A Multi-Modal Prompt Learning Framework for Early Detection of Fake News

Weiqi Hu[1], **Ye Wang**[1,2,*], **Yan Jia**[1], **Qing Liao**[1], **Bin Zhou**[2]

[1]Harbin Institute of Technology, Shenzhen, Shenzhen, China
[2]National University of Defense Technology, Changsha, China
21S151073@stu.hit.edu.cn, wangye2020@hit.edu.cn, jiayan2020@hit.edu.cn, liaoqing@hit.edu.cn, binzhou@nudt.edu.cn

Abstract

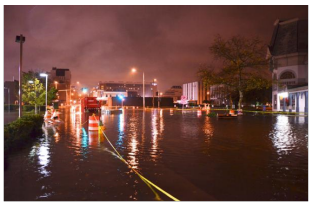
Information spreads quickly through social media platforms, especially fake news with negative or even malicious intentions. In recent years, psychological studies have found that explicit reminders of fake news would diminish its consequence. Therefore, it is crucial to identify their authenticity at an early stage to avoid serious consequences. However, existing methods for fake news detection either utilize auxiliary information including users' profiles and related events propagation networks or require sufficient and high-quality training data, which is not suitable for early fake news detection in real. An increasing number of social media news not only involves natural language content but also visual content such as images and videos, which give us a new view of fake news detection at an early stage by multi-modal data. In this paper, we propose a **M**ulti-modal **P**rompt **L**earning framework (MPL) based on the multi-modal pre-trained model CLIP for early detection of fake news. A learnable prompt module is developed to adaptively and efficiently generate prompt representations to boost the semantic context. MPL can be implemented in supervised or few-shot settings. Extensive experiments show that the proposed MPL obtains substantial performance and efficiency improvement for the early-stage fake news detection task. The results demonstrate that MPL performs considerably well compared to both the state-of-the-art supervised multi-modal models and the latest prompt-based few-shot multi-modal models. Especially, the high recall of fake news and the high precision of real news that MPL achieved compared to other baselines verify that it will better approach one of the motivations that providing early notification of "maybe real" or "maybe fake" with the release of the news.

Introduction

Social media information is multi-modal data consisting of short texts and visual data, which has become a popular way for people to receive and publish news. Moreover, due to the vivid visual perception, multi-modal information can provide an opportunity close to an immersive experience, which attracts readers to browse the news on social media platforms. In these pieces of information, fake news with negative or malicious purposes spreads rapidly using this dramatic visual context, which may cause serious consequences, especially in political situations and social disorder. For example, games between fake and real news during the recent US presidential election, the disinformation abound in the Russia-Ukraine conflict, and the intense debate and fear about wearing masks and vaccination caused by fake statements during the COVID-19 pandemic. Therefore, more and more research attention has been drawn to this task. There are many existing works proposing non-trivial solutions that consider various features, including advanced textual representation(Cheng, Nazarian, and Bogdan 2020), multi-modal information(Khattar et al. 2019), external knowledge(Hu et al. 2021), social structural context(Lu and Li 2020), and domain-specific features(Silva et al. 2021). They mostly require high-quality labelled data to train well-designed detection models, which means that computational and time consumption cannot be ignored.

Text: A powerful explosion heard from miles away happened at a chemical plant in Centerville, Louisiana #ColumbianChemicals

(a) A fake example



Text: Still experiencing signifiant flooding here in Downtown Norfolk even with tide going out #Sandy #hrsandy

(b) A real example

Figure 1: Multi-modal social message examples from Twitter dataset.

Figure 1 shows some multi-modal information collected on the real social media platform Twitter, which is related to our health and daily life. To readers who browse a large amount of news on these online platforms, it is difficult to distinguish whether a piece of news is real or fake through the brief textual content illustrated with a couple of photographs providing auxiliary context. However, the primary task of early detection of fake news is warning in time. Recent psychological studies have confirmed a situation where alerting people to misinformation helps correct their mem-

*Correspondence should be addressed to Ye Wang (wangye2020@hit.edu.cn)

ory and beliefs about fake news, reducing its negative impact in subsequent dissemination(Kemp, Loaiza, and Wahlheim 2022; Wahlheim, Alexander, and Peske 2020; Ecker, Hogan, and Lewandowsky 2017). For example, if there is a warning label of "maybe fake" displayed when people first receive the news in Figure 1a, public panic and concern will be reduced. If there is a warning label of "maybe real" displayed when people receive the news in Figure 1b, people affected by flooding may get actual help.

Therefore, to detect potential fake news as early as possible, inspired by the idea of integrating multi-modal relevant features, we propose a general framework for early detection of fake news in a multi-modal setting, typically with visual and textual features. As we emphasize, our goal is to alert people as soon as possible once a piece of news is published, auxiliary information such as the profile of users, event propagation networks and constantly changing social trends are not suitable here, although many of them are favoured by some latest works(Kang et al. 2021; Bian et al. 2020; He et al. 2021; Dou et al. 2021). Some multi-modal methods have already discussed the variety of connections between features from different modalities. Spotfake (Singhal et al. 2019) and Spotfake+ (Singhal et al. 2020) extracted visual features with a VGG encoder and encoded textual features with a text encoder, for example, BERT and XLNET, followed by concatenating them for classification. Other works follow a similar process but with improved image and text encoders as well as complicated fusion strategies (Li et al. 2022a). These methods all utilize pre-trained models for fine-tuning training, i.e., designing training targets according to fake news detection and adjusting the pre-trained model backbone. However, as the size of pre-trained models increases, the requirements for hardware and data increase, as well as practical costs for fine-tuning. Moreover, these methods assume that sufficient and high-quality labelled training data is available, while there are a large number of incomplete, imprecise, and inaccurate data in realistic scenarios, especially in early fake news detection. Fine-tuning methods can be ineffective in this situation. Therefore, it is necessary to explore fake news detection with only a small amount of labelled training data, i.e., few-shot learning.

In recent years, prompt learning has become a new paradigm for pre-trained models, exhibiting good performance in many NLP and multi-modal tasks, especially for few-shot training. Prompt learning only needs some templates to reconstruct the task into the same task as the pre-training, requiring very little parameter and training time. Inspired by this, in order to improve multi-modal fake news detection at an early stage, here we propose a **M**ulti-modal **P**rompt **L**earning (MPL) framework for fake news detection based on the multi-modal pre-trained model CLIP (Radford et al. 2021). To evaluate the performance of the proposed MPL framework, we use three publicly available real datasets: Twitter(Jin et al. 2017), Politifact(Shu et al. 2020), and GossipCop(Shu et al. 2020). The main contributions of this article can be summarized as follows:

- We propose a **M**ulti-modal **P**rompt **L**earning framework (MPL) for fake news detection based on the multi-modal pre-trained model CLIP. MPL uses a multi-modal feature fusion module to fully integrate the semantic context of different modalities and classifies by calculating the similarity distance between multi-modal features and category features.
- In order to leverage pre-trained knowledge as much as possible to assist multi-modal fake news detection, we design an adaptive prompt learning module generating continuous prompt representations to automate prompt engineering.
- Our model is evaluated on three publicly available multi-modal fake news detection benchmarks. MPL obtains substantial performance and efficiency improvement for the early-stage fake news detection task in both few-shot and supervised settings.

Related Work

Fake news detection on social media has been studied for many years. According to previous research, we define fake news as fake information spread under the guise of real news through news media or social media platforms such as the Internet, usually for political or economic gain. In the past, fake information was easily confused with other similar concepts, but recent research has sought to differentiate them, such as misinformation (Song et al. 2021; Jiang et al. 2021) and disinformation (Li et al. 2022c) these two different concepts. Misinformation refers to information with incorrect content caused by cognitive errors or biases, while disinformation is intentionally fabricated, and both are not limited to the fields of news media and social platforms (Meel and Vishwakarma 2020). The MPL proposed in this article focuses on fake news detection, but can also be extended to misinformation and disinformation detection.

Based on the availability of sufficient labelled training data, existing methods for fake news detection can be divided into two kinds: supervised methods in the case of sufficient data and weakly supervised methods in the case of few-shot learning. As this article focuses on the early detection of fake news, methods that rely on rich and complex social context and user information are not considered. In this section, we briefly review existing methods for detecting fake news, namely supervised multi-modal fake news detection methods that utilize both text and visual modalities of news and heavily rely on high-quality labelled data, as well as weakly supervised fake news detection methods that can make predictions in few-shot settings.

Multi-modal Fake News Detection

Previous research on fake news detection has shown that both text and image information are effective in fake news detection. Many methods combine textual and visual information for fake news detection. With the rise of deep neural networks and pre-trained models, there have been many powerful feature extractors, such as the text feature extractor Bert (Devlin et al. 2018), Transformers (Vaswani et al. 2017), and the visual feature extractors VGG (Simonyan and

Zisserman 2014) and Res-Net (He et al. 2016). Many studies have utilized visual feature extractors to extract visual information and text feature extractors to extract text features, and then combine and fuse the visual and text information to detect fake news (Singhal et al. 2019, 2020). EANN (Wang et al. 2018) designed an auxiliary task, event discrimination, to measure the differences between different events and further learn the invariant features of news events. This helps fake news detection by better understanding multi-modal information through the auxiliary task. MVAE (Khattar et al. 2019) proposed an end-to-end multi-modal variational autoencoder, using a bimodal variational autoencoder and binary classifier for fake news detection. HMCAN (Qian et al. 2021) sent the obtained text and image representations into a multi-modal contextual attention network to fuse intra-modality and inter-modality relationships and designed a hierarchical encoding network to capture rich semantic information in fake news. MCAN (Wu et al. 2021) extracted spatial-domain and frequency-domain features from images and textual features from the text, stacked multiple co-attention layers together to fuse multi-modal features and learned inter-dependencies between multiple modalities.

A large number of scholars believe that if the image content of news does not match the text content, it indicates that the news is fake. Based on this hypothesis, the image information and text information of the news is encoded and then the similarity between the two is calculated. If the similarity is high, it indicates that the textual information and visual information of the news match, and it is real news; if the similarity is low, it indicates that the textual information and visual information of the news do not match, and it is fake news. SAFE (Zhou, Wu, and Zafarani 2020) utilized image2text model to convert visual information into textual information, then mapped the textual and visual information into the same vector space through a fully connected layer. They then compared the similarity between the visual and textual information to detect fake news. MCNN (Xue et al. 2021) used BERT to encode textual information and Res-Net to encode visual information, calculating the similarity between the two to determine if the text and image are consistent.

Knowledge graphs contain a wealth of external knowledge which includes rich semantic information that can help us better understand news content. External knowledge also includes objective facts which can be compared to news content to detect fake news. KMGCN (Wang et al. 2020) constructed a multi-modal graph structure containing textual information, visual information, and knowledge concept. They then fused information from each modality to achieve great discrimination performance.

Few-shot Fake News Detection

Due to the high cost of data annotation in terms of manpower and resources, there are often incomplete, imprecise, and inaccurate annotations in real-world situations. Therefore, models need to be able to detect fake news in few-shot settings. Many models rely on graph structures (Guacho et al. 2018; Benamira et al. 2019) or pseudo-labelling (Helwe et al. 2019; Mansouri, Naderan-Tahan, and Rashti 2020; Konkobo et al. 2020; Meel and Vishwakarma 2021; Li et al. 2022b) to utilize incomplete labelled data for fake news detection. KPL (Jiang et al. 2022) proposed an innovative method of guiding fake news detection through prompt learning using a pre-trained language model. However, this model only relied on single-modal text for fake news detection and did not integrate visual information and textual information for multi-modal fake news detection. Building on existing work, SAMPLE (Jiang et al. 2023) recently proposed a multi-modal model for few-shot fake news detection, which combined the multi-modal features generated by the CLIP model with the text representation of the pre-trained language model to aid prompt learning. However, this method mainly aimed at the pre-trained language model Roberta for prompt learning, and CLIP, a pre-trained multi-modal model, was mainly used to supplement multi-modal semantic information. This limits the effectiveness of CLIP and does not make full use of the multi-modal knowledge it has learned during the multi-modal pre-training. Considering the progress and limitations of existing work, this article proposes the MPL framework for multi-modal fake news detection at an early stage, based on the multi-modal pre-trained model CLIP and designed with continuous prompts for prompt learning.

Preliminary

Prompt Learning

Fine-tuning and prompt learning are two typical paradigms for pre-trained models. Fine-tuning trains pre-trained models to adapt to downstream tasks, while prompt learning processes input information according to specific templates, which can reconstruct tasks into a form that can be better utilized by the pre-trained model. For example, we're going to use BERT to detect fake news in text, that is, outputting whether the information is "real" or "fake". The traditional fine-tuning paradigm trains the model parameters with a large amount of data with labels, uses the fine-tuned model to extract information features, and then inputs them into the classifier for classification. The prompt learning does not train the pre-trained model, that is, freezing model parameters. Prompt learning processes the input by adding a prompt sentence before or after the text, such as "This information is [MASK]", then allows BERT to fill in [MASK] with "real" or "fake" to classify the information. This kind of fill-in-the-blank prompt makes the task more closely related to BERT's pre-training method, that is, Masked Language Model (MLM). Prompt learning can fully utilize the knowledge learned by the model during pre-training, significantly reducing training costs, and performing well even in few-shot settings.

Prompt learning can be divided into manually designed prompts and automatically learned prompts (Liu et al. 2023). The former manually develops prompts for different tasks and datasets, but requires extra manpower and knowledge, and the results are not stable as a slight difference in a word can cause significant fluctuations in the results. The latter allows the model to automatically learn suitable prompts and can be divided into discrete prompts and continuous

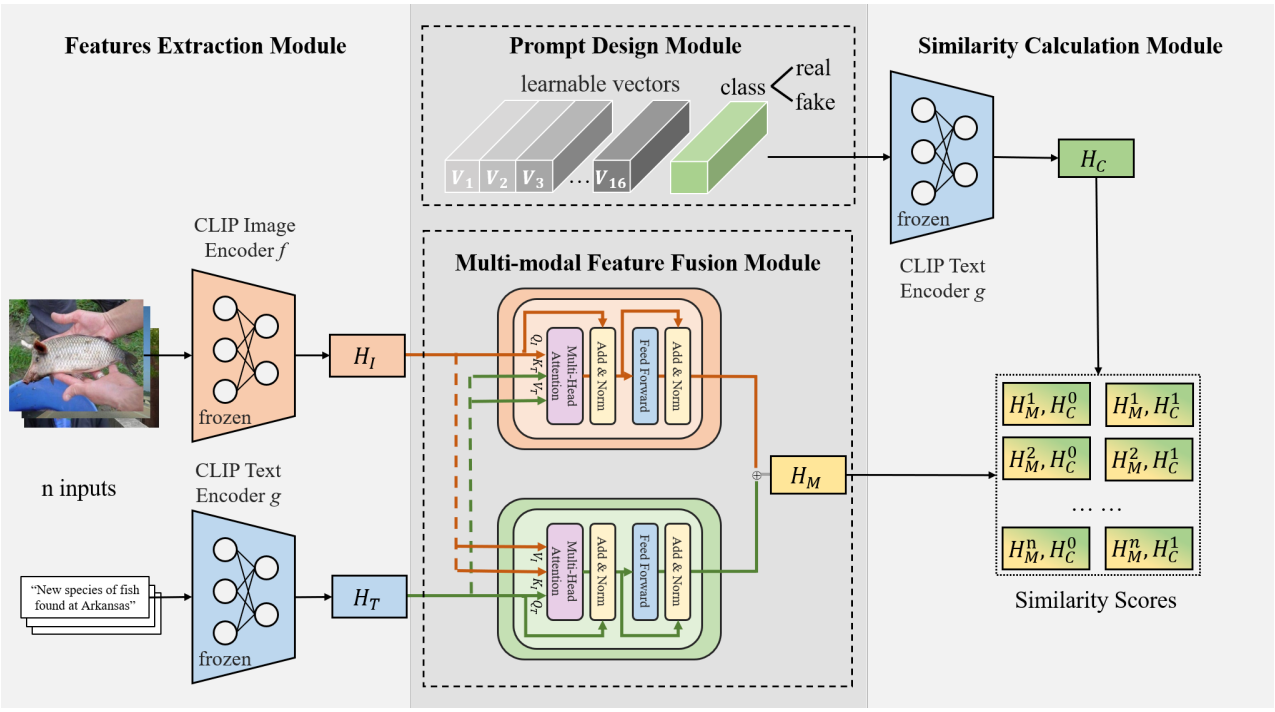


Figure 2: Overview of **M**ulti-modal **P**rompt **L**earning for fake news detection.

prompts. Discrete prompts refer to prompts generated by natural language, so the search space is discrete. Continuous prompts remove the constraints of natural language and directly search in a continuous embedding space, which makes learning prompts a sequence of vectors rather than a sentence. These learned prompts are more flexible, but because they cannot be mapped to specific natural text, they lack intuitive interpretability. Nevertheless, continuous prompts have achieved great results in many natural text classification and image classification tasks, proving the effectiveness of this method in classification tasks.

CLIP: A Multi-modal Pre-trained Model

CLIP (Radford et al. 2021) is a large-scale multi-modal pre-trained vision-language model based on contrastive learning. Here, we briefly introduce CLIP to better explain our work. Unlike the commonly used label-based representation learning methods in CV, CLIP's training data consists of image-text pairs, i.e., an image and its corresponding text description. In order to learn different concepts and make the model more applicable to downstream tasks, the CLIP team collected a large training dataset consisting of 400 million image-text pairs. CLIP consists of two encoders, Image Encoder and Text Encoder. The Image Encoder is used to extract features from the image, including different sizes of ResNet or Vision Transformer models. The Text Encoder is used to extract features from the text, using the Text Transformer model commonly used in NLP. CLIP learns by comparing the extracted text and image features. For a training batch containing N image-text pairs, the N text features and N image features are combined and the CLIP model predicts the cosine similarity between each of the N^2 possible image-text pairs. There are N positive samples, which are the real image-text pairs, and the remaining $N^2 - N$ image-text pairs are negative samples. The training objective of CLIP is to maximize the similarity of the positive samples and minimize the similarity of the negative samples. The pre-trained encoders can only extract the features of image-text pairs, so CLIP uses prompts to transfer the pre-trained model to different downstream tasks. The common practice is to pre-train and then fine-tune, but CLIP can directly achieve zero-shot image classification without any training data. First, a prompt is constructed, i.e., a descriptive text for each category is constructed based on the classification label: "A photo of $[CLASS]$", and the text is sent to the Text Encoder to extract the corresponding text features. If there are N categories, N text features will be obtained. Then, the image to be predicted is sent to the Image Encoder to extract image features, and the cosine similarity is calculated with the N text features so that the image can be classified based on the highest similarity. CLIP has become a promising method for visual representation learning and image classification recognition. In this article, multi-modal news is classified based on CLIP. To align with the pre-trained method of CLIP's contrastive learning and improve the prompts, we design contrastive learning with continuous prompts and multi-modal representation to achieve multi-modal fake news detection.

Problem Formulation and Notation

Fake news detection task is regarded as a classification problem, binary classification specifically in our work. We con-

sider two modalities, text and image which is the most prevalent information carrier across online social media platforms. A piece of multi-modal news is notated as a pair $x = (T, I)$, where T is the main text of the news that illustrated a point with an impressive image I. The goal of multi-modal fake news detection is to assign a label $y \in \{0, 1\}$ for the input news, where 0 stands for real news and 1 stands for fake news.

Methods

We propose the Multi-modal Prompt Learning framework (MPL) for fake news detection based on CLIP. The proposed framework is illustrated in Figure 2. MPL uses a co-attention layer to fuse multiple modalities and employs a learnable prompt to automate prompt engineering, thereby fully utilizing both multi-modal information and the knowledge of the pre-trained model. Specifically, for a multi-modal news article x, consisting of image I and text T, we keep the parameters of CLIP frozen. As Eq. (1) shows, we use the pre-trained CLIP Image Encoder $f(\bullet)$ to extract visual features H_I from the image, and the pre-trained CLIP Text Encoder $g(\bullet)$ to extract textual features H_T from the text.

$$H_I = f(I), H_T = g(T) \tag{1}$$

The multi-modal features of x, H_M, are obtained by fusing the two features using a multi-modal feature fusion module. Meanwhile, a set of learnable vectors is used to replace the conventional manually designed prompts and is input to the pre-trained CLIP Text Encoder to obtain category features, H_C. The similarity distance between the multi-modal features H_M and category features H_C is used to classify the multi-modal news x.

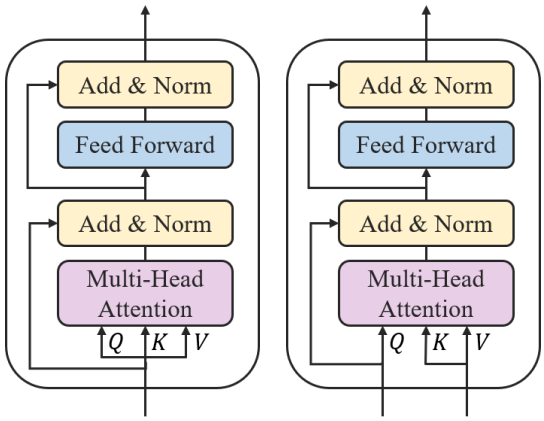
Multi-modal Feature Fusion Module

We design a multi-modal fusion module based on the co-attention block (Lu et al. 2019). In traditional self-attention, as shown in Figure 3a, the queries, keys, and values for self-attention all come from the same input. However, in co-attention, as shown in Figure 3b, the queries come from one input, while the keys and values come from another input. And the residual connection is only applied to the queries.

As shown in Figure 3b, each co-attention module consists of successively connected Multi-Heads Attention layer, Add & Norm layer, fully connected Feed Forward Network layer and Add & Norm layer. The $d \times 1$ dimension V, Q, and K from different inputs are input into the Multi-Heads Attention layer. The attention matrix is calculated based on K and Q, and V is then applied to the attention matrix to generate the output. The entire process is as Eq. (2) shows.

$$\begin{aligned} &MA(Q, K, V) = hW^O \\ &h = h_1 \oplus h_2 \oplus \ldots \oplus h_m \\ &h_i = A(Q_i, K_i, V_i) = softmax\left(\frac{Q_i K_i^T}{\sqrt{d_h}}\right) V_i \\ &Q_i = QW_i^Q, K_i = KW_i^K, V_i = VW_i^V \end{aligned} \tag{2}$$

where $W^O \in \mathbb{R}^{md_h \times 1}$, $\oplus$ denotes the concatenation of vectors,W_i^Q,W_i^K,$W_i^V \in \mathbb{R}^{1 \times d_h}$are the projection matrices for the i-th head, $d_h = d/m$ is the dimension of the output feature of each head.



(a) Self-attention block (b) Co-attention block

Figure 3: Illustration of the self-attention block and the co-attention block.

The fully connected Feed Forward Network consists of two linear transformations with a ReLU activation function in between. The calculation process is as Eq. (3) shows.

$$FFN(X) = max(0, xW_1)W_2 \tag{3}$$

where the dimension of input and output is $d \times 1$, and the inner-layer dimensionality is d_{ff}, W_1 and W_2 are linear transformation matrixs.

In order to better fuse visual feature H_I and textual feature H_T, we designs two parallel co-attention layers.

First, as Eq. (4) shows, the visual feature H_I is used as Q, the textual feature H_T is used as K and V, and a visual feature with textual information H_I^T is generated.

$$\begin{aligned} H_I' &= H_I + MA(H_I, H_T, H_T) \\ H_I^T &= H_I' + FFN(H_I') \end{aligned} \tag{4}$$

Then, as Eq. (5) shows, the textual feature H_T is used as Q, the visual feature H_I is used as K and V, and a textual feature with visual information H_T^I is generated.

$$\begin{aligned} H_T' &= H_T + MA(H_T, H_I, H_I) \\ H_T^I &= H_T' + FFN(H_T') \end{aligned} \tag{5}$$

Finally, as Eq. (6) shows, H_I^T and H_T^I are input into the fully connected layer to generate the multi-modal representation H_M that fully fuses textual and visual information for subsequent calculations.

$$H_M = \left(H_I^T \oplus H_T^I\right) W \tag{6}$$

where $W \in \mathbb{R}^{2d \times d}$ is the projection matrice.

Prompt Design Module

CLIP can be used for few-shot learning with manually-designed prompts, but manually-designed prompts require

extra knowledge and manpower, and performance is often limited by the quality of the prompts, resulting in limited success. The accuracy result may be very sensitive to certain words in manually-designed prompts, for example, "$A\ photo\ of\ a\ [CLASS]$." is better than "$A\ photo\ of\ [CLASS]$." in most cases.

In order to reduce the impact of manually-designed prompts on the results and better utilize the knowledge in the pre-trained model CLIP, we refer to CoOP (Zhou et al. 2022) and design a learnable prompt. Specifically, we use a set of learnable vectors to replace the originally manually-designed prompts. The number of learnable vectors will also affect the experimental results. We test in the experiment, and finally choose 16 learnable vectors to achieve better and more stable performance. As Eq. (7) shows, the learnable vectors are concatenated with the class token, i.e., "real" or "fake".

$$p = [V_1]\,[V_2]\,\ldots\,[V_{16}]\,[CLASS] \tag{7}$$

where each $[V_i]\,(i \in \{1, 2, \ldots, 16\})$ is a vector with the same dimension as word embeddings, i.e., 768 for CLIP, and these vectors are automatically updated during training.

The learnable prompts p is then passed through the pre-trained CLIP Text Encoder $g(\bullet)$ to obtain the category embedding H_C for classification.

$$H_C^i = g(p_i) \tag{8}$$

where the $[CLASS]$ within each prompt p_i is replaced by the corresponding word-embedding vector of the i-th class name, $i \in \{0, 1\}$.

We calculate the distance between H_M and H_C, and the probability of news x belonging to category i is computed as Eq. (9) shows.

$$p(y = i|x) = \frac{exp\left(cos\left(H_M, H_C^i\right)/\tau\right)}{\sum_{j=1}^{K} exp\left(cos\left(H_M, H_C^j\right)/\tau\right)} \tag{9}$$

where τ is a temperature parameter learned by CLIP and $\cos(\bullet, \bullet)$ denotes cosine similarity.

As Eq. (10) shows, we use the binary cross entropy loss as the loss function to update the parameters of multi-modal feature fusion module and the learnable vector.

$$L = -\left(l \cdot \log p(y = 1|x) + (1 - l) \cdot \log p(y = 0|x)\right) \tag{10}$$

where l is the ground truth.

Finally, we predict the label $y \in \{0, 1\}$ based on the probability $p(y = i|x)$, where 0 stands for real news and 1 stands for fake news.

$$y = \arg\max_i p(y = i|x) \tag{11}$$

Experiments

To validate the effectiveness of our method, we conduct experiments on three benchmark datasets and compare the results with supervised fully-trained multi-modal fake news detection methods and the latest prompt-based few-shot multi-modal fake news detection method in both supervised and few-shot settings. In this section, we first introduce the specific implementation details of the experiments. We also introduce the benchmark datasets and baselines used in experiment. Then we show the experimental results compared with the supervised and few-shot multi-modal fake news detection methods, respectively. Finally, we provide a detailed analysis and discussion to further explain our proposed method.

Experimental Setup

We use the pre-trained CLIP (ViT-L/14@336px) model as both the Text Encoder and Image Encoder and keep all its parameters frozen. The learnable vectors of continuous prompt is randomly initialized by sampling from a zero-mean Gaussian distribution with a standard deviation of 0.02. In the multi-modal feature fusion module, we set $d = 256$,$m = 4$, and $d_{ff} = 512$. We use an SGD optimizer with a learning rate of 0.001 to optimize the model parameters during training. Our model is trained for 20 epochs, and we select the checkpoint with the best validation performance for testing. To fully demonstrate the effectiveness of our method, we conduct two sets of comparative experiments with supervised fully-trained and prompt-based few-shot multi-modal fake news detection methods in both supervised and few-shot settings.

Supervised Settings For comparison with supervised multi-modal fake news detection methods, we follow the existing methods(Jin et al. 2017; Shu et al. 2020) and divide the datasets into training and testing sets in an 8:2 ratio. In the case of sufficient data, we train our model and baselines using all the training data, whereas, in the case of few-shot, we train our model on a small amount of data sampled from the training set. Specifically, we sample 16 instances from each category for training. As the few-shot training set has a significant impact on the performance of the model, we repeat data sampling with different random seeds five times and take the average score of the five experiments, excluding the highest and lowest scores, as the result of the few-shot experiment.

Few-shot Settings To further demonstrate the advantages of our method in the few-shot settings, we compare it with the state-of-the-art prompt-based few-shot fake news detection method. We randomly sample a small number of instances from the dataset for training, i.e., we sample k instances from each category, where $k \in [2, 4, 8, 16]$, and use the remaining instances for testing. In addition, we create a validation set with the same size as the training set for model selection. To reduce the influence of the training and validation sets on the model performance, we repeat data sampling with five random seeds and take the average score by excluding the highest and lowest scores as the experimental result.

Datasets

We use three benchmark fake news detection datasets with multiple modalities to evaluate the performance of our method, namely, Twitter, GossipCop, and PolitiFact. These three datasets are real datasets collected from multiple social platforms. The Twitter dataset (Boididou et al. 2015, 2018) consists of tweets that include text information, visual information, and social context information. The Poli-

tiFact and GossipCop datasets are two English datasets collected from the politics and entertainment domains, respectively, in the FakeNewsNet repository(Shu, Wang, and Liu 2017; Shu et al. 2017, 2020). The PolitiFact dataset is a dataset about political news, which experts identify as real or fake news. Meanwhile, GossipCop tells entertainment stories with scores ranging from 0 to 10, and the authors of the FakeNewsNet consider scores below 5 to be fake news. To reduce redundancy, we only keep the most relevant image for news with multiple images, which is calculated by the pre-trained CLIP model based on cosine similarity between text and image. We also exclude news without images or with invalid image URLs. Specific statistics for each dataset are shown in Table 1.

	Twitter	PolitiFact	GossipCop
# of fake news	9795	164	2,540
# of real news	4482	319	10,150
# of images	477	483	12,690

Table 1: Statistics of three datasets.

Baselines

As mentioned above, we conduct two sets of comparative experiments. In order to ensure fairness, we use the original indicators of these methods for comparison. In comparison with supervised fully-trained multi-modal fake news detection methods, we select seven classic fully-trained multi-modal models as baselines:

- The EANN (Wang et al. 2018) uses event recognizers to capture news event information and extract news features unrelated to events to aid in detecting fake news.
- The MVAE (Khattar et al. 2019) uses a variational autoencoder coupled with a binary classifier to learn shared representations of text and images.
- The SpotFake (Singhal et al. 2019)utilizes VGG and BERT to extract image and text features respectively and concatenate them for classification.
- The SAFE (Zhou, Wu, and Zafarani 2020) extracts multi-modal (text and visual) features of news content and their relationships using a similarity-aware multi-modal approach to detect fake news.
- The MCAN (Wu et al. 2021)uses frequency-domain and spatial-domain features, stacking multiple co-attention layers to fuse multi-modal features.
- The LIIMR (Singhal et al. 2022) recognizes and suppresses information from weaker modalities, extracting relevant information from stronger modalities for each sample.
- The CAFE (Chen et al. 2022)proposes an ambiguity-aware multi-modal fake news detection method to adaptively aggregate unimodal features and cross-modal correlations.

Since there is a few number research on few-shot fake news detection, and most of them are basically unimodal, we compare our model with a state-of-the-art prompt-based few-shot multi-modal fake news detection method:

- The SAMPLE (Jiang et al. 2023)leverages the multi-modal features generated by the CLIP model, fusing text representations from the pre-trained language model RoBERTa, and utilizes the soft verbalizer to assist prompt learning for detecting fake news. In addition, this method combines the manually-designed prompts and the learnable continuous prompts, and achieves good performance in the case of few-shot.

To evaluate the performance of our model and baselines, we select Accuracy, Precision, Recall, and F1 as evaluation metrics. Accuracy is the proportion of correct predictions in the whole prediction, which can visually show the performance of the model. The fake-Precision is the probability that the sample predicted to be fake news is actually fake news, representing the ability to correctly classify fake news. The fake-Recall represents how much fake news we can find, and the higher fake-Recall, the more fake news we find, which is important for early detection. It's the same for the real-Precision and real-Recall. The F1 balances Precision and Recall to maximize both. Since the SAMPLE method only provides Accuracy and macro-F1 metrics, we also adopt these two metrics when comparing with the SAMPLE method in few-shot settings, where macro-F1 is the average of fake-F1 and real-F1.

Main Results

Comparison with Supervised Methods In comparison with supervised fully-trained multi-modal fake news detection methods, Table 2 shows the results of our model under supervised and few-shot settings. MPL-Full refers to using all training set data for training, while MPL-16 refers to training using only 16 instances per category.

Compared to other models, MPL-Full performs best when data is sufficient, achieves the best Accuracy on all three datasets, and F1 is in the top-3 for both real and fake news. The improvement is particularly pronounced in the PolitiFact dataset, where almost all indicators rise by about 5 percent. More importantly, the CLIP Encoder used in MPL keeps the parameters frozen, only the parameters of the Multi-modal Feature Fusion Module and the Prompt Design Module will be updated during the training. And training parameters of MPL are much smaller than those of other models, the required training costs and time are also lower. Under the training with less parameters and less time, MPL-Full still achieve much better results than other models, which fully proved the effectiveness of MPL.

In addition, MPL-16 uses only 16 instances per category and achieves good results. It is not much worse than SOTA on Twitter and PolitiFact datasets, and even surpasses several classical models that used all training data, and its performance on the Twitter dataset is second only to the most advanced LIIMR model. This proves that MPL is advanced in the case of few-shot, and only needs to learn from a small amount of annotated data to achieve a good classification effect. On the GossipCop dataset, while the Accuracy of MPL-16 is not satisfactory, both fake-Recall and real-Precision are

Dataset	Method	Acc	fake-P	fake-R	fake-F1	real-P	real-R	real-F1
Twitter	EANN	0.715	0.822	0.638	0.719	-	-	-
	MVAE	0.745	0.801	0.719	0.785	0.689	0.777	0.730
	SAFE	0.766	0.777	0.795	0.786	0.752	0.431	0.742
	MCAN	0.809	**0.889**	0.765	0.822†	0.732	0.871†	0.795
	LIIMR	0.831†	0.836†	0.832†	**0.830**	-	-	-
	MPL-16	0.829	0.778	0.753	0.761	0.863†	**0.875**	**0.867**
	MPL-Full	**0.841**	0.729	**0.906**	0.808	**0.936**	0.804	0.865†
PolitiFact	MVAE	0.726	0.761	0.678	0.717	-	-	-
	SpotFake	0.770	0.753	0.795	0.770	-	-	-
	SAFE	0.874†	**0.851**	0.830†	0.840†	0.889	0.903	0.896
	CAFE	0.864	0.724	0.778	0.750	0.895	0.919†	0.907†
	MPL-16	0.856	0.746	0.736	0.740	0.898†	0.902	0.900
	MPL-Full	**0.923**	0.839†	**0.897**	**0.867**	**0.959**	**0.933**	**0.946**
GossipCop	MVAE	0.782	**0.802**	**0.751**	**0.776**	-	-	-
	SpotFake	0.856	-	-	-	-	-	-
	SAFE	0.838	0.758†	0.558	0.643†	0.857	0.937	0.895
	CAFE	0.867†	0.732	0.490	0.587	0.887	**0.957**	0.921†
	MPL-16	0.620	0.294	0.688†	0.411	0.891†	0.603	0.718
	MPL-Full	**0.869**	0.698	0.565	0.625	**0.901**	0.942†	**0.921**

Table 2: The results of supervised methods in supervised settings. The best score is in bold and the second best score is marked with †.

Dataset	Model	Few-shot (ACC/macro-F1)							
		k=2		k=4		k=8		k=16	
PolitiFact	SAMPLE	0.56	0.47	0.61	0.56	0.66	0.62	0.70	0.67
	MPL	**0.69**	**0.67**	**0.71**	**0.68**	**0.73**	**0.71**	**0.82**	**0.80**
GossipCop	SAMPLE	**0.53**	0.44	0.56	0.47	0.54	0.52	0.60	0.54
	MPL	0.52	**0.46**	**0.57**	**0.48**	**0.64**	**0.55**	**0.65**	**0.55**

Table 3: The results of few-shot methods in few-shot settings.

high, indicating that more fake news can be found and that the real news found are mostly credible, which is very meaningful for early few-shot fake news detection.

Comparison with Few-shot Methods To further demonstrate the advantage of our model in few-shot settings, we compare it with latest prompt-based few-shot multi-modal fake news detection method SAMPLE on PolitiFact and GossipCop datasets. We train the model using k samples per category ($k \in [2, 4, 8, 16]$) and use the same amount of data as validation, with the rest used as the test set. The results of the experiment are shown in Table 3.

On the PolitiFact dataset, all the results of MPL are optimal in all experiments at k settings, and the Accuracy and macro-F1 are greatly improved compared with SAMPLE, which proves the superiority of the method in the few-shot settings. On the GossipCop dataset, MPL achieves better results at the settings of $k = 4$, $k = 8$, and $k = 16$. At the setting of $k = 2$, the Accuracy of MPL does not exceed SAMPLE, but macro-F1 has a slight improvement.

Overall, compared with SAMPLE, MPL achieves better results on both datasets, and the improvement on the PolitiFact dataset is more obvious and stable, with all indicators improving by about 10%. At the setting of $k = 2$, the macro-F1 on the PolitiFact dataset is even improved by 20%, while the improvement on the GossipCop dataset is not significant, with only about 1% performance improvement. This result may be attributed to the fact that GossipCop presents a more complex semantic context than PolitiFact, which consists of celebrity gossip stories and is also larger in size. In scenarios where there is extreme lack of data, such as the setting of $k = 2$, SAMPLE used two pre-trained models to provide more semantic information and knowledge, and achieved better results. This also shows that there is still room for improvement in the MPL, that is, to use more knowledge information, whether it is from the pre-trained model or external knowledge bases or knowledge graphs.

Ablation Study

We conduct ablation experiments to verify the effectiveness of the components we proposed. We conduct multiple ablation experiments on three datasets in the 16-shot setting. For w/o image, we remove the visual feature H_I and only use text feature H_T and category feature H_C to calculate similarity. For w/o text, we remove text feature H_T and only use visual feature H_I and category feature H_C to calculate similarity. For w/o fusion, we remove the multi-modal feature fusion module and directly concatenate visual feature H_I and text feature H_T. For w/o learnable prompt, we remove the learnable prompt vectors and use a manually-designed

Model	Dataset(ACC/fake-F1)					
	Twitter		PolitiFact		GossipCop	
w/o image	0.584	0.578	0.631	0.538	0.469	0.371
w/o text	0.646	0.702	0.753	0.632	0.464	0.369
w/o fusion	0.798	0.759	0.612	0.536	0.485	0.371
w/o learnable prompt	0.793	0.780	0.689	0.530	0.580	0.397
w/o similarity	0.797	0.693	0.596	0.562	0.464	0.385
w/o frozen	0.668	0.778	0.420	0.455	0.525	0.387
MPL-16	**0.830**	**0.789**	**0.801**	**0.668**	**0.581**	**0.402**

Table 4: The results of ablation study on three datasets.

prompt vector "According to the image and text, this news is $[CLASS]$". For w/o similarity, we choose not to use similarity calculation for classification prediction but instead concatenate the multi-modal feature H_M and the category feature H_C to obtain feature H_F, which is input to the linear regression classifier for classification prediction. For w/o frozen, we did not freeze CLIP's parameters and fine-tuned the entire model during training. The results are shown in Table 4.

From the experimental results, as shown in Table 4 we can see that each of our modules is effective, and the absence of any module will lead to performance degradation. W/o image and w/o text demonstrate that our model fully utilizes the information from both modalities and using either modality alone will lead to performance degradation. W/o fusion shows that simply concatenating the features of the two modalities cannot fully utilize the multi-modal context. Our method of multi-modal feature fusion allows the semantic information of different modalities to learn from each other and then fuse them together, which can fully integrate the semantic information of different modalities. W/o a learnable prompt shows the improvement brought by using a learnable prompt. Compared with manually-designed prompts, allowing the model to learn prompt vectors can better capture the semantic information of categories and multi-modal features and better apply them to downstream tasks. W/o similarity demonstrates the significance of our model being consistent with the CLIP pre-training method, i.e., contrastive learning. Compared with only using the CLIP encoder to extract features and then using an extra classification network for classification, our method can more fully utilize the knowledge learned in the CLIP pre-training phase and achieve better classification results. W/o frozen shows the improvement brought by prompt learning compared to fine-tuning methods in few-shot settings, demonstrating the effectiveness of our method in few-shot settings.

Conclusion

In this article, we design a multi-modal prompt learning (MPL) framework based on the multi-modal pre-trained model CLIP for the multi-modal fake news detection task at an early stage. Since fake news usually spread more widely and faster, it is crucial to identify their authenticity at an early stage to avoid serious consequences. MPL designs a multi-modal feature fusion module based on the co-attention layer fully integrate the semantic context of different modalities. MPL also designs learnable prompts to automate prompt engineering and generate prompt representations adaptively and efficiently to boost the semantic context. After comparison testing on three public datasets, MPL obtains substantial performance and efficiency improvement for the early-stage fake news detection task in both few-shot and supervised settings. Experimental results show that MPL outperforms both the state-of-the-art supervised multi-modal models and the latest prompt-based few-shot models. We also conduct ablation experiments to verify the effectiveness of the components we proposed, which demonstrates the absence of any module will lead to performance degradation.

Overall, MPL is a concise but effective framework, and there is still room for improvement. For example, we currently do not utilize any knowledge from external knowledge bases or knowledge graphs, which will further improve the performance of the model in few-shot settings. In future work, we'll continue to explore how to integrate external knowledge. Meanwhile, we will continue to improve prompt learning methods, such as using Prefix tuning to tune the prompt token embedding of each layer, not just the input layer.

Acknowledgments

This work is suppported by the funding of Harbin Institute of Technology (Shenzhen) (No. 20210035).

Broader Perspective, Ethics and Competing Interests

This paper conducts research on the public news of online social media. The datasets and referred methods are all collected from public data sources. There is no conflict of interest.

References

Benamira, A.; Devillers, B.; Lesot, E.; Ray, A. K.; Saadi, M.; and Malliaros, F. D. 2019. Semi-supervised learning and graph neural networks for fake news detection. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 568–569.

Bian, T.; Xiao, X.; Xu, T.; Zhao, P.; Huang, W.; Rong, Y.; and Huang, J. 2020. Rumor detection on social media with bi-directional graph convolutional networks. In *Proceedings*

of the AAAI conference on artificial intelligence, volume 34, 549–556.

Boididou, C.; Andreadou, K.; Papadopoulos, S.; Dang Nguyen, D. T.; Boato, G.; Riegler, M.; Kompatsiaris, Y.; et al. 2015. Verifying multimedia use at mediaeval 2015. In *MediaEval 2015*, volume 1436. CEUR-WS.

Boididou, C.; Papadopoulos, S.; Zampoglou, M.; Apostolidis, L.; Papadopoulou, O.; and Kompatsiaris, Y. 2018. Detection and visualization of misleading content on Twitter. *International Journal of Multimedia Information Retrieval*, 7(1): 71–86.

Chen, Y.; Li, D.; Zhang, P.; Sui, J.; Lv, Q.; Tun, L.; and Shang, L. 2022. Cross-modal ambiguity learning for multimodal fake news detection. In *Proceedings of the ACM Web Conference 2022*, 2897–2905.

Cheng, M.; Nazarian, S.; and Bogdan, P. 2020. Vroc: Variational autoencoder-aided multi-task rumor classifier based on text. In *Proceedings of the web conference 2020*, 2892–2898.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dou, Y.; Shu, K.; Xia, C.; Yu, P. S.; and Sun, L. 2021. User preference-aware fake news detection. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2051–2055.

Ecker, U. K.; Hogan, J. L.; and Lewandowsky, S. 2017. Reminders and repetition of misinformation: Helping or hindering its retraction? *Journal of Applied Research in Memory and Cognition*, 6(2): 185–192.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Guacho, G. B.; Abdali, S.; Shah, N.; and Papalexakis, E. E. 2018. Semi-supervised content-based detection of misinformation via tensor embeddings. In *2018 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM)*, 322–325. IEEE.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

He, Z.; Li, C.; Zhou, F.; and Yang, Y. 2021. Rumor detection on social media with event augmentations. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020–2024.

Helwe, C.; Elbassuoni, S.; Al Zaatari, A.; and El-Hajj, W. 2019. Assessing arabic weblog credibility via deep co-learning. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, 130–136.

Hu, L.; Yang, T.; Zhang, L.; Zhong, W.; Tang, D.; Shi, C.; Duan, N.; and Zhou, M. 2021. 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)*, 754–763.

Jiang, G.; Liu, S.; Zhao, Y.; Sun, Y.; and Zhang, M. 2022. Fake news detection via knowledgeable prompt learning. *Information Processing & Management*, 59(5): 103029.

Jiang, Y.; Song, X.; Scarton, C.; Aker, A.; and Bontcheva, K. 2021. Categorising fine-to-coarse grained misinformation: An empirical study of covid-19 infodemic. *arXiv preprint arXiv:2106.11702*.

Jiang, Y.; Yu, X.; Wang, Y.; Xu, X.; Song, X.; and Maynard, D. 2023. Similarity-Aware Multimodal Prompt Learning for Fake News Detection. *arXiv preprint arXiv:2304.04187*.

Jin, Z.; Cao, J.; Guo, H.; Zhang, Y.; and Luo, J. 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM international conference on Multimedia*, 795–816.

Kang, Z.; Cao, Y.; Shang, Y.; Liang, T.; Tang, H.; and Tong, L. 2021. Fake news detection with heterogenous deep graph convolutional network. In *Advances in Knowledge Discovery and Data Mining: 25th Pacific-Asia Conference, PAKDD 2021, Virtual Event, May 11–14, 2021, Proceedings, Part I*, 408–420. Springer.

Kemp, P. L.; Loaiza, V. M.; and Wahlheim, C. N. 2022. Fake news reminders and veracity labels differentially benefit memory and belief accuracy for news headlines. *Scientific Reports*, 12(1): 21829.

Khattar, D.; Goud, J. S.; Gupta, M.; and Varma, V. 2019. Mvae: Multimodal variational autoencoder for fake news detection. In *The world wide web conference*, 2915–2921.

Konkobo, P. M.; Zhang, R.; Huang, S.; Minoungou, T. T.; Ouedraogo, J. A.; and Li, L. 2020. A deep learning model for early detection of fake news on social media. In *2020 7th International Conference on Behavioural and Social Computing (BESC)*, 1–6. IEEE.

Li, B.; Qian, Z.; Li, P.; and Zhu, Q. 2022a. Multi-modal fusion network for rumor detection with texts and images. In *MultiMedia Modeling: 28th International Conference, MMM 2022, Phu Quoc, Vietnam, June 6–10, 2022, Proceedings, Part I*, 15–27. Springer.

Li, X.; Lu, P.; Hu, L.; Wang, X.; and Lu, L. 2022b. A novel self-learning semi-supervised deep learning network to detect fake news on social media. *Multimedia Tools and Applications*, 81(14): 19341–19349.

Li, Y.; Scarton, C.; Song, X.; and Bontcheva, K. 2022c. Classifying COVID-19 vaccine narratives. *arXiv preprint arXiv:2207.08522*.

Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; and Neubig, G. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9): 1–35.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Lu, Y.-J.; and Li, C.-T. 2020. GCAN: Graph-aware co-attention networks for explainable fake news detection on social media. *arXiv preprint arXiv:2004.11648*.

Mansouri, R.; Naderan-Tahan, M.; and Rashti, M. J. 2020. A semi-supervised learning method for fake news detection in social media. In *2020 28th Iranian Conference on Electrical Engineering (ICEE)*, 1–5. IEEE.

Meel, P.; and Vishwakarma, D. K. 2020. Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities. *Expert Systems with Applications*, 153: 112986.

Meel, P.; and Vishwakarma, D. K. 2021. A temporal ensembling based semi-supervised ConvNet for the detection of fake news articles. *Expert Systems with Applications*, 177: 115002.

Qian, S.; Wang, J.; Hu, J.; Fang, Q.; and Xu, C. 2021. Hierarchical multi-modal contextual attention network for fake news detection. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, 153–162.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Shu, K.; Mahudeswaran, D.; Wang, S.; Lee, D.; and 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.

Shu, K.; Sliva, A.; Wang, S.; Tang, J.; and Liu, H. 2017. Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explorations Newsletter*, 19(1): 22–36.

Shu, K.; Wang, S.; and Liu, H. 2017. Exploiting Tri-Relationship for Fake News Detection. *arXiv preprint arXiv:1712.07709*.

Silva, A.; Luo, L.; Karunasekera, S.; and Leckie, C. 2021. Embracing domain differences in fake news: Cross-domain fake news detection using multi-modal data. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 557–565.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Singhal, S.; Kabra, A.; Sharma, M.; Shah, R. R.; Chakraborty, T.; and Kumaraguru, P. 2020. Spotfake+: A multimodal framework for fake news detection via transfer learning (student abstract). In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 13915–13916.

Singhal, S.; Pandey, T.; Mrig, S.; Shah, R. R.; and Kumaraguru, P. 2022. Leveraging Intra and Inter Modality Relationship for Multimodal Fake News Detection. In *Companion Proceedings of the Web Conference 2022*, 726–734.

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

Song, X.; Petrak, J.; Jiang, Y.; Singh, I.; Maynard, D.; and Bontcheva, K. 2021. Classification aware neural topic model and its application on a new covid-19 disinformation corpus.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wahlheim, C. N.; Alexander, T. R.; and Peske, C. D. 2020. Reminders of everyday misinformation statements can enhance memory for and beliefs in corrections of those statements in the short term. *Psychological Science*, 31(10): 1325–1339.

Wang, Y.; Ma, F.; Jin, Z.; Yuan, Y.; Xun, G.; Jha, K.; Su, L.; and 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*, 849–857.

Wang, Y.; Qian, S.; Hu, J.; Fang, Q.; and Xu, C. 2020. Fake news detection via knowledge-driven multimodal graph convolutional networks. In *Proceedings of the 2020 international conference on multimedia retrieval*, 540–547.

Wu, Y.; Zhan, P.; Zhang, Y.; Wang, L.; and Xu, Z. 2021. Multimodal fusion with co-attention networks for fake news detection. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021*, 2560–2569.

Xue, J.; Wang, Y.; Tian, Y.; Li, Y.; Shi, L.; and Wei, L. 2021. Detecting fake news by exploring the consistency of multimodal data. *Information Processing & Management*, 58(5): 102610.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.

Zhou, X.; Wu, J.; and Zafarani, R. 2020. : Similarity-Aware Multi-modal Fake News Detection. In *Advances in Knowledge Discovery and Data Mining: 24th Pacific-Asia Conference, PAKDD 2020, Singapore, May 11–14, 2020, Proceedings, Part II*, 354–367. Springer.

Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, we use the original datasets partition method and the original indicators of baselines to reduce artifacts.
   (e) Did you describe the limitations of your work? Yes, see the Main Results and Conclusion.

(f) Did you discuss any potential negative societal impacts of your work? No, but our model is difficult to have negative societal impacts.

(g) Did you discuss any potential misuse of your work? No, but our model is unlikely to be misused.

(h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? No.

(i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes.

2. Additionally, if your study involves hypotheses testing...

(a) Did you clearly state the assumptions underlying all theoretical results? NA

(b) Have you provided justifications for all theoretical results? NA

(c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA

(d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA

(e) Did you address potential biases or limitations in your theoretical framework? NA

(f) Have you related your theoretical results to the existing literature in social science? NA

(g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA

3. Additionally, if you are including theoretical proofs...

(a) Did you state the full set of assumptions of all theoretical results? NA

(b) Did you include complete proofs of all theoretical results? NA

4. Additionally, if you ran machine learning experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? No, but the datasets we used are public.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes, see the Experimental Setup.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? No, but we did set up multiple random seeds, and provided an average when compared to other methods.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? No.

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes.

(f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? No, but the cost of misclassification in the early stage of fake news detection is small because we are just giving a warning.

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...

(a) If your work uses existing assets, did you cite the creators? Yes.

(b) Did you mention the license of the assets? No, we didn't find these.

(c) Did you include any new assets in the supplemental material or as a URL? No.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? No, but the data is collected from social platforms.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? No, but social engagements and user information are not disclosed because of Twitter Policy, so it dosn't contain personally identifiable information, and few of news in datasets contain offensive content.

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? NA

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? NA

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? NA

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA

(d) Did you discuss how data is stored, shared, and deidentified? NA