Rumor Detection on Twitter with Claim-Guided Hierarchical Graph Attention Networks

Hongzhan Lin[1,2], Jing Ma[2,∗], Mingfei Cheng[1], Zhiwei Yang[3], Liangliang Chen[1], Guang Chen[1,∗]

[1]Beijing University of Posts and Telecommunications, Beijing, China
[2]Hong Kong Baptist University, Hong Kong SAR, China
[3]Jilin University, Changchun, China

{linhongzhan,mingfeicheng,outside,chenguang}@bupt.edu.cn
majing@comp.hkbu.edu.cn, yangzw18@mails.jlu.edu.cn

Abstract

Rumors are rampant in the era of social media. Conversation structures provide valuable clues to differentiate between real and fake claims. However, existing rumor detection methods are either limited to the strict relation of user responses or oversimplify the conversation structure. In this study, to substantially reinforces the interaction of user opinions while alleviating the negative impact imposed by irrelevant posts, we first represent the conversation thread as an undirected interaction graph. We then present a Claim-guided Hierarchical Graph Attention Network for rumor classification, which enhances the representation learning for responsive posts considering the entire social contexts and attends over the posts that can semantically infer the target claim. Extensive experiments on three Twitter datasets demonstrate that our rumor detection method achieves much better performance than state-of-the-art methods and exhibits a superior capacity for detecting rumors at early stages.

1 Introduction

Rumor is one type of social diseases in the era of social media. The spread of false rumors has a far-reaching destructive impact on both society and individuals (Ma et al., 2019b). For instance, the global COVID-19 pandemic has created fertile soil for the widespread of various rumors, conspiracy theories, hoaxes, and fake news, heavily disrupting people's peaceful lives and leading to unprecedented information chaos. A strange, new rumor claiming that "wearing a mask to prevent the spread of COVID-19 is unnecessary because the disease can also be spread via farts" [1] may mislead masses to belittle the importance of those potentially life-saving masks in epidemic prevention. Therefore, it is necessary to develop automatic approaches to facilitate rumor detection, especially amid crises.

Social psychology literature defines a rumor as a story or a statement whose truth value is unverified or intentionally false (DiFonzo and Bordia, 2007). Rumor detection aims to determine the veracity of a given story or statement. For automating rumor detection, previous studies focus on text mining from sequential microblog streams with supervised classifiers based on feature engineering (Castillo et al., 2011; Yang et al., 2012; Kwon et al., 2013; Liu et al., 2015; Ma et al., 2015) and feature learning (Ma et al., 2016; Yu et al., 2017). The interactions among users generally show conductive to provide useful clues for debunking rumors. Structured information is generally observed on social media platforms such as Twitter. Structure-based methods (Ma et al., 2017, 2018) are thus proposed to capture the interactive characteristics of rumor diffusion. We discuss briefly two types of state-of-the-art approaches: Transformer-based (Khoo et al., 2020; Ma and Gao, 2020) and Directed GCN-based (Bian et al., 2020) models.

Khoo et al. (2020) exploited post-level self-attention networks to model long-distance interactions between any pair of tweets even irrelevant. Ma and Gao (2020) further presented a tree-transformer model to make pairwise comparisons among the posts in the same subtree hierarchically, which better utilizes tree-structured user interactions in the conversation thread. Bian et al. (2020) utilized graph convolutional networks (GCNs) to encode directed conversation trees hierarchically. The structure-based methods however represent the conversation as a directed tree structure, following the bottom-up or top-down information flows. But such kind of structure, considering directed responsive relation, cannot enhance the representation learning of each tweet by aggregating information in parallel from the other informative tweets.

In this paper, we firstly represent the conversation thread as an undirected interaction graph, which allows full-duplex interactions between

---

*∗Corresponding authors.
[1]https://www.snopes.com/fact-check/farting-negate-covid19-masks/

(a) Sample of a conversation thread from a false rumor
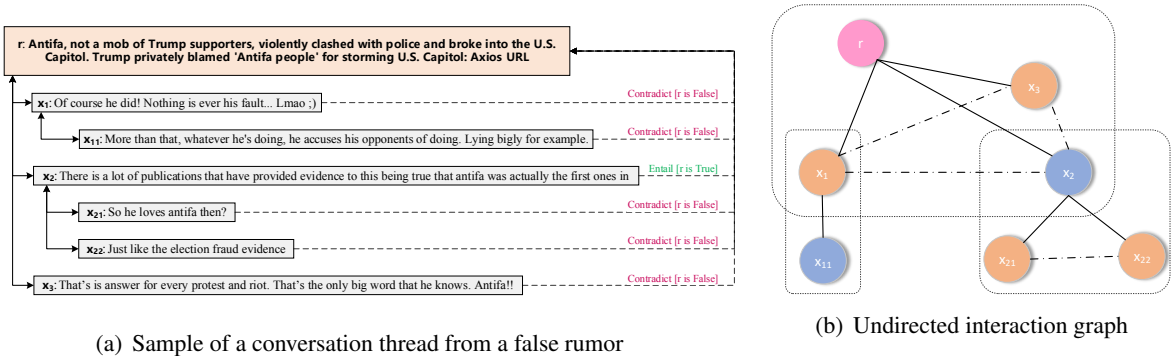
(b) Undirected interaction graph

Figure 1: (a) A motivating example: A false rumor widely spread on Twitter. (b) The undirected interaction graph for modeling the conversation thread. Blue nodes support or confirm the replied node, while orange nodes refute. For clarity's sake, we distinguish the responsive/sibling relationships between nodes with solid/chain lines.

posts with responsive parent-child or sibling relationships so that the rumor indicative features from neighbors can be fully aggregated and the interaction of user opinions can be reinforced. Intuitively, we exemplify a false rumor claim and illustrate its propagation on Twitter in Figure 1(a). We observe that a group of tweets is triggered to reply to the same post (i.e., parent post) in the conversation thread. As users share opinions, conjectures, and evidence, inaccurate information on social media can be "self-checked" by making a comparison with correlative tweets (Zubiaga et al., 2018). In order to lower the weight of inaccurate responsive information (e.g., the supportive post x_2 toward the false claim r), coherent opinions need to be captured by comparing all responsive posts toward the same post. To achieve this, our proposed interaction topology as shown in Figure 1(b) takes the correlations between sibling nodes such as the dotted box portion into account. On the other hand, by leveraging the intrinsic structural property of graph-based modeling, the undirected graph allows each tweet to learn the representation by aggregating features from all its informative neighbors. In this way, information association between nodes in the conversation can be adaptively propagated to each other along the responsive parent-child or sibling relationships while avoiding the negative impact of irrelevant interactions such as the comparison between x_{11} and x_{21} in Figure 1(a).

Moreover, previous studies show that it is critical to strengthen the semantic inference capacity between posts and the claim based on textual entailment reasoning (Ma et al., 2019a), so that we could semantically infer the claim by implicitly excavating textual inference relations such as entail, contradict, and neutral. We hypothesis that all

the informative posts should be developed and extended around the content of the claim, i.e., the potential and implicit target to be checked. Therefore, the claim content is significant to catch informative tweets, such as that in Figure 1(a), it is observed that x_{22} satirizes the opinion expressed in x_2, but its contextual information is limited. Integrating claim information for claim-aware representations could not only enrich the semantic context of x_{22}, but also enable it to better guard the consistency of topics when interacting with other nodes such as x_2 and x_{21}.

To this end, we propose a novel Claim-guided Hierarchical Graph Attention Network (ClaHi-GAT) for detecting rumors on Twitter, which not only enhances the representation learning for posts by taking the entire conversation context but also attends over the subset of informative posts. More specifically, we firstly model the conversation thread of a claim as an undirected interaction graph. To flexibly deal with the interaction of node information and the association of the global structure of the graph, we propose ClaHi-GAT to embed the undirected interaction graph. Different from standard graph attention networks (GATs) (Veličković et al., 2017), we design a claim-guided hierarchical attention mechanism at both post and event level to attend over informative posts by considering the coherent attitude and semantic inference strength toward the claim. As a result, the post-level representation is enhanced by the claim-aware attention weights obtained based on the textual content of the claim. Finally, we utilize an inference-based attention layer to implicitly capture the inference relation between the claim and the selected informative posts for rumor prediction at the event-level. We conduct extensive experiments on THREE pub-

10036

lic Twitter datasets and demonstrate that our proposed ClaHi-GAT model yields outstanding improvements over the state-of-the-art baselines with a large margin, and our method performs particularly well on early rumor detection which is crucial for timely intervention and debunking. The main contributions of this paper are three-fold:

- To our best knowledge, this is the first study of representing conversation structure as an undirected interaction graph. The graph attention-based representation achieves significant improvements over state-of-the-art methods that rely on bottom-up/top-down tree structure.

- We propose a novel ClaHi-GAT model to represent both tweet contents and the interaction graph into a latent space, which captures multi-level rumor indicative features via a claim-aware attention at the post level and an inference-based attention at the event level.

- Experimental results show that our model achieves superior performance on three real-world Twitter benchmarks for both rumor classification and early detection tasks.

2 Related Work

Pioneer studies for automatic rumor detection focus on features crafted from post contents, user profiles, and propagation patterns to learn a supervised classifier (Castillo et al., 2011; Yang et al., 2012; Liu et al., 2015). Subsequent studies were then conducted to engineer new features such as those representing rumor diffusion and cascades (Kwon et al., 2013; Friggeri et al., 2014; Hannak et al., 2014). Ma et al. (2015) extended their model with a large set of chronological social context features. These approaches typically require heavy preprocessing and feature engineering.

Zhao et al. (2015) relieved the engineering effort by using a set of regular expressions (such as "really?", "not true", etc) to find questing and denying tweets, but the oversimplified approach suffered from very low recall. Ma et al. (2016) and Yu et al. (2017) respectively utilized recurrent neural networks (RNNs) and convolutional neural networks (CNNs) to learn the representations from tweets content based on time series. Guo et al. (2018) proposed a hierarchical attention model that captures important clues from the social context of a rumorous event at the post and sub-event levels. Jin et al. (2016) exploited the conflicting viewpoints in a credibility propagation network for verifying news

stories propagated among the tweets. However, these approaches cannot embed features reflecting how posts are propagated and require careful data segmentation to prepare for time sequences.

To extract useful clues jointly from content semantics and propagation structures, Wu et al. (2015) proposed a hybrid SVM classifier to capture both flat and propagation patterns for detecting rumors on Sina Weibo. Ma et al. (2017) used Tree Kernel to capture the similarity of propagation trees in order to identify different types of rumors on Twitter. Ma et al. (2018) presented tree-structured recursive neural networks (RvNN) to jointly generate the representation of a propagation tree based on the post contents and their propagation structure. More recently, Khoo et al. (2020) proposed to model potential dependencies between any two microblog posts with the post-level self-attention networks, which is too vulnerable to avoid the negative impact of interactions among irrelevant posts. Ma and Gao (2020) treated transformer as the unit of the tree structure to further enhance the representation learning but its running time is sensitive to conversation's depth. Bian et al. (2020) used GCNs (Kipf and Welling, 2016) to encode the bi-directional conversation trees for higher-level representations.

In recent years, GATs have demonstrated superior performance in a variety of NLP tasks, such as text classification (Linmei et al., 2019), machine reading (Zheng et al., 2020), recommendation system (Wang et al., 2019), modeling knowledge graph (Cui et al., 2020) and social network bias (Yuan et al., 2019; Huang et al., 2020), etc. Different from these previous works, in this paper, we attempt to learn graph attention-based embeddings that attend to user interactions from community response for rumor detection.

3 Problem Statement

We define a Twitter rumor detection dataset as a set of events $\mathbb{C} = \{C_1, C_2, ..., C_{|\mathbb{C}|}\}$, where each event C_τ corresponds to a claim $\mathbf{c}$, composed of ideally all its relevant responsive tweets in chronological order, i.e., $C_\tau = \{\mathbf{c}, \mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_m\}$, where $\mathbf{c}$ can also be denoted as $\mathbf{x}_0$ and m is the number of responsive tweets in the conversation thread. Note that although the tweets are notated sequentially, there are connections among them based on their reply or repost relationships. So most previous works represent the conversation thread as a directed tree
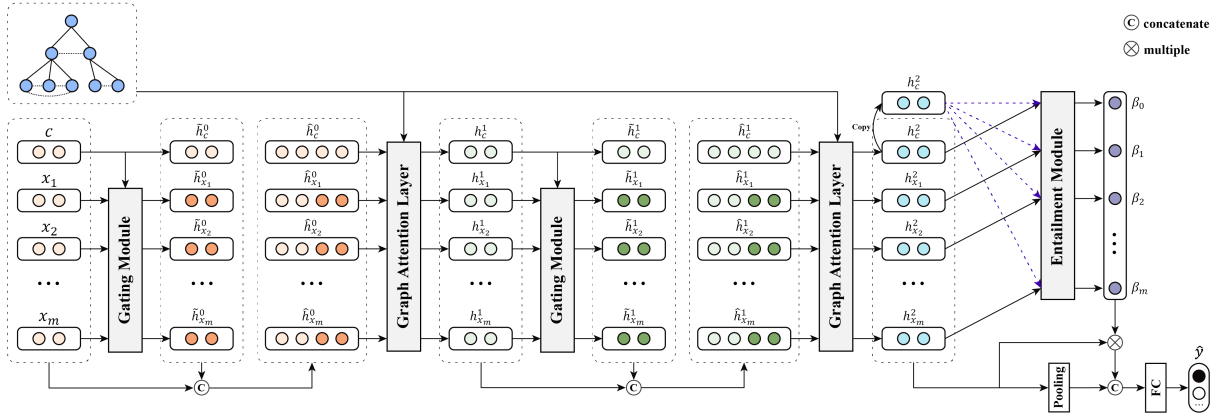
10037

Figure 2: The architecture of our proposed Claim-guided Hierarchical Graph Attention Networks.

structure (Wu et al., 2015; Ma et al., 2017, 2018; Khoo et al., 2020).

We formulate the task of rumor detection as a supervised classification problem that learns a classifier f from the labeled claims, that is, $f : C_\tau \to Y_\tau$, where Y_τ takes one of the classes defined by the specific dataset:

- Binary labels: rumor and non-rumor, which simply predicts a claim as rumor or not;
- Finer-grained labels: non-rumor, false rumor, true rumor, and unverified rumor, which makes rumor detection a more challenging classification problem (Ma et al., 2017; Zubiaga et al., 2016b).

Undirected Interaction Graphs Construction. On Twitter, each set of responsive posts triggered by the same post contains distinct rumor-indicative patterns (Ma et al., 2017). It is worth noting that we consider interactions not just between responsive parent-child nodes, but also those with the sibling relationship, for better feature aggregation from the informative tweets. To explore the full-duplex interaction patterns between responsive parent-child nodes or sibling nodes, we model the interaction topology among tweets as an undirected graph $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ for an undetermined event C_τ, as exemplified in Figure 1(b), where $\mathcal{V} = C_\tau$ that consists of all relevant posts as nodes and $\mathcal{E}$ refers to a set of undirected edges corresponding to the interactions between the nodes in $\mathcal{V}$. For example, for any $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{V}$, $\mathbf{x}_i \to \mathbf{x}_j$ and $\mathbf{x}_j \to \mathbf{x}_i$ exist if they have responsive parent-child or sibling relationships.

4 Claim-guided Hierarchical Graph Attention Networks

In this section, we introduce our Claim-guided Hierarchical Graph Attention Networks to embed the

undirected interaction graph for rumor detection. The proposed neural network consists of two attention mechanisms, i.e., a Graph Attention to capture the importance of different neighboring tweets, and a claim-guided hierarchical attention to enhance post content understanding. Figure 2 illustrates an overview of our proposed model, which will be depicted in the following subsections.

4.1 Graph Attention Networks

The core idea of GATs is to enhance the representation of responsive posts, which assign various levels of importance to neighboring posts, rather than treating all of them with equal importance, as is done in the GCN model. Our intuition for applying GATs to embed undirected interaction graphs is to reduce the weights of noisy information.

Given a tweet $\mathbf{x}_i$, we utilize a bi-directional LSTM encoder over its involved word sequence which is represented by pre-trained word embeddings. We then obtain the post-level representation using the last hidden state of the bi-directional LSTM. We thus denote the event as a matrix, i.e., $X = [c, x_1, x_2, \cdots, x_{|\mathcal{V}|-1}]^\top$, where $c, x_i \in \mathbb{R}^d$ respectively denotes the d-dimensional embedding of the claim and each responsive tweet.

In order to encode structural contexts to improve the post-level representation by adaptively aggregating more informative signals from neighboring tweets, we utilize self-attention to model the interactions between one tweet and its neighboring tweets in $\mathcal{G}$. So the attention coefficients would correlate to the impact of neighbors on the current tweet. Specifically, the input for the calculation is a set of vectors, $H^{(l)} = [h_c^{(l)}, h_{x_1}^{(l)}, h_{x_2}^{(l)}, ..., h_{x_{|\mathcal{V}|-1}}^{(l)}]^\top$ that denotes the hidden representations of nodes at the l-th layer and $h_c^{(l)}$ can also be denoted as $h_{x_0}^{(l)}$. Initially, $H^{(0)} = X$. The attention coefficient can

be computed as follows:

$$\alpha_{i,j}^{(l)} = Atten(h_{x_i}^{(l)}, h_{x_j}^{(l)})$$

$$= \frac{exp(\phi(a^\top[W^{(l)}h_{x_i}^{(l)}||W^{(l)}h_{x_j}^{(l)}]))}{\sum\limits_{j \in \mathcal{N}_i} exp(\phi(a^\top[W^{(l)}h_{x_i}^{(l)}||W^{(l)}h_{x_j}^{(l)}]))} \quad (1)$$

where $\alpha_{i,j}^{(l)}$ indicates the importance of tweet $\mathbf{x}_j$ to $\mathbf{x}_i$, a is a weight vector, $W^{(l)}$ is a layer-specific trainable transformation matrix, $||$ means "concatenate" operation, $\mathcal{N}_i$ contains $\mathbf{x}_i$'s one-hop neighbors and $\mathbf{x}_i$ itself, ϕ denotes the activation function, such as LeakyReLU(Girshick et al., 2014). Then the layer-wise propagation rule is defined as:

$$h_{x_i}^{(l+1)} = ReLU(\sum_{j \in \mathcal{N}_i} \alpha_{i,j}^{(l)} W^{(l)} h_{x_j}^{(l)}) \quad (2)$$

After that, multi-head attention is introduced to expand the channel of self-attention and stabilize the learning process (Vaswani et al., 2017). Thus Eq.2 would be extended to the multi-head attention process of concatenating K attention heads:

$$h_{x_i}^{(l+1)} = \overset{K}{\underset{k=1}{\|}} ReLU(\sum_{j \in \mathcal{N}_i} \alpha_{i,j}^{(l,k)} W_k^{(l)} h_{x_j}^{(l)}) \quad (3)$$

where $h_{x_i}^{(l+1)}$ denotes the hidden representations of the tweet $\mathbf{x}_i$ at the $(l+1)$-th layer. $\alpha_{i,j}^{(l,k)}$ is a normalized attention coefficient calculated by the k-th head at the l-th layer, and $W_k^{(l)}$ represents the corresponding linear transformation matrix. After going through an L-layer GAT, the output embedding in the final layer is calculated using averaging, instead of the concatenation operation:

$$h_{x_i}^{(L)} = ReLU(\frac{1}{K}\sum_{k=1}^{K}\sum_{j \in \mathcal{N}_i} \alpha_{i,j}^{(l',k)} W_k^{(l')} h_{x_j}^{(l')}) \quad (4)$$

where $l' = L - 1$, $h_{x_i}^{(L)}$ is the refined node representation of $\mathbf{x}_i$ after aggregating information from the other informative tweets. Here we employ mean-pooling operators to jointly capture the opinions expressed in the whole conversation, which is obtained based on the refined node representation:

$$\bar{s} = mean\text{-}pooling(H^{(L)}) \quad (5)$$

where $\bar{s}$ is the mean-pooled representation of the entire graph.

4.2 Claim-guided Hierarchical Attention

On top of the GATs, we further propose the claim-guided hierarchical attention mechanism to strengthen the topical coherence and semantic in-

ference for our model.

Post-level Attention. To make full use of abundant information in the claim and prevent off-topic coherence that deviates from the claim's focus, we exploit a gating module to endow the model with the capacity of deciding how much information it should accept from the claim for better guiding the importance allocation of the related post in the neighborhood. The claim-aware representation could be obtained as follows:

$$g_{c \to x_i}^{(l)} = sigmoid(W_g^{(l)} h_{x_i}^{(l)} + U_g^{(l)} h_c^{(l)})$$
$$\tilde{h}_{x_i}^{(l)} = g_{c \to x_i}^{(l)} \odot h_{x_i}^{(l)} + (1 - g_{c \to x_i}^{(l)}) \odot h_c^{(l)} \quad (6)$$

where $g_{c \to x_i}^{(l)}$ is the gate vector at the l-th layer, with trainable parameters $W_g^{(l)}$ and $U_g^{(l)}$. We omit the bias to avoid notation clutter. $\odot$ denotes Hadamard product. Then we concatenate the claim-aware representation with the original representation to feed into Eq.1 for a refined claim-aware attention weight:

$$\hat{h}_{x_i}^{(l)} = [\tilde{h}_{x_i}^{(l)}||h_{x_i}^{(l)}]$$
$$\hat{\alpha}_{i,j}^{(l)} = Atten(\hat{h}_{x_i}^{(l)}, \hat{h}_{x_j}^{(l)}) \quad (7)$$

Note that in this way, we update the raw representation and attention score h, α fed into Eq. 2-4 with the refined representation and attention score $\hat{h}$, $\hat{\alpha}$, so that our model can determine the verdict of a claim more reasonably with evidential posts taking the learned claim representation into account.

Event-level Attention. A natural argument against the prior GAT-mean-based model (see Section 4.1) is that mean-pooling over the node vectors does not always make sense, since some nodes are more important than others for reasoning the veracity of the rumorous event. In order to strengthen the semantic inference capacity of our model, we propose an inference module at the event level to implicitly capture the entailment relations between the posts and the claim based on the Natural Language Inference (NLI) (Bowman et al., 2015).

Inspired by the matching scheme used in classical NLI models (Mou et al., 2015; Yang et al., 2019), given the output of the last graph attention layer, we conduct each such pair by integrating three matching functions between $h_c^{(L)}$ and $h_{x_i}^{(L)}$: 1) concatenation $[h_c^{(L)}||h_{x_i}^{(L)}]$; 2) element-wise product $h_{prod}^{(L)} = h_c^{(L)} \odot h_{x_i}^{(L)}$; 3) absolute element-wise difference $h_{diff}^{(L)} = |h_c^{(L)} - h_{x_i}^{(L)}|$. Afterwards, we can obtain a joint representation as:

$$h_{x_i}^c = tanh\left(FC([h_c^{(L)}||h_{x_i}^{(L)}||h_{prod}^{(L)}||h_{diff}^{(L)}])\right) \quad (8)$$

10039

We employ an attention over the output embeddings of the last graph attention layer to select inference-based informative posts, which is guided by the joint representation $h_{x_i}^c$. This yields:

$$b_i = tanh(FC(h_{x_i}^c))$$
$$\beta_i = \frac{exp(b_i)}{\sum_i exp(b_i)} \quad (9)$$
$$\hat{s} = \sum_i \beta_i h_{x_i}^{(L)}$$

where β_i is the normalized inference-based attention weight of $\mathbf{x}_i$ for attaining the representation $\hat{s}$ of an entire graph. Lastly, we concatenate $\hat{s}$ with $\bar{s}$ and feed them into a fully-connected layer to get a low-dimensional veracity prediction vector:

$$\hat{y} = softmax(FC([\hat{s}||\bar{s}])) \quad (10)$$

where FC means a fully-connected network.

4.3 Model Training

During model training, we exploit the cross-entropy loss of the predictions $\hat{y}$ and ground truth distributions y over training data with the L2-norm. We set the number L of the graph attention layer as 2, and the head number K as 4. Parameters are updated through back-propagation (Collobert et al., 2011) with the Adam optimizer (Kingma and Ba, 2014). The learning rate is initialized as 0.0005, and the dropout rate is 0.2. Early stopping (Yao et al., 2007) is applied to avoid overfitting.

5 Experiments

5.1 Datasets

We conduct experiments on three public benchmarks, including Twitter15 (Ma et al., 2017), Twitter16 (Ma et al., 2017), and PHEME (Zubiaga et al., 2016a). The label of each event in Twitter15 and Twitter16 is annotated according to the veracity tag of the article in rumor debunking websites (e.g., snopes.com, Emergent.info, etc) (Ma et al., 2017). Moreover, the fraction of different types of rumors is imbalanced in the real-world. For example, the number of real news usually far exceeds that of false rumors. Therefore, we resort to another public benchmark rumor dataset PHEME[2], which is unbalanced and collected based on five real-world breaking news items. TWITTER (Twitter15&16) datasets contain four labels: Non-rumor (NR), False Rumor (FR), True Rumor (TR),

and Unverified Rumor (UR), while the unbalanced dataset PHEME collected based on five real-world breaking news items contains two binary labels: Rumor and Non-rumor. To evaluate the robustness of our model on complex responsive relations, we further split TWITTER datasets into TWITTER-S and TWITTER-D according to the conversation depth (TWITTER-S: ≤ 3; TWITTER-D: ≥ 4) following Ma and Gao (2020). The full statistics of datasets and implementation details are shown in the appendix.

5.2 Experimental Setup

We compare our proposed model with the following baseline and state-of-the-art models: 1) **DTR**: A Decision-Tree-based Ranking model (Zhao et al., 2015) that identifies trending rumors by searching for inquiry phrases. 2) **DTC**: A decision tree-based model (Castillo et al., 2011) that utilizes a combination of news characteristics. 3) **RFC**: A random forest classifier (Kwon et al., 2013) with a set of hand-crafted features like linguistic and structure characteristics, etc. 4) **SVM-TK**: A SVM classifier that uses a Tree Kernel (Ma et al., 2017) which try to capture propagation structure via kernel learning. 5) **GRU-RNN**: A RNN-based model that learns temporal-linguistic patterns from user comments (Ma et al., 2016). 6) **RvNN**: A rumor detection approach based on tree-structured recursive neural networks (Ma et al., 2018) with GRU units that learn rumor representations via the propagation structure. 7) **PLAN**: A transformer-based model (Khoo et al., 2020) for rumor detection to model long-distance interactions between any pair of tweets even irrelevant. 8) **HD-TRANS**: Ma and Gao (2020) proposed a model based on tree-transformer networks, which focuses on proving its effectiveness on shallow and deep conversations of datasets separately. Thus we compare it on TWITTER-S/-D. 9) **Bi-GCN**: A GCN-based model (Bian et al., 2020) on directed conversation trees to learn higher-level representations.

We use accuracy and class-specific F-measure as evaluation metrics. To make a fair comparison, we conduct five-fold cross-validation on the datasets following all baselines to obtain robust results.

5.3 Rumor Classification Performance

Table 1 and Table 2 show the performance of our proposed method versus all the compared methods on the TWITTER and PHEME datasets, where the best result of each column is bolded to indicate the

---

[2] https://figshare.com/articles/dataset/PHEME_dataset_of_rumours_and_non-rumours/4010619

Dataset	TWITTER-S					TWITTER-D				
		NR	FR	TR	UR		NR	FR	TR	UR
Method	Acc.	F_1	F_1	F_1	F_1	Acc.	F_1	F_1	F_1	F_1
DTR	0.467	0.622	0.329	0.520	0.299	0.566	0.447	0.577	0.555	0.484
DTC	0.523	0.728	0.418	0.512	0.349	0.538	0.758	0.516	0.332	0.381
RFC	0.599	0.782	0.470	0.561	0.385	0.582	0.774	0.501	0.461	0.395
SVM-TK	0.719	0.705	0.683	0.785	0.682	0.669	0.698	0.649	0.689	0.615
GRU-RNN	0.715	0.700	0.697	0.780	0.620	0.646	0.645	0.624	0.714	0.598
RvNN	0.749	0.724	0.729	0.830	0.684	0.705	0.725	0.677	0.759	0.656
PLAN	0.764	0.742	0.744	0.840	0.699	0.719	0.746	0.708	0.760	0.646
HD-Trans	0.789	0.749	0.784	0.837	0.776	0.768	0.773	0.781	0.783	0.721
Bi-GCN	0.790	0.716	0.758	0.843	0.816	0.803	0.792	0.788	0.796	0.814
ClaHi-GAT	**0.847**	**0.806**	**0.817**	**0.886**	**0.854**	**0.835**	**0.832**	**0.823**	**0.824**	**0.849**

Table 1: Rumor detection results on TWITTER datasets.

significant improvement over all baselines ($p <$ 0.05). To fairly compare with HD-Trans, our main experiments are conducted on TWITTER-S/-D and we also provide experimental results on the original TWITTER datasets in the appendix for completeness.

It is observed that the performances of the baselines in the first group based on handcrafted features are obviously poor. RFC performs relatively better because of the usage of additional temporal traits. Except for the first group, other baselines exploit the collective wisdom of the community by applying natural language processing to comments directed toward a claim without dependency on metadata and laborious feature engineering.

Among the baselines without feature engineering in the second group, due to the representation power of message-passing architectures and tree structures, PLAN, HD-Trans and Bi-GCN outperform RvNN in general. However, our aggregation-based method achieves superior performance among all the baselines on different datasets, even in the case where data is just shallow/deep conversation separately or unbalanced, which reflects its keen judgment on rumors and indicates the flexibility of our model on different types of datasets. Different from the aforementioned baselines, ClaHi-GAT is based on the interaction topology considering not only the intrinsic structural property but also the interaction between close associated posts.

The outstanding results indicate that the claim-guided hierarchical attention mechanism based on undirected interaction graphs modeling can effectively enhance the representation learning using semantic and structural information.

Method	Acc.	Non-rumor	Rumor
		F_1	F_1
DTR	0.657	0.772	0.317
DTC	0.670	0.755	0.494
RFC	0.709	0.809	0.393
SVM-TK	0.785	0.839	0.677
GRU-RNN	0.775	0.832	0.658
RvNN	0.829	0.873	0.736
PLAN	0.824	0.868	0.731
Bi-GCN	0.835	0.872	0.764
ClaHi-GAT	**0.859**	**0.893**	**0.790**

Table 2: Rumor detection results on PHEME dataset.

5.4 Ablation Study

We perform ablation studies by discarding some important components of ClaHi-GAT on Twitter15&16, and PHEME respectively, which include 1) *ClaHi-GAT/DT*: Instead of the undirected interaction graph, we use the directed trees (Ma et al., 2018; Bian et al., 2020) as the model input. 2) *GAT+EA+SC*: We simply concatenate the features of the claim with the node features at each GAT layer, to replace the claim-aware representation in Eq.6. 3) *w/o EA*: We discard the event-level (inference-based) attention as presented in Eq.9. 4) *w/o PA*: We neglect the post-level (claim-aware) attention by leaving out the gating module introduced in Eq.6. 5) *GAT*: The backbone model described in Sec.4.1. 6) *GCN*: The vanilla graph convolutional networks with no attention.

As demonstrated in Table 3, *ClaHi-GAT/DT* suffers a large decrease, indicating that our proposed undirected interaction graph modeling contributes to the final performance and its combination with claim-guided hierarchical graph attention encoding is critical. Each component of our model alone improves the model, indicating their effectiveness for embedding the interaction graph. Specifically, *GAT* makes remarkable improvements over *GCN*,

Method	Twitter15	Twitter16	PHEME
	Acc.	Acc.	Acc.
ClaHi-GAT	0.891	0.908	0.859
ClaHi-GAT/DT	0.813	0.848	0.837
GAT+EA+SC	0.853	0.866	0.846
GAT+PA(w/o EA)	0.878	0.889	0.848
GAT+EA(w/o PA)	0.847	0.864	0.845
GAT	0.835	0.854	0.840
GCN	0.825	0.820	0.832

Table 3: Ablation studies on our proposed model.

reflecting the role of naive attention in reducing the weights of noisy nodes; *w/o EA* and *w/o PA* consistently outperform *GAT*, suggesting that both levels of attention are comparably helpful; Combining them hierarchically makes further improvements and implies their complementary as represented by *ClaHi-GAT*, and replacing the claim-aware attention at the post level with simple concatenation (*GAT+EA+SC*) also leads to performance degradation, reaffirming the more effective and reasonable involvement of claims and advantages of the claim-guided hierarchical attention mechanism.

5.5 Evaluation of Undirected Interaction Graphs

We present more qualitative analyses about the undirected interaction graph and event-level attention in this section. Figure 3 provides the experimental results of ClaHi-GAT and the following models based on different modeling ways:

1. **ClaHi-GAT/DT** Utilize the directional tree applied in past influential works (Ma et al., 2018; Ma and Gao, 2020; Bian et al., 2020) as the modeling way instead of our proposed undirected interaction graph.

2. **ClaHi-GAT/DTS** Based on the directional tree structure similar to ClaHi-GAT/DT but the explicit interactions between sibling nodes are taken into account.

3. **ClaHi-GAT/UD** The modeling way is our undirected interaction topology but without considering the explicit correlations between sibling nodes that reply to the same target.

4. **ClaHi-GAT** In this paper, we propose to model the conversation thread as an undirected interaction graph for our claim-guided hierarchical graph attention networks.

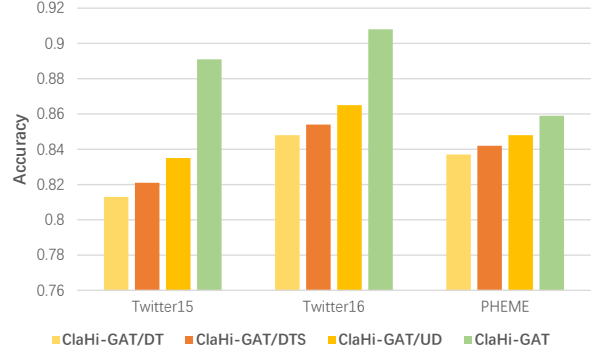From the experimental results of Figure 3, we draw the following observations:

Figure 3: The rumor classification performance of ClaHi-GAT based on different modeling ways.

Effectiveness of exploring coherent opinions among sibling nodes. Compared with *ClaHi-GAT/DT*, *ClaHi-GAT/DTS* achieves 0.8%, 0.6% and 0.5% boosts in accuracy on Twitter15, Twitter16 and PHEME respectively. Compared with *ClaHi-GAT/UD*, *ClaHi-GAT* achieves 5.6%, 4.3% and 1.1% boosts in accuracy on Twitter15, Twitter16 and PHEME respectively. It proves the effectiveness of the enhanced interaction of user opinions by exploring the correlation between sibling nodes that reply to the same target.

Effectiveness of the undirected graphs. Due to the simplex interactions between posts in the directional tree, the interaction between sibling nodes can not have a strong impact. Therefore, we propose the undirected structure to strengthen the aggregation of rumor indication features and maximize the influence of the interaction between sibling nodes. We can see that without considering the sibling relationship, *ClaiHi-GAT/UD* has better results than *ClaHi-GAT/DT*, suggesting that the combination of the undirected graph with our proposed claim-guided hierarchical graph attention mechanism is more suitable and complementary. Not only that, *ClaHi-GAT* boosts the performance as compared with *ClaHi-GAT/DTS*, showing 7.0%, 5.4% and 1.7% improvements in accuracy on the three datasets, which reveals that the undirected interaction topology does enhance semantic associations and fusion.

5.6 Early Rumor Detection

To take preventive measures to rumor spreading in a timely manner, debunking rumors at the early stage of their propagation is important. In early detection task, we compare different detection methods at a series of checkpoints of "delays" that can be measured by either the count of responsive posts received (for Twitter15&16 dataset) or the time

(a) Twitter15 (posts count)   (b) Twitter16 (posts count)   (c) PHEME (elapsed time)
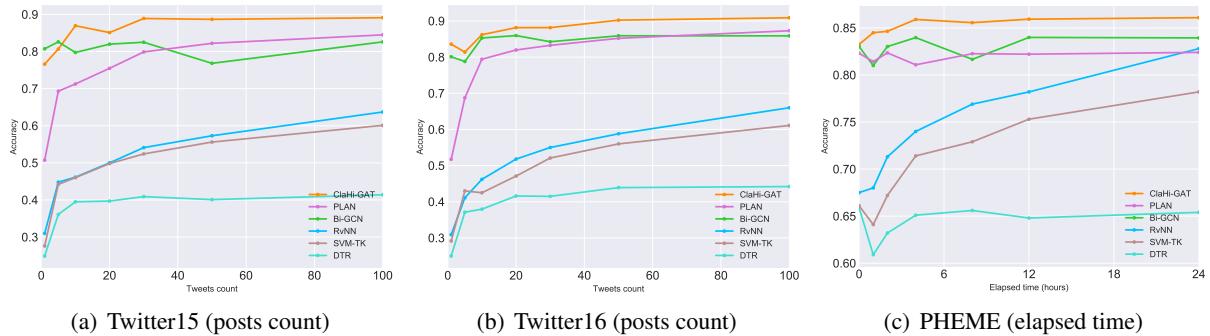
Figure 4: Early rumor detection accuracy at different checkpoints in terms of post count (or elapsed time) on Twitter15, Twitter16, and PHEME datasets.

r: Police seen in Longpont, a village near to a petrol station reportedly robbed by the two #CharlieHebdo suspects. URL
x₁: This is getting ridiculous - are there NEVER police personnel there? Send us news, not #excitoes
x₂: Get those criminals!
x₃: Keep doing Mossad/CIA work by churning out fake news.
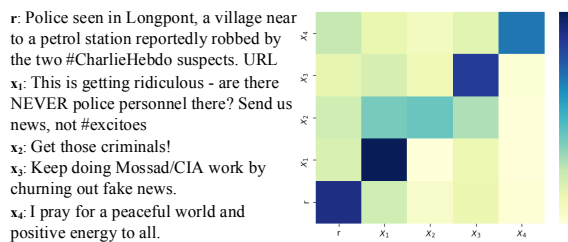**x₄**: I pray for a peaceful world and positive energy to all.

Figure 5: Example of correctly detected false rumors at early stage of our model.

elapsed since the claim was posted (for PHEME dataset). The performance is evaluated by the accuracy obtained when we incrementally scan test data in order of time until the target time delay or post volume is reached.

Figure 4 shows the performances of our ClaHi-GAT method versus PLAN, Bi-GCN, RvNN, SVM-TK, and DTR at various deadlines. It is observed that models leveraging the structural information (e.g., ClaHi-GAT method, PLAN, and Bi-GCN) reach relatively high accuracy at a very early period after the initial broadcast. One interesting phenomenon is that the early performance of all methods fluctuated more or less. We conjecture that this is because with the propagation of the claim there is more semantic and structural information, meanwhile, the noisy information is increased. Therefore, the results show that our model is insensitive to data and has better stability and robustness. ClaHi-GAT only needs about 30 posts on TWITTER and around 4 hours on PHEME, to achieve the saturated performance, which indicates remarkably superior early detection performance of our method.

To get an intuitive understanding of what is happening when we use the ClaHi-GAT model, we present an example of sibling nodes responding to the false claim r in our undirected interaction graph with a heatmap of the averaged multi-head attention score of neighbors at the last graph at-

tention layer. In Figure 5 we can see that for the false rumor, the inaccurate information like x_2 and x_4 could reduce their weights and pay more attention to the claim-related denial or questioning posts that contradict the claim, which may help us correctly predict the false rumor. Furthermore, the obtained attention scores play a crucial role in the interpretability of the prediction by the highlighted informative posts and hidden correlations.

6 Conclusion

In this paper, we propose a novel Claim-guided Hierarchical Graph Attention Network based on undirected interaction graphs to learn graph attention-based embeddings that attend to user interactions for rumor detection. Multi-level rumor indicative features could be better captured via the claim-aware attention at post level and the inference-based attention at event level. The results on three public benchmark datasets confirm the advantages of our model. Our framework is expected to provide new guidance for future rumor detection work.

Acknowledgements

References

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

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *Computer Science*.

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.

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.

Limeng Cui, Haeseung Seo, Maryam Tabar, Fenglong Ma, Suhang Wang, and Dongwon Lee. 2020. Deterrent: Knowledge guided graph attention network for detecting healthcare misinformation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 492–502.

Nicholas DiFonzo and Prashant Bordia. 2007. Rumor, gossip and urban legends. *Diogenes*, 54(1):19–35.

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

Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*.

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*.

Qi Huang, Junshuai Yu, Jia Wu, and Bin Wang. 2020. Heterogeneous graph attention networks for early detection of rumors on twitter. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Zhiwei Jin, Juan Cao, Yongdong Zhang, and Jiebo Luo. 2016. News verification by exploiting conflicting social viewpoints in microblogs. In *Thirtieth AAAI conference on artificial intelligence*.

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.

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

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

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.

Hu Linmei, Tianchi Yang, Chuan Shi, Houye Ji, and Xiaoli Li. 2019. Heterogeneous graph attention networks for semi-supervised short text classification. 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 4823–4832.

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.

Yi-Ju Lu and Cheng-Te Li. 2020. Gcan: Graph-aware co-attention networks for explainable fake news detection on social media. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 505–514.

Jing Ma and Wei Gao. 2020. Debunking rumors on twitter with tree transformer. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5455–5466.

Jing Ma, Wei Gao, Shafiq Joty, and Kam-Fai Wong. 2019a. Sentence-level evidence embedding for claim verification with hierarchical attention networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2561–2571.

Jing Ma, Wei Gao, Shafiq Joty, and Kam-Fai Wong. 2020. An attention-based rumor detection model with tree-structured recursive neural networks. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(4):1–28.

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, Zhongyu Wei, Yueming Lu, and Kam-Fai Wong. 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*, pages 1751–1754.

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.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2019b. Detect rumors on twitter by promoting information campaigns with generative adversarial learning. In *The World Wide Web Conference*, pages 3049–3055.

Lili Mou, Rui Men, Ge Li, Yan Xu, Lu Zhang, Rui Yan, and Zhi Jin. 2015. Natural language inference by tree-based convolution and heuristic matching. *Computer Science*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

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

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Xiang Wang, Xiangnan He, Yixin Cao, Meng Liu, and Tat-Seng Chua. 2019. Kgat: Knowledge graph attention network for recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 950–958.

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.

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.

Runqi Yang, Jianhai Zhang, Xing Gao, Feng Ji, and Haiqing Chen. 2019. Simple and effective text matching with richer alignment features. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4699–4709.

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*.

Chunyuan Yuan, Qianwen Ma, Wei Zhou, Jizhong Han, and Songlin Hu. 2019. Jointly embedding the local and global relations of heterogeneous graph for rumor detection. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 796–805. 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.

Bo Zheng, Haoyang Wen, Yaobo Liang, Nan Duan, Wanxiang Che, Daxin Jiang, Ming Zhou, and Ting Liu. 2020. Document modeling with graph attention networks for multi-grained machine reading comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6708–6718.

Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018. Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys (CSUR)*, 51(2):1–36.

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

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016b. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PLoS ONE*, 11(3).

A Dataset Details

We conduct experiments on three public benchmark datasets, including Twitter15 (Ma et al., 2017), Twitter16 (Ma et al., 2017), and PHEME (Zubiaga et al., 2016a). Twitter15 and Twitter16 datasets contain four labels: Non-rumor (NR), False Rumor (FR), True Rumor (TR), and Unverified Rumor (UR), while the PHEME dataset contains two binary labels: Rumor and Non-rumor. The statistics of the three datasets are shown in Table 4.

B Implementation Details

During model training, we exploit the cross-entropy loss of the predictions and ground truth distributions over training data with the L2-norm. We set the number L of the graph attention layer as 2, and the head number K as 4. Parameters are updated through back-propagation (Collobert et al., 2011) with the Adam optimizer (Kingma and Ba, 2014). The learning rate is initialized as 0.0005, 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 V100-PCIE GPU. We set the batch size to 128. Since the focus in this paper is primarily on better leveraging the graph structure and correlations between nodes, we choose the text representations widely used in previous works (Ma and Gao, 2020; Ma et al., 2020). Specifically, we use the GLOVE 300d (Pennington et al., 2014) embedding to represent each token in a tweet and get 128-dimensional contextual sentence features with a single-layer Bi-LSTM encoder. The hidden dimension of each node is set to 128. We hold out 10% of the datasets for tuning the hyperparameters and conduct 5-fold cross-validation on the rest of the datasets. We use accuracy and class-specific F-measure as evaluation metrics. The average runtime for our approach on five-fold cross-validation in one iteration is about 1.0 hours. The number of total parameters is 52,851,029 for our model. We implement our model with pytorch[3].

C Supplemental Experiments

We provide a supplemental experiment on the original version of TWITTER datasets for completeness, as depicted in Table 5. Previous works like Yuan et al. (2019), Lu and Li (2020) and Huang et al. (2020) also conducted on the original TWITTER

Statistic	Twitter15	Twitter16	PHEME
# of source tweets	1,490	818	5,802
# of tree nodes	76,351	40,867	30,376
# of non-rumors	374	205	3,830
# of false rumors	370	205	1,972
# of true rumors	372	205	–
# of unverified rumors	374	203	–
Avg. time length / tree	444 Hours	196 Hours	18 Hours
Avg. # of posts / tree	52	50	6

Table 4: Statistics of TWITTER and PHEME Dataset.

datasets leveraging the bias and social network of the source of the claim. We did not include these models in our experiments, because: 1) In this paper, we work on detecting rumors solely from the posts and comments, which takes advantage of the "wisdom of crowds" information by mining conflicting viewpoints in microblogs. In order to improve the performance of our model effectively and equitably, we do not leverage the identities of user accounts or characteristics. 2) The experimental setups for the three models are not consistent with 5-fold cross-validation and even use the pre-split train, valid and test datasets by themselves, which can not easily conduct a fair comparison with the performance on 5-fold cross-validation for all baselines and our proposed model. Here we also do not include HD-TRANS in our supplemental experiments, because it focuses on proving its effectiveness on the shallow and deep trees separately instead of the original TWITTER datasets. Our implementation of the code[4] released by Bi-GCN has a big gap compared with results reported in their paper (Bian et al., 2020), though our model still performs better due to the robustness in five-fold cross-validation. The results indicate that our proposed methods outperform all the baselines, confirming the advantages of ClaHi-GAT for rumor detection task.

D Case Study

For a more comprehensive analysis on the event-level attention, we present an example of correctly detected false rumors, whose nodes are colored with the inference-based attention scores (i.e., 'β_i' in Eq.10 of the main body of this paper) at the event level (the higher the score, the darker the color). The visualization of tweets in Figure 6 shows that the ClaHi-GAT captures informative tweets in the conversation, which have a contradiction relation towards the false claim. Hence, our event-level

---

[3]pytorch.org

[4]https://github.com/TianBian95/BiGCN

Dataset	Twitter15					Twitter16				
Method	Acc.	NR F_1	FR F_1	TR F_1	UR F_1	Acc.	NR F_1	FR F_1	TR F_1	UR F_1
DTR	0.409	0.501	0.311	0.364	0.473	0.414	0.394	0.273	0.630	0.344
DTC	0.454	0.415	0.355	0.733	0.317	0.465	0.643	0.393	0.419	0.403
RFC	0.565	0.810	0.422	0.401	0.543	0.585	0.752	0.415	0.547	0.563
SVM-TK	0.667	0.619	0.669	0.772	0.645	0.662	0.643	0.623	0.783	0.655
GRU-RNN	0.641	0.684	0.634	0.688	0.571	0.633	0.617	0.715	0.577	0.527
RvNN	0.723	0.682	0.758	0.821	0.654	0.737	0.662	0.743	0.835	0.708
Bi-GCN	0.826	0.779	0.835	0.888	0.791	0.859	0.773	0.857	0.930	0.860
PLAN	0.845	0.823	0.858	0.895	0.802	0.874	0.853	0.839	0.917	0.888
ClaHi-GAT	**0.891**	**0.878**	**0.882**	**0.931**	**0.867**	**0.908**	**0.862**	**0.916**	**0.954**	**0.901**

Table 5: Rumor detection results on original Twitter15 and Twitter16 datasets.
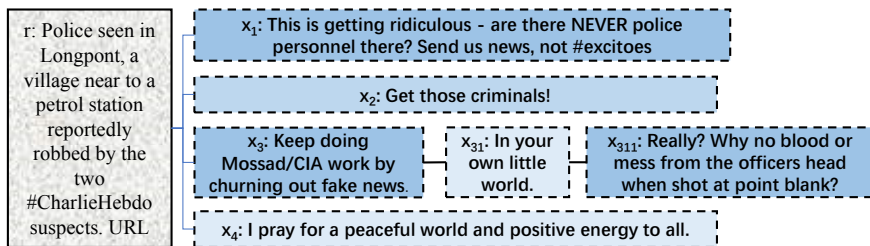

Figure 6: A sample case of correctly detected false rumors of our model. We show important tweets in the conversation and truncate others.

attention module can notice salient indicators of rumor veracity in the conversation thread, e.g., posts that contradict the false claims or entail the true claims, and then combine them to give a correct prediction.

E Future Work

We will explore the following directions in the future based on error cases where our model can not predict the correct label of the claim:

1. Traditional embedding methods like static word vectors (e.g., GloVe or Word2Vec) used in this paper cannot disambiguate homonyms, express semantic and syntactic patterns well, especially casual expression in writing on social media. Representation from Transformer pre-training may effectively help us learn more context-aware representation at the token level. We will explore how to inject the generalized contextual information via pre-trained language models into our proposed framework, to further investigate the performance improvement.

2. The event-level attention component attempts to investigate the inference relationship between a claim and its responsive post. One issue of such component is the lack of explicit supervision signal of recognizing textual inference patterns. In the future, we will utilize some existing language inference datasets with explicit labels to obtain some prior knowledge to tackle this challenge. Specifically, the knowledge of recognizing entailment relations in the trained model can be transferred to our target component.

3. In reality, some users tend to simply reshare a claim without expressing their opinions or comments. Our model cannot perfectly handle the instance that few users' engagements are available. That case is similar to the early rumor detection scenario. Although our model achieves superior performance on the early rumor detection task, it still suffers from incorrect prediction caused by the situation where users just mainly retweet the claim without more opinion expression. Also, we found an attractive point is that the same user might reply to their own claim in the propagation way. It would be heuristic for us to model novel social networks considering the special modes (e.g., retweet or reply by the node itself posting the claim) during the rumor propagation.