Unsupervised Cross-Domain Rumor Detection with Contrastive Learning and Cross-Attention

Hongyan Ran, Caiyan Jia*

School of Computer and Information Technology & Beijing Key Lab of Traffic Data Analysis and Mining
Beijing Jiaotong University, Beijing 100044, China
{hongyran,cyjia}@bjtu.edu.cn

Abstract

Massive rumors usually appear along with breaking news or trending topics, seriously hindering the truth. Existing rumor detection methods are mostly focused on the same domain, thus have poor performance in cross-domain scenarios due to domain shift. In this work, we propose an end-to-end instance-wise and prototype-wise contrastive learning model with cross-attention mechanism for cross-domain rumor detection. The model not only performs cross-domain feature alignment, but also enforces target samples to align with the corresponding prototypes of a given source domain. Since target labels in a target domain are unavailable, we use a clustering-based approach with carefully initialized centers by a batch of source domain samples to produce pseudo labels. Moreover, we use a cross-attention mechanism on a pair of source data and target data with the same labels to learn domain-invariant representations. Because the samples in a domain pair tend to express similar semantic patterns especially on the people's attitudes (e.g., supporting or denying) towards the same category of rumors, the discrepancy between a pair of source domain and target domain will be decreased. We conduct experiments on four groups of cross-domain datasets and show that our proposed model achieves state-of-the-art performance.

Introduction

Nowadays, with the rapid development of social media which has evolved into the primary source of news, more and more people are spending more time on social media platforms expressing what they see and hear, especially when it comes to hot topics or breaking events[1]. This creates a hotbed for the rumor subsisting and inspires the rumor publishers to make numerous rumors for their purposes. For instance, as rumors about the COVID-19 pandemic era spread rapidly, around 800 deaths, 5,000 hospitalizations, and 60 permanent injuries were recorded due to false claims that household bleach was an effective panacea for the virus (Coleman 2020). There are still many such rumors on social media, if not identified in time, sensational rumors may cause social panic during emergency events, and threaten the internet's credibility and trustworthiness. Thus, it has highly practical application value to detect rumors in an efficient way on social media platforms.

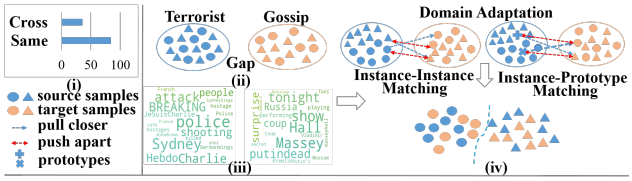


Figure 1: An illustration example of unsupervised cross-domain rumor detection.

To solve the problem, various rumor detection methods have been proposed including traditional detection models and deep learning detection models. The traditional works mostly utilize hand-crafted features to perform rumor detection (Castillo, Mendoza, and Poblete 2011; Kwon et al. 2013; Ma et al. 2015; Jin et al. 2017). The deep learning-based methods exploit the content of rumors to automatically detect rumors. These methods are developed from the original methods which view the textual information of source posts and user responses or user profiles as time series (Ma et al. 2016; Yu et al. 2017; Ma, Gao, and Wong 2019; Ma et al. 2021), to the propagation structure-based methods which model rumor content and related responses as a tree structure to learn rumor representations (Ma, Gao, and Wong 2018; Bian et al. 2020; Khoo et al. 2020; Zhang et al. 2021b), and finally to the multi-resource heterogeneous aggregation methods (Huang et al. 2020; Ran et al. 2022; Min et al. 2022) which achieve the best detection performance.

However, these methods mostly detect rumors under in-domain conditions. In practical scenarios, the real-world news platforms release various claims in different domains everyday, and newly emergent and time-critical domain events are difficult to acquire sufficient labeled data in time. If we directly use the verified posts of history domains to train these models and test them on the newly emergent domain data, we will get poor performance due to the domain shift. We take PPA-WAE model (Zhang et al. 2021b) as an example to verify its performance on cross-domain settings. Since PHEME (Zubiaga et al. 2015) dataset contains two domain events including 'terrorist' events and 'gossip' events,

*Corresponding authors

[1]https://user.guancha.cn/main/content?id=710205

we view the two domain events as a cross-domain setting (See Figure 1(ii-iii)). We compare the experimental results of PPA-WAE at the in-domain and the cross-domain settings which are shown in Figure 1(i), and observe that the accuracy of the cross-domain is far below than that of the in-domain. It demonstrates that propagation structure-based state-of-the-art detection models while they perform well for the domain they are trained on (e.g., terrorist), perform poorly in other domains (e.g., gossip). The limited cross-domain effectiveness of methods to detect rumors is mostly due to the domain-specific word usage and the writing style of rumor content (See Figure 1(iii)) making the model biased toward the training domains.

To address these challenges, some works (Lin et al. 2022; Mosallanezhad et al. 2022) propose domain-invariant feature learning algorithms for cross-domain rumor detection, while needing labeled target data as auxiliary information to train the models. Thus these models can not handle the settings where no labeled target data is available. Namely, these methods does not work at unsupervised cross-domain settings. (Min et al. 2022) utilize an adversarial topic discriminator for topic agnostic feature learning to the unsupervised cross-domain rumor detection, but it has worse performance. Therefore, an effective method is needed to alleviate the challenges of unsupervised cross-domain rumor detection. Due to the consistency of the propagation patterns for rumors, no matter which domain they come from (Ran et al. 2022), we hope that the method is able to pull closer to the same labeled samples and push apart the different labeled samples between the source domain and the target domain to decrease the domain gaps, so we believe that aligning the representation space of rumor-indicative patterns of different domains could adapt the features captured from the source data to that of the target data (See Figure 1(iv)).

In this study, inspired by self-supervised contrastive learning (He et al. 2020; Chen et al. 2020), we propose an unsupervised cross-domain rumor detection method based on contrastive learning and cross-attention. Since the same categories are shared by both domains, we build instance-wise and prototype-wise contrastive learning to align features so as to reduce the discrepancy between source data and target data. It not only performs cross-domain samples feature alignment but also enforces the target samples to be aligned with the prototype of the corresponding source domain. Since target labels are not available, we use a clustering-based approach with carefully initialized centers on batch samples of a given source domain to produce pseudo labels. Moreover, we use a cross-attention mechanism on pairs of source data and target data with the same labels to learn domain-invariant features. Because these pairs express similar semantic patterns learned on propagation paths of rumors, the discrepancy between domains will be decreased, thereby boosting the performance of our cross-domain rumor detection method.

The main contributions of this study are summarized in the following.

- We investigate the problem of cross-domain rumor detection and propose instance-wise and prototype-wise contrastive learning to align feature representations so as to reduce the discrepancy between domains.
- We construct a cross-attention mechanism between the source data and target data pairs with the same labels to learn domain-invariant features.
- We conduct experiments on four groups of cross-domain datasets and show that our proposed method achieves the best performance.

Related Work

Rumor Detection

Existing rumor detection methods mostly pay attention to the same domain rumor data and build various frameworks on it for well adapting to the tasks of rumor detection. For instance, sequence processing models leverage the textual contents from the source posts and user reply comments for rumor detection (Ma et al. 2016; Yu et al. 2017; Ma, Gao, and Wong 2019; Ma et al. 2021), propagation structure-based methods (Ma, Gao, and Wong 2018; Bian et al. 2020; Khoo et al. 2020; Zhang et al. 2021b; Liu et al. 2022; Sun et al. 2022) model the propagation paths as a tree attached with the textual content to build the semantics of posts and their propagation relationships, and some studies integrate the content of posts, relationships of user-post and user-user pairs, user profiles as a heterogeneous graph and have achieved the best performance for rumor detection (Huang et al. 2020; Li et al. 2021; Ran et al. 2022). Lately, some researchers study semi-supervised cross-domain rumor detection using supervised contrastive learning and reinforcement learning (Lin et al. 2022; Mosallanezhad et al. 2022). Different from these researches, we mainly focus on unsupervised cross-domain rumor detection, where the labels of the target domain are unavailable. Recently, (Min et al. 2022) propose an unsupervised cross-domain rumor detection method using an adversarial topic discriminator for topic-invariant feature learning with limited performance. In addition, the model demands complex multi-source information as inputs, whereas our model only considers rumor content and its social text.

Unsupervised Cross-Domain

Unsupervised cross-domain (UCD) aims to learn a model that is able to achieve good classification accuracy without any annotation in a target domain (Ben-David et al. 2010). Existing UCD methods mainly appear in the field of computer vision which includes domain-level and category-level apporaches. Domain-level approaches use the Maximum Mean Discrepancy (MMD) to mitigate the distribution divergence between the source and target domains by pulling them into the same distribution at different scale levels (Ganin and Lempitsky 2015; Long et al. 2018). Category-level methods align each category distribution between the source domain and target domain by pushing the target samples to the distribution of source samples in each category (Saito et al. 2018; Du et al. 2021; Xu et al. 2021). Recently, UCD has been applied to various applications such as text classification (Zou, Yang, and Wu 2021; Li et al. 2022) and sentiment analysis (Du et al. 2020; Ghosal et al. 2020), etc. In this work, we intent to introduce the category-level

feature alignment and domain-invariant feature learning between domains into cross-domain rumor detection tasks.

Self-Supervised Contrastive Learning

The core idea of self-supervised contrastive learning is to learn from positive samples and benefit from correcting negative ones, which has been successfully applied to many fields. In computer vision, a large collection of works (He et al. 2020; Chen et al. 2020) learn self-supervised image representation by minimizing the distance between two views of the same image. In natural language processing, studies suggest that contrastive learning is promising in the semantic textual similarity (Gao, Yao, and Chen 2021), stance detection, and short text clustering (Mohtarami, Glass, and Nakov 2019; Zhang et al. 2021a). In addition, contrastive learning has successfully promoted the development of representation learning of graph-structured data (Qiu et al. 2020; You et al. 2020; Zhu et al. 2021). Our work is inspired by self-supervised contrastive learning, but the difference is that we use supervised contrastive learning for unsupervised cross-domain rumor detection tasks.

Problem Statement

Definition Unsupervised cross-domain rumor detection aims to transfer models learned from a labeled source data to an unlabeled target data. Given a labeled rumor dataste $D^s = \{({C_i}^s, {y_i}^s)\}, i = 1, 2, 3..., n_s$ from the source domain, where ${C_i}^s$ denotes the set of post and comment contents described by ${C_i}^s = \{{s_i}^s, {r_{i1}}^s, ..., {r_{i|i|-1}}^s\}$, in which ${s_i}^s$ is a source tweet, and ${r_{ij}}^s$ is the j-th comment text and $|i|$ refers to the number of source post and comments in ${C_i}^s$, ${y_i}^s \in y^s$ denotes the corresponding label. In the target domain, also given an unlabeled dataset $D^t = \{{C_i}^t\}, i = 1, 2, ..., n_t$, where ${C_i}^t$ represents the rumor of i-th unlabeled sample in the target domain and also includes a series of reply posts. Our goal is to predict labels of testing samples in the target domain using a model $f_t : C^t \rightarrow y^t$ trained on $D^s \cup D^t$, where the target label space y^t is equal to the source label space y^s.

Data processing We build each rumor C_i whichever comes from the source domain or the target domain as a propagation tree. Since the propagation paths aggregated method (Zhang et al. 2021b) can effectively acquire the representation of a rumor, we construct the rumor C_i as a set of propagation paths $P_i = \{P_{i1}, P_{i2}, ..., P_{i|i|}\}$, where $|i|$ denotes the propagation path number of the rumor C_i, P_{ij} represents j-th path of the rumor corresponds to the path from source tweets s_i to the j-th leaf node, and it can be represented by $P_{ij} = [s_i, n_{ij1}, n_{ij1}, ..., n_{ij|n|}]$, where n_{ijk} is the k-th node of the j-th path in the i-th rumor propagation tree, $|n|$ is the number of nodes along the path, and each node includes a series of words. We use GloVe 300d word embedding vectors (Pennington, Socher, and Manning 2014) to initialize each word in a propagation path P_{ij}.

Proposed Model

In this section, we will describe our proposed model, **U**nsupervised **C**ross-**D**omain **R**umor **D**etection with Contrastive Learning and Cross-Attention (**UCD-RD**). As shown in Figure 2, the **UCD-RD** model has four main components: Rumor Representation Module (RRM) , Contrastive Learning Module (CLM), Cross-Attention Module (CAM) and Rumor Prediction Module (RPM).

Rumor Representation Module (RRM)

Since the self-attention mechanism in the transformer network (Vaswani et al. 2017) enables the model to effectively model long-range dependencies, hence we use the transformer network and its Multi-head Attention (MHA) module to learn the propagation structure information to obtain a good rumor representation.

Given a set of propagation path P_i of rumor C_i which comes from the source domain or the target domain, we apply max-pooling to each path $P_{ij} \in P_i$ in the linear structure to obtain it's path representation $X_{P_{ij}}$, the sequence of the path embedding for the rumor C_i can be represented as $X_{P_i} = (X_{P_{i1}}, X_{P_{i2}}, ..., X_{P_{i|i|}})$.

Next, we use the MHA module to learn path propagation embedding for each rumor. In detail, an MHA layer is made up of a self-attention layer and fully connected feed-forward layer. The scaled dot-product self-attention layer is the core component of the transformer. It uses an attention score to consider the relation between inputs. The inputs consist of queries and keys of dimension d_k, and values of dimension d_v. The dot products of a query (representing a specific path $X_{P_{ij}} \in X_{P_i}$) with all keys and apply a softmax function to obtain the attention score on the values. The set of queries and those of the keys and the values are packed together into matrices Q, K, and V. For each head j, these input matrices will be projected into d^k, d^k, d^v dimension subspaces as Q_j, K_j, V_j through trainable linear projections $W_j^Q \in R^{d \times d^k}$, $W_j^K \in R^{d \times d^k}$, $W_j^V \in R^{d \times d^v}$, respectively. We denote the j-th head by $head_j$ as follows:

$$head_j = f_{att}(Q_j, K_j, V_j) = softmax(\frac{Q_j K_j^T}{\sqrt{d^k}})V_j \quad (1)$$

For the rumor C_i, the final output of MHA can be calculated as a linear projection of concatenation of h heads:

$$O_i = f_{MHA} = concat(head_1, ..., head_h)W^o \quad (2)$$

where $W^o \in R^{hd^v \times d}$.

The output O_i of the self-attention layer is then passed through a fully connected feed-forward layer consisting of two linear units with the Relu activation to acquire all the path representations of the rumor C_i. And finally, we use the max-pooling operation to aggregate the path embedding to obtain the representation $\hat{O}_i$ of the rumor C_i.

Contrastive Learning Module (CLM)

Inspired by the success of self-supervised contrastive learning (He et al. 2020; Chen et al. 2020), we model cross-domain rumor detection using instance-wise and prototype-wise contrastive learning to automatically learn how to align the sample representations from both labeled source domain D^s and unlabeled target domain D^t, and use in-domain and

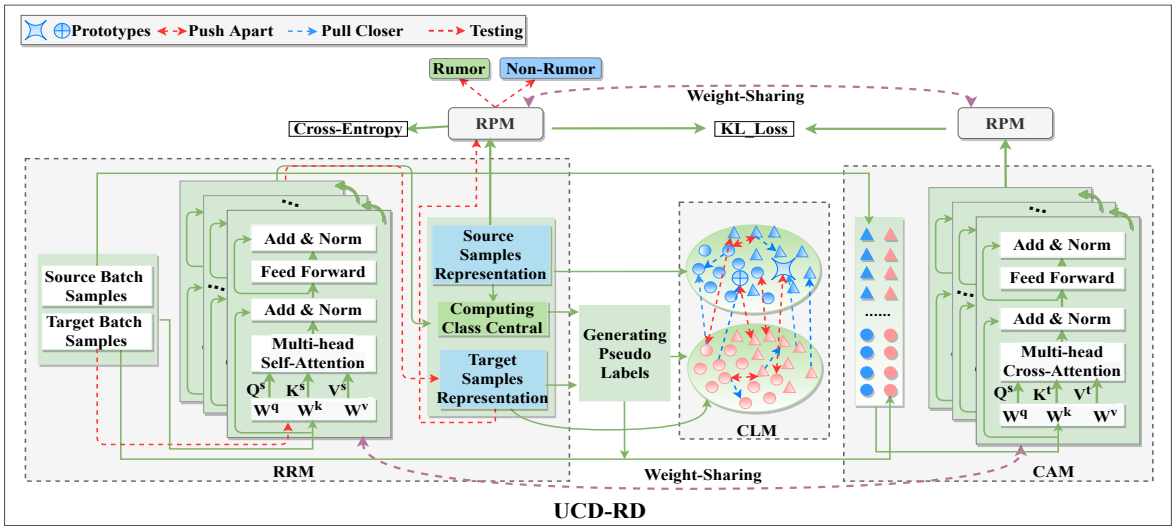


Figure 2: The proposed UCD-RD framework. It consists of four main components including RRM, CLM, CAM and RPM.

cross-domain two aspects to perform this procedure. The core idea is to make the representations of source data and target data from the same class closer while keeping representations from different classes far away. As the labels of the target data D^t are unavailable, we propose a clustering-based approach with initialized centers by batch samples of source data D^s to produce pseudo labels.

In-domain Contrastive Learning. Given a batch of rumor samples from the source domain D^s, we firstly obtain the representation for these samples according to the RRM. To make rumor representation in the source domain more discriminative, we then use an instance-wise contrastive learning objective to cluster the same class and separate different classes of samples, the objective function is computed as:

$$L^s_{SCL} = -\frac{1}{|B^s|}\sum_{i=1}^{|B^s|}\sum_{j=1}^{|B^s|}\mathbf{1}_{y^s_i=y^s_j} log\frac{exp(sim(\hat{O}^s_i,\hat{O}^s_j)/\tau)}{\sum_{k=1}^{|B^s|}exp(sim(\hat{O}^s_i,\hat{O}^s_k)/\tau)} \quad (3)$$

where $|B^s|$ is the size of a source domain batch, $\mathbf{1}$ is an indicator. $sim(\cdot)$ denotes the cosine similarity function and τ controls the temperature.

For the target data, the ground-truth labels are unavailable, so we exploit the cluster-based method to produce the pseudo labels for them (as will be introduced below), and then we also use instance-wise contrastive learning to pull closer to the same label samples and push apart the different label samples. Therefore, we can use a batch of target samples to compute L^t_{SCL}. The overall loss for in-domain is combined the two losses with different proportions α_i:

$$L_{ICL} = \alpha_1 L^s_{SCL} + \alpha_2 L^t_{SCL}, s.t. \sum\nolimits_i \alpha_i = 1 \quad (4)$$

Cross-Domain Contrastive Learning. We now introduce how to form pairs to learn domain invariant features with contrastive learning. Since samples from the source domain and target domain belong to the same set of classes, we build upon this assumption to reduce domain shift. More specifically, we hypothesize that samples within the same category are close to each other while samples from different classes lie far apart, regardless of which domain they come from. More formally, given an anchor sample in the target domain, and it forms a positive pair with a sample in the same class from the source domain, we formulate the instance-wise cross-domain contrastive loss as:

$$L^{t\to s}_{CCL} = -\frac{1}{|B^t|}\sum_{i=1}^{|B^t|}\sum_{j=1}^{|B^s|}\mathbf{1}_{y^t_i=y^s_j} log\frac{exp(sim(\hat{O}^t_i,\hat{O}^s_j)/\tau)}{\sum_{k=1}^{|B^s|}exp(sim(\hat{O}^t_i,\hat{O}^s_k)/\tau)} \quad (5)$$

The cross-domain loss forces intra-class distance to be smaller than inter-class distance for samples from different domains so as to reduce domain shift. Alternatively, we can also use source samples as anchors and compute $L^{s\to t}_{CCL}$ loss.

In order to explicitly enforce learning domain-aligned and more discriminative features in both source and target domains, we perform cross-domain prototype-wise contrastive learning. Our method discovers positive matching as well as negative matchings between instance and cluster prototypes from the target domain to the source domain. Specifically, given a feature vector $\hat{O}^t_i$ in the target domain, and prototypes $\{cen^s_j\}^{N_c}_{j=1}$ which denote sample averages with the same label in the source batch data, we use prototype-wise contrastive learning as follows to train our model.

$$L^{t\to s}_{Pro} = -\frac{1}{|B^t|}\sum_{i=1}^{|B^t|}\sum_{j=1}^{N_c}\mathbf{1}_{y^t_i=y_j} log\frac{exp(sim(\hat{O}^t_i,cen^s_j)/\tau)}{\sum_{k=1}^{N_c}exp(sim(\hat{O}^t_i,cen^s_k)/\tau)} \quad (6)$$

where N_c is the number of classes, $|B^t|$ is the size of a target domain batch.

The cross-domain contrastive loss can be denoted as:

$$L_{CCL} = L_{CCL}^{t \to s} + L_{CCL}^{s \to t} + L_{Pro}^{t \to s} \tag{7}$$

Therefore, the total loss function of the CLM can be represented as follows:

$$L_{CL} = \beta_1 L_{ICL} + \beta_2 L_{CCL}, s.t. \sum\nolimits_i \beta_i = 1 \tag{8}$$

Pseudo Labels for a Target Domain. The ground-truth labels from a target domain D^t are unavailable during training, and thus we leverage k-means clustering (Kang et al. 2019) to produce pseudo labels. Since K-means is sensitive to initialization, the correspondence is unknown between using randomly generated clusters and predefined categories. To mitigate this issue, we set the number of clusters to be the number of classes N_c and use class prototypes of batch samples from the source domain D^s as initial clusters. Formally, we first compute the centroid of each category using the sample average in the source batch and the initial cluster center cen_m^t for the m-th class is defined as:

$$cen_m^t = cen_m^s = \frac{1}{|B_m^s|} \sum_{i=1}^{|B^s|} \mathbf{1}_m \hat{O}_i^s \tag{9}$$

where $|B_m^s|$ is the sample numbers of m-th class in the batch B^s. Given a batch of features from the target domain, we perform K-means clustering using these initialized centers. Once clustering is finished, each sample in the target domain C_i^t is associated with a pseudo label y_i^t.

Cross-Attention Module (CAM)

In the previous section, we utilize self-attention in RRM for samples both in source and target domains for learning the feature representations. Since the cross-attention has been proved robust to the noisy input pairs for better feature alignment (Xu et al. 2021), owing to the noise of the pseudo labels in the target domain, we use the cross-attention mechanism to robust our model. Since each propagation path of rumors with the same labels represents similar semantic patterns, especially on the people's attitudes (e.g., supporting or denying), thus we perform the cross-attention for rumor pairs on cross-domain samples with the same labels learning domain-invariant features. Such design explicitly enforces the framework to learn discriminative domain-specific and domain-invariant representations simultaneously according to self-attention and cross-attention.

The CAM is derived from the self-attention module. The difference is that the input of cross-attention is a pair of rumors with the same labels which come from the cross-domain data, $i.e. X_{P_i}^s$ and $X_{P_j}^t$. Its query and key/value are from $X_{P_i}^s$ and $X_{P_j}^t$ respectively. The cross-attention score can be calculated as follows:

$$f_{att_{cross}}(Q^s, K^t, V^t) = softmax(\frac{Q^s (K^t)^T}{\sqrt{d^k}}) V^t \tag{10}$$

where Q^s are quries from $X_{P_i}^s$, and K^t, V^t are keys and values from $X_{P_j}^t$. For each output, it is calculated by multiplying V^t with attention weights, which comes from the similarity between the corresponding query in $X_{P_i}^s$ and all the keys in $X_{P_j}^t$. As a result, among all paths in $X_{P_j}^t$, the path that is more similar to the query of $X_{P_i}^s$ would hold a larger weight and contribute more to the output. In other words, the output of the CAM manages to aggregate the two input rumors based on their similar paths.

The CAM not only aligns distributions of two domains but is robust to the noise in the input pairs thanks to the cross-attention mechanism. Thus we use the output of the CAM to guide the model's training. We minimize the KL-divergence between predictions, which are computed by the Rumor Prediction Module (RPM) using the outputs of cross-attention and self-attention.

$$p_{cross_i} = \mathbf{RPM}(\hat{O}_{cross_i}^s), p_i = \mathbf{RPM}(\hat{O}_i^s) \tag{11}$$

$$L_{CA}^s = KL^s(p_{cross} || p) = \sum_{i=1}^{|B^s|} p_{cross_i} (log \frac{p_{cross_i}}{p_i}) \tag{12}$$

where $\hat{O}_{cross_i}^s$ is the output of the CAM.

Rumor Prediction Module (RPM)

The RPM consists of a fully collected layer with softmax. Since the labels of the source domain samples are available, we exploit the RPM to acquire their prediction and then use the cross-entropy loss function to train the RPM and RRM.

$$L_{CE}^s = -\frac{1}{|B^s|} \sum_{i=1}^{|B^s|} y_i^s log(p_i) \tag{13}$$

$$p_i = \mathbf{RPM}(\hat{O_i}^s) = softmax(FC(\hat{O_i}^s)) \tag{14}$$

We combine all of the loss functions together to jointly train our model, and can be represented as follows:

$$Loss = \gamma_1 L_{CE}^s + \gamma_2 L_{CL} + \gamma_3 L_{CA}^s, s.t. \sum\nolimits_i \gamma_i = 1 \tag{15}$$

where γ_i is hyperparameters. Once the model is well trained, the unlabeled target data first adopts the RRM to acquire its vector representation, then passes through the RPM to obtain the label distribution according to Equ.14, and finally uses the maximal assignment to the N_c values to get the final label for each rumor.

Experiments

Datasets and Settings

Datasets We evaluate the UCD-RD model on four groups of real-world cross-domain rumor datasets. The first group of data comes from PHEME (Zubiaga et al. 2015) dataset which includes terrorist domain and gossip domain, more details are listed in Supplementary Material[2]. The second group of cross-domain data is Twitter dataset (Ma, Gao, and Wong 2017) and Twitter-Covid19 dataset (Lin et al. 2022). The third group of datasets includes the Twitter15 dataset and the Twitter16 dataset (Ma, Gao, and Wong 2018). The fourth group of cross-domain data is the Chinese Weibo dataset (Ma et al. 2016) and the Weibo-Covid19 dataset (Lin et al. 2022). These cross-domain datasets contain two binary labels: Non-Rumor (N) and Rumor (R). The statistics of the four groups of cross-domain datasets are shown in Table 1.

[2]https://github.com/rhy1111/Supplementary_Material

Statistics	*source* *Terrorist*	*target* *Gossip*	*source* *Twitter15*	*target* *Twitter16*	*source* *Twitter*	*target* *Twitter-COVID19*	*source* *Weibo*	*target* *Weibo-COVID19*
# of claims	5940	485	1490	818	1154	400	4649	399
# of tree nodes	83,860	15,225	41,266	19,376	60,409	406,185	1,956,449	26,687
# of non-rumors	3907	116	374	205	579	148	2336	146
# of rumors	2033	369	1116	613	575	252	2313	253
Avg. # of posts/tree	15	32	28	24	52	1015	420	67

Table 1: Statistics of the cross-domain datasets

Methods	*Terrorist→Gossip*			*Twitter→Twitter_COVID19*			*Twitter15→Twitter16*			*Weibo→Weibo_COVID19*		
	Acc.	N.(F_1)	R.(F_1)	Acc	N.(F_1)	R.(F_1)	Acc.	N.(F_1)	R.(F_1)	Acc.	N.(F_1)	R.(F_1)
LSTM	33.08	34.56	30.31	41.23	33.96	42.58	60.78	46.15	63.25	41.57	40.25	42.79
CNN	32.57	35.62	29.68	40.58	28.47	44.95	61.25	47.32	62.85	42.09	38.21	43.83
Rumor-GAN	32.54	35.48	29.37	41.96	35.68	43.12	63.24	45.27	65.38	43.23	39.61	45.83
TD-RvNN	33.59	32.47	28.16	43.55	40.09	45.78	69.77	47.96	71.28	47.90	43.66	54.78
BU-RvNN	32.68	30.12	21.56	41.33	38.58	42.25	68.47	42.36	69.28	45.18	38.76	50.53
Bi-GCN	34.84	31.04	27.41	51.67	31.72	46.37	75.15	42.23	83.17	61.23	44.08	68.11
PLAN	30.79	40.94	16.40	45.50	**47.60**	43.23	75.04	56.92	82.43	38.44	**46.06**	28.33
PPA-WAE	41.56	38.33	45.79	46.79	43.28	48.46	74.89	60.08	77.69	57.56	35.33	65.26
UCD-CEloss	74.01	66.49	75.97	58.87	33.99	67.78	75.15	60.58	81.08	57.96	41.60	68.36
UCD-RD	**84.88**	**83.04**	**86.35**	**66.50**	45.34	**76.72**	**79.47**	**63.53**	**85.72**	**68.92**	45.13	**78.32**
↑ (%)	**14.69**	**24.92**	**13.66**	**12.96**	-4.75	**13.19**	**5.75**	**4.87**	**3.07**	**12.56**	-2.02	**14.57**

Table 2: Rumor detection results (%) on four groups of cross-domain datasets (N: Non-Rumor; R: Rumor)

Experimental Setup We compare the UCD-RD method with some state-of-the-art baselines including:

- **LSTM** (Ma et al. 2016) is an LSTM-based rumor detection model to learn feature representations of relevant posts over time.
- **CNN** (Yu et al. 2017) uses a CNN model for misinformation identification by framing the relevant posts as a fixed-length sequence.
- **Rumor-GAN** (Ma, Gao, and Wong 2019) uses a generative adversarial network (GAN) in which a generator is designed to produce conflicting voices to pressurize the discriminator to learn stronger rumor representations.
- **TD-RvNN** (Ma, Gao, and Wong 2018) exploits a top-down tree-structured recursive neural network for learning the propagation of rumors.
- **BU-RvNN** (Ma, Gao, and Wong 2018) utilizes a bottom-up tree-structured recursive neural network for learning the propagation of rumors.
- **Bi-GCN** (Bian et al. 2020) is a GCN-based model based on conversation trees to learn rumor representations.
- **PLAN** (Khoo et al. 2020) uses a transformer-based model for rumor detection to capture long-distance interactions between any pair of involved tweets.
- **PPA-WAE** (Zhang et al. 2021b) utilizes a neural topic model which is combined with a feed-forward network on propagation trees to learn the semantics of the trees and their propagation patterns.
- **UCU-CEloss** is a variant of the UCD-RD model which only uses the L_{CE} loss to train the model.

We implement LSTM and CNN models with Keras[3], other baseline methods and our model with Pytorch[4]. For these cross-domain datasets, we evaluate the Accuracy (Acc.), F1 measure (F_1) on each class. The dimension of each rumor hidden feature vector is 300. The training process is iterated upon 300 epochs. The temperature τ is 0.1.

Overall Performance

Table 2 shows the performance of the UCD-RD method and all the compared methods on the four groups of cross-domain datasets. From Table 2, the first group of experiments is based on the time series rumor detection methods, and the second group of results is based on the propagation structure rumor detection approaches. We can observe that the results of the first group of experiments are worse than those of the second group of experiments, which proves the propagation structure-based methods also excel the time series-based methods on the cross-domain datasets.

[3]https://keras.io/

[4]https://pytorch.org/

	L^s_{CE}	$+L^s_{SCL}$	$+L^t_{SCL}$	$+L^{t\to s}_{CDC}$	$+L^{s\to t}_{CDC}$	$+L^{t\to s}_{Pro}$	$+L_{CA}$
Terrorist→Gossip	74.01	74.58	75.65	77.90	79.01	79.58	**84.88**
Twitter→Twitter_Covid19	58.87	58.95	59.78	61.55	62.25	63.77	**66.50**
Twitter15→Twitter16	75.15	75.17	76.32	77.23	77.89	78.23	**79.47**
Weibo→Weibo_Covid19	57.96	58.16	60.03	63.16	65.23	65.79	**68.92**

Table 3: Ablation study results (%) on four groups of cross-domain datasets

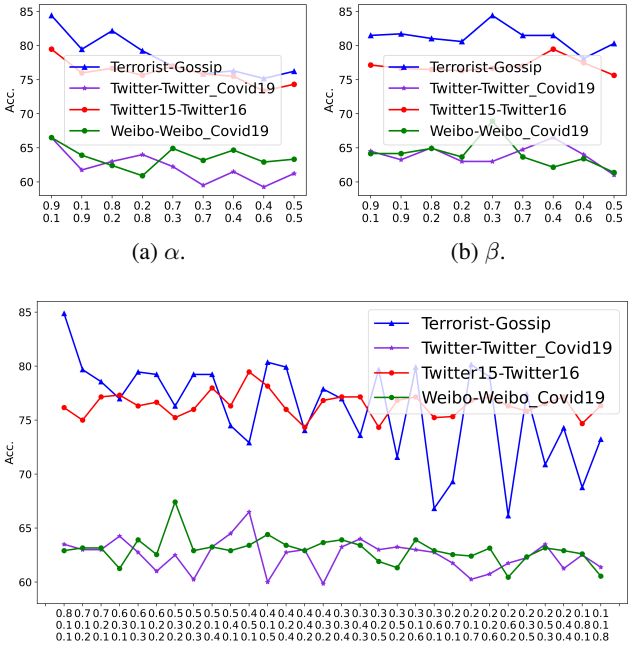
Among the propagation structure-based baselines in the second group, since these methods mainly focus on learning the rumor representation for in-domain rumor detection, the performance is worse on the cross-domain datasets due to the domain shift. Different from that, we propose the UDA-RD method to alleviate the domain shift by aligning the feature representation between a pair of source and target domains. According to Table 2, we found that UCD-CEloss gets better performance compared with these baseline models, which proves that using the self-attention mechanism to model the path embeddings of rumors enables to learn the discriminative rumor patterns. On the basis of UCD-CEloss, we add the instance-wise and prototype-wise contrastive learning and cross-attention mechanism, and the performance has significant improvement, especially on the *Terrorist→Gossip* cross-domain dataset, which achieves a performance improvement of 14.69% on accuracy. It demonstrates that our method can alleviate the problem of domain shift and gets state-of-the-art performance.

Ablation Study

We investigate the effectiveness of each loss function in UCD-RD on four groups of cross-domain datasets. Table 3 shows that adding each component contributes to the final results without any performance degradation. From Table 3, we can find that the cross-domain contrastive learning and cross-attention module play important roles in our model. The cross-domain contrastive learning can decrease the discrepancy between a source domain and a target domain, due to the instance-wise and prototype-wise contrastive learning making the representations of the source data and the target data from the same class be closer while keeping representations from different classes far away. Meanwhile, the cross-attention module makes our model more robust and alleviates the noise impact of pseudo labels in the target domain.

Effects of Hyper-parameters

We test the sensitivity of UCD-RD to the α, β, and γ on the four groups of cross-domain datasets. As shown in Figure 3, since the hyper-parameters of α, β, γ are limited to $\sum_i \alpha_i = 1$, $\sum_i \beta_i = 1$ and $\sum_i \gamma_i = 1$, α_1, α_2, β_1, β_2 and $\gamma_1, \gamma_2, \gamma_3$ are appeared in groups in our experiments respectively. Taking Figure 3a as an example, each group of values in the x-axis denotes a group of α_1, α_2 from top to bottom with fixed β, γ at their optimal values. Figure 3b and Figure 3c are set similarly. From Figure 3, we observe that UCD-RD has different sensitivity in different datasets. For instance, for the *Terrorist→Gossip* data, when these hyper parameters $\alpha_1 = 0.9$, $\alpha_2 = 0.1$, $\beta_1 = 0.7$, $\beta_2 = 0.3$, and $\gamma_1 = 0.8$, $\gamma_2 = 0.1$, $\gamma_3 = 0.1$, UCD-RD achieves the best performance.Whereas for the other datasets, the optimal hyper-parameters are at different settings.



(a) α. (b) β.

(c) γ.

Figure 3: Performance sensitivity of hyper-parameters α, β, γ in UCD-RD.

Conclusion

We propose a contrastive learning and cross-attention model for cross-domain rumor detection. We build instance-wise and prototype-wise contrastive learning to align features so as to reduce the domain discrepancy between source data and target data such that the representations of the source data and the target data from the same class are close to each other while those from different classes are far away. Moreover, we use a cross-attention mechanism on a pair of source data and target data with the same labels to learn the domain-invariant features and alleviate the noise impact of pseudo labels. Experiments on four groups of public datasets show that UCD-RD achieves the best performance.

Acknowledgments

The authors would like to thank all the anonymous reviewers for their help and insightful comments. This work is supported in part by the National Natural Science Foundation of China (61876016), the National Key R&D Program of China (2018AAA0100302) and Baidu Pinecone Program.

References

Ben-David, S.; Blitzer, J.; Crammer, K.; and Kulesza, A. 2010. A theory of learning from different domains. *Machine Learning*, 79(1): 151–175.

Bian, T.; Xiao, X.; Xu, T.; and Zhao, P. 2020. Rumor detection on social media with bi-directional graph convolutional networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 549–556.

Castillo, C.; Mendoza, M.; and Poblete, B. 2011. Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web*, 675–684.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, 1597–1607. PMLR.

Coleman, A. 2020. Hundreds dead" because of Covid-19 misinformation. *BBC News*, 12.

Du, C.; Sun, H.; Wang, J.; Qi, Q.; and Liao, J. 2020. Adversarial and domain-aware bert for cross-domain sentiment analysis. In *Proceedings of the 58th annual meeting of the Association for Computational Linguistics*, 4019–4028.

Du, Z.; Li, J.; Su, H.; and Zhu, L. 2021. Cross-domain gradient discrepancy minimization for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3937–3946.

Ganin, Y.; and Lempitsky, V. 2015. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, 1180–1189. PMLR.

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

Ghosal, D.; Hazarika, D.; Roy, A.; Majumder, N.; Mihalcea, R.; and Poria, S. 2020. Kingdom: Knowledge-guided domain adaptation for sentiment analysis. In *ACL*.

He, K.; Fan, H.; Wu, Y.; and Xie, S. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9729–9738.

Huang, Q.; Yu, J.; Wu, J.; and Wang, B. 2020. Heterogeneous graph attention networks for early detection of rumors on twitter. In *2020 International Joint Conference on Neural Networks (IJCNN)*, 1–8. IEEE.

Jin, Z.; Cao, J.; Guo, H.; and Zhang, Y. 2017. Detection and analysis of 2016 us presidential election related rumors on twitter. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, 14–24. Springer.

Kang, G.; Jiang, L.; Yang, Y.; and Hauptmann, A. G. 2019. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4893–4902.

Khoo, L. M. S.; Chieu, H. L.; Qian, Z.; and Jiang, J. 2020. Interpretable rumor detection in microblogs by attending to user interactions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 8783–8790.

Kwon, S.; Cha, M.; Jung, K.; and Chen, W. 2013. Prominent features of rumor propagation in online social media. In *2013 IEEE 13th International Conference on Data Mining*, 1103–1108. IEEE.

Li, C.; Peng, H.; Li, J.; and Sun, L. 2021. Joint Stance and Rumor Detection in Hierarchical Heterogeneous Graph. *IEEE Transactions on Neural Networks and Learning Systems*.

Li, T.; Chen, X.; Dong, Z.; Yu, W.; Yan, Y.; Keutzer, K.; and Zhang, S. 2022. Domain-Adaptive Text Classification with Structured Knowledge from Unlabeled Data. In *Proceedings of the International Joint Conference on Artificial Intelligence*.

Lin, H.; Ma, J.; Chen, L.; and Yang, Z. 2022. Detect Rumors in Microblog Posts for Low-Resource Domains via Adversarial Contrastive Learning. *arXiv preprint arXiv:2204.08143*.

Liu, B.; Sun, X.; Meng, Q.; and Yang, X. 2022. Nowhere to Hide: Online Rumor Detection Based on Retweeting Graph Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*.

Long, M.; Cao, Z.; Wang, J.; and Jordan, M. I. 2018. Conditional adversarial domain adaptation. *Advances in Neural Information Processing Systems*, 31.

Ma, J.; Gao, W.; Mitra, P.; and Kwon, S. 2016. Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 3818–3824.

Ma, J.; Gao, W.; Wei, Z.; and Lu, Y. 2015. Detect rumors using time series of social context information on microblogging websites. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, 1751–1754.

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

Ma, J.; Gao, W.; and Wong, K.-F. 2018. Rumor detection on twitter with tree-structured recursive neural networks. Association for Computational Linguistics.

Ma, J.; Gao, W.; and Wong, K.-F. 2019. Detect rumors on twitter by promoting information campaigns with generative adversarial learning. In *The World Wide Web Conference*, 3049–3055.

Ma, J.; Li, J.; Gao, W.; Yang, Y.; and Wong, K.-F. 2021. Improving Rumor Detection by Promoting Information Campaigns with Transformer-based Generative Adversarial Learning. *IEEE Transactions on Knowledge and Data Engineering*.

Min, E.; Rong, Y.; Bian, Y.; and Xu, T. 2022. Divide-and-Conquer: Post-User Interaction Network for Fake News Detection on Social Media. In *Proceedings of the ACM Web Conference 2022*, 1148–1158.

Mohtarami, M.; Glass, J.; and Nakov, P. 2019. Contrastive language adaptation for cross-lingual stance detection. In *Proceedings of the 9th International Joint Conference on Natural Language Processing*, 4442–4452.

Mosallanezhad, A.; Karami, M.; Shu, K.; and Mancenido, M. V. 2022. Domain Adaptive Fake News Detection via Reinforcement Learning. In *Proceedings of the ACM Web Conference 2022*, 3632–3640.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Qiu, J.; Chen, Q.; Dong, Y.; and Zhang, J. 2020. Gcc: Graph contrastive coding for graph neural network pre-training. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1150–1160.

Ran, H.; Jia, C.; Zhang, P.; and Li, X. 2022. MGAT-ESM: Multi-channel graph attention neural network with event-sharing module for rumor detection. *Information Sciences*, 592: 402–416.

Saito, K.; Watanabe, K.; Ushiku, Y.; and Harada, T. 2018. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3723–3732.

Sun, T.; Qian, Z.; Dong, S.; and Li, P. 2022. Rumor Detection on Social Media with Graph Adversarial Contrastive Learning. In *Proceedings of the ACM Web Conference 2022*, 2789–2797.

Vaswani, A.; Shazeer, N.; Parmar, N.; and Uszkoreit, J. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Xu, T.; Chen, W.; Wang, P.; and Wang, F. 2021. Cdtrans: Cross-domain transformer for unsupervised domain adaptation. *arXiv preprint arXiv:2109.06165*.

You, Y.; Chen, T.; Sui, Y.; Chen, T.; Wang, Z.; and Shen, Y. 2020. Graph contrastive learning with augmentations. *Advances in Neural Information Processing Systems*, 33: 5812–5823.

Yu, F.; Liu, Q.; Wu, S.; Wang, L.; Tan, T.; et al. 2017. A Convolutional Approach for Misinformation Identification. In *IJCAI*, 3901–3907.

Zhang, D.; Nan, F.; Wei, X.; and Li, S. 2021a. Supporting clustering with contrastive learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*, 5419–5430.

Zhang, P.; Ran, H.; Jia, C.; Li, X.; and Han, X. 2021b. A lightweight propagation path aggregating network with neural topic model for rumor detection. *Neurocomputing*, 458: 468–477.

Zhu, Y.; Xu, Y.; Yu, F.; Liu, Q.; Wu, S.; and Wang, L. 2021. Graph contrastive learning with adaptive augmentation. In *Proceedings of the Web Conference 2021*, 2069–2080.

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

Zubiaga, A.; Liakata, M.; Procter, R.; Bontcheva, K.; and Tolmie, P. 2015. Crowdsourcing the annotation of rumourous conversations in social media. In *Proceedings of the 24th International Conference on World Wide Web*, 347–353.