Open-Topic False Information Detection on Social Networks with Contrastive Adversarial Learning

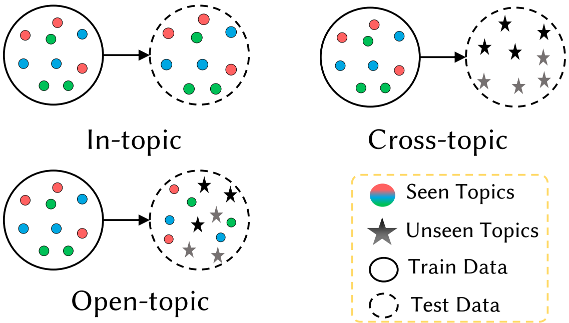
Guanghui Ma[1, 3], Chunming Hu[2, 3, *], Ling Ge[1, 3], Hong Zhang[4]
[1] School of Computer Science and Engineering, Beihang University, Beijing, China
[2] College of Software, Beihang University, Beijing, China
[3] DBDC, Beihang University, Beijing, China
[4] CNCERT/CC, Beijing, China
{maguanghui, hucm, geling}@buaa.edu.cn, zhangh@isc.org.cn

Abstract

Current works about false information detection based on conversation graphs on social networks focus primarily on two research streams from the standpoint of topic distribution: in-topic and cross-topic techniques, which assume that the data topic distribution is identical or cross, respectively. This signifies that all test data topics are seen or unseen by the model. However, these assumptions are too harsh for actual social networks that contain both seen and unseen topics simultaneously, hence restricting their practical application. In light of this, this paper develops a novel open-topic scenario that is better suited to actual social networks. In this open-topic scenario, we empirically find that the existing models suffer from impairment in the detection performance for seen or unseen topic data, resulting in poor overall model performance. To address this issue, we propose a novel Contrastive Adversarial Learning Network, CALN, that employs an unsupervised topic clustering method to capture topic-specific features to enhance the model's performance for seen topics and an unsupervised adversarial learning method to align data representation distributions to enhance the model's generalisation to unseen topics. Experiments on two benchmark datasets and a variety of graph neural networks demonstrate the effectiveness of our approach.

1 Introduction

The convenience and openness of social networks allow people to quickly engage in discussions on a wide range of topics (e.g. the COVID-19 epidemic and the war in Ukraine), and at the same time they also cause people to suffer from false information, the massive spread of which can affect the political and social order of the real world (Rao et al., 2021a; Zhang et al., 2021). Therefore, the detection of false information on social networks is increasingly attracting the attention of the research community (Ma et al., 2018; Li et al., 2019; Yu et al., 2020; Song et al., 2021; Ma et al., 2022).



(a)

Models	Seen Topics	Unseen Topics
In-topic	82.58	60.34
Cross-topic	78.41	65.06

(b)

Figure 1: (a) Comparison of topic distribution in three different scenarios. (b) The accuracy (%) of the in-topic model and cross-topic model in the open-topic scenario for seen and unseen topics, respectively.

Currently, there are two main types of detection models from the perspective of topic distribution: in-topic models and cross-topic models (Ren et al., 2021). (1) In-topic models focus on scenarios where topics of the training and test data are same (Silva et al., 2021; Song et al., 2021), that is, the topics in the test set are **seen**. (2) Cross-topic models focus on scenarios where topics of the training and test data are different (Wang et al., 2018; Ren et al., 2021), which means the topics in the test data are **unseen**. However, we argue that the real social networks contain both seen and unseen topics simultaneously, and we refer to this phenomenon as the **open-topic scenario**. As shown in Figure 1 (a), there is a significant difference among the three topic distributions.

[*] Corresponding Author

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 2911 - 2923
December 7-11, 2022 ©2022 Association for Computational Linguistics

To the best of our knowledge, there is no work on false information detection in open-topic scenarios. The intuitive idea is that we could resort to the existing models to handle detection tasks in this scenario. To implement the above conception, we conduct a preliminary validation experiment. Specifically, we train an in-topic model and a cross-topic model with the same training data, respectively. Then, we leverage an open-topic dataset, mixed seen and unseen topics data, as test dataset to evaluate the performance of the above two models (More details can be found in **Appendix A**.). The experimental results are shown in Figure 1(b). From the results, we find that the accuracy of the in-topic model on unseen topics is significantly lower than that of the cross-topic model. Although the cross-topic model improves the accuracy on unseen topics, the accuracy on seen topics is lower than that of the in-topic model.

We argue that the reasons for the above phenomenon are as follows. (1) The performance of the in-topic model benefits from prior knowledge of the data (Ren et al., 2021; Silva et al., 2021), also known as topic-specific features, such as specific topic words. This prior knowledge cannot be transferred to new topics due to differences between topic features, resulting in poor generalisation of in-topic models to unseen topics. (2) The cross-topic model mainly learns topic-invariant features, such as writing style, from the topics of the train data and transfers these features to the test data (Wang et al., 2018; Castelo et al., 2019). This learning mechanism causes the cross-topic model to discard prior topic knowledge of the data, reducing the performance of the cross-topic model for seen topics. Apparently, the deficiencies of existing methods in open topic scenarios can be alleviated if both topic-specific and topic-invariant features can be preserved.

To retain these two types of features, we face two severe challenges. (1) How to learn topic-specific features? Some works adopt user features (Silva et al., 2021) or pre-trained topic models (Ren et al., 2021) to obtain topic-specific features, which are sub-optimal since they fail to address the diversity of topics on social networks. (2) How to learn topic-invariant features? While existing work has achieved good results using adversarial learning (Wang et al., 2018), it relies on the topic labels , which is challenging for information on social networks because topics are constantly emerging.

To tackle the above problem and challenges, we propose a novel **C**ontrastive **A**dversarial **L**earning **N**etwork (**CALN**) to obtain and fuse both topic-specific features and topic-invariant features. Concretely, we use the drop edge technique (Rong et al., 2020) to generate two augmented graphs from an original conversation graph and leverage a graph neural network (GNN) to obtain their graph-level representations. Since different topic conversation graphs have different topic-specific words and topic-specific propagation patterns (Silva et al., 2021; Mosallanezhad et al., 2022), we design an unsupervised topic feature (TF) learner based on contrastive learning to obtain topic-specific knowledge. This TF learner can learn the intrinsic invariance of the data by maximizing the mutual information between the augmented graphs, ultimately achieving topic clustering. Then, we resort to classical adversarial learning (Ganin and Lempitsky, 2015) to design an unsupervised representation alignment (RA) learner to obtain topic-invariant features, which is achieved by reversing the gradient signal of contrastive learning and aligning the representation distribution of the data. Finally, we fuse the obtained topic-invariant features, topic-specific features, and graph-level representations to predict the authenticity of the conversation graph of false information.

Our contributions are summarized as follows: (1) We study a new issue of false information detection in open-topic scenarios and discover the shortcomings of existing methods in such scenarios. (2) We propose a contrastive adversarial learning network to obtain and fuse topic-specific and topic-invariant feature learning to improve the false information detection in open-topic scenarios. (3) We demonstrate the effectiveness of our method through comparison, ablation and visualization experiments on two real datasets and various GNNs.

2 Related Work

2.1 False Information Detection

Existing false information detection methods can be categorized into two types from the perspective of data distribution: data identical distribution (Rao et al., 2021b; Song et al., 2021) based approaches and non-identical distribution (Castelo et al., 2019; Han et al., 2020) based approaches. Theoretically, the setting of identical data distributions leads to training and testing data on the same topics, resulting in detection models that cannot handle the new

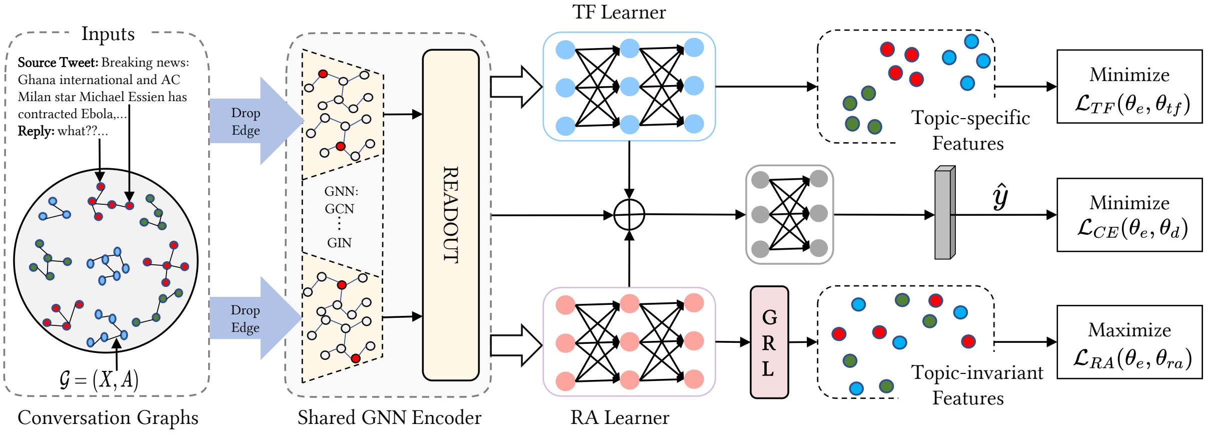

Inputs
Source Tweet: Breaking news: Ghana international and AC Milan star Michael Essien has contracted Ebola,...
Reply: what??...
$\mathcal{G} = (X, A)$
Conversation Graphs
Drop Edge
Drop Edge
GNN:
GCN
⋮
GIN
READOUT
Shared GNN Encoder
TF Learner
RA Learner
G
R
L
Topic-specific Features
Topic-invariant Features
$\hat{y}$
Minimize $\mathcal{L}_{TF}(\theta_e, \theta_{tf})$
Minimize $\mathcal{L}_{CE}(\theta_e, \theta_d)$
Maximize $\mathcal{L}_{RA}(\theta_e, \theta_{ra})$

Figure 2: An illustration of our proposed networks. The TF learner is short for topic learner and The RA learner is short for representation alignment learner.

topics, which are constantly emerging on social networks. Consequently, some researchers have started to study the case of non-identical data distribution, such as cross-topic models (Wang et al., 2018; Castelo et al., 2019; Han et al., 2020; Ren et al., 2021). The cross-topic model aims to learn topic-invariant features from source topics and generalize these features to the different new topics. However, there is a gap between the assumptions of the above works and the real social network scenario. We argue that the social networks tend to be an open-topic scenario. That is, some topics are persistent, such as conspiracy theories and racism, while some new topics are emerging, such as the COIVD-19 epidemic and the war in Ukraine. We have demonstrated through validation experiments that this open topic scenarios can lead to performance degradation of existing models.

2.2 Contrastive Learning

Contrastive learning is a representational learning method that utilizes the relevance of data as a signal for self-supervision and has demonstrated its powerful performance in various domains of natural language processing (Li et al., 2021; You et al., 2022; Ge et al., 2022). Generally, contrastive learning constructs positive samples by data augmentation methods and treats other data within the mini-batch as negative samples, forcing the model to learn the similarities or differences of the data and thus extract the intrinsic features of the data. In this paper, we explore the study of unsupervised topic clustering with contrastive learning. As there are natural differences between the conversation graphs across topics, we use contrastive learning to learn the invariance of the data and form clustering effects.

2.3 Adversarial Learning

Adversarial learning has been considered a promising solution for the topic generalization problem (Wang et al., 2021; Li et al., 2022). The basic idea (Ganin and Lempitsky, 2015) is to add an adversarial learning layer to the topic classifier to learn a topic-invariant representation (Zou et al., 2021). However, most existing adversarial approaches rely on topic labels for feature alignment (Wang et al., 2018; Han et al., 2020; Li et al., 2022), which is unrealistic for complex social networks as we have no way to obtain labels for new topics that keep emerging continuously. Based on those mentioned above, this paper explores using unsupervised adversarial learning to handle cross-topic generalization on social networks.

3 Methodology

3.1 Problem Statement

For the conversation graph detection task in open-topic scenarios, we are given n_s labeled source training examples $(\{\mathcal{G}_i^s, y_i^s\})_{i=1}^{n_s}$ from $K_s = \{K_s^1, ..., K_s^m\}$ topics where $\mathcal{G}_i^s \in \mathcal{G}_s, y_i^s \in \mathcal{Y}_s$ and n_t unlabeled target test examples $(\{\mathcal{G}_j^t\})_{j=1}^{n_t}$ from $K_t = \{K_t^1, ..., K_t^n\}$ topics where $K_s \cap K_t \neq \varnothing$. The goal of this paper is to learn a classification model to predict the conversation graph labels $\{y_j^t\}_{j=1}^{n_t}$ where $y_j^t \in \mathcal{Y}_t$ for the test dataset. We define the task as a binary classification task, where $y \in \{True, False\}$. For conversation graphs on social networks, we follow previous work (Wei et al., 2019; Li et al., 2020) and define the conversation graph as an undirected graph: $\mathcal{G} = (X, A)$, where $X \in \mathbb{R}^d$ denotes the node features and $A \in \mathbb{R}^{m \times m}$ denotes the adjacency matrix.

3.2 Overview

We propose a contrastive adversarial learning network to address the problem of false information detection in open-topic scenarios. As shown in Fig 2, our model consists of four components: a data enhancement and encoder module, a TF learner, a RA learner, and a false information classifier.

3.3 The Data Enhancement and Encoder Module

We perform data augmentation on an original graph $\mathcal{G}_i^s$ ($\mathcal{G}$ for short) to produce two augmented graphs $\hat{\mathcal{G}}_i$ and $\hat{\mathcal{G}}_j$ as positive sample pairs. The reason we adopt the drop edge method for data enhancement is that it can mitigate the influence of the echo chamber effect in false information propagation (He et al., 2021). We pass the augmented view graph to the GNN and READOUT functions to obtain a graph-level representation.

$$h = \mathbf{READOUT}(\mathbf{GNN}(\hat{\mathcal{G}})) \tag{1}$$

where $h \in \mathbb{R}^{d_1}$ and the GNN serves as the shared conversation graph encoder. We adopt global average pooling (GAP) as the READOUT function.

We denote the process above as $f_e(\cdot, \theta_e)$, where the θ_e is the parameters to be learned. Thus, we obtain two enhanced graph-level representations h_i and h_j for an original conversation graph $\mathcal{G}$, where $h_i = f_e(\hat{\mathcal{G}}_i, \theta_e)$ and $h_j = f_e(\hat{\mathcal{G}}_j, \theta_e)$.

3.4 The Topic Feature Learner

The TF learner aims to learn topic-specific features as prior knowledge for detection in an unsupervised approach. Existing methods for learning topic-specific features rely on user information and pre-trained topic models. However, they ignore the propagation patterns of conversation graphs, which are crucial for social networks. Based on the fact that different topic conversation graphs have different topic-specific words and topic-specific propagation patterns (Silva et al., 2021; Mosallanezhad et al., 2022), we propose an unsupervised topic clustering method for conversation graphs with contrastive learning to obtain topic-specific features.

We first pass the two graph-level representations, h_i and h_j, to a neural network to obtain two hidden features z_{tf}^i and z_{tf}^j. This process can be represented as follows:

$$z_{tf} = \mathbf{NN}(\mathrm{h}) \tag{2}$$

where $z_{tf} \in \mathbb{R}^{d_{ts}}$ and $\mathbf{NN}(\cdot)$ consist of two layers of perceptrons and an activation function.

Then, we introduce contrastive learning to maximize the mutual information between the two hidden features, z_{tf}^i and z_{tf}^j, to learn the intrinsic properties of the data. Since contrastive learning can capture the similarities and differences among data to form a natural clustering effect, we refer to such clustering representations as topic-specific features. Herein, the contrastive loss for the TF learner is defined as follows:

$$\begin{aligned}\mathcal{L}_{TF}(\theta_e, \theta_{tf}) = \mathbb{E}_{\mathbb{P}_{\hat{\mathcal{G}}_i}} \{ -\mathbb{E}_{\mathbb{P}_{(\hat{\mathcal{G}}_i|\hat{\mathcal{G}}_j)}} T(z_{tf}^i, z_{tf}^j) \\ + \log \mathbb{E}_{\mathbb{P}_{\hat{\mathcal{G}}_j}} e^{T(z_{tf}^i, z_{tf}^j)} \}\end{aligned} \tag{3}$$

where the θ_{tf} and θ_e denote the parameters to be learned. $\mathbb{P}_{\hat{\mathcal{G}}_i}$ and $\mathbb{P}_{(\hat{\mathcal{G}}_i|\hat{\mathcal{G}}_j)}$ are conditional and marginal distribution of augmented graphs. $T(\cdot, \cdot)$ is a learned score function: $sim(z_{tf}^i, z_{tf}^j)/\tau$, where the $sim(\cdot, \cdot)$ is a cosine similarity function, and τ is a temperature factor (You et al., 2020).

The loss function $\mathcal{L}_{TF}(\theta_e, \theta_{tf})$ can evaluate the differences among topics. The smaller the loss is, the better the clustering result is. Finally, we aim at minimizing the loss $\mathcal{L}_{TF}(\theta_e, \theta_{tf})$ ot the TF learner:

$$\hat{\theta}_{tf} = \arg\min_{\theta_{tf}} \mathcal{L}_{TF}(\theta_e, \theta_{tf}) \tag{4}$$

3.5 The Representation Alignment Learner

The RA learner seeks to align data representation distributions in an unsupervised manner to grasp topic-invariant features. Previous works rely on topic labels to perform topic-invariant learning (Wang et al., 2018). Due to the diversity and complexity of information on social networks, it is impractical to annotate each conversation graph with an explicit topic label. To learn topic-invariant features, we resort to classical adversarial learning, which achieves data representation alignment by adding a gradient reversal layer (GRL) (Ganin and Lempitsky, 2015) to fuse data features.

First, similar to the TF learner, we employ a neural network to obtain two hidden vectors, z_{ra}^i and z_{ra}^j. This process is represented as follows:

$$z_{ra} = \mathbf{NN}(\mathrm{h}) \tag{5}$$

where $z_{ra} \in \mathbb{R}^{d_{ti}}$ and $\mathbf{NN}(\cdot)$ contains two layers of perceptrons and an activation function.

Then, we pass z^i_{ra} and z^j_{ra} into GRL to obtain two vectors z^i_{rev} and z^j_{rev}. The GRL is a constant function in forward propagation, while it reverses the gradient signal by multiplying the parameter $-\lambda$ to the previous layer gradient in backward propagation:

$$z_{rev} = \mathbf{GRL}(\mathrm{z_{ra}}) \tag{6}$$

Finally, we again use contrastive learning to determine whether the two vectors come from the same original conversation graph. The loss for the RA learner is defined as follows:

$$\begin{aligned}\mathcal{L}_{RA}(\theta_e, \theta_{ra}) = \mathbb{E}_{\mathbb{P}_{\hat{\mathcal{G}}_i}}\{-\mathbb{E}_{\mathbb{P}_{(\hat{\mathcal{G}}_i|\hat{\mathcal{G}}_j)}} T(z^i_{rev}, z^j_{rev}) \\ + \log \mathbb{E}_{\mathbb{P}_{\hat{\mathcal{G}}_j}} e^{T(z^i_{rev}, z^j_{rev})}\}\end{aligned} \tag{7}$$

The goal of contrastive loss is to determine the consistency between two data through maximizing the mutual information, while the goal of the graph encoder module is exactly the opposite due to the GRL, thus forming an adversarial relationship. As training converges, the RA learner is unable to distinguish which topic the accepted features come from. Therefore, the above loss $\mathcal{L}_{RA}(\theta_e, \theta_{ra})$ is used to evaluate the differences in topics. The larger the loss, the smaller the topic difference. The goal of the RA learner is to maximize loss $\mathcal{L}_{RA}(\theta_e, \hat{\theta}_{ra})$:

$$\hat{\theta}_{rev} = \arg\max_{\theta_{ra}} \mathcal{L}_{RA}(\theta_e, \theta_{ra}) \tag{8}$$

3.6 False Information Classifier

The false information classifier is a feed-forward fusion network with a perceptron and an activation function, which is designed to predict whether a conversation graph $\mathcal{G}$ is true or false. We concatenate the augmented graph representations h_i (h_j), topic-specific representations z_{tf}, and topic-invariant representations z_{ra} (z_{rev}) and feed them into this classifier to obtain the prediction results $\hat{y}$.

The cross-entropy loss is used to optimize the classifier:

$$\begin{aligned}\mathcal{L}_{CE}(\theta_e, \theta_d)) = -\,\mathbb{E}_{(\mathcal{G},y)\sim(\mathcal{G}^s_i, \mathcal{Y}_s)}[y \log \hat{y} \\ + (1-y)\log \hat{y}]\end{aligned} \tag{9}$$

where θ_d is the parameters of the classifier.

Table 1: The statistics of datasets.

Statistics	PHEME5	PHEME9
Graphs	5802	6425
False	3830	4023
True	1972	2402
Avg comments/graph	17.8	16.3
Avg words/comment	13.6	13.6
Comments	103212	105354
Topics	5	9

The parameters θ_d can be learned by:

$$(\hat{\theta}_e, \hat{\theta}_d) = \arg\min_{\theta_e, \theta_d} \mathcal{L}_{CE}(\theta_e, \theta_d) \tag{10}$$

3.7 Training Objective and Model Analysis

Our training objective is a minimax game between the TF learner, the RA learner and the classifier. Thus, the total loss function is defined as follows:

$$\begin{aligned}\mathcal{L}_{total}(\theta_e, \theta_{tf}, \theta_{ra}, \theta_d) = \mathcal{L}_{CE}(\theta_e, \theta_d) \\ +\alpha\mathcal{L}_{TF}(\theta_e, \theta_{tf}) - \lambda\mathcal{L}_{RA}(\theta_e, \theta_{ra})\end{aligned} \tag{11}$$

where $\alpha \in [0, 1]$ and $-\lambda$ are a trade-off factor

We analyze the difference between the in-topic model (Equation 12), the cross-topic model based on adversarial learning (Equation 13) and our model (Equation 14) from the GNN encoder gradient update perspective. In more detail, the gradient update for these three models is expressed as follows, respectively:

$$\hat{\theta}_e = \theta_e - \eta\frac{\partial \mathcal{L}_y}{\partial \theta_e} \tag{12}$$

$$\hat{\theta}_e = \theta_e - \eta(\frac{\partial \mathcal{L}_y}{\partial \theta_e} - \lambda\frac{\partial \mathcal{L}_d}{\partial \theta_e}) \tag{13}$$

$$\hat{\theta}_e = \theta_e - \eta(\frac{\partial \mathcal{L}_{CE}}{\partial \theta_e} + \frac{\partial \mathcal{L}_{TF}}{\partial \theta_e} + (-\lambda\frac{\partial \mathcal{L}_{RA}}{\partial \theta_e})) \tag{14}$$

where the η is the learning rate , $\mathcal{L}_d$ is the the adversarial loss and $\mathcal{L}_y$ is the cross-entropy loss.

It can be found that the cross-topic model aligns the distinct topic representations by adding the $-\lambda\frac{\partial \mathcal{L}_d}{\partial \theta_e}$ term in an adversarial way compared to the in-topic model. Our model adds an additional $\frac{\partial \mathcal{L}_{TF}}{\partial \theta_e}$ term compared to the cross-topic model to alleviate this extreme adversarial. The above comparison of gradient update shows that the cross-topic

Table 2: Summary of false information detection results: "average ± standard deviation" and "improvement" (%).

GNN	**Method**	**PHEME5**				**PHEME9**			
		Accuracy	△	**F1**	△	**Accuracy**	△	**F1**	△
GCN	N/A	72.66$_{\pm1.43}$	-	72.35$_{\pm1.55}$	-	67.02$_{\pm1.86}$	-	64.28 $_{\pm2.11}$	-
	EANN	71.96$_{\pm1.36}$	↓0.70	71.74$_{\pm2.21}$	↓0.61	68.37 $_{\pm1.99}$	↑1.35	65.06 $_{\pm3.22}$	↑0.78
	RDEA	75.25$_{\pm1.32}$	↑2.59	72.08$_{\pm2.63}$	↓0.27	66.70$_{\pm5.77}$	↓0.32	63.40 $_{\pm5.23}$	↓0.88
	CTTM	73.71$_{\pm1.00}$	↑1.05	73.45$_{\pm1.52}$	↑1.10	69.27$_{\pm0.79}$	↑2.25	65.25$_{\pm1.54}$	↑0.97
	GACL	73.00$_{\pm1.04}$	↑0.34	72.46$_{\pm1.59}$	↑0.11	67.04$_{\pm1.29}$	↑0.02	64.38$_{\pm2.01}$	↑0.10
	CALN	**77.01**$_{\pm\mathbf{0.97}}$	**↑4.35**	**75.89** $_{\pm\mathbf{1.90}}$	**↑3.54**	**71.67**$_{\pm\mathbf{0.95}}$	**↑4.65**	**66.80**$_{\pm\mathbf{0.93}}$	**↑2.52**
SAGE	N/A	74.70$_{\pm1.04}$	-	73.91$_{\pm1.61}$	-	66.51$_{\pm3.40}$	-	63.67$_{\pm3.21}$	-
	EANN	74.41$_{\pm1.80}$	↓0.29	73.80$_{\pm1.81}$	↓0.10	67.02$_{\pm2.20}$	↑0.51	64.03$_{\pm2.10}$	↑0.36
	RDEA	75.59$_{\pm2.47}$	↑0.89	74.25 $_{\pm2.71}$	↑0.34	65.71$_{\pm5.78}$	↓0.80	63.15 $_{\pm4.63}$	↓0.52
	CTTM	75.66 $_{\pm1.96}$	↑0.96	75.30 $_{\pm1.86}$	↑1.40	69.28$_{\pm2.71}$	↑2.77	64.60$_{\pm2.03}$	↑0.93
	GACL	73.18$_{\pm1.21}$	↓1.52	72.35$_{\pm2.12}$	↓1.56	69.10$_{\pm0.59}$	↑2.59	64.83$_{\pm1.52}$	↑1.16
	CALN	**77.47** $_{\pm\mathbf{1.76}}$	**↑2.77**	**76.32** $_{\pm\mathbf{1.65}}$	**↑2.41**	**70.94**$_{\pm\mathbf{1.27}}$	**↑4.43**	**66.03** $_{\pm\mathbf{1.69}}$	**↑2.36**
GIN	N/A	73.20$_{\pm1.73}$	-	72.39$_{\pm1.79}$	-	67.36$_{\pm5.18}$	-	63.67$_{\pm4.39}$	-
	EANN	72.61$_{\pm2.23}$	↓0.59	71.68$_{\pm3.89}$	↓0.71	65.82$_{\pm3.99}$	↓1.51	62.17$_{\pm3.03}$	↓1.50
	RDEA	74.32 $_{\pm2.31}$	↑1.12	70.32 $_{\pm2.77}$	↓2.07	69.66$_{\pm2.56}$	↑2.30	65.74$_{\pm4.81}$	↑2.07
	CTTM	74.33 $_{\pm1.00}$	↑1.13	73.78$_{\pm1.17}$	↑1.39	69.13$_{\pm1.68}$	↑1.77	64.97$_{\pm2.48}$	↑1.30
	GACL	73.73$_{\pm1.72}$	↑0.53	72.35$_{\pm1.52}$	↓2.65	67.69$_{\pm2.11}$	↑0.33	64.74$_{\pm2.10}$	↑1.08
	CALN	**76.11** $_{\pm\mathbf{1.02}}$	**↑2.91**	**74.70** $_{\pm\mathbf{1.64}}$	**↑2.30**	**71.38** $_{\pm\mathbf{3.27}}$	**↑4.02**	**67.39** $_{\pm\mathbf{5.44}}$	**↑3.72**

model completely discards the prior topic knowledge of the data due to adversarial loss. In contrast, our model is a compromise between the in-topic and cross-topic models, which is consistent with our goal of improving detection performance in open-topic scenarios.

4 Experimental Studies

4.1 Datasets

We evaluate the effectiveness of our method on two publicly available benchmark datasets from the Twitter social platform. Among them, PHEME5 (Zubiaga et al., 2016a) contains five topics: the Sydney siege, the Ottawa shooting and Ferguson, etc., and PHEME9 (Zubiaga et al., 2016b) contains nine topics: the Germanwings plane crash and the Ebola virus, etc. The detailed statistics of the dataset are shown in Table 1. We reset and split the dataset to ensure that the test set contains topics out of the training set to simulate the open-topic scenario. More details on dataset pre-processing and splitting can be found in **Appendix B**.

4.2 Baseline and SOTAs

We utilize the original GNN as a baseline and select some related in-topic models and cross-topic models for comparison, including:

EANN (Wang et al., 2018): A cross-topic false information detection model with adversarial learning. To learn topic-invariant features, EANN leverages a clustering algorithm to obtain soft topic labels to perform topic adversarial learning.

RDEA (He et al., 2021): An in-topic false information detection model with contrastive learning. To enhance the model's generalisation, RDEA pretrains the GNN encoder with data augmentation and contrastive learning.

CTTM (Ren et al., 2021): A SOTA cross-topic false information detection model with the mixture of experts paradigm (MOE) (Jacobs et al., 1991). CTTM leverages the pre-trained topic model to obtain topic vectors to enhance the model's generalisation to unseen topics.

GACL (Sun et al., 2022): A SOTA in-topic false information detection model with supervised contrastive and adversarial learning. The method utilizes supervised contrastive learning to improve the model's generalization and introduces adversarial learning to boost the robustness of the model.

For a fair and extensive experimental evaluation, we utilize various GNNs as encoders, including GCN (Kipf and Welling, 2017), SAGE (Hamilton et al., 2017) and GIN (Xu et al., 2019). GCN learns the graph's multi-layer embedding representation of each node by aggregating the embeddings of adjacent nodes. SAGE expands GCN into an inductive learning task. GIN modifies the neighbour aggregation and graph readout functions so that the GNN performance approximates the upper line of the Weisfeiler-Lehman test (Leman and Weisfeiler, 1968).

4.3 Implementation Details

We chose accuracy and macro F1 as the metrics for performance evaluation. We use Adam as the optimizer (Kingma and Ba, 2015). The batch size is set to 64, and the dropout is set to 0.2 empirically. we adopt a three-layer GNN as the backbone network for all models. During the model training process, we use a grid search technique to choose the best super-parameters. The trade-off factor α are selected from $\{0.1, 0.3, 0.5, 0.7, 0.9\}$, the GRL trade-off factor λ is selected from $\{1, 0.5, 0.1, 0.05, 0.01, 0.005, 0.001\}$ and the learning rate is selected from $\{0.001, 0.005, 0.0001, 0.0005\}$. The temperature factor τ is set to 0.5. Notably, we leverage a learning rate decay strategy with linear warm-up for stable training. All calculations are done on an NVIDIA Tesla V100 GPU. More details about the super-parameter selection, baseline and SOTAs model implementation can be found in the **Appendix C**.

4.4 Performance Comparison

The experimental results are shown in Table 2, where N/A indicates the baseline and $\triangle$ represents the performance improvement. It can be observed that CALN achieves the best results on several GNN encoders and datasets, such as an accuracy improvement of more than 4% on PHEME5 compared to the baseline. RDEA and GACL are in-topic models whose performance benefits from prior knowledge of the topic, resulting in their inferior performance to CALN in open-topic scenarios. EANN obtains soft topic labels for the data with a clustering algorithm, and CTTM leverages a pre-trained model to obtain the topic vector, both of which induce additional error bias. In contrast, we naturally perform topic mining in an unsupervised manner, using the intrinsic relevance of the data as the driving signal. As a result, our approach avoids the barriers to applying the model in real-world scenarios due to the lack of topic labels. Moreover, the results on different encoders and datasets demonstrate that CALN is a model-agnostic approach.

To further evaluate the CALN performance in open-topic scenarios, we calculate the accuracy of various methods on the PHEME5 datasets with GCN as the encoder for seen and unseen topics. The experimental results are shown in Figure 3. We can find that CALN achieves the best results on both seen and unseen topics for most of the metrics. For seen topics, EANN performs the worst because it is a cross-topic model that lacks the topic prior knowledge of the data. For unseen topics, RDEA performs poor because it is an in-topic model that fails to generalize to new topics. Our model utilizes the TF learner to enable the topic distribution to directly participate in the model decision, improving the model's performance for seen topics, and utilizes the RA learner to align data representations, improving the model's generalization to unseen topics. Moreover, our contrastive clustering allows for clustering transfer to unseen topics (Section 4.6), further enhancing the model's generalization to unseen topics.

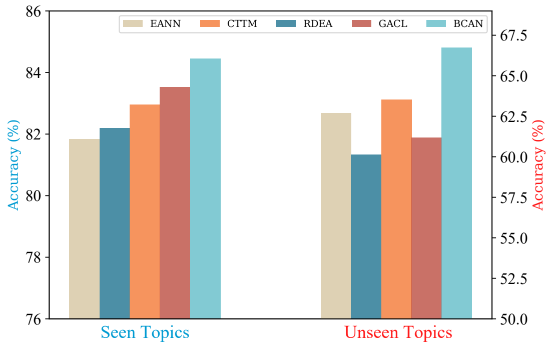


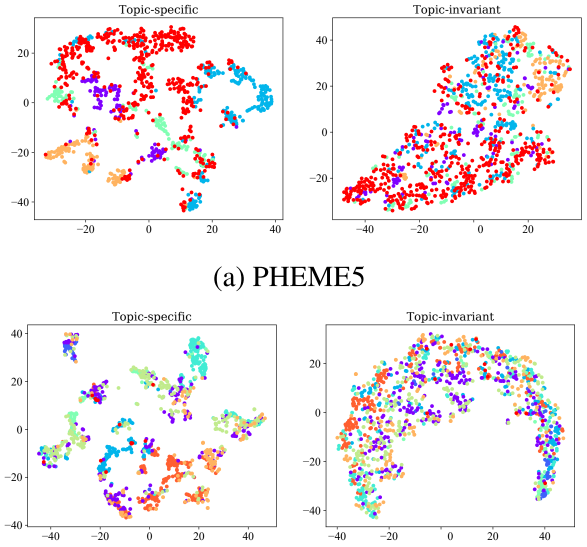
Figure 3: A comparison of the accuracy (%) of different models for seen and unseen topics in the open-topic scenario on PHEME5 dataset.

4.5 Ablation Study

To investigate the impact of each learner on model performance, we conduct an ablation study. We remove the TF learner or the RA learner to observe the model's accuracy for seen and unseen topics in open-topic scenarios, respectively. The experimental results are shown in Table 3. The "w/o TF" indicates removing the TF leaner, and the "w/o RA" indicates removing the RA leaner. Although most metrics show degradation when any leaner is removed, there is a significant difference between seen and unseen topics. For seen topics, the performance of "w/o TF" degrades more than that of "w/o RA". For example, "w/o TF" and "w/o RA" models drop 5.62% and 2.24%, respectively, compared to CALN on the PHEME5 dataset with SAGE as encoder, which indicates that the TF leaner is more

Table 3: The ablation experiment study. The "acc" denotes accuracy (%) and the △ denotes the accuracy change with respect to CALN .

Topics	Method	PHEME5						PHEME9					
		GCN		SAGE		GIN		GCN		SAGE		GIN	
		acc	△	acc	△	acc	△	acc	△	acc		acc	△
Seen	CALN	**84.45**	-	**85.76**	-	**84.14**	-	**81.06**	-	**84.80**	-	**86.13**	-
	w/o TF	81.46	↓2.99	80.14	↓5.62	81.64	↓2.50	80.53	↓0.53	82.13	↓2.67	82.13	↓ 4.00
	w/o RA	82.39	↓2.06	83.52	↓2.24	83.33	↓ 0.81	80.80	↓0.26	82.40	↓2.40	82.13	↓4.00
Unseen	CALN	66.73	-	**67.59**	-	**64.39**	-	**69.17**	-	**67.86**	-	**69.65**	-
	w/o TF	**67.16**	↑0.43	66.52	↓1.07	64.00	↓0.39	67.46	↓1.71	67.05	↓0.81	68.29	↓1.36
	w/o RA	65.45	↓1.28	63.75	↓3.84	61.83	↓2.56	65.41	↓3.76	66.12	↓1.74	63.55	↓6.10



(a) PHEME5

(b) PHEME9

Figure 4: Visualization of the representations obtained from the TF learner (topic-specific) and the RA learner (topic-invariant).

focused on seen topics. The "w/o RA" has more performance degradation for unseen topics, indicating that the RA leaner is more favourable for generalising unseen topics. The above experimental results illustrate the rationality of CALN.

4.6 Visualization Analysis

We perform visualization studies to explore the TF learner's ability to capture topic features and the RA learner's ability to align data representations. To better demonstrate the visualization results, we randomly select some test data to obtain topic-specific features z_{tf} and topic-invariant representation z_{ra} (z_{rev}) . We resort to the topic labels of the original data and use T-SNE for visualization. The experimental result is shown in Figure 4. We notice that the topic-specific features z_{tf} can form apparent topic clustering effects in the PHEME5 dataset, including for unseen topics (red sample points). It demonstrates that our model can capture topic features and has a transfer clustering effect for unseen topics. For the topic-invariant features, we observe that the distribution of the data representations presents aligned results, with no clear boundaries of distinction among different topics, compared with the topic-specific representations. It indicates that the RA learner eliminates the difference among topics. Moreover, the parameter λ is significant for the RA learner. Due to space constraints, the visualizations based on different λ can be found in the **Appendix D**.

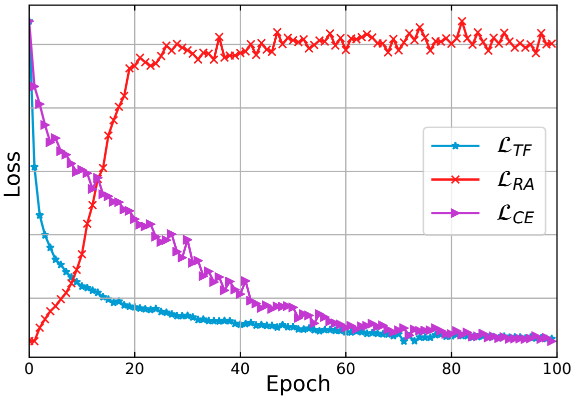


Figure 5: The training loss development.

4.7 Convergence Analysis

To investigate the stability of the training process of CALN , we collect the changes of the loss $\mathcal{L}_{TF}$ from the TF learner, the loss $\mathcal{L}_{RA}$ from the RA learner and the loss $\mathcal{L}_{CE}$ from the classifier during the training process with GCN as the GNN encoder. The experimental results are shown in Figure 5. We find that the loss $\mathcal{L}_{TF}$ and $\mathcal{L}_{CE}$ keep decreasing while the loss $\mathcal{L}_{RA}$ shows a slight decrease at the beginning and then gradually increases. The decrease in loss $\mathcal{L}_{TF}$ indicates that the TF learner gradually completes the data intrinsic feature learning and forms the clustering effect. The increase in $\mathcal{L}_{RA}$ indicates that the representation distribution of the data tends to be aligned. These three losses progressively converge to a stable level with the increase of the epoch.

5 Conclusion

This paper proposes a novel open-topic scenario containing the seen and unseen topic simultaneously for false information detection to complement existing work. We explore the shortcomings of existing models in the open-topic scenario and propose a contrastive adversarial learning network CALN, containing a topic feature learner and a representation alignment learner. The topic feature learner is an unsupervised topic-based clustering method to learn topic-specific features, improving the model's performance for seen topics. The representation alignment learner is an unsupervised adversarial learning method to learn topic-invariant features, enhancing the model's generalize for unseen topics. Experiments on various GNN encoders and two real datasets demonstrate the effectiveness of our model.

6 Limitations

There are some potential limitations in this study. First, our model is based on conversation graphs and relies on the graph structure formed by user comments, so it is weak for detecting early propagation. Early detection of false information will be the future direction of our work. Second, the open-topic scenario we constructed contains at most nine topics, which is still some gap in topic diversity from factual scenarios. Therefore, building a more diverse and large-scale dataset can further advance the field of false information detection.

Acknowledgements

We sincerely thank the reviewers for their insightful comments and valuable suggestions. Thanks for the computing infrastructure provided by Beijing Advanced Innovation Center for Big Data and Brain Computing.

References

Tian Bian, Xi Xiao, Tingyang Xu, Peilin Zhao, Wenbing Huang, Yu Rong, and Junzhou Huang. 2020. Rumor detection on social media with bi-directional graph convolutional networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 549–556.

Sonia Castelo, Thais Almeida, Anas Elghafari, Aécio Santos, Kien Pham, Eduardo Nakamura, and Juliana Freire. 2019. A topic-agnostic approach for identifying fake news pages. In *Companion proceedings of the 2019 World Wide Web conference*, pages 975–980.

Matthias Fey and Jan E. Lenssen. 2019. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*.

Yaroslav Ganin and Victor S. Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 1180–1189. JMLR.org.

Ling Ge, ChunMing Hu, Guanghui Ma, Junshuang Wu, Junfan Chen, JiHong Liu, Hong Zhang, Wenyi Qin, and Richong Zhang. 2022. E-VarM: Enhanced variational word masks to improve the interpretability of text classification models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1036–1050, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

William L. Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 1024–1034.

Yi Han, Shanika Karunasekera, and Christopher Leckie. 2020. Graph neural networks with continual learning for fake news detection from social media. *arXiv preprint arXiv:2007.03316*.

Zhenyu He, Ce Li, Fan Zhou, and Yi Yang. 2021. Rumor detection on social media with event augmentations. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2020–2024.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. 1991. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015*.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

AA Leman and Boris Weisfeiler. 1968. A reduction of a graph to a canonical form and an algebra arising during this reduction. *Nauchno-Technicheskaya Informatsiya*, 2(9):12–16.

Jia Li, Chongyang Tao, Huang Hu, Can Xu, Yining Chen, and Daxin Jiang. 2022. Unsupervised cross-domain adaptation for response selection using self-supervised and adversarial training. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 562–570.

Jiawen Li, Yudianto Sujana, and Hung-Yu Kao. 2020. Exploiting microblog conversation structures to detect rumors. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5420–5429.

Quanzhi Li, Qiong Zhang, and Luo Si. 2019. Rumor detection by exploiting user credibility information, attention and multi-task learning. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1173–1179.

Yunfan Li, Peng Hu, Zitao Liu, Dezhong Peng, Joey Tianyi Zhou, and Xi Peng. 2021. Contrastive clustering. In *2021 AAAI Conference on Artificial Intelligence (AAAI)*.

Guanghui Ma, Chunming Hu, Ling Ge, Junfan Chen, Hong Zhang, and Richong Zhang. 2022. Towards robust false information detection on social networks with contrastive learning. In *Proceedings of the 31st ACM International Conference on Information Knowledge Management*, CIKM '22, page 1441–1450, New York, NY, USA. Association for Computing Machinery.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2018. Rumor detection on twitter with tree-structured recursive neural networks. Association for Computational Linguistics.

Ahmadreza Mosallanezhad, Mansooreh Karami, Kai Shu, Michelle V Mancenido, and Huan Liu. 2022. Domain adaptive fake news detection via reinforcement learning. *arXiv preprint arXiv:2202.08159*.

Dongning Rao, Xin Miao, Zhihua Jiang, and Ran Li. 2021a. STANKER: stacking network based on level-grained attention-masked BERT for rumor detection on social media. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 3347–3363. Association for Computational Linguistics.

Dongning Rao, Xin Miao, Zhihua Jiang, and Ran Li. 2021b. Stanker: Stacking network based on level-grained attention-masked bert for rumor detection on social media. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3347–3363.

Xiaoying Ren, Jing Jiang, Ling Min Serena Khoo, and Hai Leong Chieu. 2021. Cross-topic rumor detection using topic-mixtures. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1534–1538.

Yu Rong, Wenbing Huang, Tingyang Xu, and Junzhou Huang. 2020. Dropedge: Towards deep graph convolutional networks on node classification. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Amila Silva, Ling Luo, Shanika Karunasekera, and Christopher Leckie. 2021. Embracing domain differences in fake news: Cross-domain fake news detection using multi-modal data. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021,*, pages 557–565. AAAI Press.

Yun-Zhu Song, Yi-Syuan Chen, Yi-Ting Chang, Shao-Yu Weng, and Hong-Han Shuai. 2021. Adversary-aware rumor detection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1371–1382.

Tiening Sun, Zhong Qian, Sujun Dong, Peifeng Li, and Qiaoming Zhu. 2022. Rumor detection on social media with graph adversarial contrastive learning. In *WWW '22: The ACM Web Conference 2022*.

Runchuan Wang, Zhao Zhang, Fuzhen Zhuang, Dehong Gao, Yi Wei, and Qing He. 2021. Adversarial domain adaptation for cross-lingual information retrieval with multilingual bert. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 3498–3502.

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 *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*, pages 849–857.

Penghui Wei, Nan Xu, and Wenji Mao. 2019. Modeling conversation structure and temporal dynamics for jointly predicting rumor stance and veracity. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 4786–4797. Association for Computational Linguistics.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. 2019. How powerful are graph neural networks? In *International Conference on Learning Representations*.

Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. 2020. Graph contrastive learning with augmentations. *Advances in Neural Information Processing Systems*, 33:5812–5823.

Yuning You, Tianlong Chen, Zhangyang Wang, and Yang Shen. 2022. Bringing your own view: Graph contrastive learning without prefabricated data augmentations. WSDM '22, page 1300–1309, New York, NY, USA. Association for Computing Machinery.

Jianfei Yu, Jing Jiang, Ling Min Serena Khoo, Hai Leong Chieu, and Rui Xia. 2020. Coupled hierarchical transformer for stance-aware rumor verification in social media conversations. Association for Computational Linguistics.

Xueyao Zhang, Juan Cao, Xirong Li, Qiang Sheng, Lei Zhong, and Kai Shu. 2021. Mining dual emotion for fake news detection. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, pages 3465–3476. ACM / IW3C2.

Han Zou, Jianfei Yang, and Xiaojian Wu. 2021. Unsupervised energy-based adversarial domain adaptation for cross-domain text classification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1208–1218.

Arkaitz Zubiaga, Maria Liakata, and Rob Procter. 2016a. Learning reporting dynamics during breaking news for rumour detection in social media. *arXiv preprint arXiv:1610.07363*.

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016b. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS one*.

A Validation Experiment Supplement

This section supplements the validation experiments described in the introduction.

For the experiment, we use Bian et al. (2020), a prominent architecture utilized by several models for false information detection, as in-topic model, and select the Wang et al. (2018), a traditional model with adversarial learning, as the cross-topic model. The PHEME5 is used as the dataset. We split the dataset to ensure that the test set contains topics out of the training set (Section B). In this manner, we construct a training and test data that fits the open-topic scenario. We feed the test set to the above two models and measure the accuracy of each model for seen and unseen topics detection.

B Dataset Details Supplement

In this section, we present the additional details of the datasets.

Table 4: The PHEME5 Datasets.

Topics	Graphs	Tweets	False	True
Charlie Hebdo	2079	38268	458	1621
Sydney siege	1221	23996	522	699
Ferguson	1143	24175	284	859
Ottawa shooting	890	12284	470	420
Germanwings-crash	469	4489	238	231

In the paper, we use two publicly available datasets, PHEME5 and PHEME9. These two datasets derived from the real social platform Twitter and contain the complete user comment text and reply relationships, which are accessible from https://figshare.com/articles/dataset/PHEME_dataset_for_Rumour_Detection_and_Veracity_Classification/6392078 . As shown in Table 4, the PHEME5 dataset contains five different topics. The PHEME9 is developed from PHEME5 and contains nine topics. In the data pre-processing, we replace all the user names in the text with "User", and the hyperlinks address with 'URL' to avoid user information leakage. To create an open-topic dataset, we take PHEME5 as an example. We split "Charlie Hebdo", "Sydney siege", "Ferguson" and "Germanwings crash" by 8:1:1 for the training set, validation set and test set, respectively, ensuring that the topics in both the training and validation sets are seen. To guarantee that the test set has unseen topics, we merge "Ottawa shooting" with the above test set to create a new test set as the open-topic scenario.

C Implementation Details Supplement

In this section, we present additional details on the model implementation.

we adopt a three-layer GNN as the backbone network for all models. For the Baseline implementation, we add a perceptron layer as a classifier on top of the backbone network. We employ the source code from the original paper to implement EANN [1], RDEA [2] and GACL [3]. Since the original paper of CTTM does not provide available source code, we have reproduced it as much as possible according to the description of the paper.Specifically, the CTTM contains two variants of the model, **Avg** and **Param**. As the performance of **Param** is significantly better than **Avg** in the original paper,we use the **Param** model as our comparison model. In addition, since EANN and CTTM are not graph neural network-based models, we convert their encoders to GNN encoders. We use PyTorch [4] and PyTorch-geometric [5] (Fey and Lenssen, 2019) to implement all models. In the data augmentation process,we use the standard interface of PyTorch-geometric to drop edge.The drop probability is set to 0.1 and 0.2 to obtain the two augmented graphs.

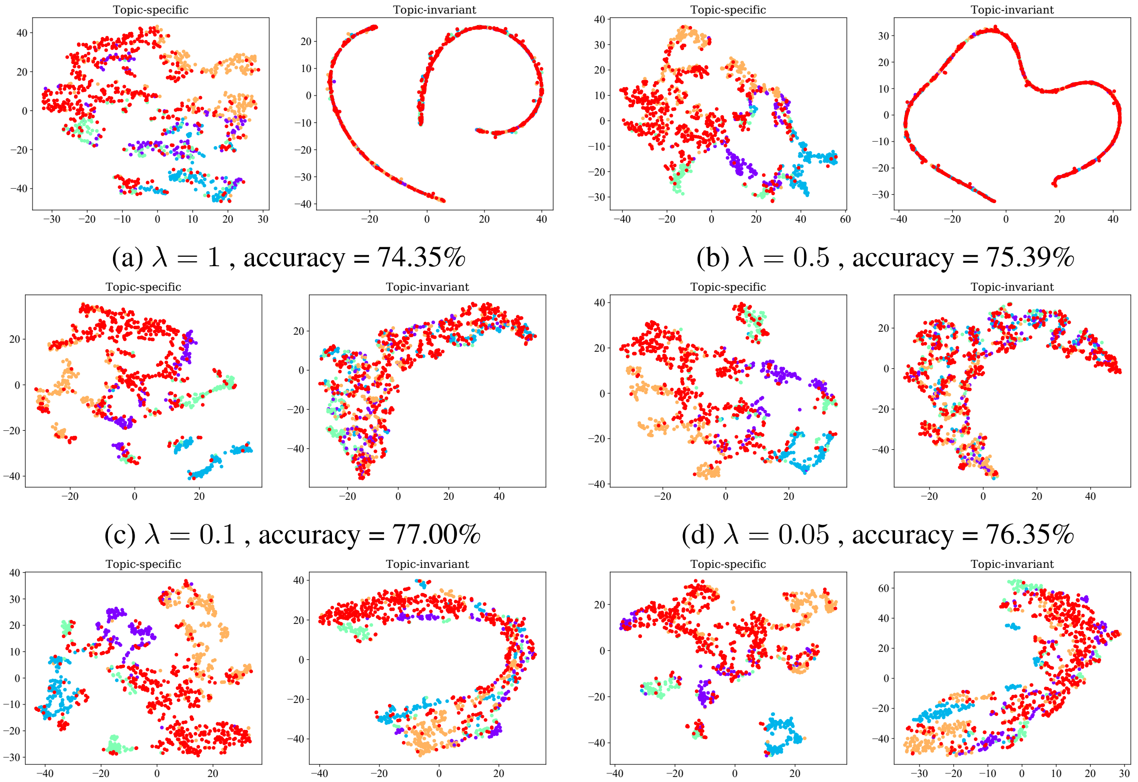
D Visualization Analysis Supplement

In this section, we supplement the visualization results of topic-specific and topic-invariant features with different λ.

We use PHEME5 as an example to illustrate the effect of λ on the model. The visualization experiment results are shown in Figure 6. We can notice that when λ=1, the excessive adversarial effect causes the data to collapse in the RA learner and lose their discriminability completely. When λ=0.0005, it causes a certain topic clustering effect in the RA learner. We explain this phenomenon by the following analysis.

The contrastive loss can be rewritten as follows

[1] https://github.com/yaqingwang/EANN-KDD18
[2] https://github.com/hzy-hzy/RDEA
[3] https://github.com/agangbe/GACL
[4] https://github.com/pytorch
[5] https://github.com/pyg-team/pytorch_geometric



(a) $\lambda = 1$, accuracy = 74.35%

(b) $\lambda = 0.5$, accuracy = 75.39%

(c) $\lambda = 0.1$, accuracy = 77.00%

(d) $\lambda = 0.05$, accuracy = 76.35%

(e) $\lambda = 0.005$, accuracy = 76.59%

(f) $\lambda = 0.0005$, accuracy = 75.37%

Figure 6: Visualizing Supplement.

[6]:

$$\mathcal{L} = -\log \frac{\exp(z_m^i \cdot z_m^j/\tau)}{\exp(z_m^i \cdot z_m^j/\tau) + \sum_N \exp(z_m^i \cdot z_n/\tau)} \tag{15}$$

where z_m^i and z_m^j denote positive sample pairs and the z_n denotes negative sample.

We notice that the contrastive loss aims to pull closer the positive pairs from the same conversation graph, and push away the negative samples from the different conversation graphs within the batch. Due to adversarial learning, the GNN encoder would pull in all negative samples, thus forming a representational alignment. When the adversarial learning is strong enough (λ is large), it causes the data to shrink completely together, making it impossible for the data to retain enough information for the classification task. When the adversarial learning is small (λ is small), the data shows a certain topic clustering effect due to contrastive learning.

As shown in the Figure 6, we can observe that when $\lambda = 1$ or $\lambda = 0.5$, the topic-invariant features collapse due to excessive adversarial, which affects the distribution of representations in the TF learner. At this point, the accuracy of the model is about 75%. When $\lambda = 0.1$ or $\lambda = 0.05$, the data representation distribution of the RA learner is well-aligned, and the TF learner captures apparent topic features. The model obtained better results at this point with an accuracy of 76.5%. When $\lambda = 0.0005$, the adversarial learning power is too small, leading to the effect of topic clustering in the RA learner, and the model's accuracy decreases at this time.

[6] For the relationship between mutual information and contrastive learning, please refer to the paper (You et al., 2020)