Efficient Cross-modal Prompt Learning with Semantic Enhancement for Domain-robust Fake News Detection

Fei Wu[1], Hao Jin[1], Changhui Hu[1], Yimu Ji[1,2], Xiao-Yuan Jing[3,4], Guo-Ping Jiang[1]

[1]College of Automation & College of Artificial Intelligence,
Nanjing University of Posts and Telecommunications, Nanjing, China
[2]The State Key Laboratory of Blockchain and Data Security, Zhejiang University, Hangzhou, China
[3]Guangdong Provincial Key Laboratory of Petrochemical Equipment Intelligent Security,
Guangdong University of Petrochemical Technology, Maoming, China
[4]School of Computer, Wuhan University, Wuhan, China

Abstract

With the development of multimedia technology, online social media has become a major medium for people to access news, but meanwhile, it has also exacerbated the dissemination of multi-modal fake news. An automatic and efficient multi-modal fake news detection (MFND) method is urgently needed. Existing MFND methods usually conduct cross-modal information interaction at later stage, resulting in insufficient exploration of complementary information between modalities. Another challenge lies in the differences among news data from different domains, leading to the weak generalization ability in detecting news from various domains. In this work, we propose an efficient Cross-modal Prompt Learning with Semantic enhancement method for Domain-robust fake news detection (CPLSD). Specifically, we design an efficient cross-modal prompt interaction module, which utilizes prompt as medium to realize lightweight cross-modal information interaction in the early stage of feature extraction, enabling to exploit rich modality complementary information. We design a domain-general prompt generation module that can adaptively blend domain-specific news features to generate domain-general prompts, for improving the domain generalization ability of the model. Furthermore, an image semantic enhancement module is designed to achieve image-to-text translation, fully exploring the semantic discriminative information of the image modality. Extensive experiments conducted on three MFND benchmarks demonstrate the superiority of our proposed approach over existing state-of-the-art MFND methods.

1 Introduction

Online social media, as the main platform for the public to obtain news and related information in the information age, has greatly facilitated our lives. However, the rapid growth of online social media has also exacerbated the spread of fake news and information. Fake news (Jiang et al., 2024a,b) will not only cause harmful public opinion, but also cause serious spiritual and even economic loss to individuals. In the multimedia era, news mainly exists in the form of multiple modalities. Therefore, it is very important to study automatic and effective multi-modal fake news detection (MFND) methods.

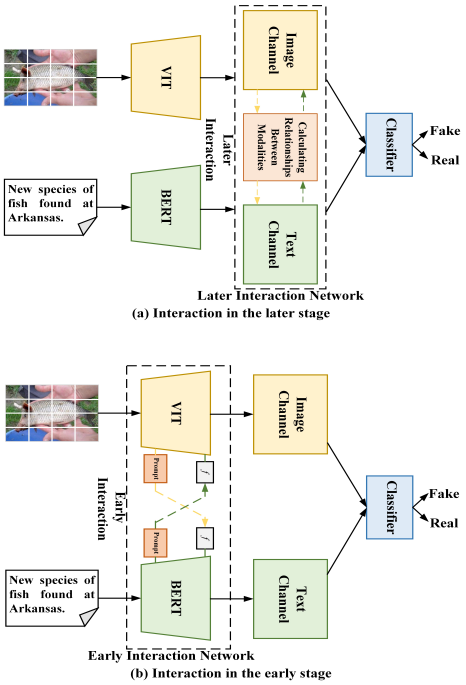


Figure 1: Different manners of modality information interaction.

In recent years, there have been many MFND methods that have achieved significant success. For example, CAFE (Chen et al., 2022) extracts uni-modal features using pre-trained BERT (Devlin et al., 2018) and ResNet-34 (He et al., 2016) models, and obtains multi-modal fused features with uni-modal features by calculating cross-modality consistency. HMCAN (Qian et al., 2021) introduces a multi-modal context attention network to integrate relationships within and between modal-

Proceedings of the 31st International Conference on Computational Linguistics, pages 4175–4185
January 19–24, 2025.

ities to get multi-modal fused features. MMFN (Zhou et al., 2023) integrates BERT, CLIP (Radford et al., 2021), and Swin-Transformer (Liu et al., 2021) models to extract textual and visual features, while utilizing a multilevel co-attention mechanism to obtain multi-modal fused features. MRHFR (Wu et al., 2023) has developed a cognitive perception fusion layer and a coherent constraint inference layer for multi-modal feature fusion.

We refer to the stage using pre-trained models to obtain basic features as the early stage, and the subsequent stage of further feature extraction as the later stage. As shown in Fig. 1 (a), these existing methods first use pre-trained models to independently extract basic features from image and text, and then perform cross-modal information interaction at later stage to realize multi-modal feature fusion. However, these approaches face three main challenges. (1) According to (Barnum et al., 2020), compared to the interaction in the later stage, the early stage is more in line with human intuition in perceiving and understanding things. (2) (Gadzicki et al., 2020) also expounds on that later interaction has a major drawback that is the very limited potential for the exploitation of cross correlation between different uni-modal data. We infer that the lack of early interaction in existing MFND methods fails to fully exploit the relationships between data of different modalities. (3) These methods require a large number of parameters, computationally intensive feature fusion networks, which reduce model efficiency (Sun et al., 2017).

Inspired by prompt tuning, we propose the efficient Cross-modal Prompt Learning with Semantic enhancement for Domain-robust fake news detection (CPLSD) approach. The contributions are:

- We introduce an efficient cross-modal prompt interaction module to facilitate cross-modal information interaction mediated by lightweight prompts at the early stage of feature extraction as shown in Fig. 1 (b), more comprehensively exploring relationships between modalities.
- The domain-general prompt generation module produces domain-general prompts, which enriches the multi-domain information contained in news features through the information interaction between prompts and feature sequences. This reduces the model's bias towards learning domain-specific information, such that its generation ability in detecting news from various domains is improved.
- The image semantic enhancement module further achieves inter-modal alignment through modality translation from image to text, and extracts the discriminative information contained in the image modality more thoroughly.
- Comprehensive experiments on three practical benchmarks Pheme (Zubiaga et al., 2016), Weibo (Jin et al., 2017) and Fakeddit (Nakamura et al., 2019) demonstrate the state-of-the-art performance of our approach over existing MFND methods.

2 Related Work

2.1 Multi-modal Fake News Detection

In recent years, many methods for multi-modal fake news detection (Dong et al., 2023) have been developed, which mainly focus on extracting discriminative information from images and text and fusing them to detect fake news. For example, MVAE (Khattar et al., 2019) adopts Bi-LSTM (Huang et al., 2015) and pre-trained VGG-19 (Simonyan and Zisserman, 2014) for intra-modal feature extraction, and utilizes bimodal variational autoencoder for the detection task. MCAN (Wu et al., 2021) detects fake news by deeply integrating spatial and frequency features of images as well as textual features. MCNN (Xue et al., 2021) focuses on the consistency of multi-modal data, extracting textual and visual semantic features for fusion. LIMR (Singhal et al., 2022) models the relationship between modalities in a multiplicative way and extracts the significant intra-modal features of fine granularity. For MRML (Peng et al., 2023), metric-based triplet learning is utilized to discover the intra-modal between-class relationships, and contrastive pairwise learning is performed for inter-modal relationship modeling. (Yang et al., 2023) uses Faster R-CNN and BERT to extract image and textual feature, then a novel deep visual-linguistic fusion network is followed to realize multi-modal feature fusion.

However, these methods merely utilize pre-trained models to extract specific basic features of each modality and then perform feature fusion. The information interaction between different modalities is ignored during the early stage of feature extraction, resulting in underutilization of the relationships between different modalities.

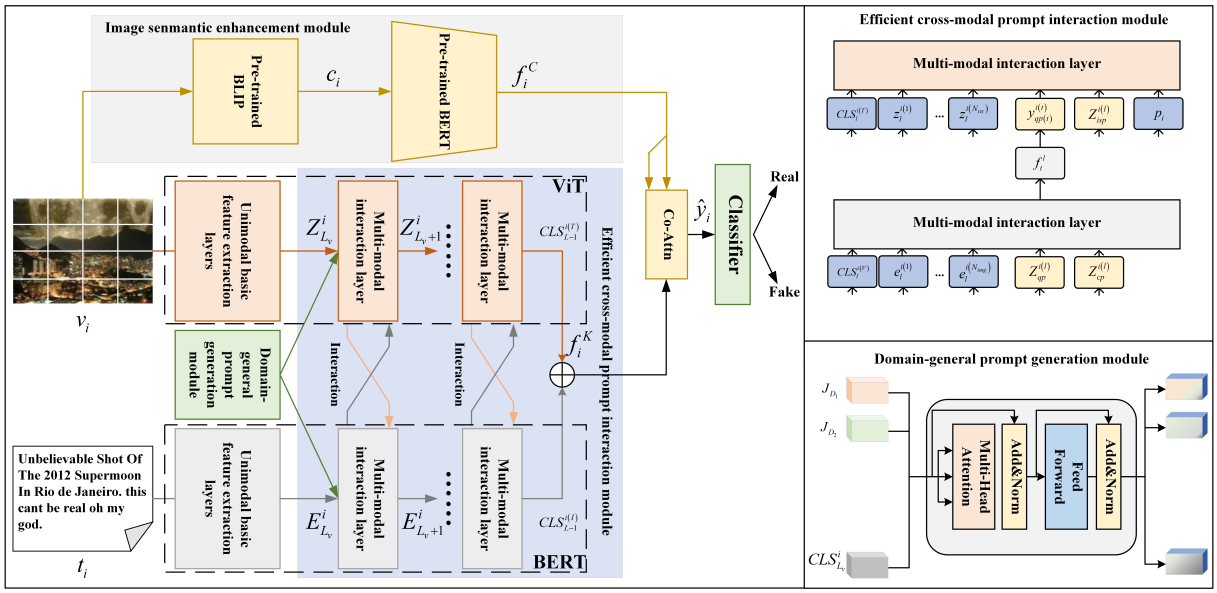


Figure 2: The architecture of the proposed CPLSD approach.

2.2 Prompt Learning

Prompt learning was first proposed in the field of natural language processing, and gradually applied to the field of computer vision and other fields (Khattak et al., 2023). PTR (Han et al., 2022) model is proposed, which constructs multi-class text classification prompts using logical rules containing multiple sub-prompts. PLST (Zhu et al., 2023) model is developed to combine text and external knowledge from open knowledge graphs for short text classification tasks. Other works adopt prompt learning technique to enhance generalization ability of vision-language models. For example, CoOp (Zhou et al., 2022b) fine-tunes CLIP for few-shot transfer by making its language branch's prompt vectors optimal. Co-CoOp (Zhou et al., 2022a) deals with the generalization problem of CoOp by generating an input-conditional token for each image. Recently, some works deal specifically with domain adaptation (Singha et al., 2023). For example, (Zhao et al., 2024) develops a domain-invariant prompt tuning method, which can improve the generalization to the invisible domain.

More recently, prompt learning has also been used in fake news detection tasks. For instance, KPL (Jiang et al., 2022) detects fake news by incorporating external knowledge into a prompt representation. SAMPLE (Jiang et al., 2023) uses three prompt templates with a soft language analyzer to detect fake news. PVCG (Zou et al., 2024) realizes multi-modal fusion by adding image information as an adjustable prompt to text embedding. However, existing single-modal or multi-modal fake news detection methods based on prompt learning either only employ prompts for the text modality or only employ prompts to unidirectionally transfer information from the image modality to the text modality, failing to realize fully and flexible cross-modal information interaction. This results in insufficient capture of complementary modality information, which affects MFND performance.

3 Proposed Method

3.1 Model Overview

The overall framework of our proposed model is shown in Fig. 2. In this work, we propose a method named efficient Cross-modal Prompt Learning with Semantic enhancement for Domain-robust fake news detection (CPLSD). The network architecture of CPLSD consists of three main parts: (a) Efficient cross-modal prompt interaction module, which introduces a few prompts to achieve information interaction between modalities in the early stage, i.e., the basic feature extraction stage. (b) Domain-general prompt generation module, which obtains the domain-general prompts by adaptively mixing the news features of each domain using transformer to enhance the domain robustness of the model. (c) Image semantic enhancement module, which obtains the image caption through pre-trained BLIP model, and then uses pre-trained BERT model to extract the text features of the image caption to obtain more discriminative information contained in the image modality.

Given a news item $o_i \in O, \forall i \in \{1, ..., N\}$,

where N represents the number of news samples, $t_i \in T, i \in \{1, ..., N\}$ and $v_i \in V, i \in \{1, ..., N\}$ represent the text and image of the news, respectively. $y_i = \{0, 1\} \in Y, i \in \{1, ..., N\}$ indicates the ground-truth category label corresponding to the news. The goal of MFND is to predict the ground-truth category label of each news.

3.2 Unimodal Basic Feature Extraction

Text Feature Sequence Extraction. For each news item o_i, we first use the pre-trained BERT model to convert news text t_i to one-hot encoding vectors and then these encoding vectors are converted to a continuous embedding sequence composed of N_{txt} vectors. Next, we use the first layer of the BERT to extract the textual feature sequence $Z_0^i = \left[CLS_0^{i(T)}, z_0^{i(1)}, ..., z_0^{i(N_{txt})}\right]$ of the news, where $CLS_0^{i(T)}$ represents the CLS token of the text. The CLS token is a special token attached to the sequence, which represents the semantic information of whole news text. Z_0^i is then sequentially processed through the subsequent $L_v - 1$ layers of BERT for further feature extraction and Z_l^i represents the output feature sequence of the l^{th} transformer encoder layer. The extraction process of features through the transformer layer of text modality is as follows:

$$Z_{l+1}^i = \text{TransLayer}_T^l\left(Z_l^i\right) \tag{1}$$

TransLayer_T^l represents the l^{th} layer of BERT.

Image Feature Sequence Extraction. For each news item o_i, firstly, we segment the news image v_i into N_{img} slices, then project them linearly into a sequence of N_{img} vectors, and use the first layer of the pre-trained transformer based encoder, i.e., ViT, to extract image representation sequence $E_0^i = \left[CLS_0^{i(I)}, e_0^{i(1)}, ..., e_0^{i(N_{img})}\right]$, where $CLS_0^{i(I)}$ represents the CLS token of the news image. The image embedding sequence also proceeds to further feature extraction by sequentially passing through the subsequent $L_v - 1$ transformer encoder layers of ViT. E_l^i represents the output feature sequence of the l^{th} transformer encoder layer. The extraction process of features through the transformer layer of image modality is as follows:

$$E_{l+1}^i = \text{TransLayer}_V^l\left(E_l^i\right) \tag{2}$$

where TransLayer_V^l represents the l^{th} layer of ViT.

The pre-trained encoders for both modalities are composed of L transformer layers. We refer to the preceding L_v transformer encoder layers of two modalities as unimodal basic feature extraction layers, while the subsequent L_f transformer encoder layers are called multi-modal interaction layers, where $L_v + L_f = L$. Unimodal basic feature extraction is followed by cross-modal interaction with the generated domain-general prompts.

3.3 Domain-general Prompt Generation

News belongs to different domains, such as politics, economy, entertainment, etc., and there are certain data differences between news of different domains. Generally, the model tends to learn specific features of the known domain, thereby affecting the detection of news from different domains. Although each news article has its corresponding and unique domain label, it may contain information or characteristics from several different news domains simultaneously. We can supplement each news with characteristics from other relevant news domains, making each news contain richer information from different news domains to suppress the model's learning of domain-specific features. This will enhance the model's generalization ability in detecting news from various domains. In order to effectively integrate information from relevant news domains into the features of each news item o_i, inspired by prompt learning, we generate domain-general prompts by mixing the average features of news from different domains through attention mechanism. Domain-general prompts will supplement each news with news information of related domains through an interactive attention mechanism, reducing the influence of domain-specific news information on the model's learning process, thus improving the model's generalization ability in detecting news from various domains.

The news dataset is composed of multiple domains, and we can represent the training dataset as $D_{train} = \{D_1, D_2, ..., D_m\}$, where $D_k = \{\left(t_j^k, v_j^k, y_j^k\right)\}, \forall j \in 1, 2, ..., n_{D_k}$ represents the k^{th} domain in the dataset and n_{D_k} represents the number of news samples in domain D_k. t_j^k and v_j^k represent the text and image of the j^{th} news article in this domain, and $y_j^k = \{0, 1\}$ represents the corresponding ground truth label. In order to effectively utilize information from known news domains, we need to represent the data of each domain D_k in an appropriate way. To this end, we take the following steps: first, for each domain, we calculate its modality-specific average features to

represent the domain. We obtain the basic image and text features for each news item o_i in the domain, which are represented by the CLS tokens output from the L_v^{th} layer of BERT and ViT, namely $f_j^T = CLS_{L_v}^{i(T)}$, $f_j^V = CLS_{L_v}^{i(I)}, j \in 1, ..., n_{D_k}$. Then, we obtain the modality-specific average features of the domain, i.e., $J_{D_k^T}$ and $J_{D_k^V}$, as follows:

$$J_{D_k^T} = \frac{\sum_{j=1}^{n_{D_k}} f_j^T}{n_{D_k}}, J_{D_k^V} = \frac{\sum_{j=1}^{n_{D_k}} f_j^V}{n_{D_k}} \quad (3)$$

We further unify the average features of different domains for information integration to obtain the domain unified feature matrices $A^T \in R^{d_v \times m} = \{J_{D_1^T}, ..., J_{D_m^T}\}$ and $A^I \in R^{d_v \times m} = \{J_{D_1^V}, ..., J_{D_m^V}\}$, where d_v represents the dimension of features. For each news item $o_i \in O, \forall i \in \{1, ..., N\}$, to integrate information from various known news domains into it, we employ a multi-head self-attention mechanism to learn the relationships between the basic features $CLS_{L_v}^{i(T)}$, $CLS_{L_v}^{i(I)}$ and $J_{D_k^T}$, $J_{D_k^V}, k \in \{1, ..., m\}$. By interacting the input news features with the features of individual news domains, we adaptively obtain the mixing proportions of information for multiple domains, resulting in final fused features of each modality. Specifically, we input $CLS_{L_v}^{i(T)}$ and A^T, $CLS_{L_v}^{i(I)}$ and A^I to modality-specific transformer encoders (Vaswani et al., 2017) with the initialized parameters θ_{Trans} following a normal distribution. The output is named as the domain-general prompts p_i^T and p_i^I. The calculation process is as follows:

$$p_i^T = \text{softmax}\left(\frac{A^T\left(CLS_{L_v}^{i(T)}\right)}{\sqrt{d_v}}\right) CLS_{L_v}^{i(T)} \quad (4)$$

$$p_i^I = \text{softmax}\left(\frac{A^I\left(CLS_{L_v}^{i(I)}\right)}{\sqrt{d_v}}\right) CLS_{L_v}^{i(I)} \quad (5)$$

Then we concatenate p_i^T and $Z_{L_v}^i$, p_i^I and $E_{L_v}^i$ together respectively to obtain the input of the subsequent L_f layers of BERT and ViT.

Incorporating domain-general prompts into the model enriches the multi-domain information contained within the news item, thereby inhibiting the model's tendency to learn domain-specific features. This manner encourages the model to learn domain-shared features, enhancing its domain generalization capability.

3.4 Efficient Cross-modal Prompt Interaction

To realize inter-modal information interaction in the early stage of feature extraction and fully explore the relationships between modalities, we design an efficient cross-modal prompt interaction module, which is composed of multiple multi-modal interaction layers. Each layer consists of an information extraction step and an information aggregation step. Specifically, the information extraction step focuses on extracting information to be transmitted to the other modality, while the information aggregation step focuses on integrating the information from both modalities and domain-general prompts.

We design three kinds of prompts to help information interaction in the multi-modal interaction layer, and name them as 'query prompt (QP, i.e., $Z_{qp}^{i(l)}$ and $E_{qp}^{i(l)}$ for text and image modalities of each news item o_i in the l^{th} layer of transformer encoder, respectively)', 'context prompt (CP, i.e., $Z_{cp}^{i(l)}$ and $E_{cp}^{i(l)}$)' and 'information supplyment prompt (ISP, i.e., $Z_{isp}^{i(l)}$ and $E_{isp}^{i(l)}$)'. QP is interacted with unimodal input sequences, and the output is transmitted to the other modality. CP provides contextual information for the optimization process of QP. ISP, on the other hand, is responsible for providing contextual information during the information aggregation step.

We use the query prompts $Z_{qp}^{i(l)}$, $E_{qp}^{i(l)}$ and the context query prompts $Z_{cp}^{i(l)}$, $E_{cp}^{i(l)}$ in the information extraction step, and use information supplyment prompts $Z_{isp}^{i(l)}$, $E_{isp}^{i(l)}$ and domain-general prompts p_i^T, p_i^I in the information aggregation step.

Information extraction step. Firstly, we concatenate the query prompts $Z_{qp}^{i(l)}$, $E_{qp}^{i(l)}$, the context prompts $Z_{cp}^{i(l)}$, $E_{cp}^{i(l)}$ with the feature sequences Z_l^i, E_l^i to get the concatenated input sequences $\left[Z_l^i||Z_{cp}^{i(l)}||Z_{qp}^{i(l)}\right]$ and $\left[E_l^i||E_{cp}^{i(l)}||E_{qp}^{i(l)}\right]$, where $||$ indicates the concatenation operation. Then, we input the concatenated sequence into the modality-specific multi-modal interaction layer as shown in the following formula:

$$\left[\hat{Z}_l^i||\hat{Z}_{cp}^{i(l)}||\hat{Z}_{qp}^{i(l)}\right] = \text{TransLayer}_T^l\left(\left[Z_l^i||Z_{cp}^{i(l)}||Z_{qp}^{i(l)}\right]\right) \quad (6)$$

$$\left[\hat{E}_l^i||\hat{E}_{cp}^{i(l)}||\hat{E}_{qp}^{i(l)}\right] = \text{TransLayer}_V^l\left(\left[E_l^i||E_{cp}^{i(l)}||E_{qp}^{i(l)}\right]\right) \quad (7)$$

To align the query prompts $\hat{Z}_{qp}^{i(l)}$, $\hat{E}_{qp}^{i(l)}$ with the input feature sequences E_l^i, Z_l^i respectively, we perform cross-modal mapping for the query prompts

$\hat{Z}_{qp}^{i(l)}$ and $\hat{E}_{qp}^{i(l)}$ through the modality-specific mapping networks, i.e., f_t^l and f_v^l, which consist of two fully connected layers and an activation function. The mapping process is as follows:

$$y_{qp(l)}^{i(t)} = f_t^l\left(\hat{Z}_{qp}^{i(l)}\right), y_{qp(l)}^{i(v)} = f_v^l\left(\hat{E}_{qp}^{i(l)}\right) \quad (8)$$

Information aggregation step. First, we concatenate the mapped prompts $y_{qp(l)}^{i(t)}$ and $y_{qp(l)}^{i(v)}$, the information supplyment prompts $Z_{isp}^{i(l)}$ and $E_{isp}^{i(l)}$, the input sequences Z_l^i and E_l^i, and the domain-general prompts p_i^T and p_i^I to obtain the input sequence of the modality-specific multi-modal interaction layer. The calculation process of the image modality is shown as follows:

$$\begin{aligned}&[Z_{l+1}^i||\hat{Z}_{isp}^{i(l)}||\hat{y}_{qp(l)}^{i(t)}||\hat{p}_i^T \\ &= \text{TransLayer}_T^l\left(\left[Z_l^i||Z_{isp}^{i(l)}||y_{qp(l)}^{i(t)}||p_i^T\right]\right)\end{aligned} \quad (9)$$

The text modality can be calculated similarly. After the information aggregation step is completed, we concatenate the CLS tokens $CLS_{L-1}^{i(T)}$ and $CLS_{L-1}^{i(I)}$ of the last transformer encoder layer from two modalities to get the fused feature f_i^K.

3.5 Image Semantic Enhancement

For each news item o_i, in order to extract rich information from the image v_i, we extract the text features of the image caption. Firstly, we input the prompt, i.e., 'This is a news picture description of' to the pre-trained BLIP (Li et al., 2022) model for extracting the image caption c_i of the news image v_i, and use the pre-trained BERT model to obtain the text feature f_i^C of the image description. Finally, we integrate f_i^C into f_i^K through the cross-attention mechanism and introduce residual connections to reduce the influence of noise information to get final feature f_i^E as follows:

$$f_i^E = f_i^K + \text{softmax}\left(\frac{f_i^K\left(f_i^C\right)'}{\sqrt{d_v}}\right) f_i^K \quad (10)$$

where $(.)'$ represents the transposition operation. By extracting the textual feature of the generated news image caption and adaptively integrating the text feature into the final feature, we conduct fine exploration of the image's semantic information and enrich the discriminative information contained in the final feature.

3.6 Classifier and model optimization

For each news item o_i, we input the final feature f_i^E to the classifier parameterized by W and b, which consists of two fully connected layers and a softmax activation function. The calculation process is as follows:

$$\hat{y}_i = \text{softmax}\left(W f_i^E + b\right) \quad (11)$$

where $\hat{y}_i$ represents the predicted label of the news item o_i. With the aim to classify the news to the correct category, we optimize the parameters by minimizing the cross entropy loss function which is defined as:

$$L_c = y_i \log\left(\hat{y}_i\right) + \left(1 - y_i\right) \log\left(1 - \hat{y}_i\right) \quad (12)$$

4 Experiments

4.1 Datasets

- The PHEME (Zubiaga et al., 2016) dataset comprises data stemming from five significant news events. It contains 1,972 fake news and 3,830 real news. Each of these news has corresponding event label. Each of these news encompasses a collection of posts, all of which have their corresponding event labels.

- The Weibo (Jin et al., 2017) dataset is collected from the Chinese social media platform Weibo. It consists of a training set that is composed of 6,151 tweets and a test set which is composed of 1,697 tweets.

- The Fakeddit (Nakamura et al., 2019) dataset is a large-scale dataset for fake news detection collected from the social news site Reddit. We randomly select 30,000 image-text pairs in the Fakeddit training set for training, and 10,000 image-text pairs from the test set for testing.

4.2 Implementation Details

We set batch size to 64 and use the adam optimizer (Kingma and Ba, 2014) to optimize the parameters and prompts of the model. The learning rate of prompts is set to 1e-4, and the learning rate of other parameters is set to 1e-3. We set weight decay to 1e-3. All prompts are initialized with Gaussian distribution (mean=0, std=0.02). Our model is trained with a maximum training epoches of 100 and we employ an early stopping strategy to prevent over-fitting. The code is run on single NVIDIA RTX 4090. For pre-trained encoder

Dataset	Method	Accuracy	Fake News			Real News		
			P	R	F1	P	R	F1
Pheme	EANN [KDD-2018]	0.648	0.810	0.498	0.617	0.584	0.759	0.660
	Spotfake [BigMM-2019]	0.823	0.743	0.745	0.744	0.864	0.863	0.863
	BDANN [IJCNN-2020]	0.766	0.619	0.531	0.571	0.816	0.864	0.839
	HMCAN [SIGIR-2021]	0.881	0.830	0.838	0.834	0.910	0.905	0.907
	CAFE [WWW-2022]	0.861	0.812	0.645	0.719	0.875	0.943	0.908
	MRML [ICASSP-2023]	0.842	0.741	0.708	0.724	0.881	0.897	0.889
	QFND [INFFUS-2024]	0.779	0.769	0.354	0.485	0.781	**0.956**	0.860
	CPLSD	**0.891**	**0.832**	**0.839**	**0.836**	**0.918**	0.934	**0.924**
Weibo	EANN [KDD-2018]	0.827	0.847	0.812	0.829	0.807	0.843	0.825
	Spotfake [BigMM-2019]	0.892	0.902	**0.964**	0.932	0.847	0.656	0.739
	BDANN [IJCNN-2020]	0.842	0.830	0.870	0.850	0.850	0.820	0.830
	HMCAN [SIGIR-2021]	0.885	0.920	0.845	0.881	0.856	0.926	0.890
	CAFE [WWW-2022]	0.840	0.855	0.830	0.842	0.825	0.851	0.837
	MRML [ICASSP-2023]	0.897	0.898	0.887	0.892	0.896	0.905	0.901
	QFND [INFFUS-2024]	0.869	0.900	0.810	0.850	0.840	0.920	0.880
	CPLSD	**0.920**	**0.927**	0.939	**0.933**	**0.910**	**0.927**	**0.915**
Fakeddit	EANN [KDD-2018]	0.724	0.727	0.719	0.723	0.722	0.729	0.726
	SpotFake [BigMM-2019]	0.819	0.801	0.848	0.824	0.839	0.790	0.813
	BDANN [IJCNN-2020]	0.812	0.836	0.776	0.805	0.791	0.847	0.818
	HMCAN [SIGIR-2021]	0.881	0.880	0.882	0.881	0.882	0.880	0.881
	CAFE [WWW-2022]	0.912	0.946	0.886	0.915	0.878	0.942	0.909
	MRML [ICASSP-2023]	0.840	0.819	0.874	0.846	0.865	0.807	0.835
	QFND [INFFUS-2024]	0.923	0.917	0.931	0.924	0.930	0.915	0.923
	CPLSD	**0.940**	**0.947**	**0.932**	**0.938**	**0.931**	**0.948**	**0.939**

Table 1: Performance comparison on three datasets, and the best results are in bold.

models, we use the 'ImageNet-21k pre-trained vit-base' model for the image encoder, and use the 'bert-base-chinese' model for Weibo and the 'bert-base-uncased' model for Pheme and Fakeddit.

4.3 Performance Comparison

To evaluate the performance of our proposed method, we select 7 state-of-the-art MFND methods for comparison, including EANN (Wang et al., 2018), Spotfake (Singhal et al., 2019), BDANN (Zhang et al., 2020), HMCAN (Qian et al., 2021), CAFE (Chen et al., 2022), MRML (Peng et al., 2023), and QFND (Qu et al., 2024).

- EANN (Wang et al., 2018) designs an adversarial network to learn event-invariant multi-modal features for fake news detecion.
- Spotfake (Singhal et al., 2019) uses pre-trained BERT and VGG models to obtain modality-specific features, which are then concatenated to derive discriminative results.
- BDANN (Zhang et al., 2020) utilizes domain-adaptive neural networks along with a domain classifier to minimize disparities across events.
- MEAN (Wei et al., 2022) uses dual discriminators to reduce modality differences and event disparities.
- HMCAN (Qian et al., 2021) employs a multi-modal contextual attention network to combine intra-modal and inter-modal relationships.
- CAFE (Chen et al., 2022) adaptively aggregates unimodal features by learning cross-modal correlation to identify fake news.
- MRML (Peng et al., 2023) derives semantic relationships within modalities using metric learning.
- QFND (Qu et al., 2024) uses quantum convolutional neural network to derive discriminative results based on obtained multi-modal fused features.

The accuracy (Acc), precision (P), recall (R), and F-measure score (F1) for fake/real news detection of the compared baselines and our proposed approach on three datasets are presented in Table 1, with the best results highlighted in bold. The results show that our method can achieve the best performance on comprehensive metrics accuracy and F1, demonstrating a noteworthy enhancement in the detection of fake news. There are three reasons for the improvement of performance: (1) In the early stage of feature extraction, the in-depth interaction of cross-modal information enables the model

to acquire richer multi-modal complementary information, thereby rendering the extracted basic features with favorable discriminative ability. (2) The inclusion of domain-general prompts enriches the multi-domain characteristics possessed by the feature, thus mitigating the influence of domain-specific information on the model's generalization ability to detect news of different domains. (3) By translating the images of news samples into text, we further strengthen modality alignment and uncover richer discriminative information in image.

4.4 Ablation Studies

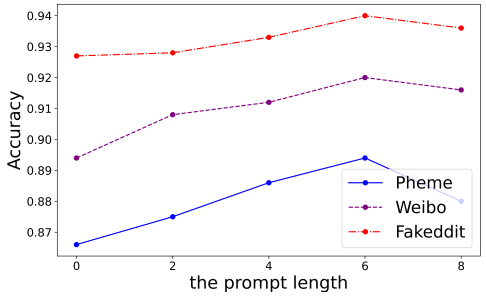
We analyze the influence of each proposed component in CPLSD. Specifically, the compared variants of CPLSD are:

- CPLSD w/o P: We remove the efficient cross-modal prompt interaction module and do not conduct cross-modal information interaction during the basic feature extraction stage.
- CPLSD w/o S: We remove the image semantic enhancement module and directly input the fused features of image and text modalities to the classifier for classification.
- CPLSD w/o E: We remove the domain-general prompt generation module and do not use domain-general prompts in information aggregation step of multi-modal interaction layers.

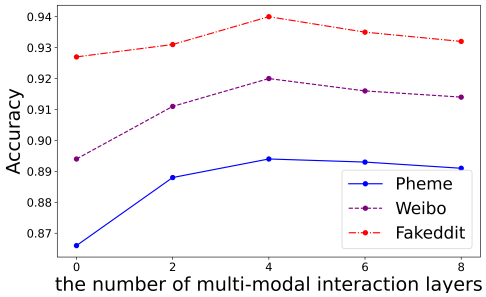
From Table 2, we have the following observations: 1) The performance of CPLSD w/o P has declined, proving that the early cross-modal interaction effectively explores diversity information and relevant information between modalities. 2) The performance of CPLSD w/o S inferiors to the results of CPLSD, showing that the text features of generated image captions provide the model with richer discriminative information, which helps fully capture semantic information in image. 3) The performance of CPLSD w/o E is also worse than the complete version of CPLSD, proving that domain-general prompts mitigate the model's tendency to learn domain-specific information, significantly improving its domain generalization capability.

4.5 Parameter Analysis

The length of prompt. To analyze the influence of prompt length on results, we conduct experiments on three datasets. In experiments, query prompts, context prompts and information supplyment prompts, are all set to the same length. The detection performance of CPLSD is shown in Fig. 3(a), and we can see that when the prompt length becomes excessively large (greater than 6), the detection accuracy begins to decrease. In fact, as the prompt length increases, the model is able to realize more comprehensive cross-modal information interaction, enabling each modality to obtain richer information from the other modality. However, as the prompt length continues to increase, it may introduce more redundancy and lead to over-fitting.



(a)



(b)

Figure 3: Performance with (a) various prompt lengths and (b) different numbers of interaction layers.

The number of multi-modal interaction layers. The depth of basic feature extraction for individual modalities and the timing of initiating inter-modal information interaction present a delicate balancing consideration. To identify the optimal balance between them, we analyze the influence of the number of multi-modal interaction layers on performance of fake news detection, and the results are shown in Fig. 3(b). Since the total number of layers of the pre-trained feature extractor, i.e., BERT or ViT, is fixed, when the number of multi-modal interaction layers increases, the number of the unimodal basic feature extraction layers decreases, resulting in an earlier start of modality information interaction. We can see that when the number is less than 4, the detection accuracy on both datasets improves as

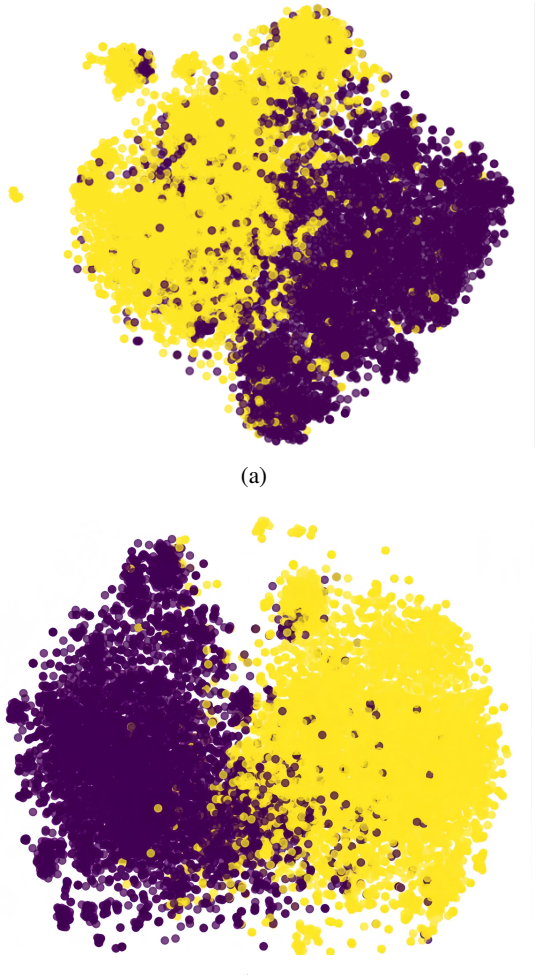
Dataset	Model	Accuracy	Fake News			Real News		
			P	R	F1	P	R	F1
Pheme	CPLSD	**0.891**	**0.832**	**0.839**	**0.836**	**0.918**	0.934	**0.924**
	w/o P	0.865	0.752	0.805	0.778	0.917	0.890	0.903
	w/o S	0.878	0.830	0.735	0.779	0.895	**0.938**	0.916
	w/o E	0.881	0.813	0.770	0.791	0.906	0.926	0.917
Weibo	CPLSD	**0.920**	**0.927**	**0.939**	**0.933**	0.910	**0.927**	**0.915**
	w/o P	0.894	0.858	0.929	0.892	**0.931**	0.862	0.895
	w/o S	0.899	0.877	0.917	0.896	0.922	0.884	0.903
	w/o E	0.902	0.878	0.920	0.899	0.925	0.885	0.905
Fakeddit	CPLSD	**0.940**	**0.947**	**0.932**	**0.938**	**0.931**	**0.948**	**0.939**
	w/o P	0.927	0.946	0.908	0.926	0.912	0.945	0.928
	w/o S	0.931	0.933	0.929	0.931	0.930	0.933	0.931
	w/o E	0.930	0.907	0.930	0.919	0.928	0.902	0.916

Table 2: Ablation analysis results on three datasets.

the number increases. However, when the number becomes excessively large (greater than 4), the detection accuracy begins to decrease. Therefore, in this paper, we set the number of multi-modal interaction layers to 4 to realize the balance of sufficient unimodal basic feature extraction and cross-modal interaction as much as possible with the lightweight prompt learning manner.

4.6 Visualization

In order to observe classification effect more intuitively, we use the TSNE (Van der Maaten and Hinton, 2008) to visualize original samples and learned features by our approach on Fakeddit, as shown in Fig. 4. For original samples, ViT features and BERT features for image and text modalities are extracted and then are concatenated for visualization. We can observe that the boundary of two categories is clearer for learned features than original samples. This indicates the learned features have more strong discriminative ability.



(a)

(b)

Figure 4: Feature distributions of (a) original samples and (b) learned features on Fakeddit. Features of fake and real news are indicated by yellow and purple dots.

5 Conclusion

In this paper, we propose a novel approach named CPLSD for MFND. The efficient cross-modal interaction module learns features well adaptive to the MFND task and enriches complementary information across modalities. The domain-general prompt generation module enhances the domain generalization capability of the model. And image semantic enhancement module provides richer discriminative information from images.

Comprehensive experiments on three widely used MFND datasets demonstrate CPLSD can significantly outperform state-of-the-art MFND methods. The component discussion results also indicate effectiveness of the designed main modules. In this paper, we consider the most common modalities of image and text. In the future, we will evaluate effectiveness of our approach with more modality forms.

Acknowledgments

This work was supported by National Natural Science Foundation of China (No. 62076139), National Key Laboratory of Information System Engineering (No. 05202305), and the Open Research Fund of The State Key Laboratory of Blockchain and Data Security, Zhejiang University.

References

George Barnum, Sabera Talukder, and Yisong Yue. 2020. On the benefits of early fusion in multimodal representation learning. *arXiv preprint arXiv:2011.07191*.

Yixuan Chen, Dongsheng Li, Peng Zhang, Jie Sui, Qin Lv, Lu Tun, and Li Shang. 2022. Cross-modal ambiguity learning for multimodal fake news detection. In *ACM Web Conference*, pages 2897–2905.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yiqi Dong, Dongxiao He, Xiaobao Wang, Yawen Li, Xiaowen Su, and Di Jin. 2023. A generalized deep markov random fields framework for fake news detection. In *International Joint Conference on Artificial Intelligence*, pages 4758–4765.

Konrad Gadzicki, Razieh Khamsehashari, and Christoph Zetzsche. 2020. Early vs late fusion in multimodal convolutional neural networks. In *IEEE International Conference on Information Fusion*, pages 1–6.

Xu Han, Weilin Zhao, Ning Ding, Zhiyuan Liu, and Maosong Sun. 2022. Prompt tuning with rules for text classification. *AI Open*, 3:182–192.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Bohan Jiang, Zhen Tan, Ayushi Nirmal, and Huan Liu. 2024a. Disinformation detection: An evolving challenge in the age of llms. In *SIAM International Conference on Data Mining*, pages 427–435.

Bohan Jiang, Chengshuai Zhao, Zhen Tan, and Huan Liu. 2024b. Catching chameleons: Detecting evolving disinformation generated using large language models. *arXiv preprint arXiv:2406.17992*.

Gongyao Jiang, Shuang Liu, Yu Zhao, Yueheng Sun, and Meishan Zhang. 2022. Fake news detection via knowledgeable prompt learning. *Information Processing & Management*, 59(5):103029.

Ye Jiang, Xiaomin Yu, Yimin Wang, Xiaoman Xu, Xingyi Song, and Diana Maynard. 2023. Similarity-aware multimodal prompt learning for fake news detection. *Information Sciences*, 647:119446.

Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *ACM International Conference on Multimedia*, pages 795–816.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. 2023. Maple: Multi-modal prompt learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122.

Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. Mvae: multimodal variational autoencoder for fake news detection. In *The World Wide Web Conference*, pages 2915–2921.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *IEEE/CVF International Conference on Computer Vision*, pages 10012–10022.

Kai Nakamura, Sharon Levy, and William Yang Wang. 2019. r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. *arXiv preprint arXiv:1911.03854*.

Liwen Peng, Songlei Jian, Dongsheng Li, and Siqi Shen. 2023. Mrml: multimodal rumor detection by deep metric learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5.

Shengsheng Qian, Jinguang Wang, Jun Hu, Quan Fang, and Changsheng Xu. 2021. Hierarchical multi-modal contextual attention network for fake news detection. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 153–162.

Zhiguo Qu, Yunyi Meng, Ghulam Muhammad, and Prayag Tiwari. 2024. Qmfnd: A quantum multimodal fusion-based fake news detection model for social media. *Information Fusion*, 104:102172.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from

natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Mainak Singha, Harsh Pal, Ankit Jha, and Biplab Banerjee. 2023. Ad-clip: Adapting domains in prompt space using clip. In *IEEE/CVF International Conference on Computer Vision*, pages 4355–4364.

Shivangi Singhal, Tanisha Pandey, Saksham Mrig, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. 2022. Leveraging intra and inter modality relationship for multimodal fake news detection. In *Companion Proceedings of the Web Conference*, pages 726–734.

Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh. 2019. Spotfake: A multi-modal framework for fake news detection. In *IEEE International Conference on Multimedia Big Data*, pages 39–47.

Xu Sun, Weiwei Sun, Shuming Ma, Xuancheng Ren, Yi Zhang, Wenjie Li, and Houfeng Wang. 2017. Complex structure leads to overfitting: A structure regularization decoding method for natural language processing. *arXiv preprint arXiv:1711.10331*.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11):2579–2605.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. In *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 849–857.

Pengfei Wei, Fei Wu, Ying Sun, Hong Zhou, and Xiao-Yuan Jing. 2022. Modality and event adversarial networks for multi-modal fake news detection. *IEEE Signal Processing Letters*, 29:1382–1386.

Lianwei Wu, Pusheng Liu, and Yanning Zhang. 2023. See how you read? multi-reading habits fusion reasoning for multi-modal fake news detection. In *AAAI Conference on Artificial Intelligence*, pages 13736–13744.

Yang Wu, Pengwei Zhan, Yunjian Zhang, Liming Wang, and Zhen Xu. 2021. Multimodal fusion with co-attention networks for fake news detection. In *Findings of the Association for Computational Linguistics*, pages 2560–2569.

Junxiao Xue, Yabo Wang, Yichen Tian, Yafei Li, Lei Shi, and Lin Wei. 2021. Detecting fake news by exploring the consistency of multimodal data. *Information Processing & Management*, 58(5):102610.

Yang Yang, Ran Bao, Weili Guo, De-Chuan Zhan, Yilong Yin, and Jian Yang. 2023. Deep visual-linguistic fusion network considering cross-modal inconsistency for rumor detection. *Science China Information Sciences*, 66(17):222102.

Tong Zhang, Di Wang, Huanhuan Chen, Zhiwei Zeng, Wei Guo, Chunyan Miao, and Lizhen Cui. 2020. Bdann: Bert-based domain adaptation neural network for multi-modal fake news detection. In *International Joint Conference on Neural Networks*, pages 1–8.

Cairong Zhao, Yubin Wang, Xinyang Jiang, Yifei Shen, Kaitao Song, Dongsheng Li, and Duoqian Miao. 2024. Learning domain invariant prompt for vision-language models. *arXiv preprint arXiv:2212.04196*.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022a. Conditional prompt learning for vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022b. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348.

Yangming Zhou, Yuzhou Yang, Qichao Ying, Zhenxing Qian, and Xinpeng Zhang. 2023. Multi-modal fake news detection on social media via multi-grained information fusion. In *ACM International Conference on Multimedia Retrieval*, pages 343–352.

Yi Zhu, Ye Wang, Jipeng Qiang, and Xindong Wu. 2023. Prompt-learning for short text classification. *arXiv preprint arXiv:2202.11345*.

Ting Zou, Zhong Qian, Peifeng Li, and Qiaoming Zhu. 2024. Pvcg: Prompt-based vision-aware classification and generation for multi-modal rumor detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 11036–11040.

Arkaitz Zubiaga, Geraldine Wong Sak Hoi, Maria Liakata, and Rob Procter. 2016. Pheme dataset of rumours and non-rumours. *University of Warwick, Department of Computer Science*.