Social Network Analysis and Mining (2023) 13:56
https://doi.org/10.1007/s13278-023-01065-0

ORIGINAL ARTICLE

FNR: a similarity and transformer-based approach to detect multi-modal fake news in social media

Faeze Ghorbanpour[1] · Maryam Ramezani[1] · Mohammad Amin Fazli[1] · Hamid R. Rabiee[1]

Received: 4 December 2022 / Revised: 10 March 2023 / Accepted: 11 March 2023

Abstract
Many people today get their news from social media. It is possible to propagate news using textual, visual, or multi-modal information. The popularity of social networks and their wide use by people make them attractive platforms for spreading fake news. Detecting fake news is essential to preventing its spread. Fake news can be a false article or a genuine article with misleading visual information. Adding an actual image to trustworthy unrelated news can also create a fake news story. In this paper, we propose a novel and efficient similarity and transformer-based detection algorithm called Fake News Revealer (FNR), which uses text and images of news to detect fake news. The algorithm uses contrastive loss to consider text and image relations and transformer models to extract contextual and semantic features. According to experiments on two public social media news data sets, the FNR algorithm competes with conventional methods and state-of-the-art fake news detection algorithms by adding a novel mechanism without adding extra parameters or weights.

Keywords Fake news · Multi-modal learning · Transfer learning · Language and vision similarity · Social media

1 Introduction

Since social media is widely accessible and interactive, it has become many individuals' primary news source. News is published on social networks daily, and people, willingly or unwillingly, share it with their friends and followers. The study in Gross (2010) shows that most Americans receive their news via the Internet rather than newspapers and radio, and three-quarters receive it through email or social media. The popularity of social media tempts criminals to pursue their immoral intentions by producing and disseminating fake news using seductive text and misleading content and images. As a result, it is essential to verify social media news and detect fake news.

Social media news differs from other news sources like news agencies or micro-blogs. In social media news, the content is usually written by ordinary people in informal language, is brief, and contains low-quality images. Social media news via platforms like Twitter, Weibo, and Facebook fills in the gaps in earlier news reports by providing information on several facets of a current news event. In this paper, we have concentrated on this category of social media news.

The ease of using, sharing, and disseminating news on social media can lead criminals to create and publish fake news. Misleading the public, harming an institution, person, or government, or harming public and private stock markets are all examples of this fraud. Fake news has two characteristics: it is intentionally written and is provable to be false, which separates it from rumors, satires, and spam (Shu et al. 2017). It is common for ordinary users to republish news without being aware of its truthfulness. As the scope of the news expands, further damage is caused, which leads to a distrust of good news and disregard for warnings. Consequently, reporters and journalists are unable to cover important and correct news.

Faeze Ghorbanpour and Maryam Ramezani contributed equally to this study.

✉ Hamid R. Rabiee
rabiee@sharif.edu

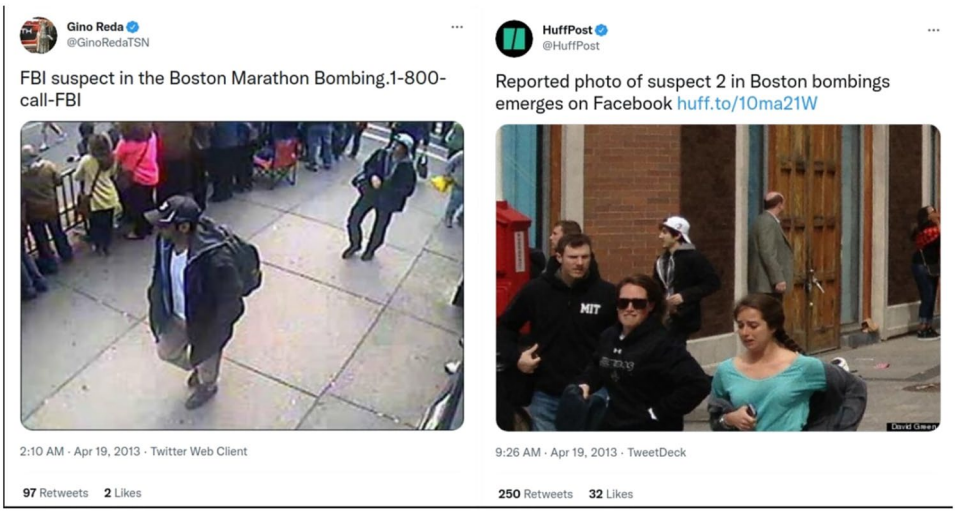
Faeze Ghorbanpour
f.gorbanpor@sharif.edu

Maryam Ramezani
maryam.ramezani@sharif.edu

Mohammad Amin Fazli
fazli@sharif.edu

[1] Department of Computer Engineering, Sharif University of Technology, Tehran, Iran

Published online: 28 March 2023



(a) Real news

(b) Fake news

The deceptions involving visual misinformation are much more straightforward. For example, it is common to use old photos and videos as evidence of recent events. In addition, news can acquire trust by using images taken out of context to illustrate events that are not relevant to the story. Fake news, for example, uses an image borrowed from a factual report to appear as factual. Images play a pivotal role in influencing public opinion and creating false perceptions. According to psychological research, Newman et al. (2012), when individuals see an image alongside a trivial statement, such as turtles being deaf, they are more likely to believe it. In a simulated social media environment, a post that incorporates photos get more likes and shares, as well as people's perception that it is factual (Fenn et al. 2019). It is, therefore, necessary to consider the visual features and the relationship between the image and text of a news report to judge its veracity.

It is not always possible to identify the news source within social media, but we can find user reactions to the news posts. Twitter, for instance, allows users to retweet or comment on tweets, and these retweets and comments are all we see, so we must determine their reliability. Fake news can result from these reactions and divert attention away from the main news, which prevents people from paying attention to the real story. For example, Fig. 1 shows four tweets related to the 2013 Boston terrorist attack. This example illustrates how fake tweets mislead readers and divert news by exploiting public attention. According to Fig. 1b, the first fake tweet attempted to portray another person as a terrorist by resembling the backpack of a terrorist. In the second fake

Fig. 1 Twitter posts about April 15, 2013, Boston terrorist attack

tweet, the image of the actual news was placed next to an article about a missing person, in which the missing person was presented as a terrorist. In this regard, we have several tweets about an event. Our objective is to identify tweets that contain false information about the event rather than determine whether the source event is correct.

To deal with these challenges, we propose an end-to-end framework referred to as *Fake News Revealer (FNR): A similarity and transformer-based approach to detect multi-modal fake news in social media*. In this approach, we embed text using BERT, a language representation model for bidirectional encoder representations from transformers (Devlin et al. 2019). Due to limitations on the length of text of BERT, it is suitable for social media news, as well as it provides a contextual representation of sentences (Peters et al. 2018). In addition, we use ViT; a vision transformer model (Wu et al. 2020) that uses patches of images as tokens in a natural language processing application and provides the sequence of linear embedding of these patches as inputs to the transformer (Dosovitskiy et al. 2021). The use of rich image feature extraction and semantic embedding is important because multi-modal detectors tend to be more susceptible to visual attacks than textual attacks (Chen et al. 2022). Even without pre-training on large data sets, transformer-based models are more robust to generalization on samples that are out of distribution (Bai et al. 2021).

Following these two pre-trained modules, we apply two projection modules to project extracted features into a similar-sized array and tune its weights based on our task. It should be noted that, in contrast to the pre-trained modules, which had their parameters frozen, these two added modules have tunable parameters, and their weights and biases are learned in the end-to-end learning process. We used two loss functions in the model's optimization: contrastive loss, a self-supervised loss between images and texts (Khosla et al. 2020), and classification loss, a cross-entropy loss between predicted labels and ground truth.

Analyzing multi-modal fake news requires comparing images with their text since some news contains unrelated images and texts. Calculating the similarity between image and text extracted feature vectors allows us to consider not only the relationship between the image and text that appear together in a news post but also the relationship between news items. Our problem involves multiple news posts about the same event that may be fake or correct about the event. As a consequence, the images and texts of news regarding an event are related to each other, which is why similarity calculations are helpful in this situation.

The main contributions we have made are as follows:

- Utilized transformer models for both textual and visual features. As a result of using ViT and BERT, it has been found that they are more effective at extracting semantic and contextual features from texts and images than other methods of detecting fake news.
- Calculated similarity among image and text of news as our contrastive loss function. In this way, it became possible to consider not only the relationship between a news post's text and image but also the relationship between news posts that were related to a particular event. This loss function does not introduce additional complexity or parameters but has improved the efficiency of our method over previous studies.
- Evaluated the proposed method on two publicly available and most-used data sets in multi-modal fake news detection. It proved to be more accurate and efficient than previous state-of-the-art approaches.

2 Literature review

The automatic detection of fake news has become increasingly important as social media has grown in popularity. The deliberate nature of fake news and its negative consequences and ramifications have prompted more researchers to focus on this subject. We categorize the relevant works in this part depending on the modalities of their inputs. Following this, we will discuss some of the most recent methods of detecting multi-modal fake news.

2.1 Single modality

In early works, only one data modal was used to detect fake news, with textual data receiving the most attention due to its prevalence in the news. Linguistic features are utilized in Shu et al. (2017) to validate news on Twitter, and structural and cognitive features are extracted to detect fake news on social networks in Kwon et al. (2013). It is impossible to generalize these methods to all topics, especially the ones based on linguistic features. Additionally, the methods in Shu et al. (2017) and Kwon et al. (2013) do not extract features automatically, resulting in insufficient and out-of-proportion solutions.

The development of deep neural networks has been shown to significantly improve the performance of detecting fake news by extracting the features automatically. In a study published by Liu and Wu (2018), both recurrent and convolutional networks are used to understand global and local differences in text. Ruchansky et al. (2017) is another study that examines textual content using a deep hybrid model based on recurrent neural networks. In another study by O'Brien et al. (2018), deep learning strategies are utilized to detect fake news. In this study, emergent representations derived from deep neural networks are shown to identify subtle differences between the language employed by fake news and real news.

Another study, FNDNet (Kaliyar et al. 2020), is developed to train the discriminating characteristics for fake news classification. It uses several hidden layers created based on a CNN-based model to extract various features at each layer. Following an extensive feature study, Kaliyar et al. (2021a) apply a tensor factorization-based approach to classify fake news based on content and context. Jain et al. (2022) is another recent study uses attention mechanisms to embed texts contextually.

Using language models and transformers have significantly improved many machine learning tasks, including detecting fake news. Bidirectional encoders from transformer modeling (BERT) are employed in the work by Jwa et al. (2019) to identify fake news in data sets of headline-body text. Another work that used BERT is called FakeBERT (Kaliyar et al. 2021b), in which the parallel blocks of the single-layer deep convolutional neural network are subjected to BERT. These experiments that included transformers fared better than earlier studies, encouraging extensive use of transformers in subsequent research.

Fewer papers focus solely on the image of fake news instead of the text. Examples of papers that consider images when evaluating the veracity of news are Gupta et al. (2012), Ping Tian et al. (2013) and Shu et al. (2020). These works examine image characteristics and their impact on social media news. Fake news detection based on visual features has recently been done by Qi et al. (2019) constructed a CNN-based network based on the frequency domain of fake news images and extracted visual features from various semantic levels in pixels using a multi-branch CNN-RNN model (Yenter and Verma 2017).

In this section, all the papers suggest a new approach, but most rely on a single modality. However, using multimedia data alongside text data provides us with more accurate and reliable performance and is closer to the reality of social media news.

2.2 Multiple modality

The publications that have worked on identifying fake news in recent years have demonstrated that using other kinds of data besides the text can be beneficial in more accurate detection. In this part, we will look at different methods for detecting fake news based on their language and vision contents.

Researchers have presented one of the earliest approaches to identifying fake news using images and news text in Jin et al. (2017) and achieved superior results. They use recurrent deep networks and present a novel data set for multimodal fake news. The authors in VQA (Antol et al. 2015) employ a visual system to answer questions via deep networks for fake news detection using multi-modal data. Alternatively, Farajtabar et al. (2017) uses subtitle texts and images to detect fake news.

Wang et al. (2018b) EANN use images and text to detect fake news using an event adversarial neural network. EANN attempts to solve the independent identification of news events challenge by reducing the impact and occurrence of the news via an adversarial mechanism. Its goal is to generalize the solution to unseen events. This model uses the pre-trained VGG19 (Simonyan and Zisserman 2015) network models for the image and a deep convolutional network for textual properties. Another approach to address emerging events is presented by the same authors in MetaFEND (Wang et al. 2021), using a few-shot learning method, encoding event names and calculating attention on the extracted features of text, image, and event name to reduce dependency over the events. This paper similarly extracts the visual and textual features of EANN.

Khattar et al. (2019) present a variational auto-encoder and an encoder-decoder network named multi-modal variational auto-encoder for fake news detection (MVAE) to detect fake news using the learned hidden vectors. This paper uses a deep bidirectional LSTM network to extract textual features, and a VGG19 network is employed to extract image features.

Singhal et al. (2019) SpotFake works by embedding text by BERT (Devlin et al. 2019) and images by VGG19 in vectors and then fusing these vectors. This approach gives Spotfake better results than previous works since it does not consider other sub-tasks. The method presented by Palani et al. (2022) is similar to SpotFake's textual embedding and fusion type. However, instead of using convolutional neural networks for the visual feature extraction, the method utilizes CapsNet (Hinton et al. 2011), which takes advantage of the presence and prediction of objects in an image.

The link between an image and text in a news item has recently been paid attention to. Cross-modal attention residual and multichannel convolutional neural networks (CARMN) (Song et al. 2021) utilizes a cross-model attention mechanism to consider the relationship between image and text. Then, It uses a self-attention mechanism to obtain the feature vectors and determines fake news using a concatenation of these feature vectors. This approach promises to fuse meaningful information across distinct modalities while maintaining each modality's unique qualities and reducing the impact of noisy information created by cross-modal fusion.

A recent work focusing on weak and strong modality issues in multi-modal fake news detection is Singhal et al. (2022), which detects fake news using inter-modal and inter-modality relationships. The inter-modality feature extractor extracts fine-grained salient text and image features.

Table 1 Related works comparison

Method	Text encoder	Image encoder	Fusion type
EANN (Wang et al. 2018b)	Text-CNN	VGG19	Concatenation
MVAE (Khattar et al. 2019)	BiLSTM	VGG19	Auto-encoder
SpotFake (Singhal et al. 2019)	BERT	VGG19	Concatenation
CARMN (Song et al. 2021)	Word level embedding	VGG19	Concatenation + attention
AMFB (Kumari and Ekbal 2021)	BiLSTM + attention	CRNN[a] + attention	Multiplication
FMFN (Wang et al. 2022b)	Roberta	VGG19	Attention
FNR (ours)	BERT	ViT[b]	Concatenation + similarity

[a]CNN-RNN blocks

[b]Vision transformer

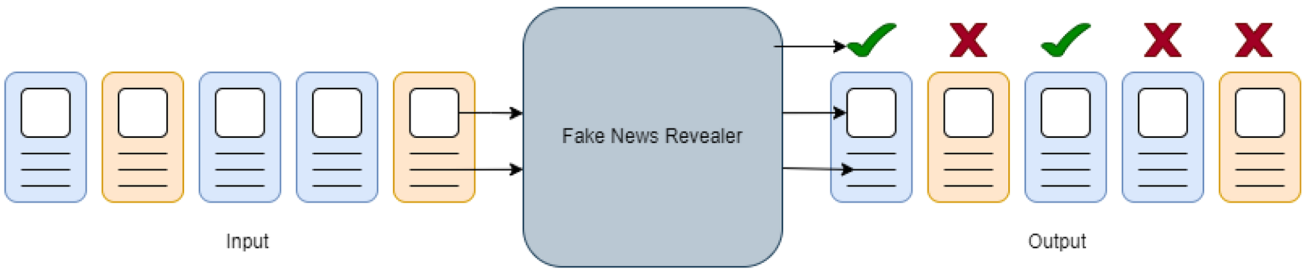


Fig. 2 A black-box diagram illustrating the problem. The system's input is a set of news posts from different news events (blue and orange represent two events), and the model assigns a fake or real label to the posts (Color figure online)

Inter-modality relationship extractors fuse multi-modal features multiplicatively, using BERT and Faster R-CNN (Ren et al. 2015) to extract text and image features. MTTV (Wang et al. 2022a) is another study that uses the Faster R-CNN model in addition to ResNet (He et al. 2016) to extract visual features and BERT to extract textual information. The retrieved features are combined, and the news is classified using a transformer encoder block.

The researchers in AMFB (Kumari and Ekbal 2021) suggest an attention-based multi-modal factorized bilinear pooling that uses attention-based bidirectional LSTM to capture textual features and attention-based CNN-RNN blocks for capturing visual features (Fukui et al. 2016). It employs a multi-modal feature fusion technique that combines information from text and images and optimizes their correlation to provide a multi-modal shared representation. Then, it uses a multi-layer perceptron to classify the calculated features.

Additionally, in FMFN (Wang et al. 2022b), the attention method enhances visual and textual features. As part of the integration process, the improved visual and textual features are fused, taking into account the dependency between them. The text and image embedding methods are Roberta and VGG19, respectively.

In this paper, we employ a transformer-based model to consider semantic and contextual features in the text, which differs from previous works that used convolutional and recurrent neural networks. Most previous works have also used convolutional models like VGG19 to extract image features, which have little capacity to extract contextual and complex features of the image. Furthermore, we addressed the connection between images and text in this study, which has been overlooked in previous studies. The loss function we added to our solution has minor complexity and outperforms other works considering image-text relationships. Importantly, prior studies have considered social media posts as a single news item without considering the relationships among related posts. Social media news, however, refers to a series of posts related to a particular news event. We calculated the similarity between news posts based on the contrastive loss function to address this. Table 1 presents a summary comparison of the proposed method with other related methods.

3 Proposed method

This section will elucidate our proposed method's methodology, logic, and rationale.

Table 2 Notations

Notation	Description
$N = (T, I, L, E)$	News set with text set T, image set I, label set L, and event set E
n, m, b	Number of news posts, Number of the news event, Batch size
k	Projection vector size
E	Ground truth of the similarity loss, which means the average similarity of both modules to themselves
F_C, F_T, F_I	Total projected vector, text projector vector, image projector vector
P, ρ	Similarity matrix of news posts, Similarity matrix of news events
F	Output vector of classifier module
ι, ι_c, ι_s	Total loss, similarity loss, classification loss

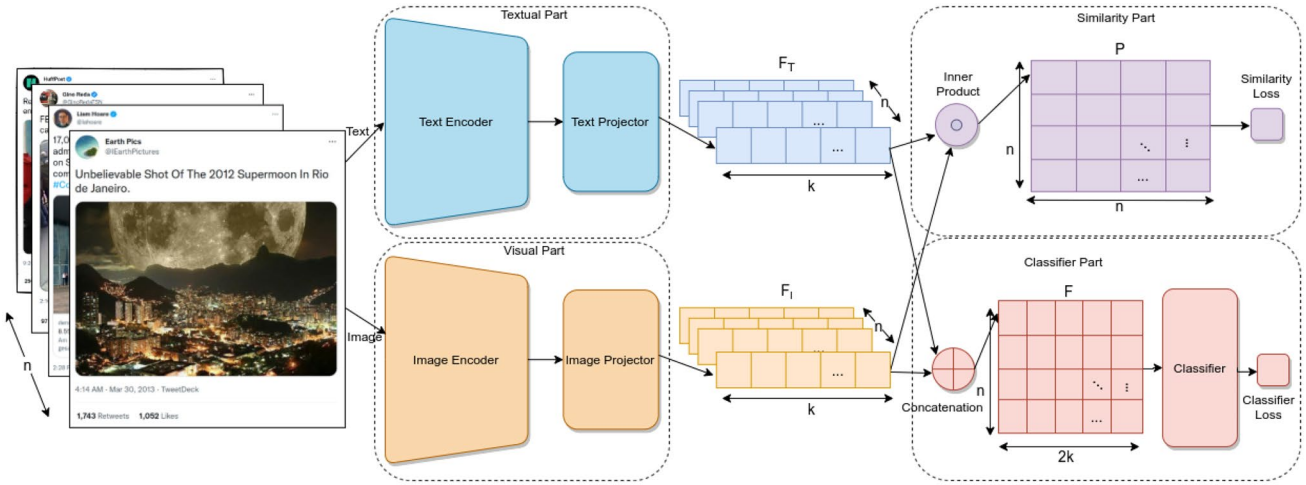


Fig. 3 An illustration of the architecture of the Fake News Revealer (FNR), the model composed of four parts: (1) textual, (2) visual, (3) similarity, and (4) classifier

3.1 Problem definition

Our objective is to take a set of news posts about an event and predict whether or not each post is fake. Each news post must include an image, text, and label. A diagram of our problem and the input and output of the model is shown in Fig. 2. For a formal definition, our data (N) consists of n news posts, and each news post (n_x) contains a text (t_x) and an image (i_x) and a label (l_x) indicating whether it is real or fake. Moreover, each news post belongs to a special event (e_x). There may be several news posts related to one event, but each news post only relates to one event, so the number of new events (m) is less than the number of news posts (n).

$$N = \{n_1, n_2, n_3, \ldots, n_n\}$$
$$n_x = (t_x, i_x, l_x, e_x) \quad (1)$$

The input of our algorithm consists of n news items, each containing textual and visual information. Our goal is to predict the label of news (l_x) using t_x and i_x information. Through training, the news is processed in batches, so we explain based on the batch size (b) rather than the whole number of news. The notation Table 2 describes the parameters and variables involved in the explanation.

3.2 Method overview

As shown in Fig. 3, our proposed framework consists of four parts: a textual feature extractor, a visual feature extractor, a similarity calculation module, and a classification module. Text and images are processed by feature extraction modules and converted into feature vectors. As part of the similarity section, visual and textual feature vectors are multiplied to compute the contrastive loss function. The fused vectors are passed through a classification module, then compared with the news post labels.

Using a linear function, our approach uses projector modules to resize the input vector inside the text and image parts. Unlike frozen encoders, projector weights are learned end-to-end during training. With this module's help, we can

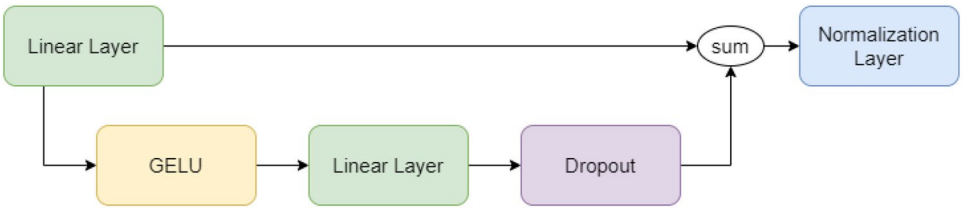


Fig. 4 Projector architecture

resize the extracted vector and consider layers that must be tuned according to our task. To avoid over-fitting, we utilize skip connections (He et al. 2015). This module produces a vector tuned based on the extracted features of each modal, and its structure is shown in Fig. 4.

3.3 Text feature extractor

This module aims to extract features from text and embed them into a vector. The text feature extractor consists of two main sub-modules; the first sub-module is an encoder that extracts representative features obtained from a pre-trained model. The second sub-module is a projector.

Pre-trained language models have accomplished cutting-edge outcomes on several natural language processing tasks. BERT (Devlin et al. 2019) and its derivatives, in particular, are frequently utilized because they exploit both left-to-right and right-to-left contextual information. BERT generates text representations that incorporate contextual information, which implies that embedding comprises information about the full-text content and may thus be regarded as a textual feature. In light of the short length of news items on social media, BERT is an appropriate choice.

This part takes a batch of texts T of size (b, τ) (τ represents the maximum text length). The result of applying BERT is a matrix (B) with size (b, 768), where 768 is the size of the last hidden layer of the BERT algorithm. Following the application of the projector, we obtain a matrix (F_T) with size (b, k) which k is the projector's final vector size: [GELU (Hendrycks and Gimpel 2016) is the Gaussian error linear unit activation function]:

$$F_T = w_2 \times (GELU(w_1 \times B + b_1)) + (w_1 \times B + b_1) + b_2 \quad (2)$$

In the text projector, w_1, w_2, b_1, and b_2 are the weights and biases of linear layers. Lastly, a fully linked layer followed by L2 normalization is used to get a normalized textual feature vector.

3.4 Image feature extractor

Two sub-modules comprise this module: an image encoder that embeds images into feature vectors based on a transformer model and a visual projector with tunable weights.

Despite the successful preservation of the spatial information in the embedding representations obtained from the final pooling layer of classical methods like VGG and CNN, the semantic relationship may be lost in the embedding representations (Wang et al. 2018a). Further, classical approaches divide an image equally on each spatial level, resulting in redundant background information (Singhal et al. 2022). We use ViT (Dosovitskiy et al. 2021) as our image encoder. ViT is a transformer encoder model (similar to BERT) that divides an image into patches to create a sequence. These fixed-size patches are linearly extracted from images and given to the transformer model.

Subsequently, this module takes a batch of images I with parameters (b, *width*, *height*, *depth*) representing images' width, height, and depth. Then, the ViT encoder processes it into the vector–matrix V with size (b, 768), where 768 is the size of the last hidden vector of the ViT. We have a projector sub-module here to serve as our visual projector, much like the text section. After applying the projector, we obtain F_I with size (b, k) with the same length as F_T obtained from the textual part. The projector works according to the following composition:

$$F_I = w_4 \times (GELU(w_3 \times V + b_3)) + (w_3 \times V + b_3) + b_4 \quad (3)$$

In this composition, $w3$, $w4$, $b3$, and $b4$ are the weights and biases of linear layers within the image projector, and the output, an image feature vector, is obtained by applying a fully connected layer, followed by L2 normalization.

3.5 Similarity calculation

This section aims to calculate the similarity between texts and images using a supervised contrastive loss algorithm (Khosla et al. 2020). A comparison of the image and text feature vectors, which are matrices of size (b, k), is made by calculating their inner products to determine whether they are similar. As a result of calculating the similarities between texts and images of b news posts, a prediction matrix (P) is calculated: (F^T means transpose of matrix F)

$$P = F_T F_I^T \quad (4)$$

The loss function considers an image and a text to be the most similar to itself. Thus, we consider the expected matrix as the average similarity between text-to-text and

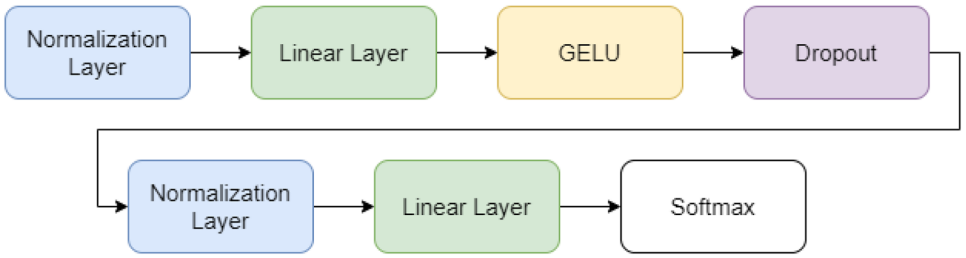


Fig. 5 Classifier architecture

image-to-image. Then we pass the average throw soft-max function to get the expected matrix (E) (Salama 2021; Radford et al. 2021), according to the following composition:

$$E = softmax\left(\frac{F_T F_T{}^T + F_I F_I{}^T}{2}\right) \tag{5}$$

As a result of calculating the expected matrix, we use cross-entropy to determine the loss. The contrastive loss is the average cross-entropy loss function of the whole similarity matrix (Radford et al. 2021).

$$\iota_s = mean(-(E \times log(P) + (1 - E) \times log(1 - P))) \tag{6}$$

The above formulas show that the entire similarity matrix is utilized rather than just the matrix's main diagonal. When the entire similarity matrix between text and image vectors of a batch is used, the similarity of text and image of a news post is calculated, and the relationship between texts and images of whole news posts is considered. This similarity between the picture of one news post and the text of another news post is compared to the average similarity of two texts and two images from these two news posts. This is because if these two news posts are about the same event, their text or image is close to each other. Thus all the images and texts should be similar, whereas if they are not about the same event, neither their text nor their image should be similar to each other, and thus their text and image should not be similar.

3.6 Classifier

The multi-modal news embedding is created in this module by concatenating the text and image feature vectors.

$$F_C = concat(F_T, F_I) \tag{7}$$

We pass the news embedding through two linear layers to classify it. After the linear function, a vector with two classes of size (b, 2) is generated. Based on Fig. 5 and assuming w_6, w_5, b_6, and b_5 represent weights and biases of linear layers in the classifier, it works as the following formula:

$$F = softmax(w_6 \times (GELU(w_5 \times F_C + b_5)) + b_6) \tag{8}$$

The loss function for classification is cross-entropy, which is calculated after calculating probabilities of predicted classes (α is the fraction of the sample which is dominant in the data set and $1 - \alpha$ denoting the fraction of the other class):

$$\iota_c = -(\alpha L \times \log(F) + (1 - \alpha)(1 - L) \times \log(1 - F)) \tag{9}$$

With λ as a trade-off parameter, the loss for the whole model can be determined from the two losses:

$$\iota = \iota_c + \lambda \cdot \iota_s. \tag{10}$$

This model is trained end-to-end to lower the loss function. An optimization algorithm reduces these loss functions, helping the model achieve its goal.

4 Experiments and evaluation

A detailed description and experimental results of applying FNR to two real-world data sets will be presented in this section, along with a comparison of this approach with state-of-the-art approaches.

4.1 Data sets

As part of our research to verify the effectiveness of the proposed model, we conducted experimental testing on two real-world data sets, which have been gathered from social media and are considered the most commonly used ones for multi-modal fake news detection. Table 3 details the number

Table 3 Class distribution

Data sets	Train		Test		Total news
	Fake news	Real news	Fake news	Real news	
Twitter	6649	4599	545	444	12,237
Weibo	3748	3758	999	995	9500

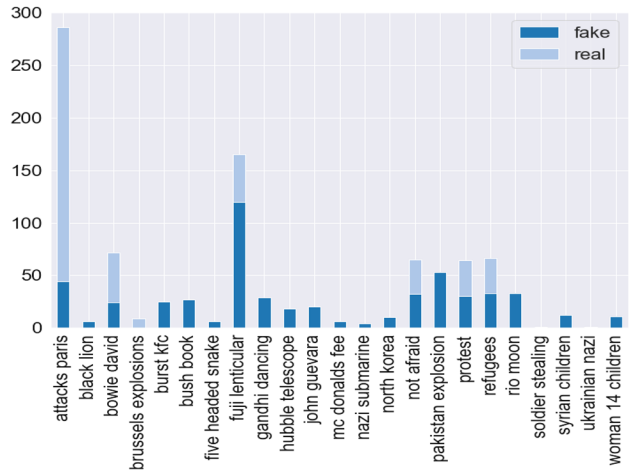


Fig. 6 Events distribution: number of real news and fake news for each event in the test division of the Twitter data set

of trains and test news posts used in our experiments and the number of fake and real news posts included in each data set.

Twitter: This data set was introduced in Boididou et al. (2015) to automatically verify multimedia tasks to distinguish fake or real news on Twitter.[1] Only tweets with both textual and visual content are kept. This data set consists mainly of tweets written in English (other languages were translated into English).

In the training and test sets, news events are not repeated. The train data set contains 15 events, and the test data set contains 23 events. In general, events are not completely fake or real; our objective is not to detect fake or real events but to detect fake or real tweets about them. Figure 6 provides an overview of the events within the data set and the number of real and fake tweets associated with each event.

Weibo: This data set (Jin et al. 2017) was collected from Weibo[2] from 2012 to 2016, and the language of comments is Chinese. The data set used in our study was processed by Wang et al. (2018b) because they removed low-quality images from the data set to ensure the overall quality and separated the training data events from the test data events. Further, text-only posts have been removed, so each post now contains text and image information.

4.2 Implementation details

The PyTorch framework builds our deep neural network with Python 3.6. We optimized the learning rate using AdamW (Loshchilov and Hutter 2017) and calculated different learning rates and weight decays for each part of our architecture to make the model converge faster (Singh et al. 2015). We have used the Optuna (Akiba et al. 2019) library to tune parameters and find best-suited values. We utilize a learning rate scheduler and an early stopping checkpoint to avoid over-fitting. The Hugging Face (Wolf et al. 2020) library was used in the encoder modules. The best parameters for each data set are obtained according to Table 4. The implementation is available in our repository.[3]

Table 4 Tuned parameters by Optuna library

	Twitter	Weibo
Dropout	0.5	0.5
Projection vector size (k)	64	64
Optimizer	AdamW	AdamW
Batch size (b)	100	100
Epochs number	300	300
Loss trade-off wight (λ)	1	1
Image learning rate	5.0e−4	1.0e−5
Text learning rate	2.0e−5	1.6e−4
Classifier learning rate	3.4e−3	1.5e−3
Weight decay	7.0e−2	1.5e−4
Text maximum length (τ)	32 words	200 characters

Raw text gathered from social media is non-standard and noisy, so normalization techniques are needed to clean it up. We performed pre-processing, which included normalizing abbreviations, removing unnecessary punctuation, and deleting non-standard characters. After pre-processing, the text is tokenized and ready to be encoded. The images were also pre-processed by deleting low-quality images, resizing them to (224×224), and converting them into appropriate input for the encoder.

The following metrics, always considered when dealing with classification problems, were considered for evaluating the proposed method: accuracy, recall, precision, micro $F1$ score, and macro $F1$ score. The AUC metric and receiver operating characteristic curve (ROC) are beneficial when binary classification consists of almost balanced classes. Due to the almost equal distribution of classes in our data sets, we also evaluated the model based on these two metrics.

4.3 Ablation study

An ablation study explored which data modal is more beneficial, why a multi-modal approach should be used, and why we should consider similarity measures. First, we used and tested each modal separately, and then we fused the modals and investigated the effectiveness of a multi-modal approach for fake news detection. Finally, we added the contrastive loss measurement to investigate its effectiveness in enhancing the results.

[1] https://twitter.com/.

[2] https://weibo.com/.

[3] http://git.dml.ir/fghorbanpoor/FakeNewsRevealer.

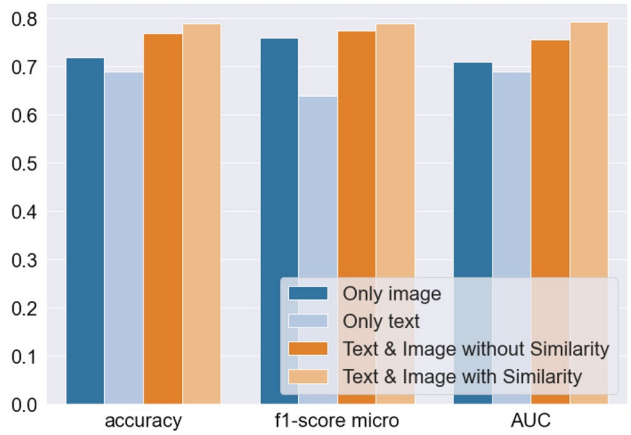


Fig. 7 Ablation study on Twitter data set

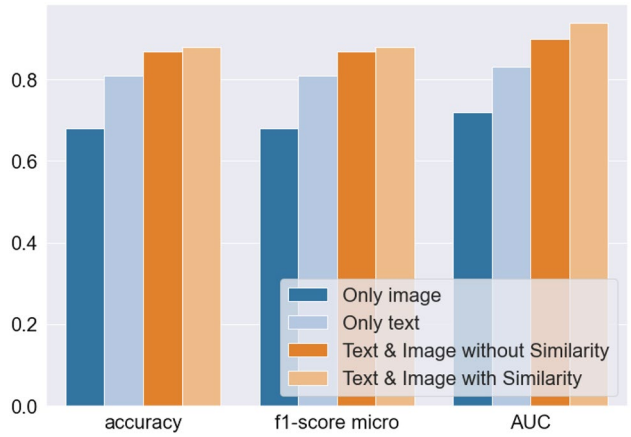


Fig. 8 Ablation study on Weibo data set

As demonstrated by Fig. 7 on Twitter, because the language of the tweets is brief, imprecise, and filthy, it is less accurate on its own, and the image modal performs better. However, the outcome improves by merging these two modalities, demonstrating that these two modals cover each other's shortcomings. Adding the similarity part, which considers the relationship between the text and the image, results in more improvements.

As illustrated in Fig. 8, the images on Weibo are not very expressive, and the actual news images are almost certainly being exploited for false news, which does not help to detect fake news on its own. Nonetheless, the text modal outperforms the visual modal. However, when these two modalities are included concurrently, the model's performance significantly improves, as these two modalities complement each other. When the relationship between text and image is considered, the accuracy also increases.

Table 5 Correlation of text and image similarity and label (Corr): with considering the similarity loss function ($\iota = \iota_c + \iota_s$), the similarity matrix is calculated, and the diagonal is averaged over fake, real, and whole news posts

	Corr
Average over real news posts	0.247
Average over whole posts	0.229
Average over fake news posts	0.215

4.4 Statistical study

This study aims to answer two questions: first, whether there is a correlation between the label of a news post and the similarity between its image and its text. Second, whether news posts related to a particular event are dissimilar to news related to other events after considering similarity loss. During this study, we calculated the similarity matrix using the dot product of visual and textual features in two models - one without considering the similarity loss function and one with consideration of the loss function.

4.4.1 Similarity and label experiment

If F_T and F_I are the extracted feature vectors from the textual and visual sections, respectively, then $P = F_T {F_I}^T$ is the dot product of these two vectors with a size of (b, b) and b here is the batch size. Assuming $b = n$ (the whole data size), we have similarity calculated throughout the complete supplied data set.

We need only consider the main diagonal (Z) of the similarity matrix (P) to answer the first question because we are attempting to determine the link between the image and text similarity and the label of each news item. As another way of putting it, we would like to find the correlation between Z and L. Consequently, we compare the average similarity of fake news posts with those of non-fake news posts over the main diagonal.

$$\begin{aligned} Z &= diag(P) \\ Corr &= correlation(Z, L) \end{aligned} \tag{11}$$

According to Table 5, after consideration of the similarity loss function, the similarity between the real news's image and text is greater than the similarity between the fake news's image and text. This shows that considering the loss function can assist in considering and controlling the similarity between visual and textual content. Due to this, fake tweets often employ unrelated images and text, and the similarity loss function is used to detect this unrelatedness and, ultimately, to detect disinformation.

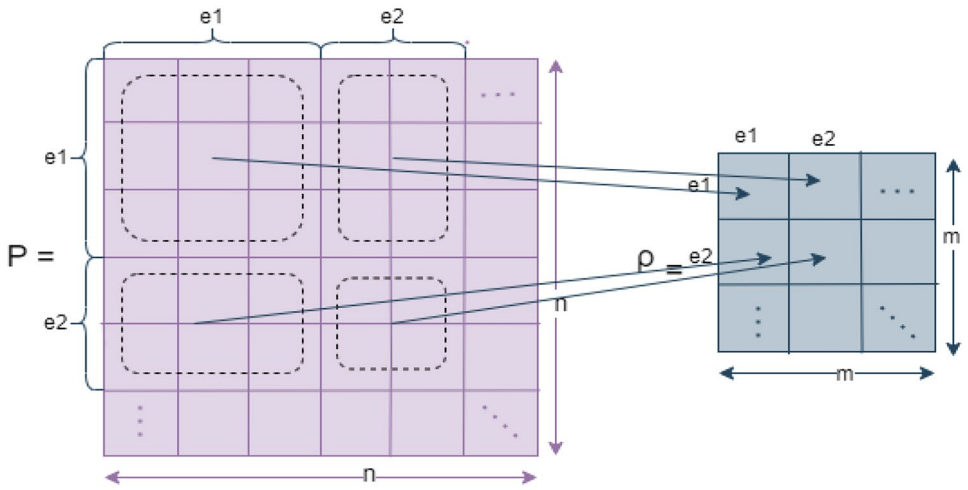


Fig. 9 An illustration of how matrix ρ is derived from matrix P

4.4.2 Similarity and event experiment

To answer the second question, this experiment is designed to find the relationship between events and the average similarity of images and text of news posts related to that event. In other words, we want to know if the similarity loss function can recognize dissimilarity between different events' contents and whether different events are considered. The similarity matrix P will also be used as in the previous experiment. Nevertheless, we group the matrix's rows and columns based on the news event, so instead of a similarity matrix of news posts (P) in size of (n, n), we have a similarity matrix of a news event (ρ) in size of (m, m) where m is the number of events.

Fig. 9 illustrates how the matrix can be obtained. According to the following formula, the (i, j) room of the matrix ρ are calculated from the average rooms of matrix P in which rows are related to the event e_i and columns are related to the event e_j.

$$\begin{aligned} &E = \{e_1, e_2, e_3, ..., e_m\} \\ &\rho(i,j) = mean(P(x,y) \| x \in e_i \wedge y \in e_j) \\ &Sim = mean(\rho(i,j) \| i \neq j \wedge i < j) \end{aligned} \tag{12}$$

In this case, we compare the ρ matrix obtained from a model without considering the similarity loss function with one that does. The average similarity between images and texts for all events is calculated by taking the average of all ρ values of the upper right of the matrix. The comparison of the similarity of different event's contents is shown in Table 6. According to this table, the similarity between sets of images of one event and those of other events has decreased due to the similarity loss function. In other words, the similarity loss function has been used to consider and control the dissimilarity between the images and texts about different events.

4.5 Performance comparison

Based on the evaluation metrics mentioned, Tables 7 and 8 provide comparisons for the Twitter and Weibo data sets, respectively. A comparison of the ROC curves of three relevant studies and our work on Twitter and Weibo data is shown in Figs. 10 and 11.

4.5.1 Baselines

The following is the list of single-modal and multi-modal benchmark methods we chose for a comparative study.

Single modality In two sets of trials, we conduct single-modal tests. The first only use the news's text, while the second only uses its image. We examine multiple text classification techniques-based purely on news text, such as logistic regression, SVM, LSTM, recurrent neural networks, and BERT. We choose a multi-filter size CNN, VGG19, and ViT for our tests on the news image.

Multi modality The works selected for multi-modality are listed in Table 1 and are discussed in the literature review section. Two versions of the proposed model are tested in this section; the first does not consider similarity measurements (FNR-WS), and the second does include similarity measurements (FNR-S).

Table 6 Similarity over different events' text and image

	Sim
Without similarity loss function ($\iota = \iota_c$)	0.607
With similarity loss function ($\iota = \iota_c + \iota_s$)	0.291

Table 7 Performance comparison for the Twitter data set

	Model name	Accuracy	AUC	*F*1 micro	Fake news			Real news		
					Precision	Recall	*F*1	Precision	Recall	*F*1
Text	LR[a]	0.626	0.623	0.626	0.69	0.55	0.61	0.56	0.70	0.62
	SVM	0.626	0.616	0.626	0.68	0.55	0.61	0.55	0.68	0.62
	BiLSTM	0.605	0.587	0.604	0.62	0.73	0.67	0.58	0.45	0.51
	BERT	0.690	0.690	0.640	0.67	0.68	0.68	0.60	0.59	0.59
Image	CNN	0.615	0.464	0.615	0.69	0.55	0.61	0.56	0.55	0.61
	VGG19	0.682	0.464	0.681	0.74	0.64	0.69	0.62	0.72	0.67
	ViT	0.720	0.710	0.760	0.74	0.86	0.80	0.79	0.63	0.70
Multi-modal	EANN (Wang et al. 2018b)	0.690	0.720	0.690	0.75	0.58	0.65	0.62	**0.76**	0.69
	MVAE (Khattar et al. 2019)	0.670	0.660	0.670	0.70	0.69	0.69	0.63	0.64	0.63
	SpotFake (Singhal et al. 2019)	0.768	0.740	0.765	0.72	**0.92**	0.81	**0.85**	0.56	0.68
	CARMN (Song et al.2021)	0.727	0.690	0.732	0.70	0.88	0.78	0.78	0.54	0.64
	AMFD (Kumari and Ekbal 2021)	0.749	0.736	0.751	0.76	0.79	0.78	0.73	0.70	0.71
	FMFN (Wang et al. 2022b)	0.629	0.525	0.629	0.64	0.76	0.69	0.61	0.47	0.53
	FNR-WS[b]	0.770	0.757	0.774	0.74	0.90	0.81	0.83	0.62	0.71
	FNR-S[c]	**0.789**	**0.793**	**0.789**	**0.78**	0.85	**0.82**	0.79	0.71	**0.75**

Bold numbers Indicate the best performance

[a]Logistics regression

[b]Fake News Revealer (FNR) without considering similarity loss

[c]Fake News Revealer (FNR) with considering similarity loss

Table 8 Performance comparison for the Weibo data set

	Model name	Accuracy	AUC	*F*1 micro	Fake news			Real news		
					Precision	Recall	*F*1	Precision	Recall	*F*1
Text	LR[a]	0.712	0.687	0.712	0.71	0.80	0.75	0.71	0.59	0.65
	SVM	0.704	0.703	0.704	0.72	0.74	0.73	0.67	0.65	0.66
	BiLSTM	0.661	0.440	0.661	0.62	0.78	0.69	0.73	0.55	0.63
	BERT	0.810	0.830	0.810	0.81	0.81	0.81	0.81	0.82	0.81
Image	CNN	0.525	0.391	0.505	0.79	0.24	0.38	0.58	0.87	0.63
	VGG19	0.602	0.473	0.602	0.60	0.61	0.60	0.60	0.59	0.59
	ViT	0.680	0.720	0.680	0.67	0.69	0.68	0.68	0.68	0.67
Multi-modal	EANN (Wang et al. 2018b)	0.810	0.860	0.810	**0.89**	0.66	0.76	0.77	**0.93**	0.85
	MVAE (Khattar et al. 2019)	0.790	0.790	0.790	**0.89**	0.65	0.75	0.74	**0.93**	0.82
	SpotFake (Singhal et al. 2019)	0.864	0.897	0.860	0.87	0.92	**0.90**	0.81	0.70	0.75
	CARMN (Song et al. 2021)	0.844	0.895	0.850	0.86	**0.93**	0.89	0.81	0.66	0.73
	AMFD (Kumari and Ekbal 2021)	0.829	0.887	0.830	0.86	0.90	0.88	0.75	0.68	0.71
	FMFN (Wang et al. 2022b)	0.871	0.932	0.871	0.86	0.88	0.87	0.88	0.85	0.86
	FNR-WS[b]	0.869	0.898	0.869	0.89	0.85	0.87	0.85	0.89	0.87
	FNR-S[c]	**0.879**	**0.938**	**0.879**	**0.87**	0.89	0.88	**0.88**	0.87	**0.88**

Bold numbers Indicate the best performance

[a]Logistics regression

[b]Fake News Revealer (FNR) without considering similarity loss

[c]Fake News Revealer (FNR) with considering similarity loss

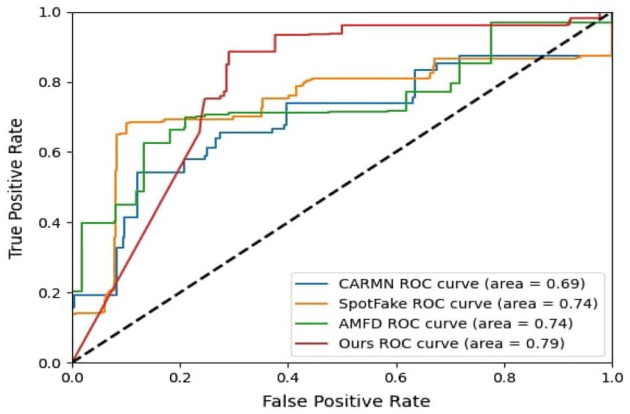


Fig. 10 ROC curve on Twitter data set

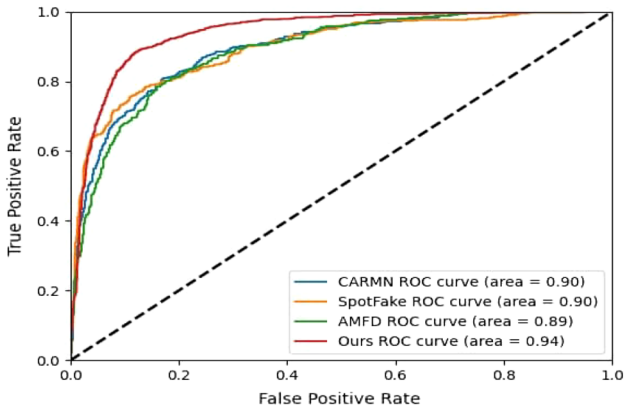


Fig. 11 ROC curve on Weibo data set

4.5.2 Results

BERT has outperformed the other text algorithms, according to the findings in these comparison Tables 7 and 8. It provides a more objective technique for identifying fake news using text alone and has suitably extracted contextual and lexical features. ViT also beats existing image-only fake news detection algorithms and extracts semantic and spatial features using a transformer-based neural network. Since our methodology enhances the AUC metric as shown in Figs. 10 and 11, it can be concluded that our classification model prioritizes fake instances over real ones and, therefore, can distinguish classes more effectively than other models.

The findings show that the existing multi-modal fake news detection models outperform single-modal approaches, proving the value of integrating data modals and extracting their inter-modal features. By extracting more meaningful information from text and images, Spotfake (Singhal et al. 2019) surpassed the competition. Additionally, CARMN (Song et al. 2021), and AMFD (Kumari and Ekbal 2021) offer enhanced attention-based processes. The Fake News Revealer (FNR) model consistently beats the opposition on various performance metrics. With our technique, each modality maintains its distinctive characteristics while smoothly incorporating similarities and complementing data from the other modalities.

5 Conclusion and future works

This paper presents a novel multi-modal framework for identifying fake news on social media that uses language and vision transformers and similarity measurements with a contrastive loss function. For detecting fake news, transformers such as BERT are commonly used to extract textual features. However, only a few works have utilized semantic and contextual features of the images extracted by the transformers. Thus, we use ViT to extract these features from the images. We have also considered the similarity between the image and the text with the least complexity and additional network. We have entered this similarity into the model with contrastive loss function and obtained excellent results. Previously, only the relationship between a news post's text and its image was considered. However, in social networks, the news posts are about a specific news event, and this aspect of social network news has not been addressed until now. Still, we have considered the similarity loss function among all news posts in our study. Consequently, we considered the relationship between several news posts and numerous experiments we have conducted to demonstrate the effectiveness of our method. Regarding identifying fake news, the proposed framework (FNR) performs better than cutting-edge techniques.

Further developments and improvements can be made to detect fake news on social media in future. For example, other modalities that news can incorporate, such as video, audio, and user information, can be employed. Furthermore, the network graph of users can be constructed and exploited to capitalize on the relationship between users and their shared news in addition to multimedia data. It is essential to build trust in the public to make fake news detection practical. Therefore, it is advisable to use trustworthy machine learning in this field. As a next step, we will use interpretable and explicable approaches to detect fake news and provide users with a reason for the label.

Author contributions BG did the conceptualization, formal analysis, Investigation, methodology, software and writing the original draft. MR did the conceptualization, formal analysis, Investigation, methodology, writing, review & editing. MFA did the supervision, conceptualization, formal analysis, review & editing. HRR did the supervision, project administration, conceptualization, methodology, validation, review & editing.

Declarations

Conflict of interest The authors declare no competing interests.

References

Akiba T, Sano S, Yanase T et al (2019) Optuna: a next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. ACM, pp 2623–2631

Antol S, Agrawal A, Lu J et al (2015) Vqa: visual question answering. In: Proceedings of the IEEE international conference on computer vision, pp 2425–2433

Bai Y, Mei J, Yuille AL et al (2021) Are transformers more robust than CNNs? In: Advances in neural information processing systems, pp 26831–26843. arXiv:2111.05464

Boididou C, Andreadou K, Papadopoulos S et al (2015) Verifying multimedia use at medieval 2015. MediaEval 3:7

Chen J, Jia C, Zheng H et al (2022) Is multi-modal necessarily better? Robustness evaluation of multi-modal fake news detection. arXiv:2206.08788

Devlin J, Chang MW, Lee K et al (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American Chapter of the Association for Computational Linguistics, vol 1. Association for Computational Linguistics, pp 4171–4186

Dosovitskiy A, Beyer L, Kolesnikov A et al (2021) An image is worth 16x16 words: transformers for image recognition at scale. Preprint at arXiv:2010.11929

Farajtabar M, Yang J, Ye X et al (2017) Fake news mitigation via point process-based intervention. In: International conference on machine learning. PMLR, pp 1097–1106

Fenn E, Ramsay N, Kantner J et al (2019) Nonprobative photos increase truth, like, and share judgments in a simulated social media environment. JARMAC 8:131–138

Fukui A, Park DH, Yang D et al (2016) Multimodal compact bilinear pooling for visual question answering and visual grounding. Preprint at arXiv:1606.01847

Gross D (2010) Survey: more Americans get news from internet than newspapers or radio. http://www.cnn.com/2010/TECH/03/01/social.network.news/index.html. Accessed 16 Jan 2020

Gupta M, Zhao P, Han J (2012) Evaluating event credibility on twitter. In: Proceedings of the 2012 SIAM international conference on data mining. SIAM, pp 153–164

He K, Zhang X, Ren S et al (2015) Deep residual learning for image recognition. Preprint at arXiv:1512.03385

He K, Zhang X, Ren S et al (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778

Hendrycks D, Gimpel K (2016) Gaussian error linear units (GELUs). Preprint at arXiv:1606.08415

Hinton GE, Krizhevsky A, Wang SD (2011) Transforming auto-encoders. In: International conference on artificial neural networks. Springer, pp 44–51

Jain V, Kaliyar RK, Goswami A et al (2022) AENeT: an attention-enabled neural architecture for fake news detection using contextual features. Neural Comput Appl 34(1):771–782

Jin Z, Cao J, Guo H, et al (2017) Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In: Proceedings of the 25th ACM international conference on multimedia. ACM, MM Í7, pp 795–816

Jwa H, Oh D, Park K et al (2019) exBAKE: automatic fake news detection model based on bidirectional encoder representations from transformers (BERT). Appl Sci 9(19):4062

Kaliyar RK, Goswami A, Narang P et al (2020) FNDNet—a deep convolutional neural network for fake news detection. Cogn Syst Res 61:32–44

Kaliyar RK, Goswami A, Narang P (2021a) EchoFakeD: improving fake news detection in social media with an efficient deep neural network. Neural Comput Appl 33(14):8597–8613

Kaliyar RK, Goswami A, Narang P (2021b) FakeBERT: fake news detection in social media with a BERT-based deep learning approach. Multimed Tools Appl 80(8):11765–11788

Khattar D, Goud JS, Gupta M et al (2019) Mvae: multimodal variational autoencoder for fake news detection. In: The world wide web conference. ACM, pp 2915–2921

Khosla P, Teterwak P, Wang C et al (2020) Supervised contrastive learning. Adv Neural Inf Process Syst 33:18661–18673

Kumari R, Ekbal A (2021) AMFB: attention based multimodal factorized bilinear pooling for multimodal fake news detection. Expert Syst Appl 184(115):412

Kwon S, Cha M, Jung K et al (2013) Prominent features of rumor propagation in online social media. In: 2013 IEEE 13th international conference on data mining. IEEE, pp 1103–1108

Liu Y, Wu YF (2018) Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In: Proceedings of the AAAI conference on artificial intelligence

Loshchilov I, Hutter F (2017) Decoupled weight decay regularization. Preprint at arXiv:1711.05101

Newman EJ, Garry M, Bernstein DM et al (2012) Nonprobative photographs (or words) inflate truthiness. Psychon Bull Rev 19:969–974

O'Brien N, Latessa S, Evangelopoulos G et al (2018) The language of fake news: opening the black-box of deep learning based detectors. In: Workshop on "AI for Social Good". NIPS

Palani B, Elango S, Viswanathan KV et al (2022) CB-Fake: a multimodal deep learning framework for automatic fake news detection using capsule neural network and BERT. Multimed Tools Appl 81:5587–5620

Peters M, Neumann M, Iyyer M, Zettlemoyer l et al (2018) Deep contextualized word representations. In: Proceedings of the 2018 conference of the North American Chapter of the Association for Computational Linguistics. Human Language Technologies, pp 2227–2237

Ping Tian D et al (2013) A review on image feature extraction and representation techniques. Int J Multimed Ubiquitous Eng 8(4):385–396

Qi P, Cao J, Yang T et al (2019) Exploiting multi-domain visual information for fake news detection. In: 2019 IEEE International Conference on Data Mining (ICDM), pp 518–527

Radford A, Kim JW, Hallacy C et al (2021) Learning transferable visual models from natural language supervision. In: International conference on machine learning. PMLR, pp 8748–8763

Ren S, He K, Girshick R, et al (2015) Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in neural information processing systems, p 28

Ruchansky N, Seo S, Liu Y (2017) Csi: a hybrid deep model for fake news detection. In: Proceedings of the 2017 ACM on conference on information and knowledge management, pp 797–806

Salama K (2021) Keras documentation: natural language image search with a dual encoder. https://keras.io/examples/nlp/nl_image_search/. Accessed 8 Nov 2021

Shu K, Sliva A, Wang S et al (2017) Fake news detection on social media: a data mining perspective. SIGKDD Explor Newsl 19:22–36

Shu K, Wang S, Lee D et al (2020) Disinformation, misinformation, and fake news in social media. Springer

Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. Preprint at arXiv:1409.1556

Singh B, De S, Zhang Y et al (2015) Layer-specific adaptive learning rates for deep networks. In: 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA). IEEE, pp 364–368

Singhal S, Shah RR, Chakraborty T et al (2019) Spotfake: a multi-modal framework for fake news detection. In: 2019 IEEE fifth international conference on multimedia big data (BigMM). IEEE, pp 39–47

Singhal S, Pandey T, Mrig S et al (2022) Leveraging intra and inter modality relationship for multimodal fake news detection. In: Companion Proceedings of the Web Conference, pp 726–734

Song C, Ning N, Zhang Y et al (2021) A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks. Inf Process Manag 58(102):437

Wang S, Chen Y, Zhuo J et al (2018a) Joint global and co-attentive representation learning for image-sentence retrieval. In: Proceedings of the 26th ACM international conference on Multimedia, pp 1398–1406

Wang Y, Ma F, Jin Z et al (2018b) 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. ACM, KDD 18, pp 849–857

Wang Y, Ma F, Wang H et al (2021) Multimodal emergent fake news detection via meta neural process networks. In: Proceedings of the 27th ACM SIGKDD conference on Knowledge Discovery & Data Mining. ACM, pp 3708–3716

Wang B, Feng Y, Xiong X et al (2022a) Multi-modal transformer using two-level visual features for fake news detection. Appl Intell 2022:1–15

Wang J, Mao H, Li H (2022b) FMFN: fine-grained multimodal fusion networks for fake news detection. Appl Sci 12(3):1093

Wolf T, Debut L, Sanh V et al (2020) Huggingface's transformers: state-of-the-art natural language processing. Preprint at arXiv:1910.03771

Wu B, Xu C, Dai X et al (2020) Visual transformers: token-based image representation and processing for computer vision. Preprint at arXiv:2006.03677

Yenter A, Verma A (2017) Deep CNN-LSTM with combined kernels from multiple branches. In: 2017 IEEE 8th annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON), pp 540–546