Information Sciences 647 (2023) 119446

Contents lists available at ScienceDirect

Information Sciences

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

INFORMATION SCIENCES
Informatics and Computer Science Intelligent Systems Applications
AN INTERNATIONAL JOURNAL

Similarity-Aware Multimodal Prompt Learning for fake news detection

Ye Jiang [a], Xiaomin Yu [a], Yimin Wang [a,*], Xiaoman Xu [a], Xingyi Song [b], Diana Maynard [b]

[a] *School of Information Science and Technology, Qingdao University of Science and Technology, China*
[b] *Department of Computer Science, University of Sheffield, United Kingdom*

ARTICLE INFO

Keywords:
Prompt learning
Fake news detection
Few-shot learning
Multimodal fusing

ABSTRACT

The standard paradigm for fake news detection relies on utilizing text information to model the truthfulness of news. However, the subtle nature of online fake news makes it challenging to solely rely on textual information for debunking. Recent studies that focus on multimodal fake news detection have demonstrated superior performance compared with text-only methods, thereby establishing a new paradigm for detecting fake news. However, this paradigm may require a large number of training instances or updating the entire set of pre-trained model parameters. Furthermore, existing multimodal approaches typically integrate cross-modal features without considering the potential introduction of noise from unrelated semantic representations. To address these issues, this paper proposes the **S**imilarity-**A**ware **M**ultimodal **P**rompt **Le**arning (SAMPLE) framework. Incorporating prompt learning into multimodal fake news detection, we used three prompt templates with a soft verbalizer to detect fake news. Moreover, we introduced a similarity-aware fusing method, which adaptively fuses the intensity of multimodal representation so as to mitigate noise injection from uncorrelated cross-modal features. Evaluation results show that SAMPLE outperformed previous work, achieving higher F1 and accuracy scores on two multimodal benchmark datasets, demonstrating its feasibility in real-world scenarios, regardless of data-rich or few-shot settings.

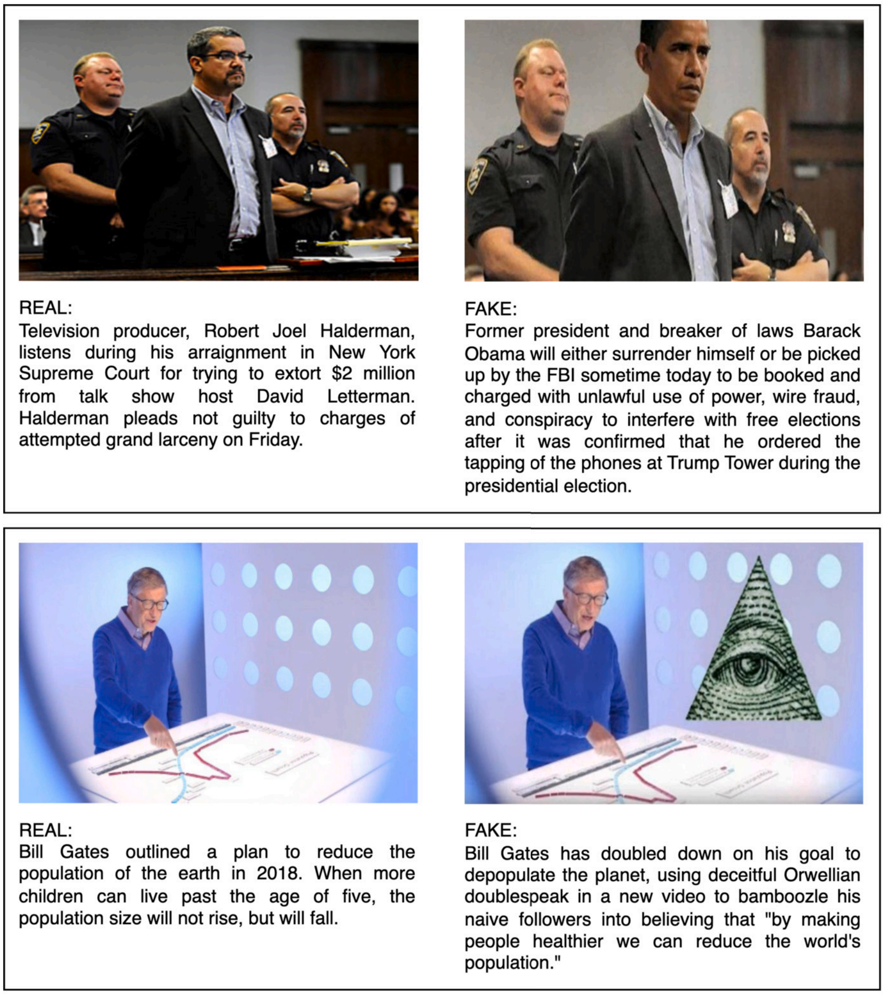
1. Introduction

The increasing prevalence of social media has significantly impacted the way information is disseminated and consumed. While social media platforms provide an efficient way for people to seek and share information, the spread of fake news has caused substantial harm to the global community. In an effort to mitigate the impacts of online fake news, academia and industry have developed various techniques. Early research [29,4] mainly focused on analyzing the textual content of fake news. However, fake news can take various forms, and verifying its truthfulness by relying only on textual information requires expertise, which can be time-consuming. For example, Fig. 1 shows two news snippets that pose a challenge in identifying their truthfulness solely through

* Corresponding author.
E-mail addresses: ye.jiang@qust.edu.cn (Y. Jiang), yuxm02@gmail.com (X. Yu), yimin.wang@qust.edu.cn (Y. Wang), xxm981215@163.com (X. Xu), x.song@sheffield.ac.uk (X. Song), d.maynard@sheffield.ac.uk (D. Maynard).

https://doi.org/10.1016/j.ins.2023.119446
Received 20 April 2023; Received in revised form 31 July 2023; Accepted 3 August 2023

Available online 9 August 2023



REAL:
Television producer, Robert Joel Halderman, listens during his arraignment in New York Supreme Court for trying to extort $2 million from talk show host David Letterman. Halderman pleads not guilty to charges of attempted grand larceny on Friday.

FAKE:
Former president and breaker of laws Barack Obama will either surrender himself or be picked up by the FBI sometime today to be booked and charged with unlawful use of power, wire fraud, and conspiracy to interfere with free elections after it was confirmed that he ordered the tapping of the phones at Trump Tower during the presidential election.

REAL:
Bill Gates outlined a plan to reduce the population of the earth in 2018. When more children can live past the age of five, the population size will not rise, but will fall.

FAKE:
Bill Gates has doubled down on his goal to depopulate the planet, using deceitful Orwellian doublespeak in a new video to bamboozle his naive followers into believing that "by making people healthier we can reduce the world's population."

Fig. 1. Two snippets of fake news and their original reports.

textual information. Therefore, multimodal Fake News Detection (FND) techniques have been developed recently to leverage both image and textual information, demonstrating promising performance as complementary benefits offered through cross-modality analysis.

Multimodal FND aims to combine features from images and texts to automatically identify fake news posts. Traditional deep learning methods, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) and Transformers, have made significant advances in modeling both textual and image representations of fake news. However, these methods are often limited by the reliance on a substantial amount of annotated data to achieve satisfactory performance. Recently, there has been an increasing interest in utilizing large pre-trained models for FND. Many studies [42,2] used pre-trained language models, such as BERT [10], and pre-trained vision models, such as ResNet [16], to encode textual and image features of news posts, respectively. However, pre-trained models are typically trained on a large, unrefined corpus that is not specific to any particular domain. Although pre-trained models can leverage external knowledge to identify fake posts, the effectiveness of an FND system is highly dependent on its focus domain [19].

Fine-tuning is a common technique for adapting pre-trained models for diverse downstream tasks. In recent research, variants of BERT, including the original pre-trained model, have been fine-tuned specifically for FND [7]. However, fine-tuning for FND typically poses difficulties in low-resource settings, due to the necessity of a significant number of labeled instances to train additional classifiers [5]. Traditional pre-trained language models are trained with a cloze-style objective, which involves predicting masked words to learn their distributions, while fine-tuning aims to identify the target label directly. Consequently, pre-trained models require a significant amount of labeled data to be fine-tuned for specific tasks. Meanwhile, fine-tuning, updating all model parameters for a single task, poses challenges on real-world FND due to the size of pre-trained models [27]. Prompt learning is an approach that aims to better utilize pre-trained knowledge by adding additional information to the input and using a cloze-style task during the tuning process, resulting in more effective application of pre-training information [37]. Furthermore, prompt learning becomes especially significant for real-world FND scenarios, where there is a scarcity of manually labeled fake news data. It enables pre-trained models to attain competitive performance even in low-resource settings with limited labeled data [5]. However, current prompt-based FND approaches [19] primarily consider textual information, and the analysis of cross-modality features in fake news posts is underdeveloped.

In contrast to fine-tuned models that directly outputs class distributions, prompt learning aligns with the language modeling objective, which generates specific answer words that are relevant to FND by adding supplementary information before the original text inputs. As shown in the news snippet on the left side of Fig. 1, by introducing a prompt preceding the original text (e.g., *"This is a piece of <mask> news. Former president and breaker of laws, Barack Obama..."*), this approach aims to retrieve the masked token of the prompt text. However, the limitation of discrete prompt is that it requires the embedding of template words to align with that of natural language words. To address this issue, continuous prompting [35] eliminates the constraint of the discrete prompt by performing prompting directly in a continuous space of the pre-trained model, for example, *"<soft><soft>...<soft><mask>. Former president and breaker of laws, Barack Obama..."*, where each <soft> can be associated with a randomly initialized trainable vector. Additionally, instead of utilizing a fully learnable prompt template, a mixed prompt [19] incorporates trainable vectors into a discrete prompt template (e.g., *"<soft> This is a piece of <mask> news <soft>. Former president and breaker of laws, Barack Obama..."*), and demonstrates superior performance to using each prompt type individually.

Previous multimodal FND methods [26,41] aimed to enhance performance by directly fusing multimodal representations. However, combining solely image and text features cannot guarantee reliable information, as the veracity of news articles is not completely associated with image-text correlation. In such cases, the correlation between text and image features tends to be weaker, leading to a noisy multimodal representation. Therefore, it is crucial for multimodal FND models to grasp the semantic correlation between different modalities and adaptively combine multimodal features to conduct accurate classification.

This paper proposes a **S**imilarity-**A**ware **M**ultimodal **P**rompt **Le**arning (SAMPLE) framework for FND. Three popular prompt learning methods (discrete prompting (DP), continuous prompting (CP), and mixed prompting (MP)) are systematically integrated into a soft verbalizer for the task of FND. In addition, the pre-trained model Contrastive Language-Image Pre-training (CLIP) [36] is applied to extract the text and image features, which are utilized to generate the multimodal representation. In order to tackle the issue of uncorrelated semantic representation between text and image, the framework calculates the semantic similarity between their features. To adjust the intensity of the aggregated multimodal representation, the semantic similarity is further normalized. To assess the performance of the proposed SAMPLE framework, two domain-specific publicly accessible datasets, PolitiFact and GossipCop [39]), are utilized. We compare SAMPLE with existing FND methods, as well as the standard fine-tuning method, under both few-shot and data-rich scenarios to simulate real-world FND settings. The experimental results demonstrate that SAMPLE significantly outperforms traditional deep learning and fine-tuned approaches in both macro-f1 and accuracy metrics, regardless of data-rich or few-shot scenarios.

The contributions of this paper are:

- We propose a framework called SAMPLE that adaptively fuses multimodal features generated by the CLIP model with textual representation from a pre-trained language model, to assist prompt learning for detecting fake news.
- The proposed framework mitigates the issue of uncorrelated cross-modal semantics by adjusting the intensity of fused multimodal features using standardized cosine similarity generated by the pre-trained CLIP model.
- SAMPLE is evaluated on two benchmark multimodal fake news detection datasets, outperforming previous approaches in both low-resource and data-rich scenarios.

2. Related work

Fake news is described as "false information that is circulated under the guise of being genuine news for political or financial gain via news outlets or the internet" [30]. In addition, many recent studies aim to differentiate false content from similar concepts, such as misinformation [20] and disinformation [43]. In this context, misinformation is false information that results from blunders or cognitive biases, whereas disinformation is intentionally fabricated, and in both cases, the formats are not limited to news outlets.

2.1. *Unimodal fake news detection*

Early research on unimodal FND often uses handcrafted features to identify anomalies in a post's text or image. Traditional methods of image manipulation detection [8] can effectively detect tampering of news images. These methods learn image forensic, semantic, statistical, and contextual features from fake news. Fake news is frequently characterized by semantic inconsistencies that violate common sense [24], as well as poor image quality [14]. In text modality, previous research [32] designed a modular tool, MedOSINT, to identify Covid-19 related fake news. Wang et al. [45] proposed a dual hierarchical contractive learning framework to include multiple data augmentation strategies and three contrastive learning tasks for FND. Khullar and Singh [23] established a federated learning framework to classify fake news as well as maintain data privacy. While unimodal FND is a robust baseline for detecting fake news, the correlation and consistency of the modalities in FND are not well established.

2.2. *Multimodal fake news detection*

Previous studies in multimodal fake news detection (FND) have typically focused on two approaches: designing complex networks or utilizing pre-trained models as feature extractors. Zhou et al. [46] proposed the SAFE model, which uses the Image2Sentence model to convert images to text captions, and extends the Text-CNN model to extract textual features from news descriptions. To detect fake news, the model computes the relevance between the textual and visual information using a slightly modified cosine similarity measure, which is then fed into a classifier. Meel and Vishwakarma [31] combines a hierarchical attention network, image captioning, and forensics analysis to detect multimodal fake news.

More recently, many studies have opted to utilize pre-trained models to extract textual and visual features in FND. For example, CAFE [9] employs BERT and ResNet-34 as pre-trained models for encoding textual and visual features, respectively. Similarly, Zhou et al. [47] proposed the FND-CLIP model, which extracts feature representations from images and text using a ResNet-based encoder, a BERT-based encoder, and two pairwise CLIP encoders simultaneously. Hua et al. [18] established a BERT-based back-Translation Text and Entire-image multimodal model with contrastive learning and data augmentation. Jing et al. [22] designed a progressive fusion network to capture each modality's feature representation at different levels and also implemented a mixer to establish the connection between modalities.

Moreover, some studies have found that fine-tuning pre-trained models can also yield competitive performance, rather than just using them as feature extractors. As an example, Ro-CT-BERT [7] expands the vocabulary with professional phrases and adapts the heated-up softmax loss for adversarial training to improve the model's robustness. Although traditional multimodal FND methods are known for accurately detecting fake news, they typically require a large amount of human-annotated data to train models effectively. Furthermore, while detecting fake news at an early stage can minimize its pernicious effects [40], FND methods are still limited by the availability of human-annotated data.

2.3. Prompt learning for fake news detection

In recent years, prompt learning has emerged as a new paradigm in Natural Language Processing (NLP) and has demonstrated comparable performance to standard fine-tuning in various NLP tasks. For example, Zhu et al. [48] developed the PLST framework, which combines both text inputs and external knowledge from open knowledge graphs in short text classification tasks. Han et al. [15] proposed the PTR model, which is designed for many-class text classification, and constructs prompts using logic rules that contain multiple sub-prompts. Prompt-based models have also been used to aid fake news detection (FND). For example, El Vaigh et al. [11] utilized the prompt-based model from DistilGPT-2 in conjunction with multitask learning to detect coronavirus-related fake news in MediaEval-2021. Jiang et al. [19] proposed KPL, which detects fake news by integrating external knowledge. However, KPL relies on human-designed prompts and verbalizers, which can be time-consuming and potentially unreliable. Besides, it does not address how the fusion of multimodal representations of news posts can enhance fake news detection.

3. Methodology

The proposed approach aims to identify the authenticity of news articles by utilizing both text and image. The main objective of multimodal FND is to assign a standard binary classification label of $y \in \{0, 1\}$, where 0 represents real news and 1 represents fake news, to a given news article that includes both text input $x = [w_1, w_2, ..., w_n]$ with n words and image input $i = [i_1, i_2, ..., i_m]$ with m images. To identify the most relevant image corresponding to a given news article's text, the pre-trained CLIP model is utilized to encode the text and image representations separately. In this process, only the image with the highest cosine similarity to the text representation is retained, while the remaining images are discarded.

In this section, we utilize discrete prompting [38], which primarily corresponds to natural language phrases and automatically searches for templates described in a discrete space. Furthermore, we introduce an extended version called continuous prompting [35], which employs prompts containing pseudo-tokens not present in the pre-trained language model vocabulary. We also employ a mixed prompt that combines both discrete and continuous prompt for FND. Finally, we standardize the semantic similarity between the text and image to adjust the fused multimodal representation. Fig. 2 illustrates the overall structure of the proposed SAMPLE.

3.1. Discrete prompting

We utilize a manually constructed discrete template as the prompting mechanism. To enable the model to retrieve the masked words, text inputs are initially masked during the prompt learning phase. The discrete prompting involves the intentional distortion of text input by means of a limited, human-designed template, with a single keyword replaced with a mask. We investigate five distinct discrete templates, as the choice of templates can potentially have a significant impact on the performance of the language model, as shown in Appendix A. The discrete template dt = "This is a news piece with $< mask >$ information", is a human-designed template. Following this, we calculate the representation of the masked word, which is tied to the target of the FND task, by the pre-trained language model. To accomplish this, we concatenate the discretionary template dt with the initial input x to generate the prompt, $x_d = [dt; x]$. Subsequently, the hidden states of the prompt x_d are calculated:

$$h_1^{dt}, ..., h_{mask}^{dt}, ...h_m^{dt} | h_1^x, ..., h_n^x = PLM(x_d) \tag{1}$$

where h_i^{dt} $(i \in [1, m])$ and h_{mask}^{dt} are the hidden vectors of length m and $< mask >$ token of the discrete template respectively. h_j^x $(j \in [1, n])$ are the hidden vectors of length n of the input text, and $PLM()$ is the masked language model output.

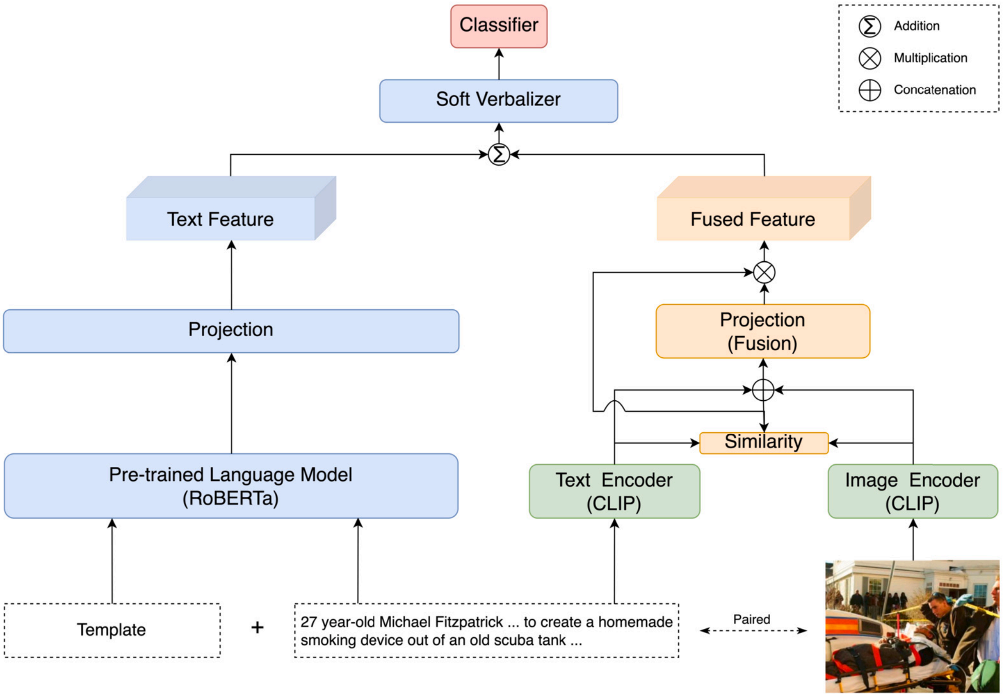


Fig. 2. The overall structure of SAMPLE for fake news detection.

3.2. Continuous prompting

Although discrete prompting naturally inherits interpretability from the task description, it is limited by the requirement of embedding template words in natural language. In addition, discrete prompts may be suboptimal because the pre-trained language model may have learned the target knowledge from substantially different contexts. Such manually designed constraints can also be applied to the verbalizer because manual verbalizers usually determine predictions based on limited information. For example, the standard verbalizer maps fake ⟶ {counterfeit, sham, ..., falsify}, meaning that only predicting those related words for the token is considered correct during inference, regardless of the predictions for other relevant words like “unreal” or “untrue” that are also informative. Such a manually designed mapping limits the coverage of label words, resulting in insufficient information for prediction, and introducing bias into the verbalizer.

To address the above issues, the discrete template was reformatted by replacing trainable tokens with the continuous template $st =$ “$< soft_1 >, < soft_2 >, \ldots < soft_t >, < mask >$” where each $< soft >$ is associated with a randomly initialized[1] trainable vector. Then, the hidden states of the continuous prompt $x_s = [st; x]$ can be calculated similarly as:

$$h_1^{st}, h_2^{st}, \ldots h_t^{st}, h_{mask}^{st} | h_1^x, \ldots, h_n^x = PLM(x_s) \tag{2}$$

where h_k^{st} $(k \in [1, t])$ and h_{mask}^{st} are the hidden vectors of length t and the $< mask >$ token of the continuous template respectively.

3.3. Mixed prompting

Recent research has demonstrated that employing mixed prompting, which blends continuous and discrete templates, exhibits superior performance compared to using them independently [15]. Building on this, we have incorporated trainable tokens into the discrete prompt template. To be specific, we have inserted two trainable tokens, h_{head}^{mt} and h_{tail}^{mt}, at the beginning and the end of the mixed template, expressed as $mt =$“$< h_{head}^{mt} >$ *This is a piece of* $< h_{mask}^{mt} >$ *news.* $< h_{tail}^{mt} >$”. Similar to discrete prompt, the new mixed prompt $x_m = [mt; x]$. We then compute its hidden states as follows:

$$h_{head}^{mt}, h_1^{mt}, \ldots, h_{mask}^{mt}, \ldots, \ldots h_m^{mt}, h_{tail}^{mt} | h_1^x, \ldots, h_n^x = PLM(x_m) \tag{3}$$

where h_i^{mt} $(i \in [1, m])$ and h_{mask}^{mt} are the hidden vectors of length m and $< mask >$ token of the mixed template respectively.

[1] Three initialization methods are compared as shown in the Appendix B. The experimental results suggest that random initialization achieves comparable performance with a slightly faster convergence of validation loss than that of the others.

3.4. Similarity-aware multimodal feature fusing

According to a previous study [47], text and image features extracted from pre-trained models exhibit large semantic gaps. As a result, direct fusion of multimodal features fails to capture intrinsic semantic correlations. The unimodal pre-trained models, such as BERT and ViT-B-32, tend to focus on trivial clues, rather than on extracting semantically meaningful information. BERT can better learn emotional features from textual inputs, whereas ViT-B-32 can capture the noise patterns in images. Thus, direct fusion of unimodal features may inject noise into the multimodal representation, even if the text and the image are semantically correlated. In contrast, pre-trained CLIP models utilize a large dataset of image-text pairs to capture semantic correlations beyond emotional features or noise patterns.

To effectively integrate image and text features, the pre-trained CLIP model is applied to extract these features independently. The CLIP model consists of a text Transformer for text encoding and employs the Vision Transformer (ViT-B-32) as the image encoder. To reduce the dimensionality of coarse features provided by the encoders and eliminate redundant information, we utilize individual projection heads P_{txt} and P_{img} to process text and image features. Each projection head features two sets of fully-connected layers (FC), followed by Batch Normalization, a Rectified Linear Unit (ReLU) activation function, and a dropout layer. Next, we measure the cosine similarity sim between P_{txt} and P_{img} to modify the intensity of the fused feature f_{fused}:

$$\begin{aligned} f_{fused} &= [P_{txt}; P_{img}] \\ sim &= \frac{P_{txt}(P_{img})^T}{||P_{txt}||\,||P_{img}||} \end{aligned} \tag{4}$$

During the experiment, we noticed that certain news posts, irrespective of their authenticity, did not exhibit explicit cross-modal semantic relationships. Consequently, concatenating the unimodal features to generate the fused feature could introduce noise, particularly in instances where the similarity was low. To remedy the issue, we apply standardization and a Sigmoid function to constrain the similarity value to the range of $[0-1]$. Standardization involves calculating the mean and standard deviation during training, subtracting the running mean from sim, and dividing the result by the running standard deviation. The standardized similarity can then be used to adjust the intensity of the final cross-modal representation, m_{fused}:

$$m_{fused} = Sigmoid(Std(sim))\, f_{fused} \tag{5}$$

3.5. Soft verbalizer

To recover masked words in the prompt template, soft verbalizer is utilized to map labels to their corresponding words. In this study, we utilize WARP [13] to identify the optimal prompt in the continuous embedding space, where a pre-trained language model predicts the masked token by conducting a search. We use soft verbalization in the three template types mentioned above to compare different prompt methods. To identify the optimal parameters $\theta = \{\theta^P, \theta^V\}$ for prompt and verbalizer embeddings, we first add the output vectors from the masked language model to the adjusted cross-modal representation using a residual connection. This combined output is then fed into an FC layer:

$$x_{fused} = FC(PLM(x') + m_{fused}) \tag{6}$$

where x' is the input sequence concatenated with one of the prompt templates from the above. The classification probability $P(y|x')$ can then be calculated as:

$$P(y|x_{fused}) = \frac{\exp \theta_y^V x_{fused}}{\sum_{i\in C} \exp \theta_i^V x_{fused}} \tag{7}$$

where C is the set of classes, θ_y^V is the embeddings of the true label and θ_i^V is the embeddings of the predicted label word. Finally, the cross-entropy loss can be minimized as:

$$\theta^* = \arg\max_{\theta}(-\log P(y|x_{fused})) \tag{8}$$

4. Experiment

We evaluate our proposed approach on two benchmark FND datasets in low-resource and data-rich scenarios. The first part of this section presents an overview of the benchmark multimodal FND datasets, including their statistics. In the second part, we explain the implementation details for both the data-rich and few-shot settings. Finally, we provide a detailed discussion and analysis of our proposed method as well as the baseline models.

4.1. Data

We use two publicly accessible datasets for detecting fake information, namely PolitiFact and GossipCop, which consist of political news and celebrity gossip, respectively, and are included in the FakeNewsNet project [39]. Using the data crawling scripts provided, we retrieve 1,056 news items from PolitiFact and 22,140 news items from GossipCop. To reduce redundancy, we only preserve the

Table 1
The statistics of the pre-processed multimodal fake news datasets.

Statistics	PolitiFact	GossipCop
Total number of news	198	6,805
Number of fake news	96	1,877
Number of real news	102	4,928
Average number of words per news	2,148	728

most relevant images based on the text and images' cosine similarity, for news with multiple images. News with no images or invalid image URLs are excluded. The resulting dataset statistics are presented in Table 1.

4.2. Implementation details

The pre-trained RoBERTa from the HuggingFace library is adopted as the main block for prompt learning. The text and image encoders from the pre-trained CLIP (ViT-B-32) model are applied to extract their respective features. The size of the hidden layer's projection layers is set to 768, and a dropout rate is 0.6. We use the AdamW optimizer to optimize the model parameters, with a learning rate of $3e{-}5$ and a decay parameter of $1e{-}3$, both of which are empirically determined. The model is trained for 20 epochs, and we choose the model checkpoints that yield the best validation performance for testing purposes. We evaluate the method in both few-shot and data-rich settings.

In the few-shot setting, our model is trained using a small number of instances (n) randomly sampled from the dataset. Specifically, we consider $n \in [2, 4, 8, 16, 100]$. The rest of the instances are used for testing. Also, a validation set of the same size as the training set is created for model selection. The PolitiFact dataset contains a limited number of news items. To address this limitation, we adopt a specific configuration known as the PolitiFact 100-shot setting. In this configuration, we allocate 100 instances for training and 50 for development purposes. Due to the significant impact of the training set and validation set quality on the model's performance, we repeat the above data sampling method five times with distinct random seeds. We then calculate the average score, excluding the highest and lowest ones, to evaluate the model's performance in the few-shot setting. For both the training and validation sets, we ensure a balanced distribution of labeled instances during the training phase.

In the data-rich setting, the two datasets are split into three parts, i.e., training set, validation set, and test set, with a split ratio of 8:1:1. In order to evaluate the stability of the proposed model, we repeat the above data sampling process five times with distinct random seeds. We report the average score, calculated as the mean of the scores after removing the highest and lowest ones from the five runs.

4.3. Baseline models

We evaluate the proposed SAMPLE model in comparison to several models that have previously achieved state-of-the-art performances on the FND dataset. Specifically, our comparison involves unimodal approaches (1-2), multimodal approaches (3-6), and the standard fine-tuning approach (7). To initialize the word embeddings, we utilize the pre-trained 100-dimensional GloVe embeddings trained on a corpus of 6 billion words [34].

(1) **LDA-HAN** [21]: This model incorporates Latent Dirichlet Allocation (LDA) [3] topic distributions into a hierarchical attention network.
(2) **T-BERT** [2]: This feature-based method uses concatenated triple BERT models to predict fake news.
(3) **SAFE** [46]: This model converts images into their text descriptions and uses the relevance between textual and visual information to detect fake news.
(4) **RIVF** [44]: This model utilizes VGG and BERT models to encode image and text features. It applies the scaled dot-product attention mechanism on fused multimodal features to capture the relationship between text and images.
(5) **SpotFake** [42]: This model uses the pre-trained image model VGG and BERT to extract respective image and text features, concatenating them to classify fake news.
(6) **CAFE** [9]: This model uses an ambiguity-aware multimodal approach to adaptively aggregate unimodal features and correlations.
(7) **FT-RoBERTa**: This is a standard, fine-tuned version of the pre-trained language model RoBERTa.

4.4. Results

Table 2 shows the overall results that compare the proposed SAMPLE frameworks with the fine-tuning approach, multimodal and unimodal FND methods.

Comparing with fine-tuning. First, we investigate the performance of the standard fine-tuned RoBERTa (FT-RoBERTa) and the proposed SAMPLE by evaluating their respective F1 scores. We calculate the average improvements of M-SAMPLE (i.e., $\frac{(0.44-0.39)+...(0.58-0.52)}{5\times2} + \frac{(0.47-0.46)+...(0.81-0.77)}{5\times2}$), C-SAMPLE and D-SAMPLE over FT-RoBERTa, and find that all the SAMPLE methods outperform FT-RoBERTa by 0.05, 0.024 and 0.035 respectively. This improvement is more significant as the number of the training samples decreases, highlighting the superiority of prompt learning in low-resource scenarios.

Table 2
The overall macro-F1 and accuracy between baselines and the multimodal prompt learning framework. D-SAMPLE, C-SAMPLE and M-SAMPLE denote discrete prompting, continuous prompting and mixed prompting in the proposed SAMPLE framework respectively.

Data	Model	Few shot (F1/Acc)										Data rich (F1/Acc)	
		2		4		8		16		100			
PolitiFact	LDA-HAN	0.37	0.39	0.42	0.43	0.44	0.47	0.48	0.52	0.61	0.63	0.70	0.74
	T-BERT	0.43	0.50	0.45	0.57	0.5	0.54	0.50	0.54	0.69	0.69	0.71	0.75
	SAFE	0.19	0.19	0.21	0.21	0.29	0.27	0.33	0.49	0.46	0.56	0.64	0.65
	RIVF	0.35	0.48	0.43	0.51	0.42	0.48	0.40	0.47	0.43	0.49	0.43	0.45
	Spotfake	0.37	0.49	0.46	0.52	0.47	0.54	0.56	0.59	0.73	0.73	0.71	0.73
	CAFE	0.30	0.39	0.37	0.47	0.45	0.46	0.47	0.49	0.52	0.61	0.67	0.67
	FT-RoBERTa	0.46	0.52	0.51	**0.63**	0.60	0.63	**0.68**	**0.70**	0.77	0.81	0.79	**0.84**
	D-SAMPLE	0.45	0.54	0.54	0.59	0.61	0.64	**0.68**	**0.70**	**0.81**	**0.82**	0.79	0.81
	C-SAMPLE	**0.49**	0.53	0.54	0.57	0.61	0.64	0.65	0.67	0.77	0.78	**0.80**	0.81
	M-SAMPLE	0.47	**0.56**	**0.56**	**0.61**	**0.62**	**0.66**	0.67	**0.70**	**0.81**	**0.82**	**0.80**	0.81
GossipCop	LDA-HAN	0.18	0.21	0.20	0.25	0.28	0.30	0.34	0.40	0.49	0.45	0.54	0.60
	T-BERT	0.38	0.48	0.38	0.57	0.45	0.66	0.45	0.71	0.52	0.61	0.61	0.74
	SAFE	0.26	0.31	0.33	0.41	0.40	0.45	0.41	0.45	0.44	0.51	0.55	0.64
	RIVF	0.24	0.29	0.24	0.29	0.24	0.29	0.27	0.31	0.29	0.31	0.51	0.61
	Spotfake	0.23	0.28	0.22	0.28	0.23	0.28	0.32	0.34	0.48	0.49	0.43	0.73
	CAFE	0.41	0.42	0.42	0.52	0.46	0.48	0.47	0.56	0.50	0.61	0.59	0.72
	FT-RoBERTa	0.39	0.41	0.33	0.46	0.44	**0.60**	0.48	**0.63**	0.52	**0.64**	0.63	0.69
	D-SAMPLE	0.42	0.47	0.44	0.50	0.50	0.58	0.51	0.59	0.57	0.62	**0.64**	**0.76**
	C-SAMPLE	**0.47**	**0.54**	0.46	**0.56**	0.45	0.52	0.46	0.53	0.52	0.58	0.63	0.75
	M-SAMPLE	0.44	0.53	**0.47**	**0.56**	**0.52**	0.54	**0.54**	0.60	**0.58**	0.62	**0.64**	0.73

However, the improvements become smaller in the data-rich setting, in which the average improvements of F1 are 0.005, 0.005 and 0.01 respectively, showing that the FT-RoBERTa is able to achieve comparable performance when the training data is sufficient. The comparison of accuracy between SAMPLE methods and FT-RoBETRa aligns with the observation mentioned above, demonstrating the superiority of the proposed method in utilizing PLM information, particularly in scenarios with scarce training data. However, in data-rich settings, the standard fine-tuning approach still serves as a robust baseline.

Comparing with multimodal methods. We evaluated the performance of SAMPLE in comparison with previous multimodal FND methods. Our results indicate that regardless of the multimodal and unimodal methods, both F1 and accuracy scores from SAMPLE outperform previous methods in all settings. For example, when evaluated on PolitiFact dataset, M-SAMPLE achieved a notable improvement of up to 0.29 s in the 100-shot setting compared to CAFE. This improvement is mainly attributed to the learning approach of the CLIP model that capitalizes on a large amount of image-text pairs to learn the extraction of multimodal semantics. On the contrary, pre-trained models like BERT and ResNet-34, commonly employed by CAFE, may not effectively capture unimodal features characterized by heterogeneous feature distributions.

SpotFake, similarly to the previously mentioned models, utilizes BERT and VGG19 to extract text and image features, correspondingly. However, our experiment results show that SpotFake performs better on the smaller dataset PolitiFact, compared to CAFE. This might be attributed to the fact that news topics in PolitiFact relate to politics, and hence, it is easier to fuse multimodal features by using pre-trained unimodality models without any ambiguous measurement. On the other hand, GossipCop presents a more complex semantic context as it consists of celebrity gossip stories. Therefore, CAFE's cross-modal ambiguity learning module performs better in handling the intricate cross-modal semantics of GossipCop.

Comparing with unimodal methods. In terms of unimodal methods, LDA-HAN performs comparably to multimodal methods when evaluated on Politifact dataset, but not on GossipCop. This disparity could be due to the variation in context length between the two datasets, as revealed by the statistics in Table 1. Specifically, PolitiFact provides a longer context length with an average of 2,148 words per news, whereas GossipCop typically contains only 728 words per news. Thus, the unimodal methods can extract richer textual features from PolitiFact compared to GossipCop. Notably, although the unimodal T-BERT performs worse than the proposed SAMPLE, it demonstrates better performance than several multimodal methods in terms of F1 score and accuracy. We attribute this to the ensemble learning of T-BERT, which stacks three BERT models and shares the same weights during training.

Analysis of different prompt templates. The results indicate that mixed prompting (M-SAMPLE) outperforms C-SAMPLE and D-SAMPLE, with averaged improvements of 0.04 and 0.02 in F1, respectively. This finding suggests that continuous prompting is inferior to the discrete and mixed prompting methods. Specifically, the use of the C-SAMPLE may not provide enough prior human knowledge to aid the verbalizer in capturing the label words from the continuous space.

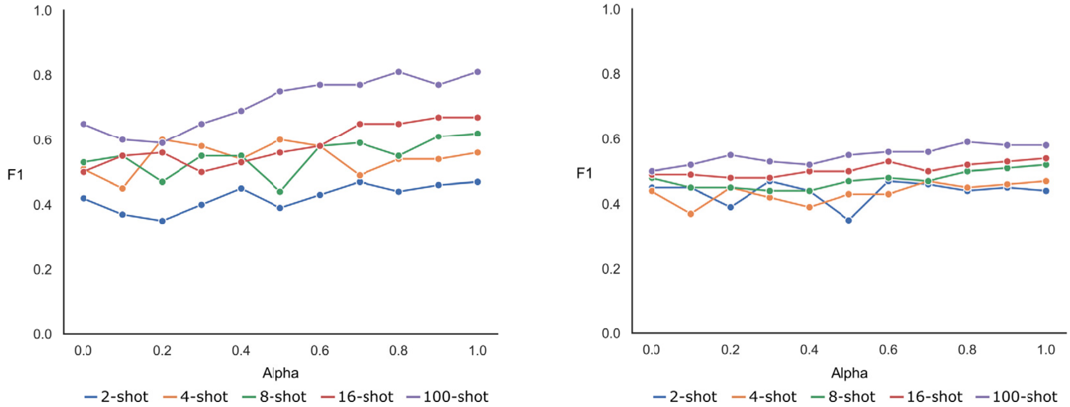
Overall, the experimental results indicate that the proposed SAMPLE method exhibits superior performance in the task of multimodal FND, regardless of whether the few-shot or data-rich setting being employed.

4.5. Analysis

This subsection provides a thorough analysis of the proposed SAMPLE method in both few-shot and data-rich settings. First, the significance of the image modality is evaluated. Next, the standard deviations of the proposed model in various data settings are presented. An ablation study further examines the key components of SAMPLE. Finally, we visualize and compare the embeddings generated by different baseline models.

4.5.1. Impact of image modality

The integration of semantic similarity between image and text features into multimodal representation in SAMPLE enables automatic adjustment of relevance across multiple modalities. However, this method does not allow direct measurement of the effectiveness of the image modality.

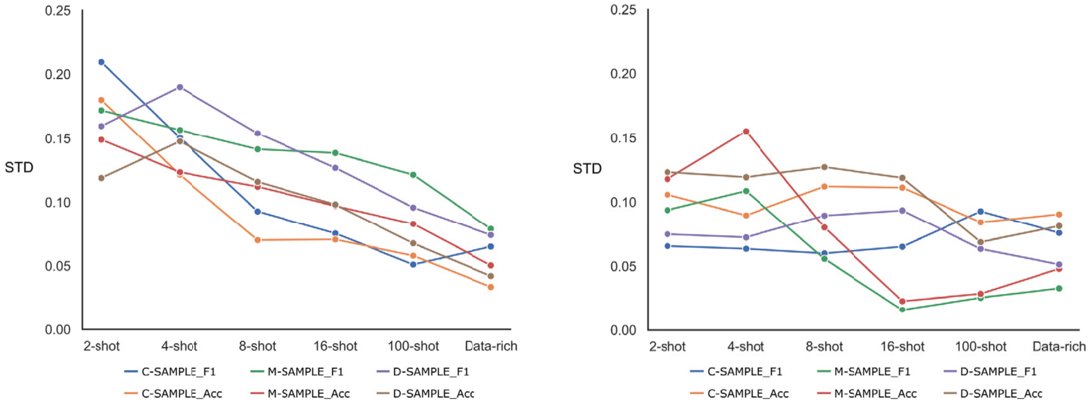


(a) PolitiFact (b) GossipCop

Fig. 3. The importance of the image modality in the proposed framework.

In order to comprehend the impact of the visual modality's contribution to the model inference, we introduce an adjustable parameter, the parameter α, to regulate the level of involvement of the visual modality in the few-shot training procedure. To be precise, the fused multimodal feature is multiplied by $\alpha \in [0, 1]$. Setting α to 0 removes the contribution of the visual modality, while setting α to 1 fully utilizes both the image and textual modalities. In this experiment, we apply M-SAMPLE, which achieves the highest F1. Based on the results depicted in Fig. 3, M-SAMPLE attains a higher F1 as α increases, suggesting that the involvement of the visual modality can enhance model performance. However, we also observe instances where the inclusion of the visual modality leads to a decrease in the F1, especially when the number of training samples is relatively small, such as in 2-shot, 4-shot, and 8-shot settings. This reveals that the presence of visual modality features may have a detrimental effect on overall performance in the few-shot settings when there is limited correlation with other modalities.

4.5.2. Stability test

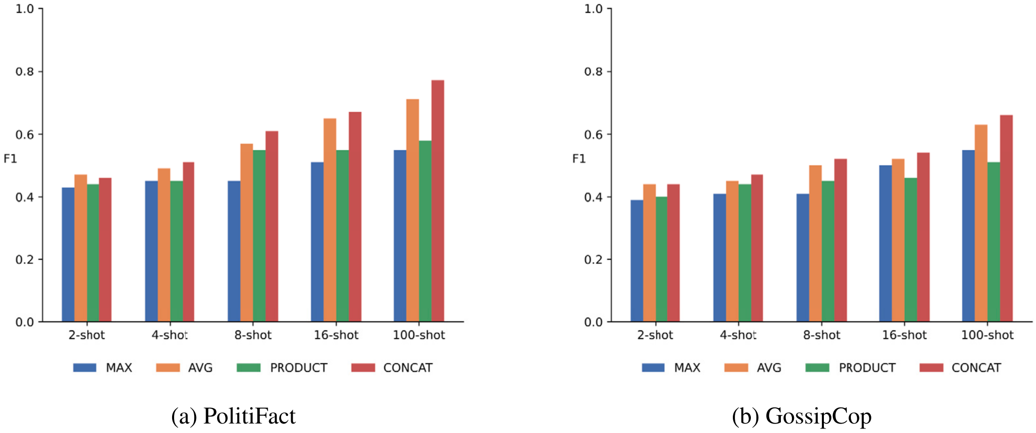


(a) PolitiFact (b) GossipCop

Fig. 4. The standard deviation of the F1 and accuracy in the proposed framework.

In this study, we evaluate the stability of the SAMPLE model by measuring the standard deviation of both F1 and accuracy in the few-shot and data-rich settings. As illustrated in Fig. 4, we present the standard deviation averaged from the aforementioned five experiments conducted for each SAMPLE model. We observe that the standard deviation decreases as the number of training samples increases, particularly in the PolitiFact dataset, as shown in Fig. 4a. Moreover, the GossipCop dataset is relatively more unstable than The PolitiFact dataset, as shown in Fig. 4b. This could be attributed to the complexity of semantics in GossipCop, which also results in lower F1-score and accuracy for all models.

4.5.3. *Multimodal fusion strategies*



(a) PolitiFact

(b) GossipCop

Fig. 5. The comparison of different multimodal fusion strategies.

As shown in Fig. 5, we employ M-SAMPLE to assess the impact of various multimodal fusion methods on model performance. Specifically, four rule-based fusion strategies [1] are compared in terms of their F1 scores. The motivation behind utilizing rule-based fusion strategies is rooted in the capability of the CLIP model to generate effective temporal alignment between unimodal features [47], as shown in Appendix C. Specifically, the notation “MAX” indicates that the multimodal features are fused using a max-pooling layer. Similarly, “AVG” denotes fusion through an average pooling layer. On the other hand, “PRODUCT” signifies that the multimodal features are formed by taking the element-wise product of all the unimodal features. Lastly, “CONCAT” means the concatenation of the unimodal features to create the multimodal features. The results suggest that in the few-shot settings, the concatenation of two unimodal features yields better F1 score compared to other fusing strategies.

4.5.4. *Trainable parameters comparisons*

We compare the number of trainable parameters between the baselines and SAMPLE, as shown in Table 3. The trainable parameters in the SAMPLE frameworks are rather small, with the majority derived from the verbalizer and templates. In contrast, when fully fine-tuning the entire model, FT-RoBERTa has the highest number of trainable parameters. As a result, prompt learning requires lower computational costs compared to fine-tuning, while still achieving comparable results to other deep learning methods.

Table 3
Trainable parameters between models. #_Para denotes trainable parameters in millions.

Model	#_Para
LDA-HAN	0.17 m
T-BERT	10 m
SAFE	0.12 m
RIVF	8 m
Spotfake	13 m
CAFE	0.95 m
FT-RoBERTa	125 m
D-SAMPLE	0.64 m
S-SAMPLE	0.66 m
M-SAMPLE	0.64 m

Table 4
Experimental results of the ablation study based on M-SAMPLE. "-SA" represents the removal of automatic similarity adjustment from M-SAMPLE. "-IF" signifies the exclusion of image features from the CLIP model. "-TF" indicates the removal of text features from the CLIP model. "-MF" indicates the exclusion of multimodal features from the CLIP model, using only the text feature from the language model RoBERTa.

Data	Method	Few shot (F1/Acc)										Data rich (F1/Acc)	
		2		4		8		16		100			
PolitiFact	M-SAMPLE	0.47	0.56	0.56	0.61	0.62	0.66	0.67	0.70	0.81	0.82	0.80	0.81
	-SA	0.44	0.55	0.55	0.57	0.58	0.65	0.65	0.65	0.75	0.79	0.76	0.81
	-IF	0.43	0.53	0.51	0.57	0.55	0.63	0.60	0.61	0.73	0.77	0.75	0.78
	-TF	0.35	0.43	0.46	0.50	0.51	0.60	0.54	0.65	0.66	0.69	0.69	0.71
	-MF	0.32	0.48	0.43	0.51	0.46	0.56	0.55	0.57	0.63	0.69	0.65	0.65
GossipCop	M-SAMPLE	0.44	0.53	0.47	0.56	0.52	0.54	0.54	0.60	0.58	0.62	0.64	0.73
	-SA	0.43	0.50	0.45	0.57	0.50	0.51	0.50	0.59	0.55	0.69	0.59	0.75
	-IF	0.39	0.43	0.43	0.50	0.48	0.55	0.50	0.55	0.53	0.65	0.53	0.70
	-TF	0.37	0.49	0.38	0.49	0.41	0.50	0.44	0.50	0.49	0.56	0.51	0.65
	-MF	0.35	0.38	0.39	0.45	0.40	0.47	0.41	0.49	0.45	0.55	0.49	0.55

4.5.5. *Ablation study*

We examine the impact of key components in the SAMPLE framework by assessing its performance under various partial configurations. We employ M-SAMPLE, remove different components in each test, and train the framework from scratch. The results presented in Table 4 show that M-SAMPLE experiences a deterioration in performance when any of its components are removed in most of the tested setups. This indicates the effectiveness of each individual key module in SAMPLE. Specifically, we find that there is a slight decrease in performance when removing the automatic similarity adjustment "-SA". This observation highlights the importance of standardizing semantic similarity in the fusion of multimodal features. By doing so, it helps to reduce uncorrelated information in the classification of fake news, while also mitigating the noise from the multimodal features of different modalities.

Furthermore, removing the text feature ("-TF") from CLIP generally results in lower F1 scores and accuracy compared to removing the image feature ("-IF") within the framework. Our findings suggest that while the image modality proves to be valuable in FND, as illustrated in Fig. 3, it is important to note that text features remain critical in the prompt learning process. This is mainly due to the training objective of prompt learning, which focuses on recovering the masked token from templates. This objective primarily align with and utilize text features extracted from pre-trained models. The extraction of two text features from different pre-trained models, namely RoBERTa and CLIP, provides the classifier with more diverse and expressive textual information. On the other hand, the primary role of image features is to minimize noise that may arise from the disparities between different modalities.

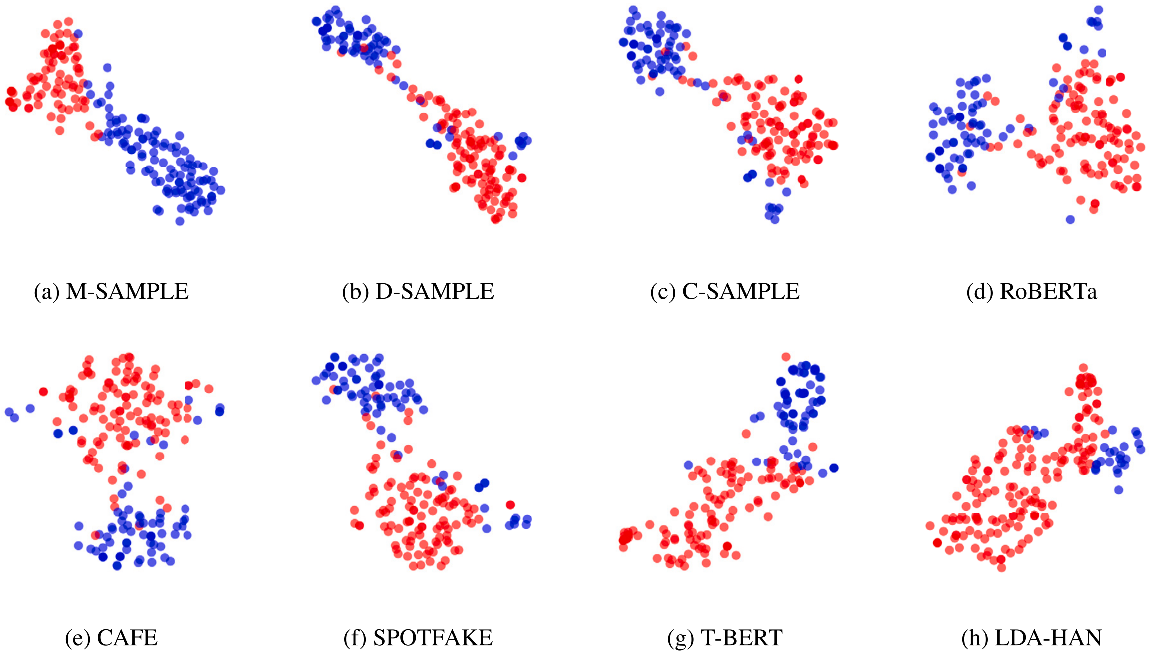
When the fused multimodal features ("-MF") obtained from the CLIP model is removed, the proposed framework comes down to the vanilla version of the prompt learning approach that leverages the pre-trained language model to directly predict FND. The analysis results reveal that even the basic prompt learning approach can outperform unimodal methods that solely rely on textual features. This observation emphasizes the superiority of prompt learning in FND.

4.5.6. *T-SNE visualization*

The features learned before the classifiers are analyzed on the test set of PolitiFact in the 2-shot setting as shown in Fig. 6. The reduced-dimensional feature representations of fake and real news are depicted by red and blue dots. From Fig. 6a, Fig. 6b, and Fig. 6c, we notice that the boundary of M-SAMPLE appears to be more sharply defined compared to that of D-SAMPLE and C-SAMPLE. This suggests that the learned feature representations in M-SAMPLE are more discriminative. While FT-RoBERTa demonstrates comparable performances in terms of F1 and accuracy, it does show some noticeable instances of misclassification within the 2-shot setting. Moreover, the learned feature representations in FT-RoBERTa tend to be sparser when compared to SAMPLE, as shown in Fig. 6d. This implies that in the few-shot scenario, the combination of multimodal features and prompt learning approach outperforms the standard fine-tuning method. We also visualize the feature representations from CAFE and SPOTFAKE, as demonstrated in Fig. 6e and Fig. 6f. The analysis reveals that the numbers of misclassified instances are significantly higher compared to those in prompt learning and fine-tuning methods. Furthermore, our findings indicate that unimodal methods like T-BERT and LDA-HAN exhibit the highest number of misclassified instances. This suggests that incorporating multimodal features can capture more expressive information compared to relying solely on textual features, as shown in Fig. 6g and Fig. 6h.

5. Discussion

The proposed SAMPLE framework integrates multiple prompt learning templates with a soft verbalizer to enable the automatic detection of fake news in few-shot and data-rich settings. Firstly, this section first analyses the relations between our approach and existing studies. Next, we elaborate on how the proposed SAMPLE approach can have a positive impact on the field of FND and offer support for real-world applications. Lastly, this paper addresses the limitations of our approach and outlines potential avenues for future work.


(a) M-SAMPLE (b) D-SAMPLE (c) C-SAMPLE (d) RoBERTa

(e) CAFE (f) SPOTFAKE (g) T-BERT (h) LDA-HAN

Fig. 6. T-SNE visualizations of features learned before the classifier from M-SAMPLE, D-SAMPLE, C-SAMPLE, RoBERTa, CAFE, SPOTFAKE, T-BERT and LDA-HAN on the test set of PolitiFact in the 2-shot setting.

5.1. Connections and comparison with previous works

SAMPLE demonstrates satisfactory performance in detecting fake news, whether in few-shot or data-rich scenarios. When comparing SAMPLE with other approaches in the FND field, traditional approaches can be classified into three categories: (1) unimodal approaches based solely on text or image features [6,8]; (2) multimodal approaches that assimilate textual and visual features via either pre-trained models or deep learning representation [44]; and (3) standard fine-tuning approaches that fine-tune pre-trained unimodality models with task-specific data [33].

In this study, SAMPLE encompasses a hybrid approaches of (2) and (3). However, it differs from the standard fine-tuning method due to its utilization of a prompt learning algorithm. Although fine-tuning has the potential to achieve optimal performance, it consumes a significantly larger amount of memory. This is because fine-tuning updates the entire set of model parameters to cater to a task-specific objective. In contrast, prompt learning, which leverages a natural language prompt to query a language model, maintains the similar objective of pre-training while shows comparable performance, particularly with limited training instances. By comparing the results of the standard fine-tuning with those of SAMPLE, the experimental findings confirm the aforementioned reasoning, as depicted in Table 2.

Prior multimodal approaches, such as CAFE and SAFE, relied on external cross-modal modules to align and measure disparate unimodality features. However, such external modules require a sufficient number of training instances to capture cross-modal correlations, which often results in inadequate performance, particularly in the few-shot setting. Our noval proposal introduces a similarity-aware multimodal feature fusion methodology that leverages the pre-training strategy of CLIP. CLIP utilizes numerous image-text pairs to learn the integration of multimodal semantics. Moreover, the standardization of cross-modal feature correlations incorporates a Sigmoid function to determine the semantic similarity between text and image inputs. An ablation study was conducted to investigate our approach in the few-shot setting, as depicted in Table 4. The results clearly demonstrate a significant improvement in few-shot performance, that is attributed to the combination of prompt learning and the proposed similarity-aware multimodal fusion process.

5.2. Contributions to future research

We introduce a novel FND framework, SAMPLE, for identifying fake news using prompt learning. Although prompt learning has demonstrated high performance in numerous classification tasks, the integration of different prompting strategies with multimodal features remains underexplored. This paper presents a promising method that achieves impressive and robust results and can serve as a significant baseline for future research in multimodal FND.

Traditional multimodal FND systems typically require substantial quantities of training data to attain satisfactory performance levels. However, the acquisition of annotated data is challenging in real-world settings. This paper demonstrates that SAMPLE offers comparable results, particularly in few-shot scenarios, indicating its capability to detect fake news in real-world situations. Moreover, the proposed approach that fuses similarity-aware multimodal features with prompt learning holds potential for future classification tasks of a similar nature.

5.3. Limitations and future work

The present study has several limitations. Firstly, SAMPLE primarily focuses on investigating the effects of the soft verbalizer, which is designed to automatically identify appropriate label words from the vocabulary. However, optimizing the soft verbalizer in a broader vocabulary under low-data conditions remains a considerable challenge. This suggests that additional adaptive modifications are necessary to improve the overall performance. Secondly, the newly proposed multimodal fusing method is based on a similarity-aware strategy that aims to reduce noise injection in cross-modal features with weaker correlations. It does not explicitly consider the uncorrelated cross-modal relations. Thirdly, there remains a need to explore further multimodal FND approaches that encompass different modalities, such as news entities and social networks.

Several studies indicated that the selection of verbalizers considerably affects performance. Manual verbalizers [38], in particular, rely on task-specific prior knowledge and require substantial labor to identify label words that represent classes. On the other hand, although the soft verbalizer [12] aims to ease this process, effectively optimizing it for a large vocabulary in low-data settings remains challenging. Moreover, the knowledgeable prompt-tuning approach [17] utilizes external knowledge bases to expand the coverage of the label words and reduce the bias associated with manual verbalizers. Investigating the impact of different verbalizers will be part of our future work. Additionally, integrating other modalities such as news entities, topics and social networks hold the potential to further expand the multimodal fusing method in the future.

6. Conclusion

This paper presents a novel similarity-aware multimodal FND framework named SAMPLE that utilizes prompt learning. To mitigate the data insufficiency issue, SAMPLE incorporates three popular prompt templates: discrete prompting, continuous prompting and mixed prompting to the original input text. The pre-trained language model RoBERTa is employed to acquire text features from the prompt. Furthermore, the pre-trained CLIP model is used to obtain the input texts, images, and their semantic similarities. To address semantic gaps and improve the collaboration between image and text modalities, we introduce a similarity-aware multimodal features fusing approach that applies standardization and a Sigmoid function to adjust the intensity of the final cross-modal representation. Finally, the multimodal features are fed into a fully-connected layer to project and obtain the word distribution that corresponds to the specific news class.

We conducted a multimodal FND experiment on two benchmark datasets to evaluate the proposed approach. SAMPLE's performance is extensively compared with unimodal, multimodal, and standard fine-tuning approaches. Our experimental results demonstrate that SAMPLE's performance is superior to previous methods, regardless of the few-shot or data-rich settings. Moreover, our results show that, although image modality provides meaningful information, the uncorrelated cross-modal features can negatively impact the FND performance, especially when the training instances are limited in quantity. Additionally, each component of our approach, particularly the standardized multimodal feature fusing module, helps unimodal features from pre-trained models collaborate more effectively in mining crucial features for FND.

CRediT authorship contribution statement

Ye Jiang: Conceptualization, Methodology, Writing – original draft. **Xiaomin Yu:** Software, Writing – review & editing. **Yimin Wang:** Project administration, Writing – review & editing. **Xiaoman Xu:** Data curation, Writing – review & editing. **Xingyi Song:** Supervision, Writing – review & editing. **Diana Maynard:** Supervision, Writing – review & editing.

Declaration of competing interest

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

Data availability

The authors do not have permission to share data.

Acknowledgements

This work has been funded by the Natural Science Foundation of Shandong Province under grant ZR2023QF151, and partially supported by a European Union Horizon 2020 Program under the scheme "INFRAIA-01-2018-2019 – Integrating Activities for Advanced Communities", Grant Agreement n.871042 ("SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics" (http://www.sobigdata.eu)).

Appendix A. Prompt engineering for discrete templates

In order to assess the impact of various templates on performance, we created discrete templates, as outlined in Table 5. Due to the time and cost involved in prompt engineering, we have restricted our study to only five discrete templates in this paper. Subsequently, we choose the discrete template that attains the highest F1 score as the final template in our experiment.

Table 5
The prompt engineering for the discrete templates. All experiments are conducted on Politifact with fixed seed in 2-shot and alpha = 0.8 settings.

Prompt	F1	Acc
This is $< mask >$.	0.41	0.43
This is $< mask >$ news.	0.41	0.47
This news is $< mask >$.	0.39	0.44
This is a piece of $< mask >$ news.	0.43	0.45
This is a piece of news with $< mask >$ information.	0.46	0.51

Appendix B. Comparing with different initialization for continuous templates

The study compares three initialization methods for the $< soft >$ token in the continuous template as demonstrated in Table 6. The "Random" initialization method initializes the $< soft >$ tokens randomly. The "FC" method reparameterizes the $< soft >$ tokens with another trainable matrix and forward propagates it through an FC layer [25]. The "LSTM" method feeds the $< soft >$ token through an LSTM layer and employs the outputs as the trainable vectors [28]. Although the performances of the three initialization methods in terms of F1 and accuracies showed slight variations, the study also observed that the "FC" and "LSTM" initializations result in later convergence of validation loss compared to the "Random" initialization. This was attributed to the need for additional training to obtain the $< soft >$ vectors.

Table 6
Different initialization for soft templates. All experiments are conducted on the Gossipcop with fixed seed in 8-shot and alpha = 1 settings.

Init methods	F1	Acc
Random	0.47	0.55
FC	0.47	0.51
LSTM	0.45	0.49

Appendix C. Comparison between CLIP and the pre-trained unimodal models for feature extraction

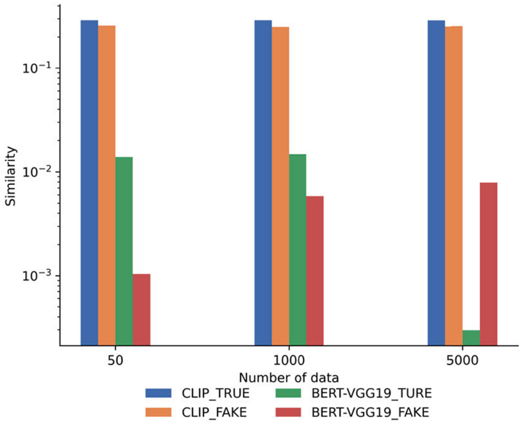


Fig. 7. Logarithmically scaled semantic similarity comparison between the pre-trained models.

We evaluated the semantic similarity between text and image features from various pre-trained models. Specifically, we applied the BERT model and VGG-19 to extract features from each training sample. Subsequently, the average similarity score was calculated in order to assess the semantic similarity between real and fake news. Similarly, we employ the CLIP text transformer and vision transformer to extract unimodal features and calculate their semantic similarity. We also increase the number of samples to observe any changes in semantic similarity. Lastly, to accurately represent the small differences between unimodal models, we logarithmically scaled the values on the axes.

Our experimental findings indicate that the text and image features extracted from the CLIP model exhibit higher consistent compared to those obtained from unimodal models, as shown in Fig. 7. This can be attributed to the capacity of CLIP to learn multimodal representations through joint training and its utilization of a contrastive loss function, which aids in distinguishing relevant pairs from irrelevant ones. As a result, the semantic similarity of real news (CLIP_TRUE) was consistently higher than that of fake news (CLIP_FAKE) regardless of the number of samples. In contrast, the BERT-VGG19 combination separately extracts features from text and images, which may introduce more noise during the feature extraction process.

References

[1] P.K. Atrey, M.A. Hossain, A. El Saddik, M.S. Kankanhalli, Multimodal fusion for multimedia analysis: a survey, Multimed. Syst. 16 (2010) 345–379.

[2] S. Bhatt, N. Goenka, S. Kalra, Y. Sharma, Fake news detection: experiments and approaches beyond linguistic features, in: Data Management, Analytics and Innovation, Springer, 2022, pp. 113–128.

[3] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.

[4] A. Bondielli, F. Marcelloni, A survey on fake news and rumour detection techniques, Inf. Sci. 497 (2019) 38–55.

[5] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Adv. Neural Inf. Process. Syst. 33 (2020) 1877–1901.

[6] J. Cao, P. Qi, Q. Sheng, T. Yang, J. Guo, J. Li, Exploring the role of visual content in fake news detection, in: Disinformation, Misinformation, and Fake News in Social Media, 2020, p. 141.

[7] B. Chen, B. Chen, D. Gao, Q. Chen, C. Huo, X. Meng, W. Ren, Y. Zhou, Transformer-based language model fine-tuning methods for Covid-19 fake news detection, in: International Workshop on Combating Online Hostile Posts in Regional Languages During Emergency Situation, Springer, 2021, pp. 83–92.

[8] X. Chen, C. Dong, J. Ji, J. Cao, X. Li, Image manipulation detection by multi-view multi-scale supervision, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 14185–14193.

[9] Y. Chen, D. Li, P. Zhang, J. Sui, Q. Lv, L. Tun, L. Shang, Cross-modal ambiguity learning for multimodal fake news detection, in: Proceedings of the ACM Web Conference 2022, 2022, pp. 2897–2905.

[10] J. Devlin, M.W. Chang, K. Lee, K. Toutanova, Bert: pre-training of deep bidirectional transformers for language understanding, arXiv preprint, 2018.

[11] C.B. El Vaigh, T. Girault, C. Mallart, D. Nguyen, Detecting fake news conspiracies with multitask and prompt-based learning, in: MediaEval 2021-MediaEval Multimedia Evaluation Benchmark, Workshop, 2021.

[12] T. Gao, A. Fisch, D. Chen, Making pre-trained language models better few-shot learners, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, vol. 1: Long Papers, 2021, pp. 3816–3830.

[13] K. Hambardzumyan, H. Khachatrian, J. May, Warp: word-level adversarial reprogramming, arXiv preprint, arXiv:2101.00121, 2021.

[14] B. Han, X. Han, H. Zhang, J. Li, X. Cao, Fighting fake news: two stream network for deepfake detection via learnable SRM, IEEE Trans. Biom. Behav. Identity Sci. 3 (2021) 320–331.

[15] X. Han, W. Zhao, N. Ding, Z. Liu, M. Sun, Ptr: prompt tuning with rules for text classification, AI Open (2022).

[16] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[17] S. Hu, N. Ding, H. Wang, Z. Liu, J. Wang, J. Li, W. Wu, M. Sun, Knowledgeable prompt-tuning: incorporating knowledge into prompt verbalizer for text classification, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, vol. 1: Long Papers, 2022, pp. 2225–2240.

[18] J. Hua, X. Cui, X. Li, K. Tang, P. Zhu, Multimodal fake news detection through data augmentation-based contrastive learning, Appl. Soft Comput. 136 (2023) 110125.

[19] G. Jiang, S. Liu, Y. Zhao, Y. Sun, M. Zhang, Fake news detection via knowledgeable prompt learning, Inf. Process. Manag. 59 (2022) 103029.

[20] Y. Jiang, X. Song, C. Scarton, A. Aker, K. Bontcheva, Categorising fine-to-coarse grained misinformation: an empirical study of Covid-19 infodemic, arXiv preprint, arXiv:2106.11702, 2021.

[21] Y. Jiang, Y. Wang, X.S.D. Maynard, Comparing topic-aware neural networks for bias detection of news, in: Proceedings of 24th European Conference on Artificial Intelligence (ECAI 2020), International Joint Conferences on Artificial Intelligence (IJCAI), 2020.

[22] J. Jing, H. Wu, J. Sun, X. Fang, H. Zhang, Multimodal fake news detection via progressive fusion networks, Inf. Process. Manag. 60 (2023) 103120.

[23] V. Khullar, H.P. Singh, f-fnc: privacy concerned efficient federated approach for fake news classification, Inf. Sci. 639 (2023) 119017.

[24] P. Li, X. Sun, H. Yu, Y. Tian, F. Yao, G. Xu, Entity-oriented multi-modal alignment and fusion network for fake news detection, IEEE Trans. Multimed. (2021).

[25] X.L. Li, P. Liang, Prefix-tuning: optimizing continuous prompts for generation, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, vol. 1: Long Papers, Association for Computational Linguistics, 2021, pp. 4582–4597 (online), https://aclanthology.org/2021.acl-long.353.

[26] Z. Lin, B. Liang, Y. Long, Y. Dang, M. Yang, M. Zhang, R. Xu, Modeling intra- and inter-modal relations: hierarchical graph contrastive learning for multimodal sentiment analysis, in: Proceedings of the 29th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 7124–7135, https://aclanthology.org/2022.coling-1.622.

[27] X. Liu, K. Ji, Y. Fu, W. Tam, Z. Du, Z. Yang, J. Tang, P-tuning: prompt tuning can be comparable to fine-tuning across scales and tasks, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, vol. 2: Short Papers, 2022, pp. 61–68.

[28] X. Liu, Y. Zheng, Z. Du, M. Ding, Y. Qian, Z. Yang, J. Tang, Gpt understands, too, arXiv preprint, arXiv:2103.10385, 2021.

[29] Y. Long, Q. Lu, R. Xiang, M. Li, C.R. Huang, Fake news detection through multi-perspective speaker profiles, in: Proceedings of the Eighth International Joint Conference on Natural Language Processing, vol. 2: Short Papers, Asian Federation of Natural Language Processing, Taipei, Taiwan, 2017, pp. 252–256, https://aclanthology.org/I17-2043.

[30] P. Meel, D.K. Vishwakarma, Fake news, rumor, information pollution in social media and web: a contemporary survey of state-of-the-arts, challenges and opportunities, Expert Syst. Appl. 153 (2020) 112986.

[31] P. Meel, D.K. Vishwakarma, Han, image captioning, and forensics ensemble multimodal fake news detection, Inf. Sci. 567 (2021) 23–41.

[32] S.M.M. Monterrubio, A. Noain-Sánchez, E.V. Pérez, R.G. Crespo, Coronavirus fake news detection via MedOSINT check in health care official bulletins with CBR explanation: the way to find the real information source through OSINT, the verifier tool for official journals, Inf. Sci. 574 (2021) 210–237.

[33] D.Q. Nguyen, T. Vu, A.T. Nguyen, Bertweet: a pre-trained language model for English tweets, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020, pp. 9–14.

[34] J. Pennington, R. Socher, C. Manning, Glove: global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532–1543.

[35] G. Qin, J. Eisner, Learning how to ask: querying LMS with mixtures of soft prompts, arXiv preprint, arXiv:2104.06599, 2021.

[36] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.

[37] T. Schick, H. Schütze, Exploiting cloze questions for few shot text classification and natural language inference, arXiv preprint, arXiv:2001.07676, 2020.

[38] T. Schick, H. Schütze, It's not just size that matters: small language models are also few-shot learners, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 2339–2352.

[39] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, H. Liu, Fakenewsnet: a data repository with news content, social context and spatialtemporal information for studying fake news on social media, arXiv preprint, arXiv:1809.01286, 2018.

[40] K. Shu, G. Zheng, Y. Li, S. Mukherjee, A.H. Awadallah, S. Ruston, H. Liu, Early detection of fake news with multi-source weak social supervision, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2021, pp. 650–666.

[41] P. Singh, R. Srivastava, K. Rana, V. Kumar, Semi-fnd: stacked ensemble based multimodal inferencing framework for faster fake news detection, Expert Syst. Appl. 215 (2023) 119302.

[42] S. Singhal, R.R. Shah, T. Chakraborty, P. Kumaraguru, S. Satoh, Spotfake: a multi-modal framework for fake news detection, in: 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM), IEEE, 2019, pp. 39–47.

[43] X. Song, J. Petrak, Y. Jiang, I. Singh, D. Maynard, K. Bontcheva, Classification aware neural topic model and its application on a new Covid-19 disinformation corpus, arXiv preprint, arXiv:2006.03354, 2020.

[44] N.M.D. Tuan, P.Q.N. Minh, Multimodal fusion with BERT and attention mechanism for fake news detection, in: 2021 RIVF International Conference on Computing and Communication Technologies (RIVF), IEEE, 2021, pp. 1–6.

[45] H. Wang, P. Tang, H. Kong, Y. Jin, C. Wu, L. Zhou, Dhcf: dual disentangled-view hierarchical contrastive learning for fake news detection on social media, Inf. Sci. (2023) 119323.

[46] X. Zhou, J. Wu, R. Zafarani, Safe: similarity-aware multi-modal fake news detection, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, 2020, pp. 354–367.

[47] Y. Zhou, Q. Ying, Z. Qian, S. Li, X. Zhang, Multimodal fake news detection via clip-guided learning, arXiv preprint, arXiv:2205.14304, 2022.

[48] Y. Zhu, X. Zhou, J. Qiang, Y. Li, Y. Yuan, X. Wu, Prompt-learning for short text classification, arXiv preprint, arXiv:2202.11345, 2022.