Graph-Rumor Aligning and Reasoning for Improving Rumor Detection in Unknown Domains

1st Xingchen Ding
School of cyber science and engineering
Wuhan University
Wuhan, China
xingos@whu.edu.cn

2nd Xianping Ma
School of Science and Engineering
The Chinese University of Hongkong
Shenzhen, China
ma.xianping125@gmail.com

3rd Chong Teng
School of cyber science and engineering
Wuhan University
Wuhan, China
tengchong@whu.edu.cn

***Abstract*—The spread of rumors on social media platforms, especially those related to incidents of urgency, poses a significant challenge to the public's pursuit of truth. Recent studies suggest that due to scarce annotation resources and unexpected events that have not been previously reported by mainstream media, detecting and verifying information can be particularly difficult. In this study, we present a new technique for Zero-shot Domain transfer using Prompt-based learning (ZDP) to detect rumors in unfamiliar domains. Our approach involves representing social media events as social rumours with propagation graphs, encoding contextual rumours and propagation structures with sentence and graph encoders, respectively. To facilitate domain adaptation, we utilize alignment techniques that align rumors and their propagation threads in order to learn cross-domain invariant features. Furthermore, we have significantly bolstered the model's cross-domain reasoning capabilities through the use of graph-rumor interaction mechanisms. In addition, we employ contrastive learning techniques to improve model training. For evaluation purposes, we extensively tested our proposed model on two real-world datasets. Results indicate that our method outperforms state-of-the-art techniques, particularly in early rumour detection.**

***Index Terms*—rumor detection, social media, domain adaption, prompt tuning, zero-shot, graph learning**

I. Introduction

The internet and social media, including Twitter and WEIBO, provide a large amount of information but have also resulted in a surge of rumors, leading to misinformation and negative outcomes. Although human fact-checkers are available, they struggle to judge rumors due to domain barriers. It is therefore crucial to develop automated methods to detect rumors during unforeseen events. Significant progress has been made in the field of rumor detection, thanks to recent advancements such as deep neural networks [1]–[3]. Nonetheless, these models rely on large amounts of annotated data for accurate training, making their application challenging in reality. There are many instances of new claims emerging daily in different domains, and acquiring sufficient labeled data in such instances, especially for time-critical domain events, is often impractical. To tackle these issues, researchers have proposed domain-invariant feature learning techniques for cross-domain rumor detection in their works [4], [5], which necessitate labeled target data for training the models. Consequently, these models are not suitable for scenarios where labeled target data is unavailable. On the other hand, [6] have put forth a domain-agnostic approach based on Generative Adversarial Networks (GANs) to enable zero-shot cross-domain rumor detection. However, the effectiveness of their model is limited. Therefore, there is a pressing need for a more robust and capable method to tackle the challenges associated with detecting rumors in unfamiliar domains.

In this paper, we investigate the effectiveness of using efficient prompt tuning and zero-shot domain transfer for detecting rumors in unknown domains. According to [7], rumors tend to exhibit consistent propagation patterns across domains, indicating the existence of domain-invariant structural features in both rumors and their propagation threads, which lays the groundwork for zero-shot domain transfer technology. Since there are no available annotations in the target domain, we utilize prompt learning mechanisms [8] based on large pre-trained language models. However, the standard prompt learning paradigm cannot handle graph data. Typically, input graphs are transformed into text sequences, but this approach may sacrifice structural information when encoding rumor propagation graphs due to the conversion rules and input length limitations.

To this end, we proposes a zero-shot domain-transfer prompt tuning (ZDP) framework for detecting unknown domain rumors on social media. The framework incorporates a text encoder and a graph encoder to encode the rumor and its associated propagation threads, correspondingly. To extract domain-invariant features, we implement a graph-rumor embedding aligning (GRA) loss, which minimizes the KL divergence between the normalized graph-rumor similarity score distributions and the normalized ground truth label matching distributions. Additionally, an implicit reasoning module is designed to build relations between graph and rumor textual representations through self- and cross-attention mechanisms. The resulting fused representation is used for prompt learning to achieve effective implicit graph-rumor relation learning, which improves the performance of the zero-shot rumor detection method. To demonstrate the effectiveness of our ZDP framework, we conducted extensive experiments on two real-world datasets related to COVID-19 rumors. Our results confirm that (1) our model substantially outperforms state-of-the-art baselines for detecting rumors in unknown

Manuscript 400 submitted to 2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC). Received April 12, 2023.

domains; and (2) our method excels in early rumor detection, which is essential for prompt intervention and debunking, especially in rapidly developing situations.

II. Related Work

Early studies on automated rumor detection primarily involved developing supervised classifiers which utilized features derived from post contents, user profiles, and propagation patterns [9]–[11]. Later research suggested additional features, such as [12] used regular expressions to identify questioning and denying tweets. Deep neural networks have also been employed to learn features from the stream of social media posts, including recurrent neural networks [13], convolutional neural networks [14], and attention mechanisms [15]. Moreover, some approaches proposed kernel-learning models [10], [16], tree-structured recursive neural networks (RvNN) [17], and self-attention models (PLAN) [18] to extract useful clues jointly from content semantics and propagation structures, and graph neural networks (BiGCN) [1] have also been employed to encode conversation threads for higher-level representations. Unlike these rumor detection methods that focus on analyzing data from the same domain and develop various frameworks to improve their ability to detect rumors, our model is specifically designed to identify rumors in unfamiliar domains.

Zero-shot Domain-transfer learning (ZDL) is a technique that aims to extract transferable features that can be shared between labeled and unseen domains, thereby being independent of specific domains [19], and prompt learning is a technique that involves converting downstream tasks into language modeling tasks through textual prompts [8]. ZDL techniques have been applied in recent studies to detect rumor [20]–[22] using pre-trained language models (PLMs) and prompt-tuning methods that are specific to downstream tasks. To adapt the model from the source domain to the target domain, [23] used PLMs and a self-training loop through a multi-step iteration, and [22] used a prompt-based tuning approach that involves linear concatenation of certain user comments. However, these approaches solely consider cross-domain text classification and might face task incompatibility issues between pre-training and fine-tuning, while ignoring domain-invariant propagation patterns from community response. In contrast, we employ ZDL prompt-based technique for zero-shot rumor detection, which leverages the alignment between rumors and their related propagation structures to effectively identify rumors in less frequently encountered domains without requiring expert annotations.

Contrastive learning has emerged as a successful approach in various fields, involving the learning from positive and negative samples. In the field of computer vision, self-supervised image representation can be learned by minimizing the distance between two views of the same image, as shown in prior works [24], [25]. The effectiveness of contrastive learning has also been observed in natural language processing tasks such as semantic textual similarity [26], and short text clustering [4]. While drawing inspiration from self-supervised contrastive learning, our work employs supervised contrastive prompt learning to achieve zero-shot rumor detection in unknown domains.

III. Proposed Method

In this section, we present our proposed method ZDP, as depicted in Figure 1. Our approach incorporates implicit reasoning by aligning rumor graphs and rumor texts, which enables more accurate detection of social rumors in previously unexplored domains.

A. Sentence Encoder

To facilitate the mapping of a post in an event to a shared semantic space, whereby the source and target domains align semantically, we employ XLM-RoBERTa [27] as our Sentence Encoder. The Sentence Encoder employs contextual interactions among tokens in the sequence to obtain the sentence-level representation, as expressed by the equation below:

$$x = \textit{Sentence Encoder}(T), \tag{1}$$

here, T denotes the original text, which serves as input to the Sentence Encoder. We obtain the resulting post-level representation x by using the output state of the s token. Texts in event C_i are represented in matrix form, $C_i = [x_0^i, x_1^i, x_2^i, ..., x_m^i]^\top$, where x_0^i represents event i, and m corresponds to the number of posts in C_i. The dimension of this matrix is $\mathbb{R}^{(m+1)\times d}$, where d represents the dimension of the output state of the Sentence Encoder.

B. Graph Encoder

To capture the propagation of events effectively, we utilize a Graph Convolutional Network (GCN) technique, namely BiGCN, as our graph encoder.. Specifically, we represent the conversation thread of each event, denoted by $C_i = \{x_0^i, x_1^i, \ldots, x_m^i\}$, as a graph structure $\mathcal{G} = \langle V, E \rangle$, where the set of nodes V comprises the event claim and its relevant corresponding posts. Meanwhile, the set of directed edges E refers to the response relation among the nodes in V. For example, if post x_2^i responds to post x_1^i, there will be a directed edge $x_1^i \longrightarrow x_2^i$ (i.e., e_{12}^i). The adjacency matrix $A^i \in \{0,1\}^{|V_i|\times|V_i|}$, where $fA_{n,m}^i = 1$ if x_n^i responds to x_m^i or $n = m$, and $A_{n,m}^i = 0$ otherwise, summarizes these relationships.

Two graphs are constructed in this study using retweet and response relationships: the propagation graph and citation graph, illustrated in Fig. 1. The propagation graph portrays the primary conversation thread and highlights the direction of information diffusion through its edges, and the citation graph displays responsive nodes that refer to their corresponding responded nodes, similar to a citation network. A l_{th} GCN module processes the propagation graph and generates a hidden representation as a matrix $H_{l+1} \in \mathbb{R}^{|V_i|\times|V_i|}$. The GCN model can be formulated as follows:

$$H_{l+1} = \mathrm{ReLU}\left(\tilde{D}_i^{-\frac{1}{2}}\tilde{A}_i\tilde{D}_i^{-\frac{1}{2}}H_l\theta_l\right), \tag{2}$$

here, the adjacency matrix with self-loop is $\tilde{A}_i = A_i + I$, where A_i is the adjacency matrix of the propagation graph,

Manuscript 400 submitted to 2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC). Received April 12, 2023.

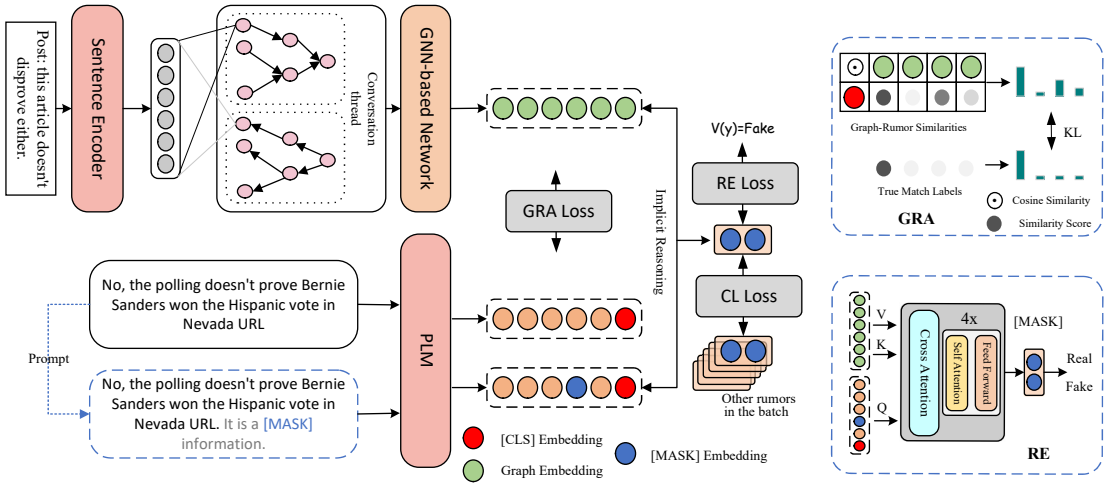

Fig. 1. An illustration of the proposed ZDP framework for rumor detection.

I is the identity matrix, and $\tilde{D}_i$ is the degree matrix of $\tilde{A}_i$. Moreover, $\theta_l \in \mathbb{R}^{|V_i|\times|V_i|}$ is a trainable weight matrix. Upon obtaining the final node representation H^{PG} for a GCN model with L layers, we use the same equation to update the hidden representation of nodes for the citation graph and obtain the output node states H^{CG} at the L_{th} graph convolutional layer.

To combine the information from the propagation and citation graphs, we concatenate their respective representations, and use the mean function to compute their final graph representation $h^g \in \mathbb{R}^{2d^{(L)}}$ at the event level, where $d^{(L)}$ is the output dimension of the GCN, as:

$$h^g = \text{MEAN}([H^{PG}, H^{CG}]). \tag{3}$$

It is important to note that the selection of the graph encoder is independent of our proposed framework and can be replaced easily with any existing structure-based models.

C. Graph-Rumor Embedding Aligning

We introduces a newly proposed loss function, Graph-Rumor Embedding Aligning (GRA), which is designed to align rumor and graph embeddings. This is achieved by utilizing the cosine similarity distributions of embeddings pairs and computing KL divergence to establish connections between representations across different embedding spaces. Aligning graph and rumor embeddings with the GRA loss function bridges the gap between structured knowledge and unstructured texts, thus enhancing the model's aptitude to comprehend and categorize novel instances, regardless of their inclusion in the initial training data.

To compute the cosine similarity between the $\mathcal{L}_2$ normalized representations of rumors and graphs, we begin by creating a set of graph-rumor representation pairs $\{(h_i^r, h_j^g), y_{i,j}\}_{j=1}^N$ for each rumor representation h_i^r in a mini-batch of N graph-rumor pairs. Here, $y_{i,j}$ denotes the true matching label, where a value of 1 for $y_{i,j}$ represents a matched pair from the same event, while a value of 0 indicates an unmatched pair. Afterward, the similarity can be computed using the below formula:

$$sim_{r2g} = h_i^r \cdot (h_j^g)^T / (|h_i^r|\,|h_j^g|), \tag{4}$$

Next, the softmax function can be utilized to compute the probability of matching pairs as:

$$p_{i,j} = \exp(sim_{r2g}/\tau) / \sum_{k=1}^{N} \exp(sim_{r2g}/\tau), \tag{5}$$

here, the temperature hyperparameter τ modulates the probability distribution peaks. $p_{i,j}$ indicates the matching probability, which is the ratio of the cosine similarity score between h_i^r and h_j^g, and the total cosine similarity scores between h_i^r and all $\{h_j^g\}_{j=1}^N$ in a mini-batch. The computation of the GRA loss from rumor to graph in a mini-batch can be conducted using the formula below:

$$\mathcal{L}_{r2g} = KL(\mathbf{p_i}|\mathbf{q_i}) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} p_{i,j}\log\left(\frac{p_{i,j}}{q_{i,j}+\epsilon}\right), \tag{6}$$

here, the parameter ϵ is introduced to avoid numerical issues, and the true matching probability is defined as $q_{i,j} = y_{i,j} / \sum_{k=1}^{N} y_{i,k}$. Likewise, the GRA loss from the graph to the rumor $\mathcal{L}_{g2r}$ can be obtained by swapping h^g and h^r in (4), (5), and (6). The bi-directional GRA loss is computed as follows:

$$\mathcal{L}_{\text{GRA}} = \mathcal{L}_{r2g} + \mathcal{L}_{g2r}. \tag{7}$$

D. Graph-Rumor Implicit Reasoning

To improve generalization in zero-shot learning and reduce the risk of overfitting to a specific domain, we employ prompt tuning to implicitly extract fine-grained relationships and acquire discriminative global features.

Prompt Encoding. Prompt-based tuning involves transforming each input sample, denoted as (T_i, y_i), into a Pattern-Verbalizer Pair (PVP) [28]. This transformation is achieved

through the pattern mapping function, $p(\cdot)$, which generates a cloze question with masks from the input T_i. For example, given a single sentence represented as 'T_i = [CLS] *News.* [SEP], ' the corresponding cloze question is '$\mathrm{p}(T_i)$ = [CLS] *News.* It was a [MASK] information.[SEP]', where [CLS] and [SEP] serve as special markers for the start and end of the sentence. In prompt-based tuning, the verbalizer function $v(\cdot)$, maps the label y_i to tokens representing its semantic meaning. For instance, this study maps labels such as "true/false" to tokens like "real/fake". The input representation T_i is obtained by taking the token embedding h_i^{MASK} corresponding to the [MASK] token from the PVP.

Graph-Rumor Interaction Reasoning. To optimize the graph-rumor interaction, we propose an efficient graph-rumor interaction reasoning module (RE) that unites the embeddings of the graph and rumor. This module comprises a multi-head cross-attention layer and four-layer transformer blocks. Given an input text description T, the prompt-encoded text $\hat{T}$ is defined and processed by the Sentence Encoder, which is explained in Section III-A. Subsequently, the last hidden states of the text encoder, $h_i^{\hat{r}}$, and the graph embedding, h_i^g, are jointly inputted into the graph-rumor interaction reasoning module. To enhance fusion between the graph and prompt-encoded text representations, the query ($\mathcal{Q}$), key ($\mathcal{K}$), and value ($\mathcal{V}$) are $h_i^{\hat{r}}$, h_i^g, and h_i^g, respectively. The fused output h_i^m is achieved as follows:

$$h_i^m = \mathrm{Transformer}(softmax(\frac{\mathcal{Q}\mathcal{K}^\top}{\sqrt{d}})\mathcal{V}), \tag{8}$$

where d is the embedding dimension of masked tokens, h_i^m is the fused graph and prompt-encoded text contextualized representation, and $\mathrm{Transformer}(\cdot)$ denotes the 4-layer transformer blocks.

When using prompt-encoded text, a multi-layer perception classifier predicts the probability of the original tokens h_i^{MASK} for the [MASK] position. The probability of the true label y_i given T_i is estimated through the predicted class $\hat{y}_i$, which represents a probability distribution over all feasible class labels. The following equation formulates $q\left(y_i \mid T_i\right)$:

$$q\left(y_i \mid T_i\right) = \frac{\exp\left(w_{v(y_i)}^\top \cdot h_i^{\mathrm{MASK}}\right)}{\sum_{y_j \in \mathcal{Y}} \exp\left(w_{v(y_j)}^\top \cdot h_i^{\mathrm{MASK}}\right)}, \tag{9}$$

here, w_v represents the logit vector of token v in the vocabulary, and $\mathcal{Y}$ refers to the set of all feasible class labels. We define y_i as a one-hot vector with all components set to 0, except for the one corresponding to the true class label $y_i \in \{1, \ldots, C\}$. The reasoning objective, denoted as $\mathcal{L}_{\mathrm{RE}}$, is formulated as:

$$\mathcal{L}_{\mathrm{RE}} = \sum_{i=1}^{N} -\log\left(\hat{y}_i\right)^\top y_i, \tag{10}$$

here, $(\cdot)^\top$ denotes the transpose operation, and N represents the total number of samples in the training set.

E. Contrastive Training

To further improve the generalization of zero-shot learning in social rumor detection, we suggest incorporating a supervised contrastive learning objective. This objective groups together samples of the same class while also distinguishing those of different classes in a mini-batch. The definition of the supervised contrastive learning objective is as follows:

$$\mathcal{L}_{CL} = -\frac{1}{N}\sum_{i=1}^{N}\frac{1}{N_{y_i}-1}\sum_{j=1}^{N} log\frac{\exp(cos(h_i^m, h_j^m)/\tau)}{\sum\limits_{k=1}^{N}\exp(cos(h_i^m, h_k^m)/\tau)} \tag{11}$$

here, N_{y_i} denotes the number of rumor examples with the same label y_i, while $\cos(\cdot)$ denotes the cosine similarity function, and τ regulates the contrastive loss temperature. We use the supervised contrastive objective in conjunction with RE and GRA losses for joint training of the model. The total loss function is defined as follows:

$$\mathcal{L} = \mathcal{L}_{\mathrm{RE}} + \alpha\mathcal{L}_{\mathrm{GRA}} + \beta\mathcal{L}_{\mathrm{CL}},$$

here, α and β are hyperparameters that adjust the respective contributions of GRA and supervised contrastive objectives.

IV. Experiments

A. Datasets and Experiment Settings

Four publicly available datasets, Twitter and WEIBO [13], Twitter-COVID19 and WEIBO-COVID19 [4], are utilized in our experiments. Twitter and Twitter-COVID19 are English-language datasets that contain threads of conversations in tweets, while WEIBO and WEIBO-COVID19 are Chinese-language datasets that share a similar structural composition. A detailed description of these datasets can be found in Table I. In our experiments, we employed a zero-shot setting to assess the most challenging scenario, detecting rumors from new domains. We used Twitter and WEIBO datasets as the source data and Twitter-COVID19 and WEIBO-COVID19 as the target data.

The evaluation metrics used in our study are accuracy and F1 score, which align with those of previous studies [4], [17]. The values of hyperparameters α and β were selected from the range [0, 0.2, 0.5, 1], and the batch size was chosen from the range [4, 8, 16] for our model. To ensure a fair comparison, all baselines utilized the same sentence encoder as input for our framework.

TABLE I
STATISTICS OF DATASETS IN CROSS-DOMAIN SETTINGS.

Statistics	Source TWITTER	Target Twitter-COVID19	Source WEIBO	Target WEIBO-COVID19
# of events	1154	400	4649	399
# of graph nodes	60409	406185	1956449	26687
# of non-rumors	579	148	2336	146
# of rumors	575	252	2313	253
Avg. time length/tree	389 Hours	2497 Hours	1007 Hours	248 Hours
Avg. depth/tree	11.67	143.03	49.85	4.31
Avg. # of posts/tree	52	1015	420	67
Domain	Open	COVID-19	Open	COVID-19

TABLE II
TEST PERFORMANCE (%) MEASURED ON ZERO-SHOT CROSS-DOMAIN DATASETS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Target (Source)	Twitter-COVID19 (TWITTER)				WEIBO-COVID19 (WEIBO)			
Model	Acc.	Mac-F_1	Rumor F_1	Non-rumor F_1	Acc.	Mac-F_1	Rumor F_1	Non-rumor F_1
CNN	0.405	0.367	0.449	0.284	0.421	0.410	0.438	0.382
LSTM	0.412	0.382	0.425	0.339	0.415	0.415	0.427	0.402
RGAN	0.419	0.394	0.431	0.356	0.432	0.427	0.458	0.396
RvNN	0.413	0.403	0.422	0.385	0.451	0.446	0.387	0.505
PLAN	0.455	0.454	0.432	0.476	0.384	0.372	0.283	0.460
BiGCN	0.516	0.390	0.463	0.317	0.612	0.561	0.681	0.441
UCDRD	0.569	0.508	0.586	0.429	0.616	0.415	0.577	0.252
CD-Prompt	0.689	0.616	0.647	0.585	0.693	0.628	0.712	0.545
ZDP-PLAN	0.724	0.707	0.741	0.673	0.739	0.670	0.766	0.574
ZDP-BiGCN	**0.773**	**0.751**	**0.784**	**0.717**	**0.795**	**0.724**	**0.803**	**0.645**

B. Baseline Model

We conducted a comprehensive comparison of our proposed model with several state-of-the-art baseline methods in the field of rumor identification. The following are brief descriptions of the compared models: 1)**CNN** [14]: uses a convolutional neural network (CNN) to detect misinformation in relevant posts as a fixed-length sequence; 2)**LSTM** [13]: employs a long short-term memory (LSTM) network to learn feature representations of posts over time; 3)**RGAN** [17]: utilizes a generative adversarial network (GAN) to learn stronger rumor representations. It produces conflicting voices to pressure the discriminator; 4)**RvNN** [16]: based on tree-structured recursive neural networks and learns rumor representations guided by the propagation structure; 5)**PLNN** [18]: based on the transformer architecture, it captures long-distance interactions between any two tweets; 6)**BiGCN** [1]: based on GCN and uses conversation trees to learn rumor representations; 7)**UCDRD** [29]: utilizes contrastive learning and cross-attention on a pair of source and target data with the same labels to learn domain-invariant representations for misinformation identification; 8)**CD-Prompt** [22]: uses a prompt-based tuning technique with propagation structure for cross-domain rumor detection; 9)**ZDP-***: our proposed model based on graph-rumor aligning and implicit reasoning leverages a contrastive prompt learning framework for zero-shot rumor detection in unknown domains. The * in ZDP- represents different graph encoders, including PLNN and BiGCN.

C. Rumor Detection Performance

Table II displays the performance metrics of both ZDP and the compared methods for the Twitter-COVID19 and WEIBO-COVID19 datasets. The initial set of baselines in previous studies showed poor performance due to neglecting intrinsic structural patterns, resulting in limited generalization abilities for zero-shot settings. Among the structure-based baselines, PLAN and BiGCN outperformed RvNN when trained with zero-shot labeled target data, utilizing the message-passing architectures and graph structures' representation power. Meanwhile, the UCDRD method generally improves the performance of structure-based baselines by extracting cross-domain features via generative adversarial nets.

The third group evaluated prompt-based tuning techniques, which exhibited improved performance compared to the previous baselines due to their ability to extract rich semantic features from the pre-trained language models. However, our proposed approach, which employs graph modeling to utilize the structural property for conversation threads fully, outperformed CD-Prompt. This is because CD-Prompt employs only a limited number of response posts from restricted rumor conversation threads due to the input length limit of the PLM. In comparison, our proposed ZDP approaches demonstrated superior performance, achieving accuracy score improvements ranging from 8.4% (10.2%) to 36.8% (37.4%) on the Twitter-COVID19 (WEIBO-COVID19) dataset. It highlights the strong generalization capabilities of our method for zero-shot transfer between different domains. The results demonstrate that aligning the rumor and rumor propagation graph can effectively learn invariant knowledge across domains. The GRA and contrastive prompt learning usage also significantly enhanced the model's reasoning ability when verifying rumors in unfamiliar domains.

D. Model Analysis

TABLE III
ABLATION STUDIES ON OUR PROPOSED MODEL.

Target (Source)	WEIBO-COVID19(WEIBO)		Twitter-COVID19 (Twitter)	
Model	Acc.	Mac-F_1	Acc.	Mac-F_1
ZDP	0.795	0.724	0.773	0.751
w/o CL	0.782	0.711	0.763	0.724
w/o RE	0.687	0.617	0.671	0.606
w/o GRA	0.726	0.691	0.715	0.713
w/o Text	0.612	0.561	0.516	0.390

We performed a range of ablation studies on critical components of ZDP to assess their relative importance. Tabulated in III, the experimental outcomes of this comparison revealed that the removal of GRA and the ZDP model, utilizing rumor propagation graphs and prompt texts 'this is a [MASK] information' excluding the rumor texts, respectively, led to about 7% and 18.4% deterioration in results. This drop in performance can be attributed to the dissimilar representation space of the graph and prompt-text embedding. Therefore, aligning the rumor and rumor graph to enhance the zero-shot rumor detection in the target domain is a feasible hypothesis. Moreover, we discovered that eradicating the graph-rumor

implicit reasoning module resulted in a performance reduction of around 10.8%, demonstrating the critical impact of reasoning based on prompt learning with rumor and graph interaction on model performance. Furthermore, we observed that discarding the CL module caused a 1.3% decline in model performance. This finding proposes that training the model with a contrastive learning objective leads to a positive influence for more accurate rumor predictions in cross-domain situations.

E. Early Detection

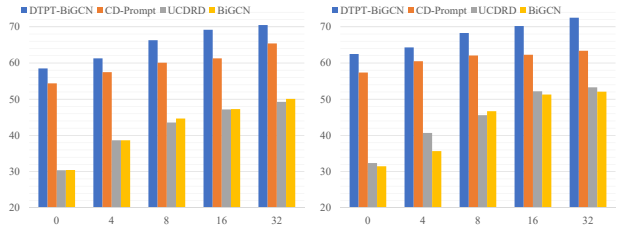

Fig. 2. Accuracy (%) of early detection at various time intervals (in hours) on the Twitter-COVID19 (left) and WEIBO-COVID19 (right) datasets.

Early detection of rumors is critical to prevent their rapid dissemination. Figure 2 displays the early detection performance of our approach, CD-Prompt, UCDRD, and BiGCN at various deadlines. To ensure fair comparisons, we encode the inputs of all baselines using the same PLM. Our proposed ZDP approach outperforms baselines throughout the entire lifecycle, yielding high Accuracy scores in the initial broadcast's early stages. Notably, our approach achieves saturated performance in about 8 hours for WEIBO-COVID19 and Twitter-COVID19, indicating its fusion strategy's advanced response and remarkable early detection proficiency.

V. Conclusion

In this paper, we present a new framework called ZDP, which is a zero-shot prompt-based tuning framework that aligns rumor and rumor propagation threads to improve cross-domain reasoning performance for rumor detection by modeling invariant features across domains. Our findings from extensive experiments on two real-world datasets demonstrate the effectiveness of our proposed model. For future research, we plan to explore specialized zero-shot learning strategies for rumor detection to leverage data from multiple modalities.

References

[1] T. Bian, X. Xiao, T. Xu, P. Zhao, W. Huang, Y. Rong, and J. Huang, "Rumor detection on social media with bi-directional graph convolutional networks," Jan 2020.

[2] H. Lin, J. Ma, M. Cheng, Z. Yang, L. Chen, and G. Chen, "Rumor detection on twitter with claim-guided hierarchical graph attention networks."

[3] D. Rao, X. Miao, Z. Jiang, and R. Li, "Stanker: Stacking network based on level-grained attention-masked bert for rumor detection on social media," Nov 2021.

[4] H. Lin, J. Ma, L. Chen, Z. Yang, M. Cheng, and G. Chen, "Detect rumors in microblog posts for low-resource domains via adversarial contrastive learning," Apr 2022.

[5] A. Mosallanezhad, M. Karami, K. Shu, M. Mancenido, and H. Liu, "Domain adaptive fake news detection via reinforcement learning."

[6] E. Min, T. Xu, P. Zhao, S. Ananiadou, Y. Rong, Y. Bian, and J. Huang, "Divide-and-conquer: Post-user interaction network for fake news detection on social media."

[7] H. Ran, C. Jia, P. Zhang, and X. Li, "Mgat-esm: Multi-channel graph attention neural network with event-sharing module for rumor detection," *Information Sciences*, vol. 592, pp. 402–416, 2022.

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

[9] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," in *Proceedings of the 20th international conference on World wide web*, Mar 2011.

[10] K. Wu, S. Yang, and K. Q. Zhu, "False rumors detection on sina weibo by propagation structures," in *2015 IEEE 31st International Conference on Data Engineering*, Jun 2015.

[11] X. Liu, A. Nourbakhsh, Q. Li, R. Fang, and S. Shah, "Real-time rumor debunking on twitter," in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, Dec 2016.

[12] Z. Zhao, P. Resnick, and Q. Mei, "Enquiring minds: Early detection of rumors in social media from enquiry posts," in *Proceedings of the 24th International Conference on World Wide Web*, Feb 2016.

[13] J. Ma, W. Gao, P. Mitra, S. Kwon, B. Jansen, K.-F. Wong, and M. Cha, "Detecting rumors from microblogs with recurrent neural networks," Jul 2016.

[14] F. Yu, Q. Liu, S. Wu, L. Wang, and T. Tan, "A convolutional approach for misinformation identification," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, Jul 2017.

[15] H. Guo, J. Cao, Y. Zhang, J. Guo, and J. Li, "Rumor detection with hierarchical social attention network," in *the 27th ACM ICKM*, Oct 2018.

[16] J. Ma, W. Gao, and K.-F. Wong, "Detect rumors in microblog posts using propagation structure via kernel learning," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Jul 2017.

[17] J. Ma, W. Gao, and K.-F. Wong, "Rumor detection on twitter with tree-structured recursive neural networks," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Jun 2019.

[18] L. Khoo, H. Chieu, Z. Qian, and J. Jiang, "Interpretable rumor detection in microblogs by attending to user interactions," Jan 2020.

[19] W. Wang, V. W. Zheng, H. Yu, and C. Miao, "A survey of zero-shot learning: Settings, methods, and applications," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 2, pp. 1–37, 2019.

[20] S. Schwarz, A. Theophilo, and A. Rocha, "Emet: Embeddings from multilingual-encoder transformer for fake news detection," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr 2020.

[21] A. De, D. Bandyopadhyay, B. Gain, and A. Ekbal, "A transformer-based approach to multilingual fake news detection in low-resource languages," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 21, p. 1–20, Nov 2021.

[22] H. Lin, P. Yi, J. Ma, H. Jiang, Z. Luo, S. Shi, and R. Liu, "Zero-shot rumor detection with propagation structure via prompt learning," *arXiv preprint arXiv:2212.01117*, 2022.

[23] L. Tian, X. Zhang, and J. H. Lau, *Rumour Detection via Zero-Shot Cross-Lingual Transfer Learning*, p. 603–618. Sep 2021.

[24] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *2020 CVPR*, Aug 2020.

[25] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*, pp. 1597–1607, PMLR, 2020.

[26] T. Gao, X. Yao, and D. Chen, "Simcse: Simple contrastive learning of sentence embeddings," *arXiv preprint arXiv:2104.08821*, 2021.

[27] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," *arXiv preprint arXiv:1911.02116*, 2019.

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

[29] H. Ran and C. Jia, "Unsupervised cross-domain rumor detection with contrastive learning and cross-attention," Mar 2023.

Manuscript 400 submitted to 2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC). Received April 12, 2023.