Detect Low-Resource Rumors in Microblog Posts via Adversarial Contrastive Learning

Anonymous ACL submission

Abstract

Massive false rumors emerging along with breaking news or trending topics severely hinder the truth. Exiting rumor detection approaches achieve promising performance on the yesterday's news, since there is enough corpus collected from the same domain for model training. However, they are poor at detecting rumors about unforeseen events such as COVID-19 due to the lack of training data and prior knowledge (i.e., low-resource rumors). In this paper, we propose an adversarial contrastive learning framework to detect low-resource rumors by adapting the features learned from well-resourced rumor data to that of the low-resourced. Our model explicitly overcomes the restriction of both domain and language usage via language alignment and contrastive training. Moreover, we develop an adversarial augmentation mechanism to further enhance the robustness of low-resource rumor representation. Extensive experiments conducted on two low-resource datasets collected from real-world microblog platforms demonstrate that our framework achieves much better performance than state-of-the-art methods and exhibits a superior capacity for detecting rumors at early stages.

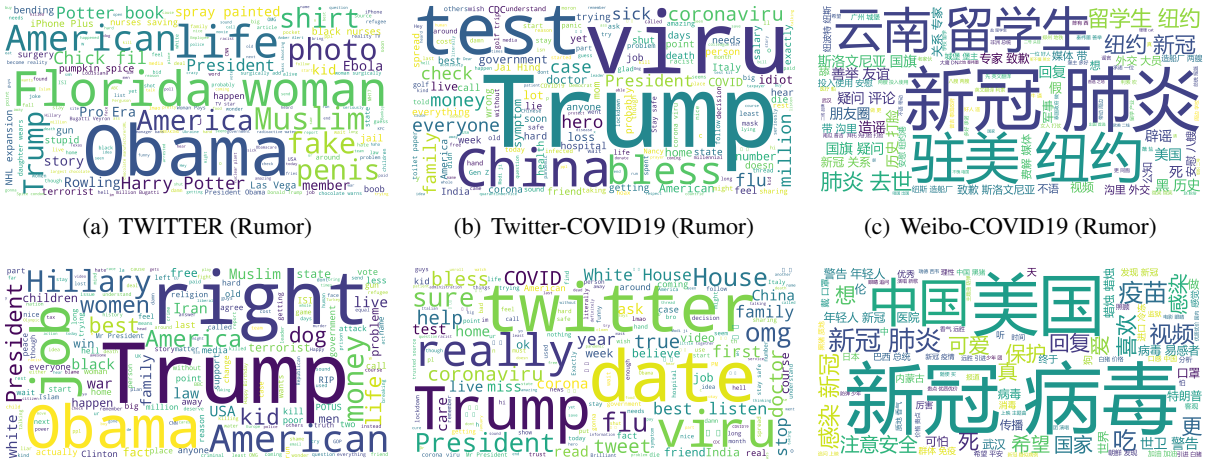
1 Introduction

With the proliferation of social media such as Twitter and Weibo, the emergency of breaking events is richly endowed by nature for the breeding and spreading of rumors, which is difficult to be identified due to limited domain expertise and relevant data. For instance, along with the unprecedented COVID-19 emergency, a false rumor claims that "everyone who gets the vaccine will die or suffer from auto-immune diseases."[1]. Such rumor was translated into many languages and spread at lightning speed on social media, which seriously confuse the public and destroy the achievements of epidemic prevention in related countries or regions of the world. Although some recent work focuses on collecting social media posts corresponding to COVID-19 (Chen et al., 2020a; Zarei et al., 2020; Alqurashi et al., 2020), existing rumor detection methods perform poorly without a large-scale qualified training corpus. Thus it is urgent to develop automatic approaches to identify such low-resource rumors especially amid breaking events.

Social psychology literature defines a rumor as a story or a statement whose truth value is unverified or deliberately false (Allport and Postman, 1947). Recently, techniques using deep neural networks (DNNs) (Ma et al., 2018; Khoo et al., 2020; Bian et al., 2020) have achieved promising results for detecting rumors on microblogging websites by learning rumor-indicative features from sizeable rumor corpus with veracity annotation. However, existing DNN-based models are purely data-driven and demonstrate state-of-the-art performance when the domains and languages used of the detected rumors are covered by the training data. On another hand, rumors emerging along with breaking news are low-resourced which may concern unprecedented domains and/or be presented in different languages. Previous studies have shown that cross-domain datasets have distinctive topic coverage and word distribution (Silva et al., 2021). Therefore, existing rumor detection models that are well-trained on public benchmarks (Ma et al., 2016; Zubiaga et al., 2016; Ma et al., 2017) generally struggle with emerging events in low-resource regimes (Janicka et al., 2019).

In this paper, we assume that the close correlations between the well-resourced rumor and the low-resourced rumor could break the barriers of domain and language, substantially boosting low-resource rumor detection. Taking the breaking event COVID-19 as an example, we collect rumor and non-rumor claims corresponding to COVID-19 from Twitter and Sina Weibo which are the

[1] https://www.bbc.com/news/uk-wales-58103604



(a) TWITTER (Rumor) (b) Twitter-COVID19 (Rumor) (c) Weibo-COVID19 (Rumor)

(d) TWITTER (Non-rumor) (e) Twitter-COVID19 (Non-rumor) (f) Weibo-COVID19 (Non-rumor)

Figure 1: Word clouds of rumor and non-rumor claims generated from TWITTER, Twitter-COVID19, and Weibo-COVID19 datasets, where the size of terms corresponds to the word frequency. Both TWITTER and Twitter-COVID19 are presented in English while Weibo-COVID19 in Chinese.

most popular microblogging websites in English and Chinese, respectively. Figure 1 illustrates the word clouds of rumor and non-rumor claims from an open domain benchmark (i.e., TWITTER (Ma et al., 2017)) and two COVID-19 datasets (i.e., Twitter-COVID19 and Weibo-COVID19). It can be seen that both TWITTER and Twitter-COVID19 contain denial opinions towards rumors, e.g., “fake", “joke", “stupid" in Figure 1(a) and “wrong symptom", “exactly sick", “health panic" in Figure 1(b). In contrast, supportive opinions towards non-rumors can be drawn from Figure 1(d)–1(e). Moreover, considering that COVID-19 is a global disease, massive misinformation could be widely propagated under different languages such as Arabic (Alam et al., 2020), Indic (Kar et al., 2020), English (Cui and Lee, 2020) and Chinese (Hu et al., 2020). Similar identical patterns can be observed in Chinese on Weibo from Figure 1(c) and Figure 1(f). Although the COVID-19 data tend to use expertise words or language-related slang, we argue that aligning the representation space of identical patterns of different domains and/or languages could adapt the features captured from well-resourced rumor data to that of the low-resourced data.

To this end, inspired by contrastive learning (He et al., 2020; Chen et al., 2020b,c), we propose an Adversarial Contrastive Learning approach for low-resource rumor detection (ACLR), to encourage effective alignment of rumor-indicative features in the well-resourced and low-resourced data. More specifically, we first transform each microblog post into a language-independent vector by semantically aligning the source and target language in a shared vector space. The diffusion of rumors generally follows a propagation tree that provides valuable clues on how a claim is transmitted. We thus resort to a structure-based neural network (Ma et al., 2018; Bian et al., 2020) for representation learning. We then propose a supervised contrastive learning framework to minimize the intra-class variance of source and target instances with same veracity, and maximize inter-class variance of instances with different veracity. To further enhance the adaption of feature learning, we exploit adversarial attacks (Kurakin et al., 2016) to plenish noise to the original event, forcing the model to learn non-trivial but effective features. Extensive experiments conducted on two real-word low-resource datasets confirm that (1) our model yields outstanding performances for detecting rumors of low-resourced domains and/or languages over the state-of-the-art baselines with a large margin; and (2) our method performs particularly well on early rumor detection which is crucial for timely intervention and debunking especially for breaking events. The main contributions of this paper are of three-fold:

- To our best knowledge, we are the first to present a radically novel adversarial contrastive learning method to study the low-resource rumor detection on social media.
- We propose a supervised contrastive learning framework for feature adaption between different domains and languages. We further employ an adversarial augmentation mechanism to enhance the model generation.

- We constructed two low-resourced microblog datasets corresponding to COVID-19 with propagation tree structure, respectively gathered from English tweets and Chinese microblog posts. Experimental results show that our model achieves superior performance for both rumor classification and early detection tasks under low-resource settings. We will make our resources publicly available.

2 Related Work

Pioneer studies for automatic rumor detection focus on learning a supervised classifier utilizing features crafted from post contents, user profiles, and propagation patterns (Castillo et al., 2011; Yang et al., 2012; Liu et al., 2015). Subsequent studies then propose new features such as those representing rumor diffusion and cascades (Kwon et al., 2013; Friggeri et al., 2014; Hannak et al., 2014). Zhao et al. (2015) alleviate the engineering effort by using a set of regular expressions to find questing and denying tweets. DNN-based models such as recurrent neural networks (Ma et al., 2016), convolutional neural networks (Yu et al., 2017), and attention mechanism (Guo et al., 2018) are then employed to learn the features from the stream of social media posts. However, these approaches simply model the post structure as a sequence while ignoring the complex propagation structure.

To extract useful clues jointly from content semantics and propagation structures, some approaches propose kernel-learning models (Wu et al., 2015; Ma et al., 2017) to make a comparison between propagation trees. Tree-structured recursive neural networks (RvNN) (Ma et al., 2018) and transformer-based models (Khoo et al., 2020) are proposed to generate the representation of each post along a propagation tree guided by the tree structure. More recently, graph neural networks (Bian et al., 2020) have been exploited to encode the conversation thread for higher-level representations. Despite the apparent success of structure-based models, they fail in the low-resource rumor detection task. In this paper, we propose a novel framework considering the effective structural features to adapt existing models for detecting rumors from different domains and/or languages.

To facilitate low-resource rumor detection or few-shot fact-checking tasks, domain adaption techniques are utilized to detect fake news (Wang et al., 2018; Yuan et al., 2021; Zhang et al., 2020; Silva et al., 2021) by learning features from multi-modal data such as texts and images. Lee et al. (2021) proposed a simple way of leveraging the perplexity score obtained from pre-trained language models (LMs) for the few-shot fact-checking task. Different from existing works of adaption on multi-modal data and transfer learning of LMs, we focus on language and domain adaptation to detect low-resourced rumors on social media corresponding to breaking events.

Contrastive learning aims to enhance representation learning by maximizing the agreement among the same types of instances and distinguishing from the others with different types (Wang and Isola, 2020). In recent years, contrastive learning has achieved great success in unsupervised visual representation learning (Chen et al., 2020b; He et al., 2020; Chen et al., 2020c). Besides computer vision, recent studies suggest that contrastive learning is promising in the semantic textual similarity (Gao et al., 2021; Yan et al., 2021), stance detection (Mohtarami et al., 2019), abstractive summarization (Liu and Liu, 2021), out-of-domain detection (Tan et al., 2019; Lin et al., 2021) and short text clustering (Zhang et al., 2021), etc. Inspired by their works, we propose a supervised contrastive learning framework to model adaptive features of the conversation structure for low-resource rumor detection.

3 Problem Statement

In this work, we define the low-resource rumor detection task as: Given a well-resourced dataset as source, classify each event in the target low-resourced dataset as a rumor or not, where the source and target data are from different domains and languages. Specifically, we define a well-resourced source dataset for training as a set of events $\mathcal{D}_s = \{C_1^s, C_2^s, \cdots, C_M^s\}$, where M is the number of source events. Each event $C^s = (y, c, \mathcal{T}(c))$ is a tuple representing a given claim c which is associated with a veracity label $y \in \{\text{rumor}, \text{non-rumor}\}$, and ideally all its relevant responsive microblog post in chronological order, i.e., $\mathcal{T}(c) = \{c, x_1^s, x_2^s, \cdots, x_{|C|}^s\}$[2], where $|C|$ is the number of responsive tweets in the conversation thread. For the target dataset with low-resourced domains and/or languages, we consider a much smaller dataset for training $\mathcal{D}_t = \{C_1^t, C_2^t, \cdots, C_N^t\}$, where $N (N \ll M)$

[2] c is equivalent to x_0^s.

is the number of target events and each $C^t = (y, c', \mathcal{T}(c'))$ has the similar composition structure of the source dataset.

We formulate the task of low-resource rumor detection as a supervised classification problem that trains a domain/language-agnostic classifier $f(\cdot)$ adapting the features learned from source datasets to that of the target events, that is, $f(C_j^t|\mathcal{D}_s) \rightarrow \hat{y}_j^t$. Note that although the tweets are notated sequentially, there are connections among them based on their responsive relationships. So most previous works represent the conversation thread as a directed tree structure (Ma et al., 2017, 2018; Bian et al., 2020).

4 Our Approach

In this section, we introduce our adversarial contrastive learning framework to adapt the features captured from the well-resourced data to detect low-resourced rumors, which considers cross-lingual and cross-domain alignment. Figure 2 illustrates an overview of our proposed model, which will be depicted in the following subsections.

4.1 Cross-lingual Sentence Encoder

Given a post in an event that could be either from source or target data, to map it into a shared semantic space where the source and target languages are semantically aligned, we utilize XLM-RoBERTa (Conneau et al., 2019) (XLM-R) to model the context interactions among tokens in the sequence for the sentence-level representation:

$$\bar{x} = \textit{XLM-R}(\mathbf{x}) \tag{1}$$

where $\mathbf{x}$ is the original post, and we obtain the post-level representation $\bar{x}$ using the output state of the <*s*> token in XLM-RoBERTa. We thus denote the representation of posts in the source event C^s and the target event C^t as a matrix X^s and X^t respectively: $X^* = [\bar{x}_0^*, \bar{x}_1^*, \bar{x}_2^*, ..., \bar{x}_{|X^*|-1}^*]^\top; * \in \{s, t\}$, where $X^s \in \mathbb{R}^{m \times d}$ and $X^t \in \mathbb{R}^{n \times d}$, d is the dimension of the output state of the XLM-RoBERTa encoder.

4.2 Propagation Structure Representation

On top of the sentence encoder, we represent the propagation of each claim with the graph convolutional network (GCN) (Kipf and Welling, 2016), which achieves state-of-the-art performance on capturing both structural and semantic information for rumor classification (Bian et al., 2020). It is worth noting that the choice of propagation structure representation is orthogonal to our proposed framework that can be easily replaced with any existing structure-based models without any other change to our contrastive learning architecture.

Given an event and its initialized embedding matrix $C^*, X^*; * \in \{s, t\}$, We model the conversation thread of the event as a tree structure $\mathcal{T} = \langle V, E \rangle$, where V consists of the event claim and all its relevant responsive posts as nodes and E refers to a set of directed edges corresponding to the response relation among the nodes in V. Inspired by Ma et al. (2018), here we consider two different propagation trees with distinct edge directions: (1) *Top-Down tree* where the edge follows the direction of information diffusion. (2) *Bottom-Up tree* where the responsive nodes point to their responded nodes, similar to a citation network.

Top-Down GCN. We treat the Top-Down tree structure as a graph and transform the edge E into an adjacency matrix $\mathbf{A} \in \{0, 1\}^{|V| \times |V|}$, where $\mathbf{A}_{i,j} = 1$ if $\mathbf{x}_i$ has a response to $\mathbf{x}_j$ or $i = j$, else $\mathbf{A}_{i,j} = 0$. Then we utilize a layer-wise propagation rule to update the node vector at the l-th layer:

$$H^{(l+1)} = ReLU(\hat{\mathbf{A}} \cdot H^{(l)} \cdot W^{(l)}) \tag{2}$$

where $\hat{\mathbf{A}} = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$ is the symmetric normalized adjacency matrix, $\mathbf{D}$ denotes the degree matrix of $\mathbf{A}$. $W^{(l)} \in \mathbb{R}^{d^{(l)} \times d^{(l+1)}}$ is a layer-specific trainable transformation matrix. $H^{(0)}$ is initialized as X^*. For a GCN model with L-layers, we obtain the final node representation H_{TD} w.r.t $H^{(L)}$.

Bottom-Up GCN. We also leverage the structure of Bottom-Up tree to encode the informative posts. Similar to Top-Down GCN, we update the hidden representation of nodes in the same manner as Eq. 2 and finally get the output node states H_{BU} at the L-th graph convolutional layer.

The Overall Model. Finally, we concatenate H_{TD} and H_{BU} via mean-pooling to jointly capture the opinions expressed in both Top-Down and Bottom-Up trees:

$$o = \textit{mean-pooling}([H_{TD}; H_{BU}]) \tag{3}$$

where $o \in \mathbb{R}^{2d^{(L)}}$ is the event-level representation of the entire propagation thread, $d^{(L)}$ is the output dimension of GCN and $[\cdot;\cdot]$ means concatenation.

4.3 Contrastive Training

To align the representation space of rumor-indicative signals from different domains and languages, we present a novel training paradigm to

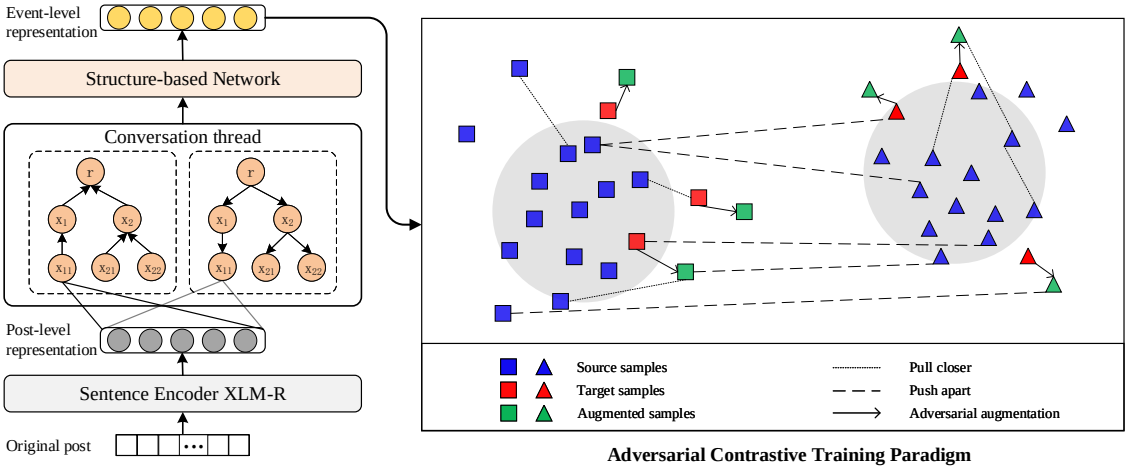


Figure 2: The overall architecture of our proposed method. For source and small target training data, we first obtain post-level representations after cross-lingual sentence encoding, then train the structure-based network with the adversarial contrastive objective. For target test data, we extract the event-level representations to detect rumors.

exploit the labeled data including rich sourced data and small-scaled target data to adapt our model on target domains and languages. The core idea is to make the representations of source and target events from the same class closer while keeping representations from different classes far away.

Given an event C_i^s from the source data, we firstly obtain the language-agnostic encoding for all the involved posts (see Eq. 1) as well as the propagation structure representation o_i^s (see Eq. 3) which is then fed into a *softmax* function to make rumor predictions. Then, we learn to minimize the cross-entropy loss between the prediction and the ground-truth label y_i^s:

$$\mathcal{L}_{CE}^s = -\frac{1}{N^s}\sum_{i=1}^{N^s} log(p_i) \quad (4)$$

where N^s is the total number of source examples in the batch, p_i is the probability of correct prediction. To make rumor representation in the source events be more dicriminative, we propose a supervised contrastive learning objective to cluster the same class and separate different classes of samples:

$$\mathcal{L}_{SCL}^s = -\frac{1}{N^s}\sum_{i=1}^{N^s}\frac{1}{N_{y_i^s}-1}\sum_{j=1}^{N^s}\mathbb{1}_{[i\neq j]}\mathbb{1}_{[y_i^s=y_j^s]} log\frac{exp(\text{sim}(o_i^s,o_j^s)/\tau)}{\sum_{k=1}^{N^s}\mathbb{1}_{[i\neq k]}exp(\text{sim}(o_i^s,o_k^s)/\tau)} \quad (5)$$

where $N_{y_i^s}$ is the number of source examples with the same label y_i^s in the event C_i^s, and $\mathbb{1}$ is the indicator. $\text{sim}(\cdot)$ denotes the cosine similarity function and τ controls the temperature.

For an event C_i^t from the target data, we also compute the classification loss $\mathcal{L}_{CE}^t$ in the same manner as Eq. 4. Although we projected the source and target languages into the same semantic space after sentence encoding, rumor detection not only relies on post-level features, but also on event-level contextual features. Without constraints, the structure-based network can only extract event-level features for all samples based on their final classification signals while these features may not be critical to the target domain and language. We make full use of the minor labels in the low-resource rumor data by parameterizing our model according to the contrastive objective between the source and target instances in the event-level representation space:

$$\mathcal{L}_{SCL}^t = -\frac{1}{N^t}\sum_{i=1}^{N^t}\frac{1}{N_{y_i^t}}\sum_{j=1}^{N^s}\mathbb{1}_{[y_i^t=y_j^s]} log\frac{exp(\text{sim}(o_i^t,o_j^s)/\tau)}{\sum_{k=1}^{N^s}exp(\text{sim}(o_i^t,o_k^s)/\tau)} \quad (6)$$

where N^t is the total number of target examples in the batch and $N_{y_i^t}$ is the number of source examples with the same label y_i^t in the event C_i^t. As a result, we project the source and target samples belonging to the same class closer than that of different categories, for feature alignment with minor annotation at the target domain and language.

4.4 Adversarial Data Augmentation

Data augmentation has been previously shown improvements for contrastive learning models (Chen et al., 2020b). However, there is no simple and

Algorithm 1 Adversarial Contrastive Learning

Input: A small set of events C_i^t in the target domain and language; A set of events C_i^s in the source domain and language.
Output: Assign rumor labels y to given unlabeled target data.

for each mini-batch N^t of the target events C_i^t **do**:
  for each mini-batch N^s of the source events C_i^s **do**:
    Pass C_i^* to the sentence encoder and then structure-based network to obtain its event-level feature o_i^*, where $* \in \{s, t\}$.
    Compute the classification loss $\mathcal{L}_{CE}^*$ for source and target data, respectively.
    Adversarial augmentation for target data and update $\mathcal{L}_{CE}^t$.
    Compute the supervised contrastive loss $\mathcal{L}_{SCL}^*$.
    Compute the joint loss $\mathcal{L}^*$ as Eq. 8.
    Jointly optimize all parameters of the model using the average loss $\mathcal{L} = \text{mean}(\mathcal{L}^s + \mathcal{L}^t)$.

effective augmentation strategy for event-level features in rumor detection and related research fields, which requires massive handcrafted features or rules. In this section, we introduce adversarial attacks to generate pseudo target samples at the event-level latent space to increase the diversity of views for model robustness in the contrastive learning manner. Specifically, we apply Fast Gradient Value (FGV) (Miyato et al., 2016; Vedula et al., 2020) to approximate a worst-case perturbation as a noise vector:

$$\tilde{\boldsymbol{o}}_{noise}^t = \epsilon \frac{g}{||g||}; \text{where } g = \nabla_{o^t} \mathcal{L}_{CE}^t \quad (7)$$

where the gradient is the first-order differential of the classification loss $\mathcal{L}_{CE}^t$ for a target sample, i.e., the direction that rapidly increases the classification loss. We perform normalization and use a small ϵ to ensure the approximate is reasonable. Finally, we can obtain the pseudo augmented sample $o_{adv}^t = o^t + \tilde{\boldsymbol{o}}_{noise}^t$ in the latent space to enhance our model.

4.5 Model Training

We jointly train the model with the cross-entropy and supervised contrastive objectives:

$$\mathcal{L}^* = (1-\alpha)\mathcal{L}_{CE}^* + \alpha \mathcal{L}_{SCL}^*; * \in \{s, t\} \quad (8)$$

where α is a trade-off parameter, which is set to 0.5 in our experiments. Algorithm 1 presents the training process of our approach. We set the number L of the graph convolutional layer as 2, the temperature τ as 0.1, and the adversarial perturbation norm ϵ as 1.5. Parameters are updated through back-propagation (Collobert et al., 2011) with the Adam optimizer (Loshchilov and Hutter, 2018). The learning rate is initialized as 0.0001, and the dropout rate is 0.2. Early stopping (Yao et al., 2007) is applied to avoid overfitting.

5 Experiments

5.1 Datasets

To our knowledge, there are no public benchmarks available for detecting low-resource rumors with propagation tree structure. In this paper, we consider breaking events about COVID-19 and collect relevant rumors and non-rumors respectively from Twitter in English and Sina Weibo in Chinese. For Twitter-COVID19, we resort to a raw COVID-19 rumor dataset (Kar et al., 2020) which only contains the textural claim without its propagation thread. We then collected all the propagation threads using the Twitter academic API with the twarc2 package[3]. For Weibo-COVID19, we gather a set of rumorous claims from the Sina community management center[4] and non-rumorous claims by randomly filtering out the posts that are not reported as rumors. We then utilize Weibo API to collect all the repost/reply messages towards each claim. The full statistics of the resulting datasets are shown in Appendix.

5.2 Experimental Setup

We compare our model and several state-of-the-art baseline methods described below. 1) **CNN**: A CNN-based model for misinformation identification (Yu et al., 2017) by framing the relevant posts as a fixed-length sequence; 2) **RNN**: A RNN-based rumor detection model (Ma et al., 2016) with GRU for feature learning of relevant posts over time; 3) **RvNN**: A rumor detection approach based on tree-structured recursive neural networks (Ma et al., 2018) that learn rumor representations guided by the propagation structure; 4) **PLAN**: A transformer-based model (Khoo et al., 2020) for rumor detection to capture long-distance interactions between any pair of involved tweets; 5) **BiGCN**: A GCN-based model (Bian et al., 2020) based on directed conversation trees to learn higher-level representations (see Section 4.2); 6) **DANN-***: We employ and extend an existing domain-adversarial neural network (Ganin et al., 2016) based on the structure-based model where * could be RvNN, PLAN, and BiGCN; 7) **ACLR-***: our proposed adversarial contrastive learning framework on top of RvNN, PLAN, or BiGCN.

To facilitate real-world low-resource rumor detection, we adopt the cross-domain and cross-

[3] https://twarc-project.readthedocs.io/en/latest/twarc2_en_us/
[4] https://service.account.weibo.com/

Target (Source)	Weibo-COVID19 (TWITTER)				Twitter-COVID19 (WEIBO)			
Model	Acc.	Mac-F_1	Rumor F_1	Non-rumor F_1	Acc.	Mac-F_1	Rumor F_1	Non-rumor F_1
CNN	0.445	0.402	0.476	0.328	0.498	0.389	0.528	0.249
RNN	0.463	0.414	0.498	0.329	0.510	0.388	0.533	0.243
RvNN	0.514	0.482	0.538	0.426	0.540	0.391	0.534	0.247
PLAN	0.532	0.496	0.578	0.414	0.573	0.423	0.549	0.298
BiGCN	0.569	0.508	0.586	0.429	0.616	0.415	0.577	0.252
DANN-RvNN	0.583	0.498	0.591	0.405	0.577	0.482	0.648	0.317
DANN-PLAN	0.601	0.507	0.606	0.409	0.593	0.471	0.574	0.369
DANN-BiGCN	0.629	0.561	0.616	0.506	0.618	0.510	0.676	0.344
ACLR-RvNN	0.778	0.716	0.843	0.589	0.653	0.616	0.710	0.521
ACLR-PLAN	0.824	0.769	0.842	0.696	0.709	0.648	0.752	0.544
ACLR-BiGCN	**0.873**	**0.861**	**0.896**	**0.827**	**0.765**	**0.686**	**0.766**	**0.605**

Table 1: Rumor detection results on the target test datasets.

lingual settings concurrently for model training. When the target dataset is Weibo-COVID19, we use the well-resourced TWITTER dataset (Ma et al., 2017) as the source data. When the target dataset is Twitter-COVID19, we use the well-resourced WEIBO dataset (Ma et al., 2016) as the source data. We use accuracy and macro-averaged F1 score, as well as class-specific F1 score as the evaluation metrics. To conduct five-fold cross-validation on the target datasets in our low-resource settings, we use each fold (about 80 samples) in turn for training, and test on the rest data. More implementation details are provided in Appendix.

5.3 Rumor Detection Performance

Table 1 shows the performance of our proposed method versus all the compared methods on the Weibo-COVID19 and Twitter-COVID19 test sets with pre-determined training datasets. It is observed that the performances of the baselines in the first group are obviously poor due to ignoring intrinsic structural patterns. To make fair comparisons, all baselines are employed with the same cross-lingual sentence encoder of our framework as inputs. Other state-of-the-art baselines exploit the structural property of community wisdom on social media, which confirms the necessity of propagation structure representations in our framework.

Among the structure-based baselines in the second group, due to the representation power of message-passing architectures and tree structures, PLAN and BiGCN outperform RvNN with only limited labeled target data for training. The third group shows the results for DANN-based methods. It improves the performance of structure-based baselines in general since it extracts cross-domain features between source and target datasets via generative adversarial nets (Goodfellow et al., 2014). Different from that, we use the adversarial attacks to improve the robustness of our proposed contrastive training paradigm, which explicitly encourages effective alignment of rumor-indicative features from different domains and languages.

In contrast, our proposed ACLR-based approaches achieve superior performances among all their counterparts ranging from 21.8% (13.4%) to 30.0% (17.7%) in terms of Macro F1 score on Weibo-COVID19 (Twitter-COVID19) datasets, which suggests their strong judgment on low-resource rumors from different domains/languages. ACLR-BiGCN performs the best among the three ACLR-based methods by making full use of the structural property via graph modeling for conversation threads. This also justifies the good performance of DANN-BiGCN and BiGCN. The results also indicate that the adversarial contrastive learning framework can effectively transfer knowledge from the source to target data at the event level, and substantiate our method is model-agnostic for different structure-based networks.

5.4 Ablation Study

We perform ablation studies based on our best-performed approach ACLR-BiGCN. As demonstrated in Table 2, the first group shows the results for the backbone baseline BiGCN. We observe that the model performs best if pre-trained on source data and then fine-tuned on target training data (i.e., BiGCN(S,T)), compared with the poor performance when trained on either minor labeled target data only (i.e., BiGCN(T)) or well-resourced source data (i.e., BiGCN(S)). This suggests that our hypothesis of leveraging well-resourced source data to improve the low-resource rumor detection

Model	Weibo-COVID19		Twitter-COVID19	
	Acc.	Mac-F_1	Acc.	Mac-F_1
BiGCN(T)	0.569	0.508	0.616	0.415
BiGCN(S)	0.578	0.463	0.611	0.425
BiGCN(S, T)	0.693	0.472	0.617	0.471
DANN-BiGCN	0.629	0.561	0.618	0.510
CLR-BiGCN	0.844	0.804	0.719	0.618
ACLR-BiGCN	0.873	0.861	0.765	0.686

Table 2: Ablation studies on our proposed model.

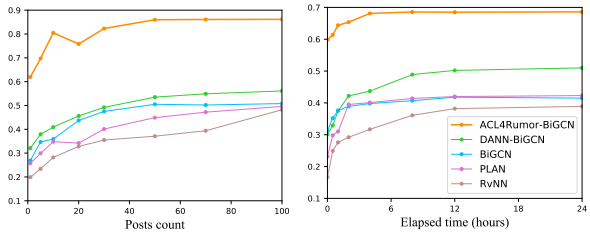


Figure 3: Early detection performance at different checkpoints of posts count (or elapsed time) on Weibo-COVID19 (left) and Twitter-COVID19 (right) datasets.

on target data is feasible. In the second group, the DANN-based model makes better use of the source data to extract domain-agnostic features, which further leads to performance improvement. Our proposed contrastive learning approach CLR without adversarial augmentation mechanism, has already achieved outstanding performance compared with other baselines, which illustrates its effectiveness on domain and language adaptation. We further notice that our ACLR-BiGCN consistently outperforms all baselines and improves the prediction performance of CLR-BiGCN, suggesting that training model together with adversarial augmentation on target data provide positive guidance for more accurate rumor predictions, especially in low-resource regimes. More qualitative analyses of hyper-parameters, training data size and alternative source datasets are shown in Appendix.

5.5 Early Detection

Early alerts of rumors can prevent the widespreading of rumorous contents. By setting detection checkpoints of "delays" that can be either the count of corresponding posts or the time elapsed since the first posting, only contents posted no later than the checkpoints is available for model evaluation. The performance is evaluated by Macro F1 score obtained at each checkpoint. To satisfy each checkpoint, we incrementally scan test data in order of time until the target time delay or post volume is reached.

Figure 3 shows the performances of our approach versus DANN-BiGCN, BiGCN, PLAN, and RvNN at various deadlines. Firstly, we observe that our proposed ACLR-based approach outperforms other counterparts and baselines throughout the whole lifecycle, and reaches a relatively high Macro F1 score at a very early period after the initial broadcast. One interesting phenomenon is that the early performance of some methods may fluctuate more or less. It is because with the propagation of the claim there is more semantic and structural information but the noisy information is increased simultaneously. Our method only needs about 50 posts on Weibo-COVID19 and around 4 hours on Twitter-COVID19, to achieve the saturated performance, indicating the remarkably superior early detection performance of our method.

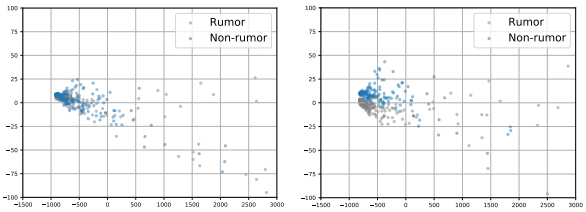


Figure 4: Visualization of target event-level representation distribution.

5.6 Feature Visualization

Figure 4 shows the PCA visualization of learned target event-level features on BiGCN (left) and ACLR-BiGCN (right) for analysis. The left figure represents training with only classification loss, and the right figure uses ACLR for training. We observe that (1) due to the lack of sufficient training data, the features extracted with the traditional training paradigm are entangled, making it difficult to detect rumors in low-resource regimes; and (2) our ACLR-based approach learns more discriminative representations to improve low-resource rumor classification, reaffirming that our training paradigm can effectively transfer knowledge to bridge the gap between source and target data distribution resulting from different domains and languages.

6 Conclusion

In this paper, we propose a novel Adversarial Contrastive Learning framework to bridge low-resource gaps for rumor detection by adapting features learned from well-resourced data to that of the low-resourced breaking events, without restrictions on specific domain/language usage. The results on two real-world benchmarks confirm the advantages of our model in low-resource rumor detection task.

References

Firoj Alam, Shaden Shaar, Fahim Dalvi, Hassan Sajjad, Alex Nikolov, Hamdy Mubarak, Giovanni Da San Martino, Ahmed Abdelali, Nadir Durrani, Kareem Darwish, et al. 2020. Fighting the covid-19 infodemic: modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society. *arXiv preprint arXiv:2005.00033*.

Gordon W Allport and Leo Postman. 1947. The psychology of rumor.

Sarah Alqurashi, Ahmad Alhindi, and Eisa Alanazi. 2020. Large arabic twitter dataset on covid-19. *arXiv preprint arXiv:2004.04315*.

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.

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 675–684.

Emily Chen, Kristina Lerman, and Emilio Ferrara. 2020a. Covid-19: The first public coronavirus twitter dataset.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020b. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. 2020c. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(ARTICLE):2493–2537.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Limeng Cui and Dongwon Lee. 2020. Coaid: Covid-19 healthcare misinformation dataset. *arXiv preprint arXiv:2006.00885*.

Adrien Friggeri, Lada Adamic, Dean Eckles, and Justin Cheng. 2014. Rumor cascades. In *Eighth international AAAI conference on weblogs and social media*.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Han Guo, Juan Cao, Yazi Zhang, Junbo Guo, and Jintao Li. 2018. Rumor detection with hierarchical social attention network. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 943–951.

Aniko Hannak, Drew Margolin, Brian Keegan, and Ingmar Weber. 2014. Get back! you don't know me like that: The social mediation of fact checking interventions in twitter conversations. In *ICWSM*.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738.

Yong Hu, He-Yan Huang, Anfan Chen, and Xian-Ling Mao. 2020. Weibo-cov: A large-scale covid-19 social media dataset from weibo. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*.

Maria Janicka, Maria Pszona, and Aleksander Wawer. 2019. Cross-domain failures of fake news detection. *Computación y Sistemas*, 23(3).

Debanjana Kar, Mohit Bhardwaj, Suranjana Samanta, and Amar Prakash Azad. 2020. No rumours please! a multi-indic-lingual approach for covid fake-tweet detection. *arXiv preprint arXiv:2010.06906*.

Ling Min Serena Khoo, Hai Leong Chieu, Zhong Qian, and Jing Jiang. 2020. Interpretable rumor detection in microblogs by attending to user interactions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8783–8790.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. 2016. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*.

Sejeong Kwon, Meeyoung Cha, Kyomin Jung, Wei Chen, and Yajun Wang. 2013. Prominent features of rumor propagation in online social media. In *2013*

IEEE 13th International Conference on Data Mining, pages 1103–1108. IEEE.

Nayeon Lee, Yejin Bang, Andrea Madotto, and Pascale Fung. 2021. Towards few-shot fact-checking via perplexity. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1971–1981.

Hongzhan Lin, Yuanmeng Yan, and Guang Chen. 2021. Boosting low-resource intent detection with in-scope prototypical networks. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7623–7627. IEEE.

Xiaomo Liu, Armineh Nourbakhsh, Quanzhi Li, Rui Fang, and Sameena Shah. 2015. Real-time rumor debunking on twitter. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 1867–1870.

Yixin Liu and Pengfei Liu. 2021. Simcls: A simple framework for contrastive learning of abstractive summarization. *arXiv preprint arXiv:2106.01890*.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. In *International Joint Conference on Artificial Intelligence*.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2017. Detect rumors in microblog posts using propagation structure via kernel learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 708–717.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2018. Rumor detection on twitter with tree-structured recursive neural networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1980–1989.

Takeru Miyato, Andrew M Dai, and Ian Goodfellow. 2016. Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725*.

Mitra Mohtarami, James Glass, and Preslav Nakov. 2019. Contrastive language adaptation for cross-lingual stance detection. 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)*, pages 4442–4452.

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 557–565.

Ming Tan, Yang Yu, Haoyu Wang, Dakuo Wang, Saloni Potdar, Shiyu Chang, and Mo Yu. 2019. Out-of-domain detection for low-resource text classification tasks. 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)*, pages 3566–3572.

Nikhita Vedula, Nedim Lipka, Pranav Maneriker, and Srinivasan Parthasarathy. 2020. Open intent extraction from natural language interactions. In *Proceedings of The Web Conference 2020*, pages 2009–2020.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR.

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.

Ke Wu, Song Yang, and Kenny Q Zhu. 2015. False rumors detection on sina weibo by propagation structures. In *2015 IEEE 31st international conference on data engineering*, pages 651–662. IEEE.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. Consert: A contrastive framework for self-supervised sentence representation transfer. *arXiv preprint arXiv:2105.11741*.

Fan Yang, Yang Liu, Xiaohui Yu, and Min Yang. 2012. Automatic detection of rumor on sina weibo. In *Proceedings of the ACM SIGKDD workshop on mining data semantics*, pages 1–7.

Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. 2007. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315.

Feng Yu, Qiang Liu, Shu Wu, Liang Wang, and Tieniu Tan. 2017. A convolutional approach for misinformation identification. In *Twenty-Sixth International Joint Conference on Artificial Intelligence*.

Hua Yuan, Jie Zheng, Qiongwei Ye, Yu Qian, and Yan Zhang. 2021. Improving fake news detection with domain-adversarial and graph-attention neural network. *Decision Support Systems*, page 113633.

Koosha Zarei, Reza Farahbakhsh, Noel Crespi, and Gareth Tyson. 2020. A first instagram dataset on covid-19. *arXiv preprint arXiv:2004.12226*.

Dejiao Zhang, Feng Nan, Xiaokai Wei, Shang-Wen Li, Henghui Zhu, Kathleen McKeown, Ramesh Nallapati, Andrew O Arnold, and Bing Xiang. 2021. Supporting clustering with contrastive learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5419–5430.

Tong Zhang, Di Wang, Huanhuan Chen, Zhiwei Zeng, Wei Guo, Chunyan Miao, and Lizhen Cui. 2020. Bdann: Bert-based domain adaptation neural network for multi-modal fake news detection. In *2020 international joint conference on neural networks (IJCNN)*, pages 1–8. IEEE.

Zhe Zhao, Paul Resnick, and Qiaozhu Mei. 2015. Enquiring minds: Early detection of rumors in social media from enquiry posts. In *Proceedings of the 24th international conference on world wide web*, pages 1395–1405.

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

A Datasets

To our knowledge, there is no public dataset available for classifying propagation trees in tweets about COVID-19, where we need the tree roots together with the corresponding propagation structure, to be appropriately annotated with ground truth. In this paper, we organize and construct two datasets Weibo-COVID19 and Twitter-COVID19 for experiments. For Twitter-COVID19, the original dataset (Kar et al., 2020) of tweets was released with just the source tweet without its propagation thread. So we collected all the propagation threads using the Twitter academic API with the twarc2 package[5] in python. Finally, we annotated the source tweets by referring to the labels of the events they are from. For Weibo-COVID19, we gather a set of rumorous claims from the Sina community management center[6] and non-rumorous claims by randomly filtering out the posts that are not reported as rumors. Both Weibo-COVID19 and Twitter-COVID19 contain two binary labels: Rumor and Non-rumor. For Weibo-COVID19 as the target dataset, we use the TWITTER dataset (Ma et al., 2017) as the source data in our low-resource (i.e., cross-domain and cross-lingual) settings; In terms of Twitter-COVID19 as the target dataset, we use Weibo (Ma et al., 2016) as the source data. We will release all the datasets and make codes available after the anonymity period. The statistics of the four datasets are shown in Table 3.

B Implementation Details

We set the number L of the graph convolutional layer as 2, the trade-off parameter α as 0.5, and the adversarial perturbation norm ϵ as 1.5. The temperature τ is set to 0.1. Parameters are updated through back-propagation (Collobert et al., 2011) with the Adam optimizer (Loshchilov and Hutter, 2018). The learning rate is initialized as 0.0001, and the dropout rate is 0.2. Early stopping (Yao et al., 2007) is applied to avoid overfitting. We run all of our experiments on one single NVIDIA Tesla T4 GPU. We set the total batch size to 64, where the batch size of source samples is set to 32, the same as target samples. The hidden and output dimensions of each node in the structure-based network are set to 512 and 128, respectively. Since the focus in this paper is primarily on better

[5] https://twarc-project.readthedocs.io/en/latest/twarc2_en_us/

[6] https://service.account.weibo.com/

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

Table 3: Statistics of Datasets in Cross-Domain and Cross-Lingual Settings.

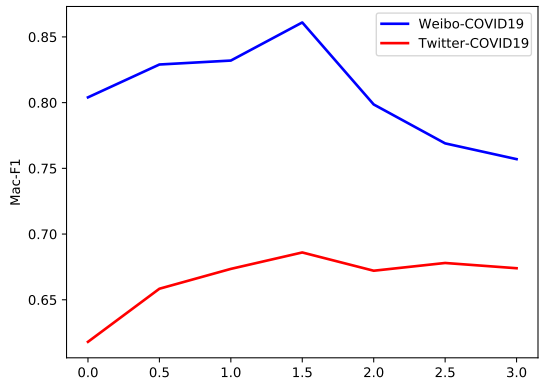


Figure 5: Effect of Adversarial Perturbation Norm ϵ.

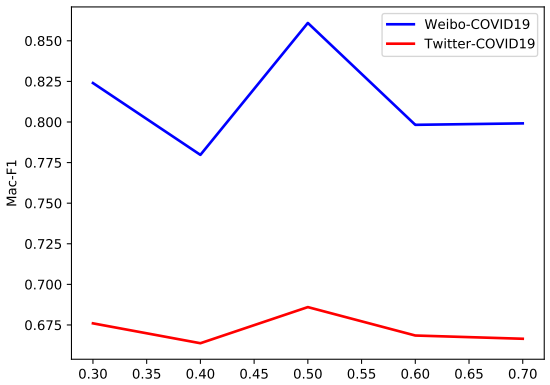


Figure 6: Effect of trade-off parameter α.

leveraging the contrastive learning for domain and language adaptation on top of event-level representations, we choose the XLM-R$_{Base}$ (Layer number = 12, Hidden dimension = 768, Attention head = 12, 270M params) as our sentence encoder for language-agnostic representations at the post level. We use accuracy and macro-averaged F1 score, as well as class-specific F1 score as the evaluation metrics. Unusually, to conduct five-fold cross-validation on the target dataset in our low-resource settings, we use each fold (about 80 claim posts with propagation threads in the target data) in turn for training, and test on the rest of the dataset. The average runtime for our approach on five-fold cross-validation in one iteration is about 3 hours. The number of total parameters is 1,117,954 for our model. We implement our model with pytorch[7].

C Qualitative Analysis

C.1 Effect of Adversarial Perturbation Norm

Figure 5 shows the effect of adversarial perturbation norm on rumor detection performance. The X-axis denotes the value of ϵ, where $\epsilon = 0.0$ in the line means no adversarial augmentation. In general, the adversarial augmentation contributes to the improvements and $\epsilon \in [1.0, 2.0)$ achieves better performances. For the Weibo-COVID19 dataset, our proposed approach ACLR with a smaller adversarial perturbation can still obtain better results but lower than the results with an optimal range of perturbation, while large norms tend to damage the effect of ACLR. In terms of Twitter-COVID19, our method still performs well with a broad range of adversarial perturbations and the performance tends to stabilize as the norm value increases.

C.2 Effect of Trade-off Parameter between Classification and Contrastive Objectives

To study the effects of the trade-off hyper-parameter in our training paradigm, we conduct ablation analysis under ACLR architecture (Figure 6). We can see that $\alpha = 0.5$ achieves the best performance while the point where $\alpha = 0.3$ also has good performance. Looking at the overall trend, the performance fluctuates more or less as the value of α grows. We conjecture that this is because the supervised contrastive objective, while optimizing the representation distribution, compromises the mapping relationship with labels. Multitask means optimizing two losses simultaneously. This setting

[7] pytorch.org

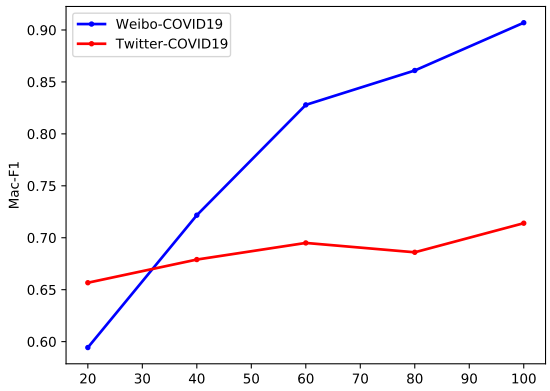


Figure 7: Effect of target training data size.

leads to mutual interference between two tasks, which affects the convergence effect. This phenomenon points out the direction for our further research in the future.

C.3 Effect of Target Training Data Size.

Figure 7 shows the effect of target training data size. We randomly choose training data with a certain proportion from target data and use the rest set for evaluation. We use the cross-domain and cross-lingual settings concurrently for model training, the same as the main experiments. Results show that with the decrease of training data size, the performance gradually decreases. Especially for Weibo-COVID19, it will be greatly affected. However, even when only 20 target data are used for training, our model can still achieve more than approximately 60% and 65% rumor detection performance (Macro F1 score) on two target data sets Weibo-COVID19 and Twitter-COVID19 respectively, which further proves ACLR has strong applicability for improving low-resource rumor detection on social media.

C.4 Discussion about Low-Resource Settings

In this section, we evaluate our proposed framework with different source datasets to discuss the low-resource settings in our experiments. Considering the cross-domain and cross-lingual settings in the main experiments, we also conduct an experiment in cross-domain settings. Specifically, for the Weibo-COVID as the target data, we utilize the WEIBO dataset as the source data with rich annotation. In terms of Twitter-COVID19, we set the TWITTER dataset as the source data. Table 4 depicted the results in different low-resource settings. It can be seen from the results that our model performs generally better in cross-domain

Target	Weibo-COVID19		Twitter-COVID19	
Settings	Acc.	Mac-F_1	Acc.	Mac-F_1
Cross-D&L	0.873	**0.861**	**0.765**	**0.686**
Cross-D	**0.884**	0.855	0.737	0.623

Table 4: Rumor detection results of our proposed framework in different low-resource settings. Cross-D&L denotes the cross-domain and cross-lingual settings and Cross-D denotes the cross-domain settings.

and cross-lingual settings concurrently than that only in cross-domain settings, which demonstrates the key insight to bridge the low-resource gap is to relieve the limitation imposed by the specific language resource dependency besides the specific domain. Our proposed adversarial contrastive learning framework could alleviate the low-resource issue of rumor detection as well as reduce the heavy reliance on datasets annotated with specific domain and language knowledge.

D Future Work

We will explore the following directions in the future:

1. We are going to explore the pre-training method with contrastive learning and then finetune the model with classification loss, which may further improve the performance and stability of the model.

2. Considering that our model has explicitly overcome the restriction of both domain and language usage in different datasets, we plan to evaluate our model on the datasets about more breaking events in low-resource domains and/or languages by leveraging existing datasets with rich annotation. We believe that our work could provide new guidance for future rumor detection about breaking events on social media.