84.6% on the Gossip and Politifact datasets, respectively, which is quite promising. The accuracy



(a) The confusion matrix of QMFND on Gossip.

(b) The confusion matrix of QMFND on Politifact.

Fig. 10. The confusion matrices of QMFND on the two datasets.

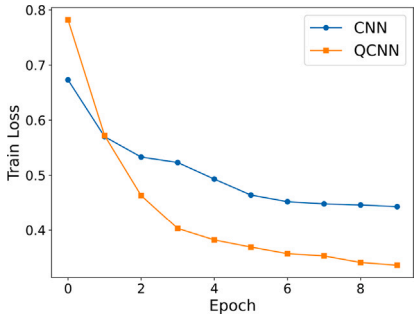


Fig. 11. The comparison of loss function between QCNN and CNN.

of the QMFND model is 0.5% lower than XLNet on the Gossip dataset and 0.1% lower than XLNet on the Politifact dataset. However, on the Politifact dataset, QMFND not only outperforms XLNet in other performance metrics but also outperforms the other six baseline models on both datasets.

In addition, with the integration of a QCNN, QMFND outperforms the structurally equivalent CNN, with a detection accuracy about 7% higher than that of CNN. Moreover, it performs better than most of the baselines. It can be concluded that the improvement in detection performance represents that QCNNs are sharper and more accurate in representing image features. They can capture subtle details that may have been missed by traditional CNNs, thus improving the classification performance.

QCNN is also more efficient. As shown in Fig. 11, the loss function of QCNN decreases faster. This indicates that QMFND utilizes the training data more efficiently within a given number of training iterations, helping to reduce training time and computational cost. Considering the need for efficient processing in big data era, using QCNN implies quicker identification of fake news.

4.6. Parameter effects on the model performance

4.6.1. Effect of different data modalities on detection performance

The performances of unimodal and multimodal on QMFND are compared in this section. As shown in Table 4, on both datasets, QMFND's detection performance is best for text, followed by multimodal, and performs worst on images. In addition, the detection accuracies of text and images in multimodal scenarios are slightly lower than those of unimodal text-only scenarios by 0.5% and 3.8%, respectively, on two datasets.

It is easy to know that, textual information plays a key role in detecting fake news as compared to image information. It can precisely convey the entire content of the news. On the other hand, images alone may not be helpful enough to detect fake news. Although images can present rich information, the key information such as time and location of the events in news, as well as the participants, is usually impossible to be accurately obtained by only images. Therefore, using multimodal data processing methods for detecting fake news becomes more practical and effective by nature.

Table 4
The effect of modal type on detection performance.

DataSet	Modal	Accuracy	Recall	Precision	TNR	F1
Gossip	Image	0.802	0.929	0.842	0.749	0.883
	Text	0.884	0.942	0.917	0.897	0.929
	Multimodal	**0.879**	**0.958**	**0.899**	**0.882**	**0.928**
Politifact	Image	0.692	0.733	0.821	0.719	0.774
	Text	0.884	0.973	0.879	0.882	0.924
	Multimodal	**0.846**	**0.853**	**0.927**	**0.904**	**0.888**

Table 5
The effect of real and fake data proportion on detection performance. "R" represents real news, and "F" represents fake news.

	DataSet	Accuracy	Recall	Precision	TNR	F1
R: F=1:1	Gossip	0.849	0.649	0.940	0.892	0.768
	Politifact	0.840	0.583	0.978	0.932	0.73
R: F=9:1	Gossip	0.879	0.958	0.899	0.882	0.928
	Politifact	0.846	0.853	0.927	0.904	0.888

4.6.2. Effect of data proportion on detection performance

This section explores the effect of the proportion of data labeled as "real" and "fake" in the training dataset on the detection results.

As shown in Table 5, for the real to fake data proportion of 1:1, the accuracies on the Gossip and Politifact datasets are 3% and 0.06% lower compared to those of the 9:1 proportion, respectively. However, in terms of the recall and F1 score, the performance is significantly lower when the real-to-fake data proportion is 1:1 versus 9:1. Specifically, for both datasets, recall is lower by 30.9% and 27%, and the F1 score is lower by 16% and 15.8% for the 1:1 proportion compared to the 9:1 proportion, respectively. This indicates that although the model accuracy on data with a 9:1 real-to-fake proportion does not show an obvious advantage, other performance metrics exhibit clear advantages. This can be explained by the fact that the model focuses more on those classes with a higher percentage of samples during the training process, and learns relatively little about the lower percentage classes. In practical applications, it is necessary to weigh these evaluation metrics according to the specific tasks and goals to deal with the proportion imbalance in datasets.

4.6.3. Effect of the number of qubits and barren plateau phenomenon

The number of qubits represented has an effect on detection performance and the barren plateau phenomenon. With the limited resources available, we compared the performance of QMFND using 4, 8, and 16 qubits for representation. As shown in Table 6, the performance is best when using 8 qubits for representation, followed by 4 qubits, and finally 16 qubits. On both datasets, the accuracy using 8 qubits is 1%

Table 6
The effect of the number of qubits on the detection performance.

DataSet	Qubits	Accuracy	Recall	Precision	TNR	F1
Gossip	4	0.869	0.781	0.9	0.882	0.836
	16	0.827	0.719	0.869	0.835	0.787
	8	**0.879**	**0.958**	**0.899**	**0.882**	**0.928**
Politifact	4	0.836	0.753	0.868	0.844	0.806
	16	0.797	0.71	0.832	0.797	0.766
	8	**0.846**	**0.853**	**0.927**	**0.904**	**0.888**

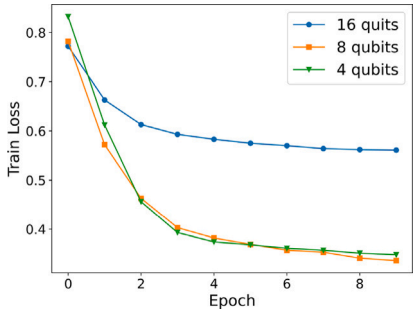


Fig. 12. The loss functions for 4, 8, and 16 qubits quantum circuits.

higher than when using 4 qubits, and with 8 qubits, the recall is higher by 17.7% and 10% compared to that of 4 and 16 qubits, respectively. This indicates that using 8 qubits is the most suitable option for our detection.

The "barren plateau" refers to a phenomenon that the training of QNN becomes inefficient when the number of qubits in quantum computing is large. The objective function tends to flatten, resulting in excessively long training times or training failures [44].

In initial experiments, 16 qubits are used in QCNN. During training, the gradient of the QCNN descends very slowly. As depicted in Fig. 12, a 16-qubit circuit underperforms the 8-qubit or 4-qubit circuit in terms of loss function reduction over the same number of iterations. This is possibly due to the effect of quantum noise. In a limited simulation environment, the noise may make the gradient information lost as the number of qubits increases, making training more difficult. Therefore, to mitigate the effect of the barren plateau phenomenon, 8 qubits are used for representation in QMFND.

4.7. The expressibility and entangling capability of QCNN

The structures of VQCs in QCNNs play a crucial role in the performance of quantum models. Expressibility and entangling capability can be used to assess the quality of VQCs [45].

Expressibility refers to the ability of a circuit to generate states representing the Hilbert space. The quantification of expressibility is defined as the deviation of the distribution of states generated by the set of Haar random states from the states obtained by uniformly sampling the VQC parameters. $\hat{P}_{\mathrm{VQC}}(F;\theta)$ represents the set of states obtained by uniformly sampling parameters. $P_{\mathrm{Haar}}(F)$ represents the set of states obtained from a uniform distribution. $P_{\mathrm{Haar}}(F) = (N-1)(1-F)^{N-2}$, where F corresponds to fidelity, and N is the dimension of the Hilbert space. D_{KL} is for calculating the KL divergence between two sets [46]. Then, the expressibility of a VQC can be expressed as

$$\mathrm{Expr} = D_{\mathrm{KL}}\left(\hat{P}_{\mathrm{VQC}}(F;\theta) \parallel P_{\mathrm{Haar}}(F)\right). \tag{16}$$

In VQCs, a higher degree of entanglement in the circuit implies a better representation of the solution spaces for tasks such as data classification, as well as better capture of quantum data correlations. The Meyer–Wallach (MW) measurement is used to quantify the entangling capability of the VQC. For a given VQC, this value is estimated by sampling the circuit parameters and calculating the sample average

Table 7
The expressibility and entangling capability of the designed QCNN circuit.

	Conv1	Conv2	Pool1	Pool2	QCNN
Entangling capability	0.498	0.497	0.5	0.25	**0.819**
Expressibility	2.311	1.472	2.31	1.479	**0.00015**

of the MW measurements of the output states. Here, θ represents the circuit parameters. $|\vec{\theta}|$ is the number of test parameters, n is the number of parameters, and ρ_k is the partial trace. The MW measurement formulation is expressed as follows:

$$Q^{MW} = \frac{2}{|\vec{\theta}|}\sum_{\theta_i\in\vec{\theta}}\left(1-\frac{1}{n}\sum_{k=1}^{n}\mathrm{Tr}\left(\rho_k^2\left(\theta_i\right)\right)\right). \tag{17}$$

The expressibility and entangling capability of the designed QCNN circuit are listed in Table 7. The entangling capability of quantum circuits is from 0 to 1, with the higher value being preferable. On the contrary, for quantum circuits, lower values of expressibility are desirable. It is evident that the VQC of QCNN in this paper exhibits excellent expressibility and entangling capability.

4.8. Robustness analysis against quantum noise

In the noisy intermediate-scale quantum (NISQ) era, the implications of quantum noise must be considered when developing QML models. Noise originating from the quantum channels frequently disrupts the quantum entanglement state, which can adversely affect the performance of QML models. Quantum noise not only undermines model precision but can potentially lead to a barren plateau phenomenon. Quantum noise associated with a single qubit can be described using the Kraus matrices [47]. The Kraus matrices of bit flip (BF) are

$$K_0 = \sqrt{1-p}\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}, K_1 = \sqrt{p}\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix} \tag{18}$$

The Kraus matrices of phase flip (PF) noise are

$$K_0 = \sqrt{1-p}\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}, K_1 = \sqrt{p}\begin{bmatrix}1 & 0\\ 0 & -1\end{bmatrix} \tag{19}$$

The Kraus matrices of amplitude damping (AD) noise are

$$K_0 = \begin{bmatrix}1 & 0\\ 0 & \sqrt{1-p}\end{bmatrix}, K_1 = \begin{bmatrix}0 & \sqrt{p}\\ 0 & 0\end{bmatrix} \tag{20}$$

The Kraus matrices of depolarization noise (DN) noise are

$$K_2 = \sqrt{p/3}\begin{bmatrix}0 & -i\\ i & 1\end{bmatrix}, K_3 = \sqrt{p/3}\begin{bmatrix}1 & 0\\ 0 & -1\end{bmatrix} \tag{21}$$

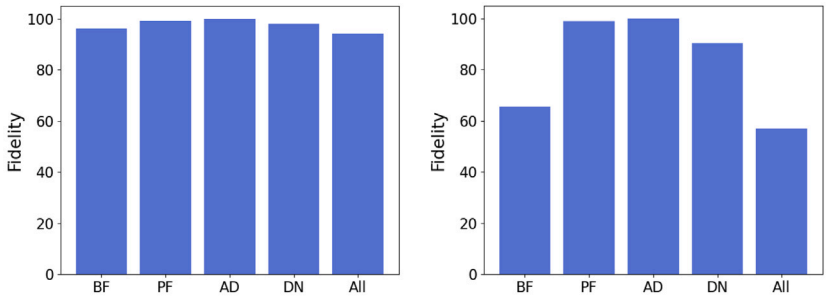
where p denotes the probability of noise occuring on a single qubit. $p \in \{0, 1\}$.

The robustness of quantum algorithms against noise is assessed using the fidelity metric. Fidelity serves as a measure of the closeness between two quantum states, representing the probability of one quantum state being recognized as another after testing. A high-quality quantum algorithm possesses high fidelity. Thus, fidelity can be perceived as an indicator of the extent of the effect of quantum noise on algorithmic outcomes. The lower the effect of noise on the results is, the higher its fidelity will be. For two mixed states, the formulation for fidelity is expressed as follows:

$$F(\rho,\sigma) = Tr\left(\sqrt{\sqrt{\rho}\sigma\sqrt{\rho}}\right)^2. \tag{22}$$

where ρ and σ are the density matrices of the two mixed states.

In this study, the fidelities of the QCNN circuit are calculated in four single noise channels and a noise combination channel of all noise types with $p = 0.1$ and $p = 0.01$ respectively. "All" represents the noise



(a) The fidelities in noise channels with $p = 0.01$. (b) The fidelities in noise channels with $p = 0.1$.

Fig. 13. The fidelities of the QCNN circuit in noise channels with different p values.

Table 8
The fidelities of our QCNN circuit in different noise channels with $p = 0.1$ and $p = 0.01$.

Noise type	$p = 0.01$	$p = 0.1$
BF	96.06	65.6
PF	99.12	99.01
AD	99.99	99.99
DN	98.01	90.37
All	94.2	57.03

Table 9
The comparison of time complexity between QMFND and the baselines.

Model	Time complexity
BERT [42]	$O(n)$
XLNet [41]	$O(n)$
MCAN/MCAN-A [43]	$O(n)$
SpotFake [17]/SpotFake+ [18]	$O(n)$
CNN	$O(n)$
QMFND	$O\left(\log_2 n\right)$

combination channel by the superposition of all noise operators (Kraus matrices) on single qubits. The results are shown in Table 8 and Fig. 13.

It is not difficult to find that the BF noise has a greater impact on the fidelity of QCNN. In QCNNs, qubits are usually used to represent the features of the input data. If a qubit flips, the feature information originally encoded can be lost. In contrast, other noises may affect the phase or amplitude of the qubits without directly causing the loss of information. Also, when the noise increases, the effect of the BF noise will be obvious. This is possible because, in a QCNN, information will pass through different layers. If a qubit flips, its information in the subsequent layers will also be affected, which can lead to cumulative errors throughout the network. Although the QCNN circuit in this study is somewhat affected by the BF noise when $p = 0.1$, it consistently upholds high fidelity. Notably, for both PF and AD noises, the fidelities are nearly 1, suggesting a minimal influence of these noise types on performance. This attests to the robustness of the QCNN against quantum noise.

4.9. Complexity analysis

The complexity mainly includes quantum circuit complexity and time complexity. Quantum circuit complexity can be calculated by the number of quantum gates used in the QCNN. As depicted in Figs. 7 and 8, eight quantum gates are required in a quantum convolutional layer, and six quantum gates are required in the quantum pooling layer. Suppose N qubits satisfy $N = 2^a$ (a is an integer greater than or equal to 1), then, the number of quantum gates required in the $i-$th layer is $14N/2^i$, where $i \leqslant a$. And the total number of quantum gates required by the QMFND model is calculated as:

$$2 \cdot \left(\frac{N}{2} \cdot 7 + \frac{N}{2^2} \cdot 7 + \cdots + \frac{N}{2^{a-1}} \cdot 7\right) + 14 = 7N\left(2 - \frac{1}{2^{a-2}}\right) + 14 = 14\left(N - 1\right). \tag{23}$$

Therefore, the total circuit complexity is $O(N)$.

A lower time complexity in a CNN or QCNN model typically indicates higher computational efficiency of its network. In other words, for a given task, the model's computational cost is relatively low. Comparing the QCNN in QMFND with that in baselines, as shown in Table 9, the complexity of executing a CNN to generate n-dimensional classical data is $O(n)$. However, for a QCNN using quantum amplitude encoding, only $\lceil \log_2 n \rceil$ qubits are required to represent an n-dimensional vector. Therefore, the time complexity of QCNN models can be minimized as $O\left(\log_2 n\right)$. This emphasizes the benefits of employing qubits for feature representation. They have the capability to capture a greater number of features while utilizing fewer resources compared to classical computer data, which is more efficient.

5. Conclusion

For quicker and more accurate detection of fake news on social media, the QMFND model is proposed by combining quantum perspectives. The experimental results show that QMFND achieves high accuracies of 87.9% and 84.6% on two datasets. QMFND outperforms the six baselines and has only a slight disadvantage compared to XLNet. The designed QCNN circuit has good expressibility, entangling capability, and robustness against quantum noise. The complexity of QMFND is much lower than the classical models by combining amplitude encoding to encode multimodal data into quantum states. However, the performance of QMFND can be compromised by current hardware limitations and substantial BF noise in the running environment of quantum computers. In future, We can introduce quantum error-correcting codes like Shor codes to minimize the effect of noise. Also, to improve the accuracy, we can consider quantum fuzzy neural networks to increase the network's ability to adapt to uncertainty. These considerations will help to improve the performance of fake information detection and develop more efficient and reliable quantum-based models.

CRediT authorship contribution statement

Zhiguo Qu: Conceptualization, Formal analysis, Methodology, Visualization, Writing – original draft. **Yunyi Meng:** Data curation, Formal analysis, Methodology, Supervision, Writing – original draft. **Ghulam Muhammad:** Funding acquisition, Investigation, Project administration, Supervision, Writing – review & editing. **Prayag Tiwari:** Data curation, Formal analysis, Investigation, Project administration, Software, Validation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

https://github.com/olivia-2333/QMFND_quantum_fake_news_detection.

Acknowledgment

This work was supported by the Deputyship for Research and Innovation, "Ministry of Education," Saudi Arabia, under Grant IFKSUOR3-283-2.

References

[1] E.C. Tandoc Jr., The facts of fake news: A research review, Sociol. Compass 13 (9) (2019) e12724.
[2] T. Meng, X. Jing, Z. Yan, W. Pedrycz, A survey on machine learning for data fusion, Inf. Fusion 57 (2020) 115–129.
[3] L. Wei-bin, Z. Zhi-yuan, X. Wei-wei, Feature fusion methods in pattern classification, J. Beijing Univ. Posts Telecommun. 40 (4) (2017) 1.
[4] U.G. Mangai, S. Samanta, S. Das, P.R. Chowdhury, A survey of decision fusion and feature fusion strategies for pattern classification, IETE Tech. Rev. 27 (4) (2010) 293–307.
[5] J. Yao, V.V. Raghavan, Z. Wu, Web information fusion: A review of the state of the art, Inf. Fusion 9 (4) (2008) 446–449.
[6] H.T. Phan, N.T. Nguyen, D. Hwang, Fake news detection: A survey of graph neural network methods, Appl. Soft Comput. (2023) 110235.
[7] N.R. de Oliveira, P.S. Pisa, M.A. Lopez, D.S.V. de Medeiros, D.M. Mattos, Identifying fake news on social networks based on natural language processing: Trends and challenges, Information 12 (1) (2021) 38.
[8] R. Oshikawa, J. Qian, W.Y. Wang, A survey on natural language processing for fake news detection, 2018, arXiv preprint arXiv:1811.00770.
[9] Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su, J. Gao, Eann: Event adversarial neural networks for multi-modal fake news detection, in: Proceedings of the 24th Acm Sigkdd International Conference on Knowledge Discovery & Data Mining, 2018, pp. 849–857.
[10] G.L. Long, L. Xiao, Parallel quantum computing in a single ensemble quantum computer, Phys. Rev. A 69 (5) (2004) 052303.
[11] M. Sasaki, A. Carlini, Quantum learning and universal quantum matching machine, Phys. Rev. A 66 (2) (2002) 022303.
[12] H. Neven, V.S. Denchev, G. Rose, W.G. Macready, Training a large scale classifier with the quantum adiabatic algorithm, 2009, arXiv preprint arXiv:0912.0779.
[13] D. Anguita, S. Ridella, F. Rivieccio, R. Zunino, Quantum optimization for training support vector machines, Neural Netw. 16 (5–6) (2003) 763–770.
[14] E. Ovalle-Magallanes, J.G. Avina-Cervantes, I. Cruz-Aceves, J. Ruiz-Pinales, Hybrid classical–quantum convolutional neural network for stenosis detection in X-ray coronary angiography, Expert Syst. Appl. 189 (2022) 116112.
[15] Z. Qu, X. Liu, M. Zheng, Temporal-spatial quantum graph convolutional neural network based on Schrödinger approach for traffic congestion prediction, IEEE Trans. Intell. Transp. Syst. (2022).
[16] J. Ma, W. Gao, K.-F. Wong, Detect rumors on twitter by promoting information campaigns with generative adversarial learning, in: The World Wide Web Conference, 2019, pp. 3049–3055.
[17] S. Singhal, R.R. Shah, T. Chakraborty, P. Kumaraguru, S. Satoh, Spotfake: A multi-modal framework for fake news detection, in: 2019 IEEE Fifth International Conference on Multimedia Big Data, BigMM, IEEE, 2019, pp. 39–47.
[18] S. Singhal, A. Kabra, M. Sharma, R.R. Shah, T. Chakraborty, P. Kumaraguru, Spotfake+: A multimodal framework for fake news detection via transfer learning (student abstract), in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, (no. 10) 2020, pp. 13915–13916.
[19] I. Segura-Bedmar, S. Alonso-Bartolome, Multimodal fake news detection, Information 13 (6) (2022) 284.
[20] N. Jayakody, A. Mohammad, M.N. Halgamuge, Fake news detection using a decentralized deep learning model and federated learning, in: IECON 2022–48th Annual Conference of the IEEE Industrial Electronics Society, IEEE, 2022, pp. 1–6.
[21] A.M. Luvembe, W. Li, S. Li, F. Liu, G. Xu, Dual emotion based fake news detection: A deep attention-weight update approach, Inf. Process. Manage. 60 (4) (2023) 103354.
[22] V.H. Nguyen, K. Sugiyama, P. Nakov, M.Y. Kan, FANG: Leveraging social context for fake news detection using graph representation, 2020.
[23] S. Raza, C. Ding, Fake news detection based on news content and social contexts: A transformer-based approach, Int. J. Data Sci. Anal. 13 (4) (2022) 335–362.
[24] J. Li, S. Ni, H.-Y. Kao, Meet the truth: Leverage objective facts and subjective views for interpretable rumor detection, 2021, arXiv preprint arXiv:2107.10747.
[25] L. Hu, T. Yang, L. Zhang, W. Zhong, D. Tang, C. Shi, N. Duan, M. Zhou, Compare to the knowledge: Graph neural fake news detection with external knowledge, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 754–763.
[26] B. Narottama, S.Y. Shin, Quantum neural networks for resource allocation in wireless communications, IEEE Trans. Wireless Commun. 21 (2) (2021) 1103–1116.
[27] P. Tiwari, L. Zhang, Z. Qu, G. Muhammad, Quantum fuzzy neural network for multimodal sentiment and sarcasm detection, Inf. Fusion (2023) 102085.
[28] V. Rajesh, U.P. Naik, et al., Quantum convolutional neural networks (QCNN) using deep learning for computer vision applications, in: 2021 International Conference on Recent Trends on Electronics, Information, Communication & Technology, RTEICT, IEEE, 2021, pp. 728–734.
[29] Y. Li, R.-G. Zhou, R. Xu, J. Luo, W. Hu, A quantum deep convolutional neural network for image recognition, Quantum Sci. Technol. 5 (4) (2020) 044003.
[30] G. Chen, Q. Chen, S. Long, W. Zhu, Z. Yuan, Y. Wu, Quantum convolutional neural network for image classification, Pattern Anal. Appl. 26 (2) (2023) 655–667.
[31] C.-H.H. Yang, J. Qi, S.Y.-C. Chen, P.-Y. Chen, S.M. Siniscalchi, X. Ma, C.-H. Lee, Decentralizing feature extraction with quantum convolutional neural network for automatic speech recognition, in: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2021, pp. 6523–6527.
[32] Z. Qu, Y. Li, P. Tiwari, QNMF: A quantum neural network based multimodal fusion system for intelligent diagnosis, Inf. Fusion 100 (2023) 101913.
[33] Z. Qu, W. Shi, B. Liu, D. Gupta, P. Tiwari, IoMT-based smart healthcare detection system driven by quantum blockchain and quantum neural network, IEEE J. Biomed. Health Inform. (2023).
[34] I. Cong, S. Choi, M.D. Lukin, Quantum convolutional neural networks, Nat. Phys. 15 (12) (2019) 1273–1278.
[35] W. Jiang, J. Xiong, Y. Shi, A co-design framework of neural networks and quantum circuits towards quantum advantage, Nat. Commun. 12 (1) (2021) 579.
[36] D. Kartsaklis, I. Fan, R. Yeung, A. Pearson, R. Lorenz, A. Toumi, G. de Felice, K. Meichanetzidis, S. Clark, B. Coecke, Lambeq: An efficient high-level python library for quantum NLP, 2021, arXiv preprint arXiv:2110.04236.
[37] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.
[38] M. Schuld, V. Bergholm, C. Gogolin, J. Izaac, N. Killoran, Evaluating analytic gradients on quantum hardware, Phys. Rev. A 99 (3) (2019) 032331.
[39] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, H. Liu, FakeNewsNet: A data repository with news content, social context and dynamic information for studying fake news on social media, 2018, arXiv preprint arXiv:1809.01286.
[40] Tampa Bay Times, Politifact, 2007, https://github.com/KaiDMML/FakeNewsNet. (Accessed 2007).
[41] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R.R. Salakhutdinov, Q.V. Le, Xlnet: Generalized autoregressive pretraining for language understanding, Adv. Neural Inf. Process. Syst. 32 (2019).
[42] I. Turc, M.-W. Chang, K. Lee, K. Toutanova, Well-read students learn better: On the importance of pre-training compact models, 2019, arXiv preprint arXiv:1908.08962v2.
[43] Y. Wu, P. Zhan, Y. Zhang, L. Wang, Z. Xu, Multimodal fusion with co-attention networks for fake news detection, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 2021, pp. 2560–2569.
[44] J.R. McClean, S. Boixo, V.N. Smelyanskiy, R. Babbush, H. Neven, Barren plateaus in quantum neural network training landscapes, Nat. Commun. 9 (1) (2018) 4812.
[45] S. Sim, P.D. Johnson, A. Aspuru-Guzik, Expressibility and entangling capability of parameterized quantum circuits for hybrid quantum-classical algorithms, Adv. Quantum Technol. 2 (12) (2019).
[46] K. yczkowski, H.-J. Sommers, Average fidelity between random quantum states, Phys. Rev. A 71 (3) (2005) 32313.
[47] M.A. Nielsen, I. Chuang, Quantum Computation and Quantum Information, American Association of Physics Teachers, 2002.

ABOUT UMT FACULTY

SDI

Selective Dissemination of Information (SDI) service is a current-awareness service offered by the PSNZ for UMT Faculty Members. The contents selection criteria include current publications (last 5 years), highly cited and most viewed/downloaded documents. The contents with pdf full text from subscribed databases are organized and compiled according to a monthly theme which is determined based on the topics of specified interest.

For more information or further assistance, kindly contact us at 09-6684185/4298 or email to psnz@umt.edu.my/sh_akmal@umt.edu.my

Thank you.

Perpustakaan Sultanah Nur Zahirah
Universiti Malaysia Terengganu
21030 Kuala Nerus, Terengganu.

Tel. : 09-6684185 (Main Counter)
Fax : 09-6684179
Email : psnz@umt.edu.my