
Article:
Jiang, Y. orcid.org/0000-0002-6683-0205, Yu, X. orcid.org/0000-0003-4846-3162, Wang, Y. orcid.org/0000-0002-8835-3825 et al. (3 more authors) (2023) Similarity-aware multimodal prompt learning for fake news detection. Information Sciences, 647. 119446. ISSN 0020-0255

https://doi.org/10.1016/j.ins.2023.119446

Similarity-Aware Multimodal Prompt Learning for Fake News Detection

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

Online fake news explosion has posed significant challenges to academics and industries by overloading fact-checkers and social media. The standard paradigm for fake news detection relies on utilizing text information to model news' truthfulness. However, the subtle nature of online fake news makes it challenging to use textual information alone to debunk it. Recent studies, focusing on multimodal fake news detection, have outperformed text-only methods. Deep learning approaches, primarily utilizing pre-trained models, to extract unimodal features, or fine-tuning the pre-trained model, has become a new paradigm for detecting fake news. Nevertheless, this paradigm may require a large number of training instances or updating the entire set of pre-trained model parameters, making it impractical for real-world fake news detection. In addition, traditional multimodal methods directly fuse the cross-modal features without considering that the uncorrelated semantic representation may introduce noise into the multimodal features. To address these issues, this paper proposed 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. Additionally, we introduced the similarity-aware fusing method, which adaptively fuses the intensity of multimodal representation and mitigates noise injection via uncorrelated cross-modal features. Evaluation results show that SAMPLE outperformed previous works by achieving higher F1 and accuracy scores on two benchmark multimodal datasets, demonstrating its feasibility in real-world scenarios, regardless of data-rich or few-shot settings.

1. Introduction

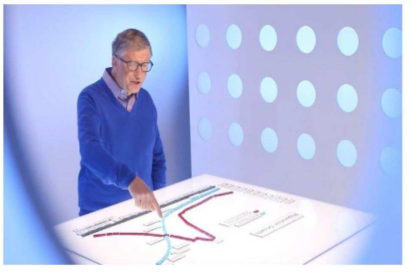
The increasing prevalence of social media has significantly impacted the way information is disseminated and consumed. While social media platforms provide an efficient way for people to seek and share information, the spread of fake news has caused substantial harm to the global community. In an effort to mitigate the impacts of online fake news, academia and industry have developed various techniques. For example, previous research (Ma et al., 2017; Bahad et al., 2019; Shu et al., 2019; Dun et al., 2021) has mainly focused on 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, Figure 1 shows two news snippets that are difficult to identify their truthfulness if only by looking at their text information. Therefore, multimodal Fake News Detection (FND) techniques (Wang et al., 2018; Khattar et al., 2019; Chen et al., 2022) have been developed recently to combine 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 modelling both textual and image representations from fake news. However, these methods are limited by the fact that they often need a considerable amount of annotated data to achieve good performance. Recently, there has been an increasing interest in FND for using large pre-trained models. Much research (Singhal et al., 2019, 2020; Bhatt et al., 2022; Zhou et al., 2020) has used pre-trained language models, such as BERT (Devlin et al., 2018a) and RoBERTa (Liu et al., 2019), and pre-trained vision models, such as ResNet (He et al., 2016) and Vision-Transformer (Dosovitskiy et al., 2020), to encode the textual and image features

*Corresponding author
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)
ORCID(s): 0000-0002-6683-0205 (Y. Jiang); 0000-0003-4846-3162 (X. Yu); 0000-0002-8835-3825 (Y. Wang); 0000-0001-7448-4587 (X. Xu); 0000-0002-4188-6974 (X. Song); 0000-0002-1773-7020 (D. Maynard)

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."

Figure 1: Two snippets of fake news and their original reports.

of the 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 (Jiang et al., 2022).

Fine-tuning is a common technique for adapting pre-trained models to different downstream tasks. Recent studies have already fine-tuned variants of BERT, including the pre-trained BERT itself, for use in FND (Aggarwal et al., 2020; Chen et al., 2021a; Kumar et al., 2021). However, fine-tuning requires a significant number of labeled instances to train additional classifiers, making it difficult to use in low-resource settings (Brown et al., 2020). 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. Additionally, fine-tuning updates all model parameters for a single task, creating challenges for real-world FND due to the size of pre-trained models (Liu et al., 2022). 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 (Schick and Schütze, 2020). Furthermore, prompt learning allows pre-trained models to achieve competitive performance in low-resource settings with limited labeled data (Brown et al., 2020), which is particularly important for real-world FND, where manually labeled fake news is scarce. However, current prompt-based FND approaches (Jiang et al., 2022) primarily consider textual information, and the analysis of cross-modality features in fake news posts is underdeveloped.

Compared to the fine-tuned model that outputs class distributions directly, prompt learning aligns with the language modeling objective, which generates specific answering words that are relevant to fake news detection by adding additional information before the original text inputs. For example, the news snippet on the left in Figure 1 can utilize a discrete prompt by adding a prompt before the original text (e.g., *"This is a piece of <mask> news. Former president and breaker of laws, Barack Obama..."*), with the goal of recovering the masked token from the prompt text. However, the discrete prompt has limitations since the embedding of template words must be the embedding of natural language words, and the template can only be parameterized by the pre-trained LM parameters instead of the LM parameters that can be tuned via a specific task such as fake news detection (Lester et al., 2021; Liu et al., 2021b). To address this issue, continuous prompting (Qin and Eisner, 2021; Zhong et al., 2021) 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 (Gu et al., 2022; Jiang et al., 2022) 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 (Wang et al., 2018; Singhal et al., 2019) 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 is less, resulting in the multimodal representation being noisy. 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 in the task of FND. In addition, the pre-trained model Contrastive Language-Image Pre-training (CLIP) (Radford et al., 2021) is applied to extract the text and image features, which are utilized to generate the multimodal representation. Meanwhile, the semantic similarity between the text and image features is also calculated to address the issue of uncorrelated semantic representation between image and text. To adjust the intensity of the aggregated multimodal representation, the semantic similarity is normalized. To assess the performance of the proposed SAMPLE framework, two domain-specific publicly accessible datasets, PolitiFact and GossipCop (Shu et al., 2018)), 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 from the pre-trained CLIP model.
- SAMPLE are evaluated on two benchmark multimodal fake news detection datasets and demonstrates that it outperforms previous approaches in both low-resource and data-rich scenarios.

2. Related work

In previous research, FND has been extensively examined (Zhang and Ghorbani, 2020). Specifically, 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 (Shu et al., 2017; Meel and Vishwakarma, 2020). In addition, many recent studies aim to differentiate false content from similar concepts, such as misinformation (Song et al., 2020; Jiang et al., 2021) and disinformation (Li et al., 2022). However, 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 (Meel and Vishwakarma, 2020). In this paper, we propose SAMPLE, which focuses primarily on FND but has the potential to extend to detecting misinformation and disinformation.

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 (Chen et al., 2021b) 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 (Li et al., 2021), as well as poor image quality (Han et al., 2021). In text modality, textual features are commonly coupled with social features based on statistical information to detect fake news content in the news description (Castillo et al., 2011; Kwon et al., 2013). Previous research (Ajao et al., 2019; Zhang et al., 2021) has investigated the connection between a publisher's emotions and social sentiment, which is closely tied to the accuracy of the news. In addition, logical soundness (Guo et al., 2018), grammatical errors (Potthast et al., 2017), and rhetorical structure (Conroy et al., 2015) are critical elements of FND. 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. The EANN (Wang et al., 2018) combines textual and visual features using a multi-task learning framework, which is designed to handle event classification and fake news detection simultaneously. One key aspect of this framework is that the event classification task removes post-specific information from fake news, while keeping invariant features that are useful for identifying rumors. Zhou et al. (2020) proposed the SAFE model, which uses the Image2Sentence model (Vinyals et al., 2016) to convert images to text captions, and extends the Text-CNN model (Kim, 2014) 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.

More recently, many studies have opted to utilize pre-trained models to extract textual and visual features in FND. For example, the CAFE (Chen et al., 2022) employs BERT and ResNet-34 (He et al., 2016) as pre-trained models for encoding textual and visual features, respectively. Similarly, Tuan and Minh (2021) utilized pre-trained BERT and VGG-19 models for extracting unimodal textual and visual features, respectively, and then applied a scaled dot-product attention mechanism to fuse the multimodal features. Qi et al. (2021) proposed entity-enhanced multimodal fusion framework to capture entity inconsistency, mutual enhancement, and text complementation in the fake news. Zhou et al. (2022) 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.

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, Yang et al. (2019) fine-tuned the pre-trained XLNet model for multi-class and binary class FND. The Ro-CT-BERT (Chen et al., 2021a) 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 (Shu et al., 2021), 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. (2022) developed PLST, which combines both text inputs and external knowledge from open knowledge graphs in short text classification tasks. Han et al. (2022) 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. (2021) utilized the prompt-based model from DistilGPT-2 in conjunction with multitask learning to detect coronavirus-related fake news in MediaEval-2021. Moreover, Jiang et al. (2022) proposed KPL, which detects fake news by integrating external knowledge. Nevertheless, KPL relies on human-designed prompts and verbalizers, which can be time-consuming and potentially unreliable. Additionally, it is not yet well understood 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 determine the image that is most relevant to a given news article's text, the pre-trained CLIP model is utilized to encode both the text representation and image representation separately. To achieve this, only the image with the highest cosine similarity to the text is kept while the rest are discarded.

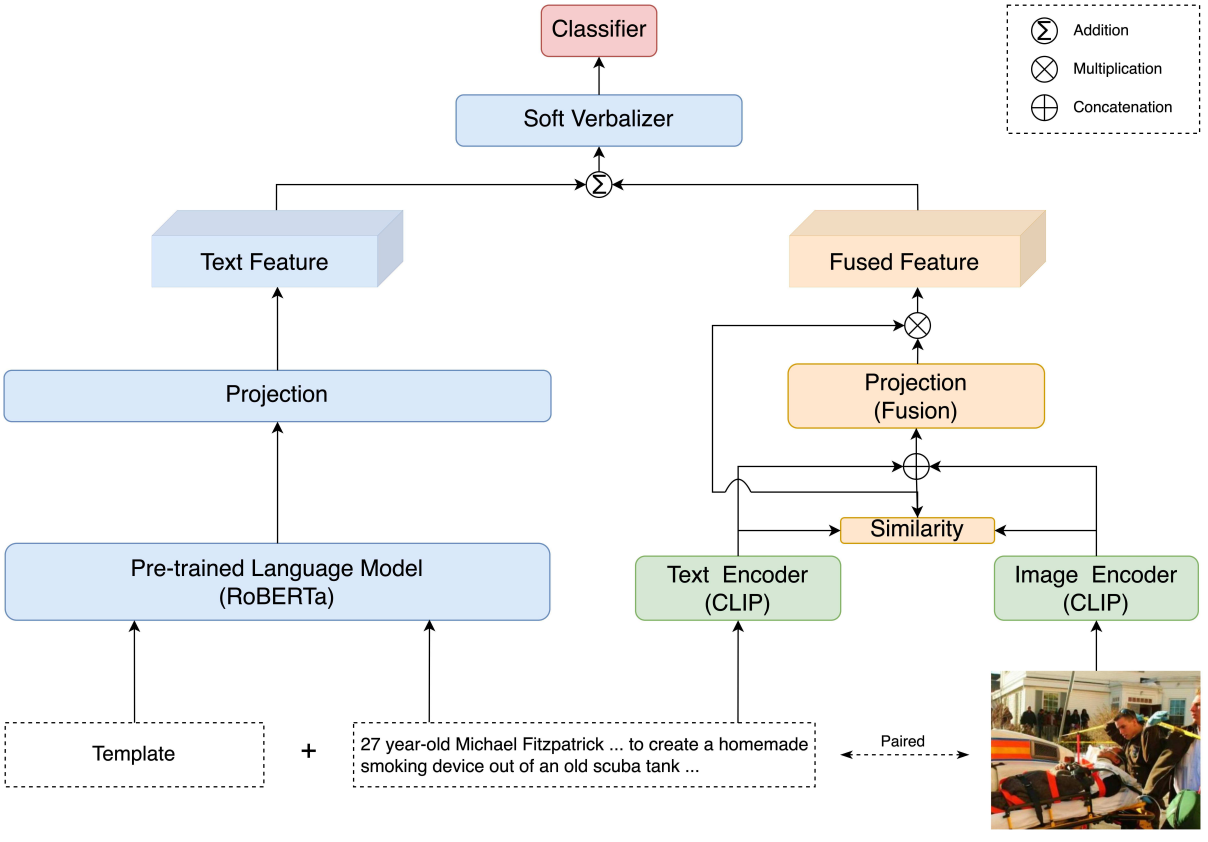


Figure 2: The overall structure of SAMPLE for fake news detection.

In this section, we utilize discrete prompting (Schick and Schütze, 2021; Tam et al., 2021), which primarily corresponds to natural language phrases and automatically searches for templates described in a discrete space. Additionally, we discuss the extended version called continuous prompting (Qin and Eisner, 2021; Zhong et al., 2021), which employs prompts containing pseudo-tokens not present in the pre-trained language model vocabulary. We further present 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. Figure 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 allow 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 singular keyword replaced with a mask. We investigate five discrete templates since the different templates might have a great 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 the sum of the human-designed templates. After which, we can calculate the representation of the masked word, related to the FND task target, 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, we calculate the hidden states of prompt x_d:

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

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

3.2. Continuous prompting

Although discrete prompting naturally inherits interpretability from the task description, the discrete prompt is limited because the embedding of template words is required to be natural language words, and the template is parameterized by the pre-trained language model's parameters (Liu et al., 2021a). 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 $\longrightarrow$ {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 $st =$ "$< soft_1 >, < soft_2 >, ... < soft_t >, < mask >$" 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:

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

where h^{st} and h_{mask}^{st} are the hidden vectors in the t length and the $< mask >$ token of 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 (Liu et al., 2021b; Han et al., 2022). 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 end of the mixed template $mt =$"$< h_{head}^{mt} > This\ is\ a\ piece\ of < h_{mask}^{mt} > news. < h_{tail}^{mt} >$ ".Like the discrete prompt, the new mixed prompt, x_m, can be expressed as $x_m = [mt; x]$. We then compute the hidden states as follows:

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

where h^{mt} and h_{mask}^{mt} are the hidden vectors in the m length and $< mask >$ token of the mixed template respectively.

3.4. Similarity-aware multimodal feature fusing

According to a previous study (Zhou et al., 2022), 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. 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. Conversely, pre-trained CLIP models utilize a large dataset of image-text pairs to capture semantic correlations beyond emotional features or noise patterns.

[1] We compares three initialization methods as shown in Appendix B, and the experimental results suggest that the random initialization achieves comparable performance with a slightly faster convergence of validation loss than that of the others.

To effectively integrate image and text features, we initially use the pre-trained CLIP model to extract these features separately. The CLIP model includes a text Transformer and 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}, 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 observed that certain news posts lacked explicit cross-modal semantic relations, regardless of whether they were real or fake. As a result, concatenating the unimodal features to create the fused feature could add noise in instances where 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, the soft verbalizer is utilized to map labels to their corresponding words. We use WARP (Hambardzumyan et al., 2021) in this study to identify the optimal prompt for which a pre-trained language model predicts the masked token by searching for a prompt in the continuous embedding space. We use soft verbalization in the three template types above to compare prompt methods. To identify the optimal parameter $\theta = \{\theta^P, \theta^V\}$ 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 concatenate with one of the prompt template from the above, and then the classification probability $P(y|x')$ can 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 consisted of political news and celebrity gossip, respectively, and are included in the FakeNewsNet project (Shu et al., 2018). Using the data crawling scripts provided, we retrieve 1,056 news items in PolitiFact and 22,140 news items in GossipCop. To reduce redundancy, we only preserve the relevant image, calculated by the pre-trained CLIP model based on the text and the images' cosine similarity, for news with multiple images. We also exclude news with no images or invalid image URLs. The resulting dataset statistics are presented in Table 1.

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

Statistics	PolitiFact	GossipCop
Total news	198	6,805
Fake news	96	1,877
Real news	102	4,928
Words per news	2,148	728

4.2. Implementation details

We adopt the pre-trained RoBERTa from the HuggingFace library (Wolf et al., 2020) as the main block for prompt learning. We extract text and image features using the text encoder and image encoder, respectively, from the pre-trained CLIP (ViT-B-32) model. The size of the hidden layer's projection layers is assigned as 768, and the dropout rate is 0.6. We use the AdamW optimizer to optimize the model parameters, and the learning rate 3e−5 and the decay parameter 1e−3 are empirically set. The model is trained in 20 epochs, and we choose the model checkpoints that obtain the best validation performance to test. We evaluate the method in both few-shot and data-rich settings.

In the few-shot setting, our model is trained with a small number of instances randomly sampled from the dataset, i.e., $k \in [2, 4, 8, 16, 100]$, and the rest of the instances are used for testing. Also, a validation set the same size as the training set is created for model selection. The PolitiFact dataset contains a limited number of news items, and to compensate for it, we adopt a special configuration called the PolitiFact 100-shot setting, where we use 100 instances for training and 50 for developing. As the quality of the training set and validation set has a significant impact on the model's performance, we repeat the above data sampling method five times with different random seeds and report the average score, excluding the highest and lowest ones, for the few-shot setting. For the few-shot setting, we report the average score of our model, computed by the mean of the scores without the maximum and the minimum ones. Additionally, we balance the number of labelled instances across the training and validation sets during the training phase.

In the data-rich setting, we split the two datasets 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 different 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 compare the proposed SAMPLE to several models that have previously achieved state-of-the-art performances in FND dataset. Specifically, we compare unimodal approaches (1-2), multimodal approaches (3-6), and the standard fine-tuning approach (7). To initialize our word embeddings, we exploit the pre-trained 100-dim 6 billion Glove embeddings (Pennington et al., 2014).

(1) **LDA-HAN** (Jiang et al., 2020): This model incorporates Latent Dirichlet Allocation (LDA) (Blei et al., 2003) topic distributions into a hierarchical attention network.
(2) **T-BERT** (Bhatt et al., 2022): This feature-based method uses concatenated triple BERT models to predict fake news.
(3) **SAFE** (Zhou et al., 2020): This model converts images into their text descriptions and uses the relevance between textual and visual information to detect fake news.

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

(4) **RIVF** (Tuan and Minh, 2021): 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** (Singhal et al., 2019): This model uses pre-trained image model VGG and BERT to extract respective image and text features, concatenating them to classify fake news.

(6) **CAFE** (Chen et al., 2022): 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 fine-tuning approach, multimodal and unimodal FND methods.

Comparing with fine-tuning. First, we investigate the performances between the standard fine-tuned RoBERTa (FT-RoBERTa) and the proposed SAMPLE by evaluating the F1 score. 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 to FT-RoBERTa, and all SAMPLE methods outperform FT-RoBERTa in 0.05, 0.024 and 0.035 respectively. This improvement is more significant with the decrease in the training samples, showing the superiority of prompt learning in low-resource scenarios.

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 accuracies between SAMPLE methods and FT-RoBETRa is similar to the above observation, showing the superiority of the proposed method in utilizing PLM information, especially when the training data is scarce, but the standard fine-tuning can still be a strong baseline in a data-rich setting.

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 outperformed previous methods in all settings. For example, with PolitiFact dataset,

M-SAMPLE achieved a maximum 0.29 improvement in the 100-shot setting compared to CAFE. This observation 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. In contrast, pre-trained models for unimodal tasks, such as BERT and ResNet-34 used by CAFE, might not be effective in capturing unimodal features with heterogeneous feature distributions.

The above characteristics are also applicable to SpotFake. SpotFake extracts text and image features using BERT and VGG19, respectively. However, our experiments demonstrate that SpotFake performs better on our 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 exhibits a performance on par with multimodal methods when tested on Politifact, but not on GossipCop. This disparity could be due to the variance in context length between the two datasets, as revealed by the data in Table 1. Specifically, PolitiFact contains a longer context length with 2,148 words per news than GossipCop, which contains only 728 words per news. Thus, the unimodal methods can extract more rich 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. Despite the potentially higher computational cost associated with stacking multiple pre-trained language models, our findings also demonstrate the effectiveness of ensemble learning in FND.

Analysis of different prompt templates. The results indicate that mixed prompting (M-SAMPLE) outperforms C-SAMPLE and D-SAMPLE, with an average improvement 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 demonstrate that the proposed SAMPLE method exhibits superior performance in the task of multimodal FND, regardless of whether the few-shot or data-rich setting is employed.

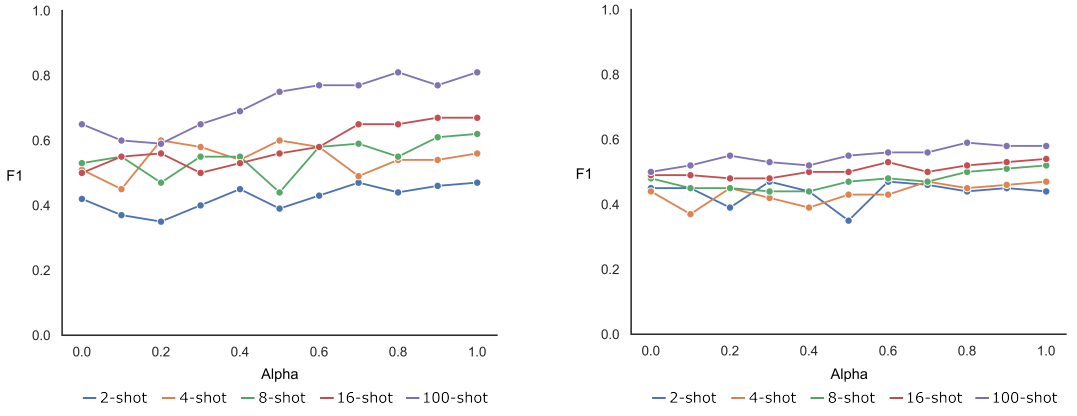
4.5. Analysis

This subsection provides a thorough analysis of the proposed SAMPLE method in both few-shot and data-rich settings. First, we evaluate the significance of the image modality. Next, we present the standard deviations of the proposed model in various data settings. An ablation study further examines the key components of SAMPLE. Finally, we visualize and compare the embeddings from different baselines.

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.

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. Precisely, the fused multimodal feature is multiplied by $\alpha \in [0, 1]$. By setting α to 0, we eliminate the involvement of the visual modality. Conversely, if we set α to 1, both image and textual modalities are fully utilized. In this experiment, we apply M-SAMPLE, which achieves the highest F1. Based on the results depicted in Figure 3, M-SAMPLE attains a higher F1 as α increases, thereby indicating that the involvement of the visual modality can improve model performance. Nevertheless, we also observe that, in some instances, the inclusion of the visual modality leads to a decrease in the F1, especially when the number of training samples available is relatively small, such as in 2-shot, 4-shot, and 8-shot settings. This reflects that the features from the visual modality might harm the overall performance when there is less correlation with other modalities in the few-shot settings.

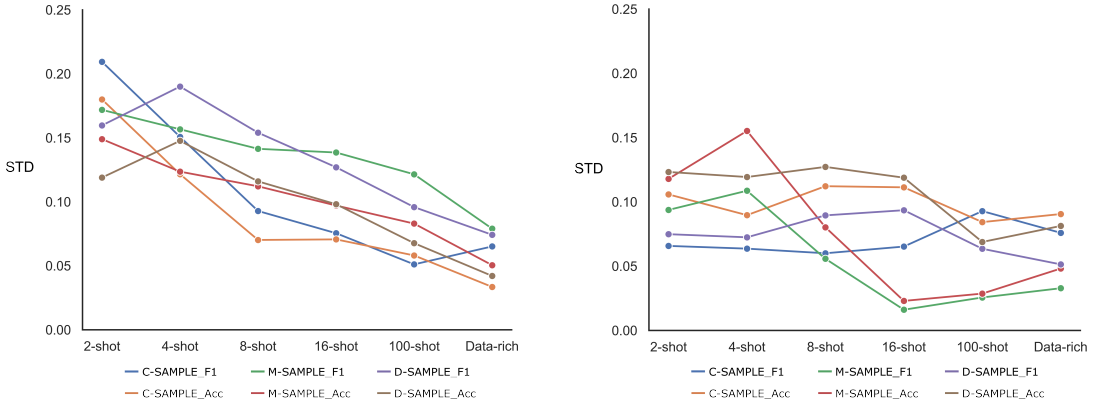


(a) PolitiFact (b) GossipCop

Figure 3: The importance of the image modality in the proposed framework.

4.5.2. Stability test

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 Figure 4, we present the standard deviation of 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 Figure 4a. Moreover, the GossipCop dataset is relatively more unstable than The PolitiFact dataset, as shown in Figure 4b. This could be attributed to the complexity of semantics in GossipCop, which results in lower F1-score and accuracy for all models.



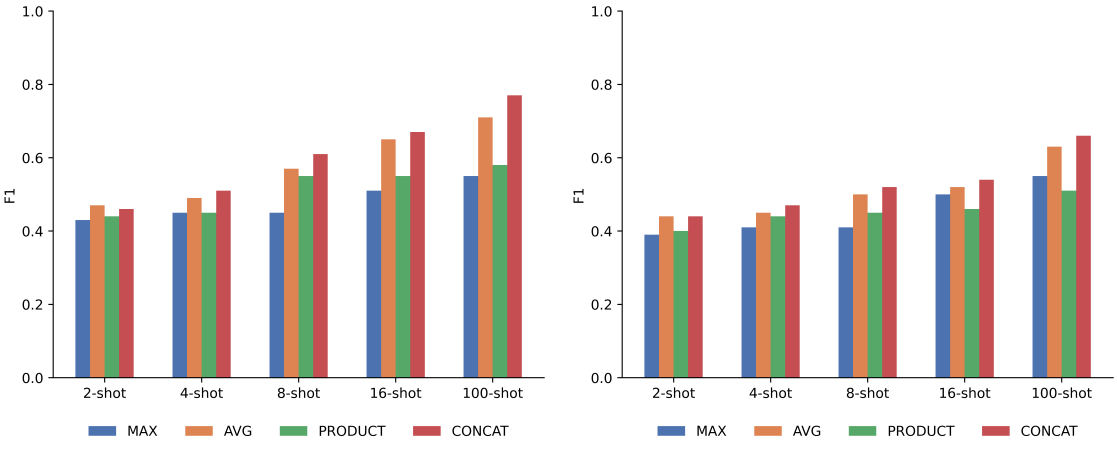
(a) PolitiFact (b) GossipCop

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

4.5.3. Multimodal fusion strategies

The pre-trained clip model generates the unimodal features that normally come with better feature alignment than that generated from the unimodal pre-trained models (Zhou et al., 2022). In this paper, we apply the M-SAMPLE and compare four rule-based fusion strategies(Atrey et al., 2010) to evaluate how the multimodal fusion would affect the performances in terms of F1 score, as shown in Figure 5. Specifically, "MAX" denotes the multimodal features are fused by a max-pooling layer, "AVG" denotes the multimodal features are fused by an average pooling layer, "PRODUCT" means multimodal features are made up of the output product of all unimodal features, "CONCAT"

means the multimodal features are concatenated from the unimodal features. The results indicate that the concatenation of two unimodal features yields better f1 than other fusing strategies in the few-shot settings.



(a) PolitiFact (b) GossipCop

Figure 5: The comparison of different multimodal fusion strategies.

4.5.4. Trainable parameters comparisons

We also compare the numbers of trainable parameters between baselines and SAMPLE as shown in Table 3. The trainable parameters in SAMPLE frameworks are rather small and most of those come from the verbalizer and templates. In contrast, the FT-RoBERTa has the largest trainable parameter if fully fine-tuning the entire model. As a result, prompt learning is less computational cost than fine-tuning, and also can achieve comparable results compare with other deep learning methods.

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

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

4.5.5. Ablation study

We investigate the influence of key components in SAMPLE by evaluating the framework's performance through various and partial setups. In each test, we employ M-SAMPLE, remove different components, and train the framework from scratch. The results are presented in Table 4 show that M-SAMPLE experiences a performance decay without each component in most setups, indicating the effectiveness of each key module in SAMPLE. Specifically, we observe a slight decrease in performance when removing the automatic similarity adjustment "-SA". This demonstrates that the

standardizing of semantic similarity in fusing multimodal features helps reduce uncorrelated information in classifying fake news while mitigating the noise from different modalities' multimodal features.

Table 4
The experimental results of ablation study based on M-SAMPLE. "-SA" denotes the automatic similarity adjustment is removed from M-SAMPLE, "-IF" means we remove the image feature from CLIP model, "-TF" means we remove the text feature from CLIP model, "-MF" means the multimodal feature from CLIP model is removed and only use the text feature from 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

Furthermore, removing the text feature from CLIP "-TF" resulted in worse F1 and accuracy compare to the framework that remove the image feature from CLIP-"-IF" in general. Our findings indicate that, although the image modality provides valuable information for FND (as shown in Figure 3), text features are still critical in prompt learning. This is mainly due to prompt learning's training objective which is to recover the masked token from the templates, primarily aligning text features extracted from pre-trained models. Additionally, as two text features are extracted from different pre-trained models (RoBERTa and CLIP, respectively), they give the classifier more expressive textual information. However, image features are mainly utilized to mitigate noise between different modalities.

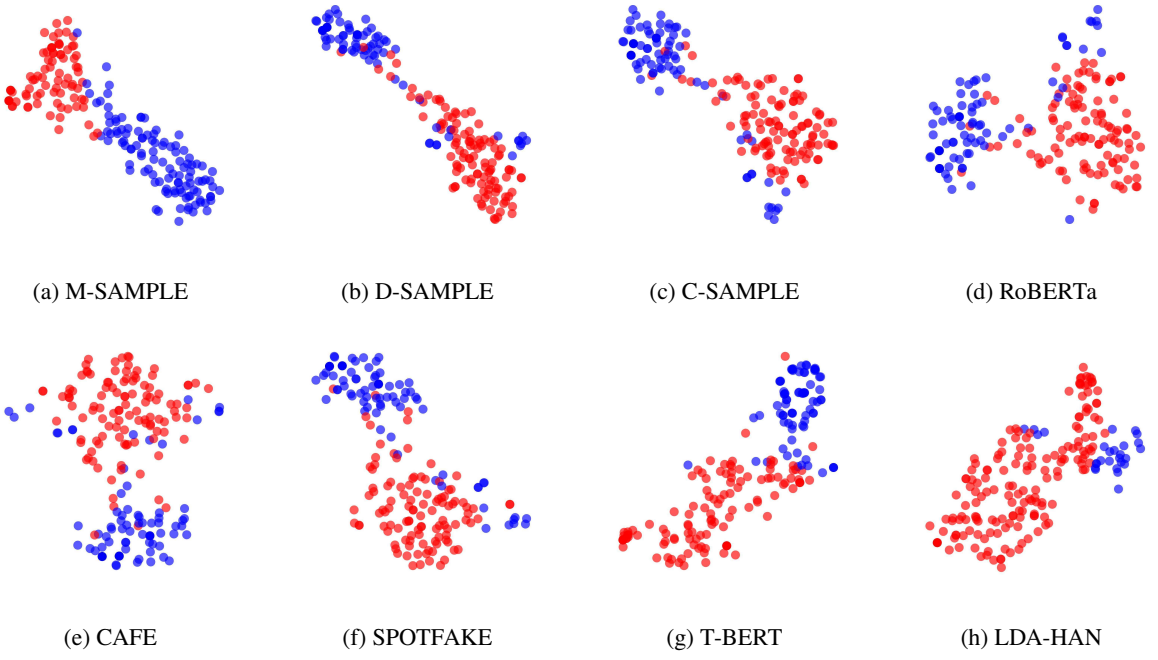
We also removed fused multimodal features "-MF" generated from the CLIP model. In this partial setup, the proposed framework is the vanilla version of the prompt learning approach that leverages the pre-trained language model to predict FND directly. We discovered that vanilla prompt learning can still yield better performance than unimodal methods that utilize textual features only, highlighting the superiority of prompt learning in FND.

4.5.6. T-SNE visualization

We analyze the learned features of the classifiers on the test set of PolitiFact in the 2-shot setting as shown in Figure 6. The reduced dimension learned feature representations of fake and real news are represented by red and blue dots. We observe that the boundary of M-SAMPLE is more sharply defined than that of D-SAMPLE and C-SAMPLE as seen from Figure 6a, Figure 6b and Figure 6c. This suggests that the learned feature representations are more discriminative. Although FT-RoBERTa produces comparable performances in F1 and accuracy, it exhibits some clear misclassified instances in the 2-shot setting. Additionally, the learned feature representations are sparser compared to SAMPLE as shown in Figure 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 Figure 6e and Figure 6f. It is observed that the numbers of misclassified instances are significantly higher than that in prompt learning and fine-tuning methods. Moreover, we find that unimodal methods such as T-BERT and LDA-HAN have the highest number of misclassified instances, suggesting that the multimodal feature can capture more expressive information than the text feature alone, as shown in Figure 6g and Figure 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. This section initially analyses the relationships between our approach and existing studies. Additionally, we detail how our proposed SAMPLE approach can positively



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

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

Figure 6: T-SNE visualizations of features learned before 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.

impact the FND and provide support for real-world applications. Finally, this paper discusses the limitations and future work.

5.1. Connections and comparison with previous works

SAMPLE demonstrates satisfactory performance in detecting fake news in both few-shot and data-rich scenarios. To compare with other approaches in the Fake News Detection (FND) field, traditional approaches fall into one of three categories: (1) unimodal approaches based solely on text or image features (Ajao et al., 2019; Cao et al., 2020; Chen et al., 2021b); (2) multimodal approaches that assimilate textual and visual features via either pre-trained models or deep learning representation (Kim, 2014; Tuan and Minh, 2021); and (3) standard fine-tuning approaches that fine-tune pre-trained unimodality models with task-specific data (Devlin et al., 2018b; Nguyen et al., 2020).

In this study, SAMPLE encompasses a hybrid of approaches (2) and (3), although different from the standard fine-tuning method due to its prompt learning algorithm. Fine-tuning may achieve optimal performance, but consumes a significant amount of memory as it updates the entire set of model parameters for a task-specific objective. Conversely, prompt learning, which leverages a natural language prompt to query a language model, maintains similarity with pre-training and shows comparable performance, particularly with limited training instances. By contrasting 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. Nevertheless, such external modules require an adequate quantity of training instances to capture cross-modal correlations and lead to inadequate performance in the few-shot setting. Consequently, our new proposal offers a similarity-aware multimodal feature fusion methodology that exploits CLIP's pre-training strategy. 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 investigated our approach in the few-shot setting, as depicted in Table 4, and the resulting data indicates a significant enhancement in few-shot performance due 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, integrating different prompting strategies with multimodal features remains underexplored. This paper presents a promising method that achieves strong results and can serve as a significant baseline for future multimodal FND research.

In contrast, traditional multimodal FND systems typically necessitate large amounts of training data to attain satisfactory performance levels. However, obtaining 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 approach that fuses similarity-aware multimodal features with prompt learning holds potential for future similar classification tasks.

5.3. Limitations and future work

The present study has several limitations. Firstly, SAMPLE solely focuses on investigating the effects of the soft verbalizer that is designed to identify appropriate label words from the vocabulary automatically. Nevertheless, optimizing the soft verbalizer in a broader vocabulary under low-data conditions remains a considerable challenge, indicating that further adaptive modifications are required to enhance the overall performance. Secondly, the newly proposed multimodal fusing method is based on a similarity-aware strategy that aims to reduce noise injection in less correlated cross-modal features. Nonetheless, it does not explicitly address the uncorrelated cross-modal relations. Thirdly, there is still a need for discovering more multimodal FND approaches to other modalities like news entities and social networks.

Several studies have indicated that the selection of verbalizers considerably affects performance. Notably, manual verbalizers (Schick and Schütze, 2021) rely on task-specific prior knowledge and intensive labour work to identify label words representing classes. On the other hand, while the soft verbalizer (Shin et al., 2020; Gao et al., 2021) attempts to ease this process, it remains challenging to optimize it adequately for a large vocabulary in low-data settings. Moreover, the knowledgeable prompt-tuning approach (Hu et al., 2022) utilizes external knowledge bases to expand the coverage of the label words and reduce the bias associated with manual verbalizers. Investigating the impact of verbalizers will be part of our future work. Additionally, integrating other modalities such as news entities, topics and social networks can extend 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 insufficient issue, SAMPLE incorporates three popular prompt templates: discrete prompting, continuous prompting and mixed prompting to the original input text, and employs the pre-trained language model RoBERTa to acquire text features from the prompt. Furthermore, the pre-trained CLIP model is used to obtain the input texts, input images, and their semantic similarities. To address semantic gaps and improve the collaboration between image and text modalities, we introduce a similarity-aware multimodal feature fusing approach that applies standardization and a Sigmoid function to adjust the intensity of the final cross-modal representation. Finally, by feeding the multimodal feature into a fully-connected layer, we project the feature to obtain the word distribution mapped to the specific news class.

We evaluate the proposed approach by conducting a multimodal FND experiment on two benchmark datasets, and extensively compare SAMPLE's performance 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 might harm the FND performances especially when the training instances are small. 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.

References

Aggarwal, A., Chauhan, A., Kumar, D., Verma, S., Mittal, M., 2020. Classification of fake news by fine-tuning deep bidirectional transformers based language model. EAI Endorsed Transactions on Scalable Information Systems 7, e10–e10.

Ajao, O., Bhowmik, D., Zargari, S., 2019. Sentiment aware fake news detection on online social networks, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 2507–2511.

Atrey, P.K., Hossain, M.A., El Saddik, A., Kankanhalli, M.S., 2010. Multimodal fusion for multimedia analysis: a survey. Multimedia systems 16, 345–379.

Bahad, P., Saxena, P., Kamal, R., 2019. Fake news detection using bi-directional lstm-recurrent neural network. Procedia Computer Science 165, 74–82.

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

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent dirichlet allocation. Journal of machine Learning research 3, 993–1022.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al., 2020. Language models are few-shot learners. Advances in neural information processing systems 33, 1877–1901.

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

Castillo, C., Mendoza, M., Poblete, B., 2011. Information credibility on twitter, in: Proceedings of the 20th international conference on World wide web, pp. 675–684.

Chen, B., Chen, B., Gao, D., Chen, Q., Huo, C., Meng, X., Ren, W., Zhou, Y., 2021a. 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. pp. 83–92.

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

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

Conroy, N.K., Rubin, V.L., Chen, Y., 2015. Automatic deception detection: Methods for finding fake news. Proceedings of the association for information science and technology 52, 1–4.

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

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

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 .

Dun, Y., Tu, K., Chen, C., Hou, C., Yuan, X., 2021. Kan: Knowledge-aware attention network for fake news detection, in: Proceedings of the AAAI conference on artificial intelligence, pp. 81–89.

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

Gao, T., Fisch, A., Chen, D., 2021. 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 (Volume 1: Long Papers), pp. 3816–3830.

Gu, Y., Han, X., Liu, Z., Huang, M., 2022. Ppt: Pre-trained prompt tuning for few-shot learning, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 8410–8423.

Guo, H., Cao, J., Zhang, Y., Guo, J., Li, J., 2018. Rumor detection with hierarchical social attention network, in: Proceedings of the 27th ACM international conference on information and knowledge management, pp. 943–951.

Hambardzumyan, K., Khachatrian, H., May, J., 2021. Warp: Word-level adversarial reprogramming. arXiv preprint arXiv:2101.00121 .

Han, B., Han, X., Zhang, H., Li, J., Cao, X., 2021. Fighting fake news: two stream network for deepfake detection via learnable srm. IEEE Transactions on Biometrics, Behavior, and Identity Science 3, 320–331.

Han, X., Zhao, W., Ding, N., Liu, Z., Sun, M., 2022. Ptr: Prompt tuning with rules for text classification. AI Open .

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

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

Jiang, G., Liu, S., Zhao, Y., Sun, Y., Zhang, M., 2022. Fake news detection via knowledgeable prompt learning. Information Processing & Management 59, 103029.

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

Jiang, Y., Wang, Y., Maynard, X.S.D., 2020. 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).

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

Kim, Y., 2014. Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882 .

Kumar, A., Trueman, T.E., Cambria, E., 2021. Fake news detection using xlnet fine-tuning model, in: 2021 International Conference on Computational Intelligence and Computing Applications (ICCICA), IEEE. pp. 1–4.

Kwon, S., Cha, M., Jung, K., Chen, W., Wang, Y., 2013. Prominent features of rumor propagation in online social media, in: 2013 IEEE 13th international conference on data mining, IEEE. pp. 1103–1108.

Lester, B., Al-Rfou, R., Constant, N., 2021. The power of scale for parameter-efficient prompt tuning. arXiv preprint arXiv:2104.08691 .

Li, P., Sun, X., Yu, H., Tian, Y., Yao, F., Xu, G., 2021. Entity-oriented multi-modal alignment and fusion network for fake news detection. IEEE Transactions on Multimedia .

Li, X.L., Liang, P., 2021. 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 (Volume 1: Long Papers), Association for Computational Linguistics, Online. pp. 4582–4597. URL: https://aclanthology.org/2021.acl-long.353, doi:10.18653/v1/2021.acl-long.353.

Li, Y., Scarton, C., Song, X., Bontcheva, K., 2022. Classifying covid-19 vaccine narratives. arXiv preprint arXiv:2207.08522 .

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G., 2021a. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. arXiv preprint arXiv:2107.13586 .

Liu, X., Ji, K., Fu, Y., Tam, W., Du, Z., Yang, Z., Tang, J., 2022. 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 (Volume 2: Short Papers), pp. 61–68.

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

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 .

Ma, J., Gao, W., Wong, K.F., 2017. Detect rumors in microblog posts using propagation structure via kernel learning, Association for Computational Linguistics.

Meel, P., 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.

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

Pennington, J., Socher, R., Manning, C., 2014. Glove: Global vectors for word representation, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532–1543.

Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., Stein, B., 2017. A stylometric inquiry into hyperpartisan and fake news. arXiv preprint arXiv:1702.05638 .

Qi, P., Cao, J., Li, X., Liu, H., Sheng, Q., Mi, X., He, Q., Lv, Y., Guo, C., Yu, Y., 2021. Improving fake news detection by using an entity-enhanced framework to fuse diverse multimodal clues, in: Proceedings of the 29th ACM International Conference on Multimedia, pp. 1212–1220.

Qin, G., Eisner, J., 2021. Learning how to ask: Querying lms with mixtures of soft prompts. arXiv preprint arXiv:2104.06599 .

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, PMLR. pp. 8748–8763.

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

Schick, T., Schütze, H., 2021. 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, pp. 2339–2352.

Shin, T., Razeghi, Y., Logan IV, R.L., Wallace, E., Singh, S., 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 4222–4235.

Shu, K., Cui, L., Wang, S., Lee, D., Liu, H., 2019. defend: Explainable fake news detection, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 395–405.

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

Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H., 2017. Fake news detection on social media: A data mining perspective. ACM SIGKDD explorations newsletter 19, 22–36.

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

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

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

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

Tam, D., Menon, R.R., Bansal, M., Srivastava, S., Raffel, C., 2021. Improving and simplifying pattern exploiting training, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 4980–4991.

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

Vinyals, O., Toshev, A., Bengio, S., Erhan, D., 2016. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. IEEE transactions on pattern analysis and machine intelligence 39, 652–663.

Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., Su, L., Gao, J., 2018. Eann: Event adversarial neural networks for multi-modal fake news detection, in: Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining, pp. 849–857.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al., 2020. Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations, pp. 38–45.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V., 2019. Xlnet: Generalized autoregressive pretraining for language understanding. Advances in neural information processing systems 32.

Zhang, X., Cao, J., Li, X., Sheng, Q., Zhong, L., Shu, K., 2021. Mining dual emotion for fake news detection, in: Proceedings of the Web Conference 2021, pp. 3465–3476.

Zhang, X., Ghorbani, A.A., 2020. An overview of online fake news: Characterization, detection, and discussion. Information Processing & Management 57, 102025.

Zhong, Z., Friedman, D., Chen, D., 2021. Factual probing is [mask]: Learning vs. learning to recall, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 5017–5033.

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

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

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

Appendix

A. Prompt engineering for discrete templates

In order to assess the impact of various templates on performance, we have created discrete templates, which are illustrated 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 the 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

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 Figure 5. 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 (Li and Liang, 2021). The "LSTM" method feeds the $< soft >$ token through an LSTM layer and employs the outputs as the trainable vectors (Liu et al., 2021b). Although the performances of the three initialization methods in terms of F1 and accuracies are slightly affected, the study also noted that the "FC" and "LSTM" initializations result in later convergence of validation loss than the "Random" initialization as they require 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

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

We evaluated the semantic similarity between text and image features from various pre-trained models. We use the BERT model and VGG-19 to extract features from each training sample and then calculate the average to determine semantic similarity for 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. Finally, we scale the values on the axes logarithmically to represent differences since unimodal model differences are small.

Our experimental findings indicate that the text and image features extract from the CLIP model are more consistent than those from unimodal models, as shown in Figure C. This can be attributed to the capacity of CLIP to learn

multimodal representations through joint training and its utilization of a contrastive loss function to distinguish 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 could lead to more noise in feature extraction.

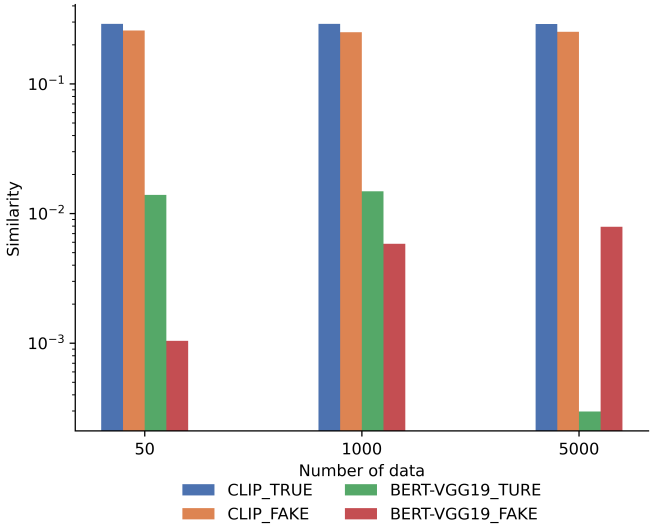


Figure 7: Logarithmically scaled semantic similarity comparison between the pre-trained models.