A Weakly Supervised Propagation Model for Rumor Verification and Stance Detection with Multiple Instance Learning

Ruichao Yang
Hong Kong Baptist University
Hong Kong SAR, China
csrcyang@comp.hkbu.edu.hk

Jing Ma
Hong Kong Baptist University
Hong Kong SAR, China
majing@comp.hkbu.edu.hk

Hongzhan Lin
Beijing University of Posts and Telecommunications
Beijing, China
linhongzhan@bupt.edu.cn

Wei Gao
Singapore Management University
Singapore
weigao@smu.edu.sg

ABSTRACT

The diffusion of rumors on microblogs generally follows a propagation tree structure, that provides valuable clues on how an original message is transmitted and responded by users over time. Recent studies reveal that rumor detection and stance detection are two different but relevant tasks which can jointly enhance each other, e.g., rumors can be debunked by cross-checking the stances conveyed by their relevant microblog posts, and stances are also conditioned on the nature of the rumor. However, most stance detection methods require enormous post-level stance labels for training, which are labor-intensive given a large number of posts.

Enlightened by Multiple Instance Learning (MIL) scheme, we first represent the diffusion of claims with bottom-up and top-down trees, then propose two tree-structured weakly supervised frameworks to jointly classify rumors and stances, where only the bag-level labels concerning claim's veracity are needed. Specifically, we convert the multi-class problem into a multiple MIL-based binary classification problem where each binary model focuses on differentiating a target stance or rumor type and other types. Finally, we propose a hierarchical attention mechanism to aggregate the binary predictions, including (1) a bottom-up or top-down tree attention layer to aggregate binary stances into binary veracity; and (2) a discriminative attention layer to aggregate the binary class into finer-grained classes. Extensive experiments conducted on three Twitter-based datasets demonstrate promising performance of our model on both claim-level rumor detection and post-level stance classification compared with state-of-the-art methods.

CCS CONCEPTS

• **Computing methodologies → Natural language processing**.

KEYWORDS

MIL, Rumor Verification, Stance Detection, Propagation Tree, Hierarchical Attention Mechanism

ACM Reference Format:
Ruichao Yang, Jing Ma, Hongzhan Lin, and Wei Gao. 2022. A Weakly Supervised Propagation Model for Rumor Verification and Stance Detection with Multiple Instance Learning. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/XXXXXX.XXXXXX

1 INTRODUCTION

The rapid development of social networks has spawned a large number of rumors, which jeopardize the environment of online community and result in harmful consequences to the individuals and our society. For instance, during the COVID-19 pandemic, a false rumor claimed that "magnetism will be generated in the body after the injection of coronavirus vaccine"[1] went viral and shared hundreds of thousands times on Twitter, which causes vaccination hesitation to the public and delays the establishment of the immune barrier of the whole society. It is meanwhile noteworthy that a wide variety of online opinions spread about a rumor can be considerably useful for us to understand the collective wisdom of crowd, thus further conducive to the improvement of our capability in recognizing some very challenging rumors [52]. In recent years, this has inspired researchers to develop automatic rumor verification approaches by leveraging large-scale analysis of online posts to mitigate the harm of rumors.

Rumor verification is a task to determine the veracity of a given claim about some subject matter [21]. Most of rumor verification methods focused on training supervised model utilizing pre-defined features [5, 24, 46] or rules [50] over the claim and its responding posts. To avoid tedious manual effort on feature engineering, data-driven methods such as recurrent neural networks (RNNs) [31] and convolutional neural networks (CNNs) [47] are proposed to learn rumor-indicative features from the sequential structure of rumor propagation. More recently, to further capture the complex propagation patterns, kernel learning algorithms are designed to compare propagation trees [32, 41, 45]. Propagation trees are also utilized to guide feature learning for classifying different types

[1] https://www.bbc.com/news/av/57207134

SIGIR '22, July 11–15, 2022, Madrid, Spain.

ACM ISBN 978-1-4503-8732-3/22/07.
https://doi.org/10.1145/XXXXXX.XXXXXX

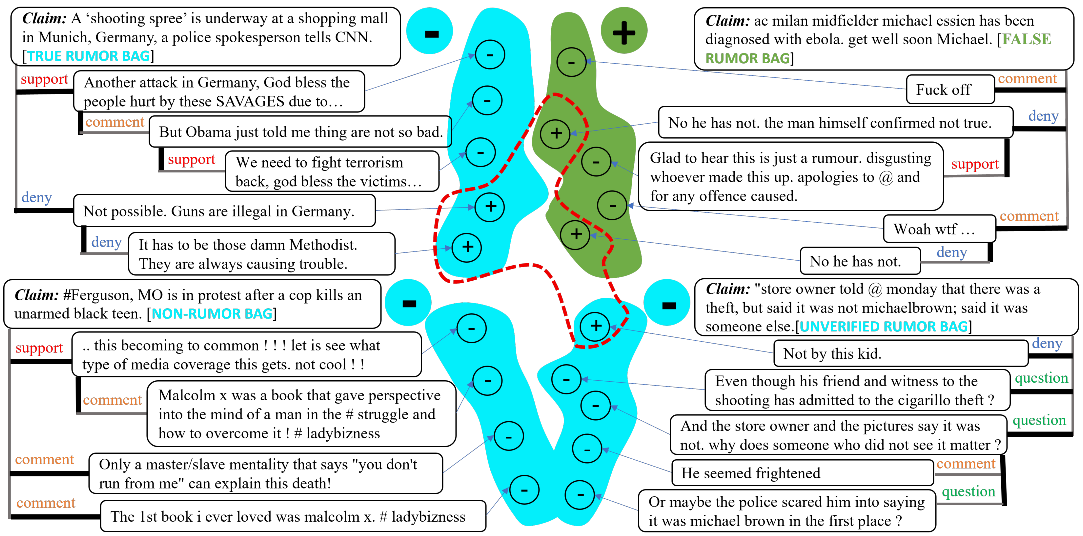


Figure 1: An illustration of tree-based MIL binary classification for simultaneous rumor and stance detection.

of rumors on Twitter based on recursive neural networks [34] or transformer model [16, 30].

Stance detection in social media aims to determine the attitude expressed in a post towards a specific target. Previous studies conducted massive manual analysis on stances pertaining to different rumor types [36]. Subsequent methods leveraged temporal traits to classify rumor stances with Gaussian Process [28] and Hawkes Process [29]. Some studies propose to train supervised models based on hand-crafted features [3, 48, 54]. To alleviate feature engineering, Zhang et al. [49] proposed a hierarchical representation of stance classes to overcome the class imbalance problem, and the multi-task learning framework is utilized to mutually reinforce stance detection and rumor classification task [20, 33]. However, they generally require large-scale annotated corpus for model training, which is a daunting task especially on social media. Some unsupervised methods are thus proposed to solve this issue [1, 18] but they achieve poor performance due to the vulnerable pre-defined features or pre-trained models. To extract useful clues from the responsive relations, a CRF-tree stance classifier [53] and a tree-structured multi-task model for the joint classification of stance and rumor [44] are proposed based on propagation tree structure.

Previous studies reveal the close correlations between rumor verification and stance classification task, e.g., the voices of opposition and doubt always appear along with the spread of rumors [36], on the other hand, sufficient skepticism and enquiries motivate finding out the claim veracity [55]. Moreover, for rumor and stance analysis from microblog posts, propagation structure provides valuable clues on how a claim is transmitted and opinioned by users over time [30, 44]. Figure 1 exemplifies the dissemination of four types of rumors, where each claim is represented as a tree by harvesting all the user responses. For rumor verification, we observe that: (1) a response express identified stance to the responded post instead of the source claim directly; (2) when a post denies the false claim, it tends to trigger supportive replies to confirm the objection; and (3) a post denies the true claim will spark denial replies. So the tree-structured stances expressed in the responsive posts can be hidden clues for rumors. For detecting stances, the false claim contains more paths such as "$supp \rightarrow deny$" and "$deny \rightarrow supp$" than that in the true claim, the unverified claim sparks more "$comm \rightarrow comm$" and "$comm \rightarrow ques$" than the other three claims. Hence, the rumor veracity can somehow help to stance detection task [20, 33].

However, post stance annotation is a daunting task, in this paper, we propose a weakly supervised model with a variant of multiple instance learning (MIL) [13] to detect stance and verify rumors simultaneously only with rumor label. The training object of original MIL is divided into two levels: bag and instance with only bag (e.g., a document) label information, there is no instance (e.g., a sentence) label information so we call MIL a weakly supervised learning framework. MIL is usually used to classify instances in the bag, then aggregate the prediction instance results as the final prediction result of the whole bag mere with bag-level annotated corpus. In addition, MIL-based models are often proposed for specific tasks without complex structure consideration, and assumes that the classes defined at the bag and instance levels are binary and should be semantically compatible. In our work, however, representation learning and classification for rumor and stance is guided by propagation tree, and both rumor and stance has multi-class and different category labels. Therefore, we propose to model the hidden correlations between rumor veracity and stances based on propagation tree in a novel way.

To this end, we firstly convert the multi-class classification problem into a multiple MIL-based binary problem with tree structure. For example, as shown in Figure 1, we treat False rumor and Deny stance (i.e., the F-D pair) as the positive class at the bag and instance level respectively, while the rest of the classes as negative. Then we develop a bottom-up and a top-down MIL-based model for stance representation learning and classification, corresponding to the propagation tree w.r.t different edge directions. Considering that the rumor types contain false (F), true (T), unverified (U), and non-rumor (N), stance types contain Deny (D), Support (S), Question(Q) and Comment (C), there will be a few possible veracity-stance pairings, and each pair tends to reach at an independent classification decision boundary. Finally, we propose a novel hierarchical attention mechanism to (1) aggregate the obtained tree-structured stances for rumor verification from each binary MIL model; and (2) combine the multiple binary results into a unified result. In this way, the stance of each post is obtained by attending over the post-level results of all the binary classifiers. Similarly, we predict the rumor class by attending over the bag-level results following the weighted collective assumption [13], which is a variant of standard MIL. Extensive experiments conducted on three real-world Twitter benchmarks demonstrate that our MIL-based methods achieve promising results for both rumor verification and stance detection tasks.

2 RELATED WORK

In this section, we provide a review of the research work on three different topics that are related to our study.

Rumor Verification. Pioneer research on automatic rumor verification focus on pre-defined features or rules crafted from texts, users, and propagation patterns to train supervised classification models [24, 40, 46]. Jin et al. [15] exploited the conflicting viewpoints in a credibility propagation network for verifying news stories propagated among the tweets. To avoid tedious and bias effort on feature engineering, subsequent studies propose data-driven methods such as recurrent neural networks (RNNs) [31], convolutional neural networks (CNN) [47] to automatically capture rumor-indicative patterns. Considering the close correlations among rumor and stance categories, multi-task learning framework are thus utilized to mutually reinforce rumor verification and stance detection tasks [20, 33]. In recent years, some approaches are proposed to model the propagation of rumors, including kernel-based method [32, 45], tree-structured recursive neural networks (RvNN) [34, 44], RNN-CNN-based method [26], transformer mechanisms [16, 30], Graph-aware co-attention networks [27] and graph neural networks [4, 22]. Inspired by the success of propagation structure, in this work, we focus on rumor verification and stance detection tasks with tree structure.

Stance Detection. Manual analysis on stance revealed some close correlations between specific veracity categories and stance [36]. A wide range of hand-crafted features are defined in the follow-up studies to train stance detection model [3, 48, 54], as well as temporal traits [28, 29]. Deep neural networks are recently utilized for stance representation learning and classification that alleviate the burden on feature engineering, such as bidirectional RNNs [2] and two-layer neural networks that learns hierarchical representation of stance classes [49]. Some studies take conversation structure into account, such as detecting stances with tree-based LSTM model [19, 53] and detecting rumors and stances jointly via a tree-structured multi-task framework [44]. However, a fundamental issue is that they require annotated corpus for model training which is an expensive and daunting task. Unsupervised methods are thus proposed to tackle this issue with pre-defined rules [18] or pre-trained models [1], but they perform poorly on stance detection. In this paper, we propose a weakly supervised propagation model to predict the rumor and stance simultaneously only with rumor label, which alleviate the predicament of post-level stance annotation.

Multiple Instance Learning (MIL). MIL is a type of weakly supervised learning which aims to learn a classifier with coarse-grained (bag-level) annotation to assign labels to instances, where instances are arranged in the bag [12]. Some follow-up researchers further propose more variants of MIL via MIL assumptions extension such as threshold-based, count-based, and weighted collective MIL assumption [13]. In recent years, MIL has been successfully applied to amounts of applications in the field of Natural Language Processing, such as a unified MIL framework that simultaneously classifies news articles and extracts sentences [43], a MIL-based model for user personalized satisfaction prediction [6], and an attention-based MIL network to recommend fashion outfit [23], etc. However, the original MIL is specifically designed for binary classification for the instances without complex structures, and the bag level labels are compatible with instance labels. In this paper, we design a tree-based MIL framework to convert the multi-class problem into multiple binary classifiers and solve the incompatible labels issue between different level of sets.

3 PROBLEM STATEMENT

We define a rumor dataset as $\{C\}$, where each training instance $C = (c, X, y)$ is a tuple representing a claim c, a sequence of relevant tweets $X = (t_1, t_2, \cdots, t_T)$ and a veracity label y of the claim. Note that although the tweets are presented in order, there are explicit connections such as response or repost relations between them. Inspired by Ma et al. [34], here we represent each claim as two different propagation trees with distinct edge directions: (1) *Bottom-up tree* where the responsive nodes point to their responded nodes, similar to a citation network; and (2) *Top-down tree* where the edge follows the direction of information diffusion by reversing the Bottom-up tree. In this paper, we consider the following two tasks:

- **Stance Detection:** To determine the post-level stance p_i for a microblog post t_i expressed towards the veracity of a claim c. That is, $f(t_1 t_2 \dots t_T | c) \rightarrow p_1 p_2 \cdots p_T$, where p_i is the stance label that takes one of Support (S), Deny (D), Question (Q) or Comment (C). Here C is assigned to tweets that do not have clear orientations to the claim veracity.
- **Rumor Verification:** To classify the claim c on top of the post stances as one of four possible veracity labels y: Non-rumor (N), True rumor (T), False rumor (F) or Unverified rumor (U). That is, $g(p_1 p_2 \cdots p_T) \rightarrow y$, here $\{p_1 p_2 \cdots p_T\}$ have a similar top-down or bottom-up tree structure.

4 OUR APPROACH

We hypothesis that the rumor diffusion process can be modeled with a bottom-up and top-down tree following the weighted collective assumption of MIL [13], where bags (i.e., claims) are labeled with the "most likely" class according to the tree-structured distribution of instances (i.e., posts) labels. In this section, we will describe our extension to the original MIL framework for verifying rumors and stance simultaneously based on the bottom-up and top-down

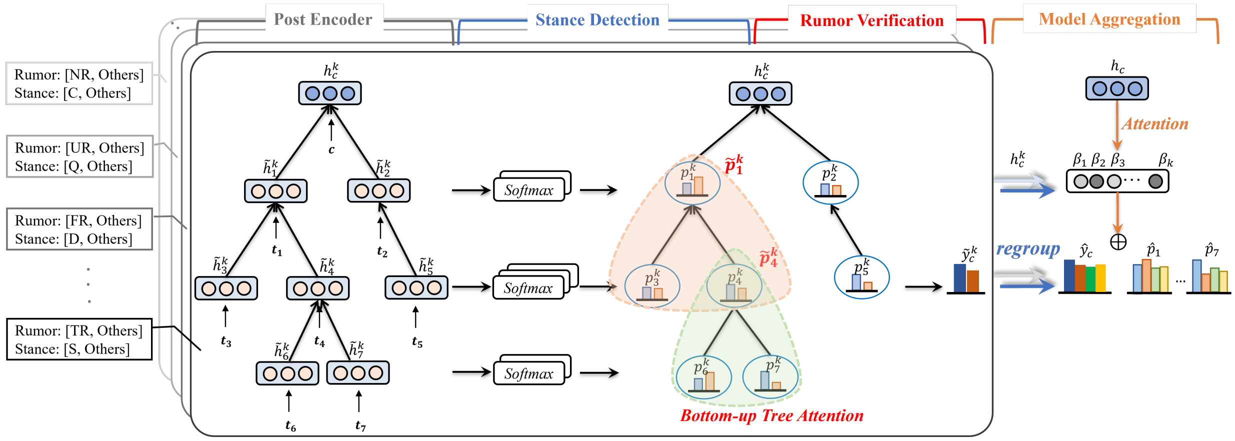


Figure 2: A framework of MIL-based model with bottom-up tree. The edge direction in the tree corresponds to stance feature aggregation recursively from bottom to up. $\tilde{p}_i^k$ denotes the aggregated stance in a subtree rooted at t_i.

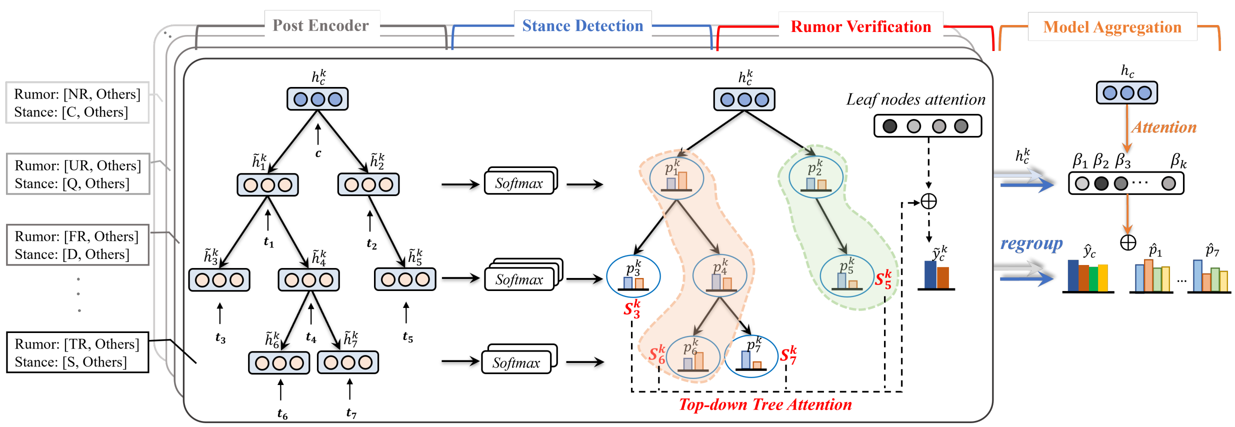


Figure 3: A framework of MIL-based model with top-down tree. S_l^k denotes the aggregated stance for specific path from c to t_l. So attending over the updated representation of all Leaf nodes correspond to selecting more informative propagation paths.

architectures presented in Section 3. Due to the fine-grained categories of stance and veracity, we convert the multi-class problem into a multiple binary classification problem at first. Assuming the rumor class number is N_r and the stance class number is N_s, then there will be $N = N_r * N_s$ possible veracity-stance pairings to get N binary classifiers.

4.1 Post Encoding

Each tweet can be represented as a word sequence $t_i = \{w_{i,1} w_{i,2} \cdots w_{i,|t_i|}\}$, where $w_{i,j} \in \mathbb{R}^d$ is a d-dimensional vector that can be initialized with pre-trained word embeddings. We map each $w_{i,j}$ into a fixed-sized hidden vector using standard GRU [7], then obtain the post-level vector for c and t_i by two GRU-based encoders[2]:

$$\begin{aligned} h_c &= h_{|c|} = GRU(w_{|c|}, h_{|c|-1}, \theta_c) \\ h_i &= h_{|t_i|} = GRU(w_{|t_i|}, h_{|t_i|-1}, \theta_X) \end{aligned} \tag{1}$$

[2]Here we choose GRU-based encoder because LSTM and GRU were widely used to learn textual representation in rumor detection task [34, 35]. But the GRU-based encoder can be replaced with pre-trained language models such as ELMO [39], BERT [11], Roberta [25], and BERTweet [37].

where $GRU(\cdot)$ denote standard GRU transition equations, $|\cdot|$ is the number of words, $w_{|c|}$ and $w_{|t_i|}$ is the last word of c and t_i, $h_{|c|-1}$ and $h_{|t_i|-1}$ denote the hidden unit for the previous word, θ_c and θ_X contain all the parameters inside the claim and post encoder.

4.2 MIL-based Bottom-up Model

Subtree structure embeds relative stance patterns which are closely correlated to the claim (i.e., root node) veracity, e.g., "*supp* $\rightarrow$ *deny*" may be more frequently observed in a false claim than that in a true claim. The core idea of MIL-based bottom-up model is to infer the post-level stance with the help of claim-level veracity label, while the prediction of both stance and veracity takes bottom-up tree structure. The overall structure of our proposed bottom-up model is illustrated in Figure 2, consisting of stance detection and rumor verification model.

Stance Detection. For a binary classifier k, we assume that the posts with similar contexts in a subtree express similar stances and the context features can be obtained by aggregating relevant branches in the subtree. So we obtain the stance representation for each post by synchronously aggregating the information from all its child nodes, following the similar bottom-up tree representation learning algorithm proposed in [34]. Using RvNN in the bottom-up manner, we map each node h_j into a context vector $\tilde{h_j}$ by recursively combining the information of its children node $C(j)$ and itself in each subtree:

$$\tilde{h}_j^k = RvNN(h_j^k, h_{C(j)}^k, \theta_j^k) \tag{2}$$

where $RvNN(\cdot)$ denote the bottom-up-based RvNN transition function [34], and θ_j represent all the parameters of RvNN.

We then use a fully-connected softmax layer to predict the stance probability of t_j towards the claim vector h_c^k related to classifier k:

$$p_j^k = softmax(W_o^k \tilde{h}_j^k + W_c^k h_c^k + b_o^k) \tag{3}$$

where W_o^k, W_c^k and b_o^k are the weights and bias in the prediction layer. Note that the stance probabilities for all posts $\{p_1^k, p_2^k, \cdots, p_T^k\}$ can be constructed as a similar bottom-up tree where each node is a stance probability.

Rumor Verification. On top of the obtained bottom-up stance tree, we define a function to aggregate the stances to predict the veracity of the claim. To this end, we propose a *bottom-up Tree Attention Mechanism* to selectively attend over specific stances expressed in more important posts from bottom to up recursively. In each recursive step, let $\mathcal{S}(i)$ denotes the set of subtree nodes rooted at i, the stance of node i is updated as the aggregated stance in a subtree:

$$\begin{aligned} \alpha_j^k &= \frac{exp(\tilde{h}_j^k \cdot h_c^{k\top})}{\sum_{j \in \mathcal{S}(i)} exp(\tilde{h}_j^k \cdot h_c^{k\top})} \\ \tilde{p}_i^k &= \sum_{j \in \mathcal{S}(i)} \alpha_j^k \cdot \tilde{p}_j^k \end{aligned} \tag{4}$$

where α_j^k denotes the attention coefficient for each node $j \in \mathcal{S}(i)$, $\tilde{h}_j^k$ and h_c^k respectively denote the hidden vector of post j and claim c, and $\tilde{p}_j^k$ is the aggregated subtree stance. After the stance aggregation from bottom to up, the updated stance of root node (i.e., claim) is equivalent to the claim-level veracity.

4.3 MIL-based Top-Down Model

The structure of top-down tree can capture complex stance patterns that model how information flows from source post to the current node, e.g., "*supp* → *comm* → *supp*" path may be more common in true rumors than that in false rumors. The core idea of MIL-based top-down model is to infer the post-level stance along propagation path based on claim-level veracity label, while the prediction of both stance and veracity takes top-down structure. The overall structure of our proposed top-down model is shown in Figure 3.

Stance Detection. In binary classifier k, the non-leaf stance features can be delivered synchronously to all its child nodes until reach at the leaf nodes. So we can obtain the stance representation for each post by aggregating all the information along the propagation path from the source, following the similar top-down representation learning algorithm proposed in [34], we map each node h_j into a context vector $\tilde{h}_j$ by recursively combining information of its parent node $P(j)$ and itself in each step[3]:

$$\tilde{h}_j^k = RvNN'(h_j^k, h_{P(j)}^k, \theta_j^k) \tag{5}$$

where $RvNN'(\cdot)$ denote the top-down-based RvNN transition function [34], and θ_j represent all the corresponding parameters.

We then use a fully-connected softmax layer to predict the stance probability towards the claim vector h_c^k related to classifier k:

$$p_j^k = softmax(W_o^k \tilde{h}_j^k + W_c^k h_c^k + b_o^k) \tag{6}$$

where $W_o^k, W_c^k h$ and b_o^k are the weights and bias in the prediction layer. Specifically, here the stance probabilities for all posts $\{p_1^k, p_2^k, \cdots, p_T^k\}$ can be constructed as a similar top-down tree where each node is a stance probability.

Rumor Verification. To aggregate the top-down stance tree into claim veracity, we propose to attend over evidential stances along each propagation path as well as selecting more evidential paths for rumor prediction. For this purpose, we design a *top-down Tree Attention Mechanism* to aggregate the stances. Firstly, our model selectively attends on the evidential stance nodes in a path expressing specific attitude towards a claim. Let r_l denote a propagation path from c to t_l (i.e., a leaf node), $\mathcal{P}(l)$ denotes node set along r_l, the aggregated stance for r_l is obtained as followed:

$$\begin{aligned} \alpha_j^k &= \frac{exp(\tilde{h}_j^k \cdot h_c^{k\top})}{\sum_{i \in \mathcal{P}(l)} exp(\tilde{h}_i^k \cdot h_c^{k\top})} \\ s_l^k &= \sum_{j \in \mathcal{P}(l)} \alpha_j^k \cdot p_j^k \end{aligned} \tag{7}$$

where α_j^k denote the attention coefficient of each node along r_l.

Secondly, for each path r_l, the information is eventually embedded into the hidden vector of the leaf nodes $\tilde{h}_l^k$. To further aggregate the path stance, we again adopt the tree attention mechanism to select more informative paths based on the leaf nodes. Let $\mathcal{K}(l)$ represents the leaf node set, the aggregated path stance representing the claim veracity can be computed as: $\tilde{y}_c^k = f'(\tilde{h}_l^k, \mathcal{K}(l), s_l^k)$, where the function $f'(\cdot)$ is a shorthand of Eq. 7.

Although the discriminative tree attention for both MIL-based models aim to predict the claim veracity by recursively aggregating all the post stance, we can conjecture that the top-down model would be better. The hypothesis is that in the bottom-up case the stance is aggregated from local subtree, and the context information is not fully considered compared with that in the top-down case where node stance is firstly aggregated through path locally then aggregate all paths globally.

[3]Note that the RvNN-based representation learning in Eq. 2 and Eq. 5 can be easily replaced with other state-of-the-art tree-based algorithms such as GCN [4], tree-LSTMs [42, 51], and PLAN [16]

4.4 Binary Models Aggregation

It is intuitive that each binary classifier contributes differently to the final prediction, according to the different strength of the veracity-stance correlation it can capture. So we design an attention mechanism to attend on the most reliable binary classifiers:

$$h_a = GRU(w_{|c|}, h_{|c|-1}, \theta_a)$$
$$\beta_k = \frac{exp(h_a \cdot h_c^{k\top})}{\sum_k exp(h_a \cdot h_c^{k\top})} \quad (8)$$

where θ_a represents all the parameters inside the GRU encoder[4] and ${h_c}^k$ is the claim representation directly obtained from the k-th classifier.

Stance Detection. We regroup all the classifiers with the same stance type target l_s into one set and then compute the final stance probability by:

$$\hat{p}_{i,l'} = \sum_{k \in U(l')} \beta_k \cdot p_i^k, \quad (9)$$

where $U(l'_s)$ represents the indicator set of the binary classifiers with $l'_s \in [S, D, Q, C]$ as the target class, p_i^k is the predicted stance probability of the post from classifier k, therefore, $\hat{p}_{i,l'_s}$ indicates the probability that the post t_i should be classified as stance l'_s. Thus, the predicted probability distribution over all the stances can be obtained, i.e., $\hat{p}_i = [\hat{p}_{i,S}, \hat{p}_{i,D}, \hat{p}_{i,Q}, \hat{p}_{i,C}]$.

Rumor Verification. We regroup all the classifiers and put the classifiers with the same rumor type l_r into one set. And then the claim-level veracity probability can computed as the weighted sum of all the classifiers' outputs: $\hat{y}_{l'_r} = g'(\beta_k, \tilde{y}^k)$, where the function $g'(\cdot)$ is a shorthand of Eq. 9, $l'_r \in [N, T, F, U]$ and $\tilde{y}^k$ is the predicted veracity class probability of the claim from classifier k. Thus, the predicted probability distribution over the veracity classes can be represented as $\hat{y} = [\hat{y}_N, \hat{y}_T, \hat{y}_F, \hat{y}_U]$.

4.5 Model Training

To train each binary classifier, we transform the finer-grained veracity and stance labels into binary labels for ground-truth representation. For example, for the classifier with [T, S] veracity-stance pair as target, the label of claim y is represented as either "T" or "others", and the model output the stance for each post represented by a probability being type "S". Similar settings are applicable to all the binary classifiers. This yields:

$$y^k = \begin{cases} 1 & \text{if the target of classifier } k \text{ is the same as } y \\ 0 & \text{others} \end{cases} \quad (10)$$

where $y \in [N, T, F, U]$ in rumor detection task, and the target refers to the veracity class instead of stance due to unavailability of label at post level.

Binary MIL-based Classifiers Training We use the negative log likelihood as loss function:

$$L_{bin} = -\sum_{k=1}^{K}\sum_{n=1}^{N} y_n^k * \log \hat{y}_n^k + (1 - y_n^k) * \log(1 - \hat{y}_n^k) \quad (11)$$

where $y_n^k \in [0, 1]$ indicating the ground truth of the n-th claim obtained in Eq. 10, $\hat{y}_n^k$ is the predicted probability for the n-th claim in classifier k, N is the total number of claims, and K is the number of binary classifiers.

Aggregation Model Training we also utilize negative log likelihood loss function to train the aggregation model:

$$L_{agg} = -\sum_{n=1}^{N}\sum_{m=1}^{M} y_{m,n} * \log \hat{y}_{m,n} + (1 - y_{m,n}) * \log(1 - \hat{y}_{m,n}) \quad (12)$$

where $y_{m,n} \in [0, 1]$ is the binary value indicating if the ground-truth veracity class of the n-th claim is m, $\hat{y}_{m,n}$ is the predicted probability the n-th claim belonging to class m, and M is the number of veracity classes.

All the parameters are updated by back-propagation [8] with Adam [17] optimizer. We use pre-trained GloVe Wikipedia 6B word embeddings [38] present on input words, set d to 100 for word vectors and model dimension, and empirically initialize the learning rate as 0.001. The training process ends when the loss value converges or the maximum epoch number is met[5]. The weighted aggregation model is trained after all the weak classifiers are well trained. And only the parameter θ_a is updated while the other parameters remain unchanged in training process.

[4]Note that the GRU have a similar yet different set of parameters with Eq. 1

Table 1: Statistics of rumor datasets for model training.

Statistics	Twitter15	Twitter16	PHEME
# of claim	1,308	818	6,425
# of Non-rumor	374 (28.6%)	205 (25.1%)	4,023 (62.6%)
# of False-rumor	370 (28.3%)	207 (25.3%)	638 (9.9%)
# of True-rumor	190 (14.5%)	205 (25.1%)	1,067 (16.6%)
# of Unverified-rumor	374 (28.6%)	201 (24.5%)	697 (10.8%)
# tree nodes	68026	40867	383569
# of Avg. posts/tree	52	50	6
# of Max. posts/tree	814	757	228
# of Min. posts/tree	1	1	3

Table 2: Statistics of the datasets for model testing.

Statistics	RumorEval2019-S	SemEval8
# of claim	425	297
# of Non-rumor	100 (23.53%)	——
# of False-rumor	74 (17.41%)	62 (20.8%)
# of True-rumor	145 (34.12%)	137 (46.1%)
# of Unverified-rumor	106 (24.94%)	98 (33.0%)
# posts of Support	1320 (19.65%)	645 (15.1%)
# posts of Deny	522 (7.77%)	334 (7.8%)
# posts of Question	531 (7.90%)	361 (8.5%)
# posts of Comment	4,345 (64.68%)	2,923 (68.6%)
# tree nodes	6,718	4,263
# Avg. posts/tree	16	14
# Max. posts/tree	249	228
# Min. posts/tree	2	3

5 EXPERIMENTS AND RESULTS

5.1 Datasets and Setup

For experimental evaluation, we refer to rumor and stance dataset with propagation structure. For model training, since only rumor labels at claim-level are required, we refer to three public tree benchmarks for detecting rumors from Twitter, namely Twitter15, Twitter16 [32] and PHEME[6]. In each dataset, each claim is annotated with one of the FOUR veracity classes (i.e., Non-rumor, True, False and Unverified) and the post-level stance label is not available. We filter out the retweets since they simply repost the claim text.

For model testing, since both post-level stance and claim level veracity are required, we resort to two rumor stance datasets collected from Twitter. namely RumorEval2019-S [14][7] and SemEval8 [9, 56]. The original datasets were used for jointly detecting rumors and stances where each claim is annotated with one of the THREE veracity classes (i.e., true-rumor, false-rumor and unverified-rumor), and each responsive post is annotated as the attitude expressed towards the claim (i.e., agreed, disagreed, appeal-for-more-information, comment). We further augmented RumorEval2019-S dataset by collecting additional 100 non-rumor claims together with their relevant posts following the method described by Zubiaga et al. [57]. We asked three annotators independent of this work to annotate the stance of each post. For post stance, we convert the original labels into [S, D, Q, C] set based on a set of rules proposed in [29]. Table 1–2 display the statistics of our datasets.

More specifically, When testing on RumorEval2019-S dataset, we train 16 binary classifiers in total considering that there are 4 veracity and 4 stance categories to be determined. And we train 12 binary stance classifiers for testing on SemEval-8 dataset since it contains 3 veracity and 4 stance categories. We hold out 20% of the test datasets as validation datasets for tuning the hyper-parameters. Due to the imbalanced rumor and stance class distribution, accuracy is not sufficient for evaluation [53]. We use AUC, micro-averaged and macro-averaged F1 score, and class-specific F-measure as evaluation metrics. We implement all the neural models with Pytorch.

5.2 Stance Detection Performance

Since our stance detection model is weakly supervised by coarse label (i.e., claim veracity) instead of explicit post-level stance label, we choose to compare with both unsupervised methods and supervised methods as followed: (1) **Zero-Shot** [1]: A pre-trained stance detection method that captures relationships between topics. (2) **Pre-Rule** [18]: An unsupervised method designed for detecting support and deny stance referring to some pre-defined rules. (3) **C-GCN** [44]: An unsupervised graph convolutional network that classifies the stances by modeling tweets with conversation structure. (4) **BrLSTM** [19]: An LSTM-based model that models the conversational branch structure of tweets to detect stance. (5) **BiGRU** [2]: A bidirectional RNN-based tweet stance model which considered the bidirectional contexts between target and tweet. We replaced the original LSTM units with GRU for fair comparisons. (6) **MT-GRU** [33]: A multi-task learning approach based on GRU to jointly detect rumors and stances by capturing the both shared and task-specific features. Here **TD/BU-MIL(DateSet)** is our proposed MIL-based top-down or bottom-up model with DateSet as the benchmark for the weak supervision.[8]

In Table 3, we use the open source of Zero-Shot and Pre-rule, this is the reason why AUC is not reported. Zero-shot, Pre-Rule and C-GCN are trained without the need of annotated data for stance detection, while BrLSTM, BiGRU and MT-GRU[9] are three supervised models for stance detection task. To train the supervised baseline systems, we use validation datasets leaved out from RumorEval2019-S and SemEval-8, the same as TD/BU-MIL(V) does.

The first group refers to unsupervised baselines. Zero-Shot and Pre-rule perform worse than other methods, because they are pre-trained models that cannot generalize well to our Twitter datasets. The results on Q and C achieved by Pre-Rule are absent since the pre-defined linguistic rules are designed for identifying the stance of Support and Deny only. Propagation-based structured method C-GCN performs better because it capture additional structural information by modeling all neighbors of each tweet.

The second group considers supervised baselines. BrLSTM improve unsupervised baselines at a large margin in terms of Micro-F1, because BrLSTM focus on modeling propagation structure while both BiGRU and MT-GRU are sequential models. But BrLSTM is poor at classifying denial stance since it is data-driven but the proportion of "D" is small in the training data. Our method TD-MIL(V) gets comparable Micro-F1 and Macro-F1 score than BiGRU, indicating that our MIL-based method has the potential to surpass supervised models. Because TD-MIL(V) considers the information propagation patterns in the whole propagation tree while BiGRU only make limited comparisons between the target and tweet.

Our MIL-based method outperforms all the baselines when training data is large enough, e.g., BU/TD-MIL(Phe) perform better than that trained on Twitter15/16 datasets. This is because PHEME datasets contains more claims than those in Twitter15/16 datasets for weak supervision. We conject that our methods will be further enhanced when training on large-scale datasets.

5.3 Rumor Verification Performance

We compare our methods with the following state-of-the-art rumor verification baselines. (1) **TD-RvNN** and **BU-RvNN** [34]: A tree-structured recursive neural networks for rumor verification with top-down and bottom-up propagation structure. (3) **H-GCN** [44]: A hierarchical multi-task learning framework for jointly predicting rumor and stance with graph convolutional network. (4) **GCAN** [27]: A graph-aware co-attention model utilizing retweet structure to verify the source tweet. (5) **PPC** [26]: a propagation-based early detection model utilizing user information and retweets. (6) **MT-GRU** [33]: A multi-task learning approach to jointly detect rumors and stances by capturing the both shared and task-specific features.

[5]The number of maximum epoch can be set by yourself, we set 150 in our experiment.
[6]This PHEME dataset is not the one widely used for stance detection [57] [10]. It is defined specifically for rumor detection: https://figshare.com/articles/PHEME_dataset_of_rumours_and_non-rumours/4010619.
[7]Here we only use the posts on Twitter from PhemeEval2019 and discard Reddit data. Note that although the original RumorEval2019-S and SemEval8 datasets contain some same claims, we further extend RumorEval2019-S dataset to a more challenging but general case to evaluate the generalization power of our methods.
[8]T15,T16 and Phe is short for Twitter15, Twitter16 and PHEME dataset respectively, and V denotes validation set (i.e., the training set for supervised baselines).
[9]Note that our stance baselines are proposed based on GRU, so we conduct experiment based on GRU encoder in our method to make fair comparisions.

Table 3: Results on stance detection: our methods achieve p-value < 0.05 under t-test for Robustness consideration.

Dataset	RumourEval2019-S							SemEval8						
Method				S	D	Q	C				S	D	Q	C
	AUC	MicF	MacF	F_1	F_1	F_1	F_1	AUC	MicF	MacF	F_1	F_1	F_1	F_1
Zero-Shot	–	0.369	0.324	0.301	0.168	0.342	0.486	–	0.383	0.344	0.278	0.162	0.480	0.456
Pre-Rule	–	0.605	0.478	0.657	0.419	——	——	——	0.429	0.389	0.432	0.644	——	——
C-GCN	0.633	0.629	0.416	0.331	0.173	0.429	0.730	0.610	0.625	0.411	0.327	0.161	0.430	0.728
BrLSTM(V)	0.71	0.66	0.42	0.460	0	0.391	0.758	0.676	0.665	0.401	0.493	0	0.381	0.730
BiGRU(V)	0.7	0.63	0.417	0.392	0.162	0.360	0.754	0.660	0.633	0.416	0.460	0.168	0.328	0.708
MT-GRU(V)	0.714	0.636	0.432	0.313	0.156	0.506	0.748	0.669	0.630	0.413	**0.498**	0.116	0.312	0.729
TD-MIL(V)	0.712	0.65	0.432	0.438	0.156	0.408	0.688	0.668	0.626	0.416	0.473	0.127	**0.463**	0.602
BU-MIL(V)	0.71	0.63	0.431	**0.485**	0.166	0.396	0.688	0.669	0.623	0.415	0.470	0.128	0.460	0.602
TD-MIL(T15)	0.706	0.668	0.427	0.339	0.173	0.444	0.752	0.663	0.642	0.418	0.330	0.174	0.420	0.750
TD-MIL(T16)	0.713	0.665	**0.436**	0.350	**0.182**	0.446	0.758	0.660	**0.671**	0.421	0.334	0.173	0.422	0.754
TD-MIL(PHE)	**0.722**	**0.691**	0.434	0.344	0.179	**0.467**	**0.767**	**0.669**	0.651	**0.426**	0.335	**0.175**	0.430	**0.763**
BU-MIL(T15)	0.706	0.662	0.428	0.341	0.173	0.436	0.756	0.661	0.638	0.415	0.326	0.168	0.420	0.748
BU-MIL(T16)	0.701	0.66	0.426	0.340	0.170	0.438	0.749	0.659	0.637	0.416	0.324	0.169	0.419	0.753
BU-MIL(PHE)	0.707	0.665	0.432	0.344	0.174	0.445	0.762	0.666	0.642	0.420	0.329	0.169	0.423	0.758

Table 4: Results on Rumor Verification: our methods achieve p-value < 0.05 under t-test for Robustness consideration.

Dataset	RumorEval2019-S							SemEval8					
Method				T	F	U	N				T	F	U
	AUC	MicF	MacF	F_1	F_1	F_1	F_1	AUC	MicF	MacF	F_1	F_1	F_1
GCAN	0.693	0.645	0.253	0.249	0.31	0.113	0.339	0.688	0.645	0.255	0.241	0.326	0.198
PPC	0.672	0.632	0.25	0.244	0.296	0.114	0.346	0.673	0.642	0.249	0.237	0.289	0.221
TD-RvNN	0.88	0.743	0.699	0.713	0.631	0.660	0.792	0.882	0.728	0.689	0.702	0.619	0.745
BU-RvNN	0.865	0.720	0.723	0.746	0.641	0.696	0.806	0.870	0.708	0.684	0.708	0.620	0.723
H-GCN	0.69	0.534	0.418	0.712	0.180	0.371	0.409	0.675	0.530	0.413	0.355	0.16	0.724
MTL2 (V)	0.683	0.653	0.43	0.622	0.279	0.352	0.457	0.680	0.651	0.433	0.640	0.289	0.372
MT-GRU (V)	0.704	0.768	0.452	0.462	0.298	0.373	0.452	0.701	0.761	0.428	0.639	0.254	0.391
TD-MIL (V)	0.685	0.678	0.45	0.667	0.329	0.376	0.428	0.680	0.621	0.436	0.650	0.274	0.384
BU-MIL (V)	0.682	0.679	0.448	0.668	0.326	0.373	0.428	0.680	0.645	0.427	0.631	0.292	0.360
TD-MIL (T15)	**0.919**	0.793	**0.79**	0.822	**0.762**	0.716	0.818	**0.913**	0.771	0.730	0.679	**0.689**	0.823
TD-MIL (T16)	0.914	0.792	0.764	0.796	0.740	0.719	0.812	0.899	0.785	0.725	0.668	0.682	0.825
TD-MIL (Phe)	0.917	**0.809**	0.776	**0.826**	0.659	0.669	**0.852**	0.908	**0.798**	**0.741**	**0.741**	0.672	0.810
BU-MIL (T15)	0.899	0.769	0.78	0.794	0.688	0.770	0.819	0.887	0.752	0.724	0.670	0.680	0.822
BU-MIL (T16)	0.902	0.776	0.76	0.780	0.664	**0.780**	0.810	0.893	0.756	0.721	0.663	0.676	**0.826**
BU-MIL (Phe)	0.904	0.776	0.763	0.793	0.666	0.770	0.833	0.902	0.763	0.729	0.728	0.649	0.809

(7) **MTL2** [20]: A sequential approach sharing a LSTM layer between the tasks, which is followed by a number of task-specific layers for multi-task outputs. Here **TD/BU-MILDateSet** is Our MIL-based methods for rumor verification.

In Table 4, we only report the best result of supervised rumor verification methods in the first group across different training datasets. MT-GRU, MTL2 are multi-task models which required to be trained on corpus with both claim and post labels. So in our case, MT-GRU, MTL2 are trained on the validation datasets with both claim and stance label. For fair comparisons, TD/BU-MIL(V) are trained with the same validation datasets.

The first group relates to structured supervised baselines. We observe that GCAN, PPC and H-GCN perform worse than the other systems, because they only consider local structure such as directly connected neighborhood. PPC perform poor because the number of user nodes in their model include both reply and retweet, which is far higher than that of the reply nodes in our model. In comparison, TD-RvNN and BU-RvNN perform better because they model the global propagation contexts by aggregating the entire propagation information recursively. However, they perform worse than our MIL methods because they aim at tweets representation in propagation process with attention mechanism, while our methods (TD-MIL(*) and BU-MIL(*)) not only use propagation information towards tweets representation, but also aggregate stances with tree-based and MIL-based attention mechanism, which reduces the role of noise stance.

The second group consists of non-structured multi-task frameworks utilizing both post-level and claim-level labels. MT-GRU(V) and MTL2(V) get higher Micro-F1 and Macro-F1 score than ours on validation dataset. This is because they are both trained under the supervision of veracity and stance annotation whereas our method only utilizes veracity labels. Moreover, TD-MIL(V) shows comparable AUC and MacF score with MTL2(V), with the increasing rumor dataset for modeling training, our methods precede MTL2(V), suggesting the potential of our weak supervised model than the supervised baselines.

Among the models jointly detecting stance and rumor, we observe that our tree-based models are more effective than the non-structured baselines (e.g., MTL2, MT-GRU), because our MIL-based propagation models capture the rumor-indicative structural features. MIL-TD (*) outperforms MIL-BU (*), because MIL-TD (*) consider both local and global contexts during stance aggregation, which verifies our assumptions in Section 4.3.

Table 5: Ablation Study Results

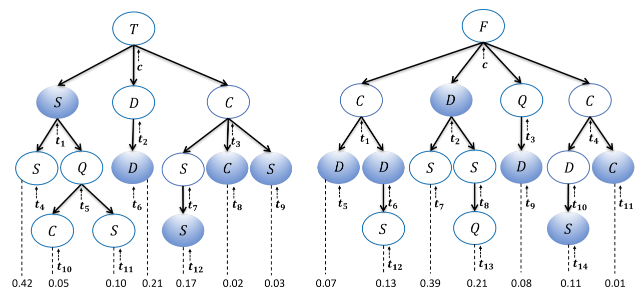
	Rumor Result			Stance Result		
Method	AUC	MicF	MacF	AUC	MicF	MacF
MIL-a	0.892	0.759	0.736	0.672	0.643	0.43
TD-MIL-b	0.912	0.802	0.746	0.701	0.658	0.426
TD-MIL-c	0.903	0.805	0.738	0.696	0.653	0.42
BU-MIL-b	0.901	0.752	0.743	0.698	0.647	0.419
BU-MIL-c	0.903	0.749	0.742	0.687	0.645	0.419
TD-MIL	0.917	0.809	0.776	0.722	0.691	0.434
BU-MIL	0.904	0.776	0.763	0.707	0.665	0.432

5.4 Ablation Study

To evaluate the impact of each component, we perform ablation tests based on the best performed BU- and TD-MIL (Phe) on RumorEval2019-S dataset minus some component(s): 1) **MIL-a**: replace all tree-based post encoder with non-structured post encoder and remove tree-based stance aggregation mechanism; 2) **TD/BU-b**: replace top-down (or bottom-up) post encoder with non-structured post encoder; 3) **TD/BU-c**: replace top-down (or bottom-up) tree attention mechanism with general attention for stance aggregation; As illustrated in Table 5, MIL-a get the lowest criteria scores, AUC/MicF/-MacF decrease about 2.5%/5.4%/4% for Rumor Result and 5%/4.8%/0.4% for Stance Result, which demonstrates top-down/bottom-up tree structure is vital to our methods. Besides, BU/TD-MIL-c variant version drops the largest percentage in both top-down and bottom-up for rumor verification and stance detection, indicating that discriminative tree attention mechanisms for stance aggregation play an important role in our methods.

5.5 Case Study

To get an intuitive understanding of the tree attention mechanism, we design an experiment to show the behavior of TD-MIL(Phe), due to its superior performance compare with the other settings. Specifically, We sample two trees from RumorEval2019-S that the source claims have been correctly classified as "true" and "false" rumor, and display the posts' predictable stance results. We compute the average path/leaf nodes attention scores over all binary classifiers, mark the most important stance with solid blue oval for each propagation path and show the leaf nodes attention scores corresponding to the importance of each propagation path in Figure 4. We observe that: 1) The "supp" posts mostly play an important role along each propagation path with "true" rumor as the target. 2) The "deny" posts contribute more in each propagation path with "false" rumor target. 3) Model with true rumor target attends more on "*supp* → *supp*" and "*deny* → *deny*" propagation patterns, that are ended with t_4 and t_6 respectively. 4) False rumor target model captures "*deny* → *supp*" and "*comm* → *deny* → *supp*" propagation patterns ended with t_7 and t_6 separately.

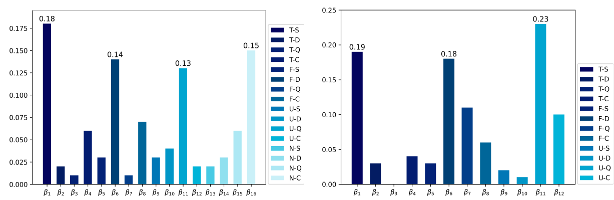


(a) True rumor case

(b) False rumor case

Figure 4: Case Study for Tree Attention Mechanism.

We also conduct experiments to show why the aggregation model can simultaneously enhance rumor verification and stance detection tasks. We randomly sample 100 claims from PHEME dataset, and then disclose the attention scores of all the binary classifiers obtained during the evaluations on RumorEval2019-S and SemEval8 datasets. The average attention scores over all the claims are shown in Figure 5. We observe that: 1) The top attention scores indicate a close correlation between the specific rumor veracity and stance category, which is compatible with previous findings [36]. Take β_1 and β_6 in Figure 5(a) for instance, they mean supportive posts indicate true rumor and denial posts can indicate false rumor. 2) The classifiers with lower attention suggest that there is a weak correlation between rumor and stance of the current target. For example, β_3 in Figure 5(b) demonstrates T-D veracity and stance pair has low correlation, which can be verified in Table 2, the true rumors has lower proportion of deny posts. 3) The rumor veracity can be generally better determined based on a combination of comprehensive stances instead of one-sided stance. For instance, among all the rumor classifiers with true rumor as the target, both T-S and T-C seem to be more important since comment stances are widely observed across all types of rumors. 4) Similarly, among the stance classifiers with question stance as the target, T-Q classifier is generally less important than the other three, which indicates the lower proportion of question posts in true rumor.



(a) RumorEval2019-S (b) SemEval8

Figure 5: Average Attention Score for Binary Classifiers from Eq. 8). RumorEval2019-S dataset has 16 binary classifiers and SemEval8 dataset has 12 binary classifiers.

6 CONCLUSION

We propose two structure-based weakly supervised propagation frameworks with Multiple Instance Learning (MIL) for detecting rumorous claims and the stances of their relevant posts simultaneously. Our models are trained only with coarse labels (i.e., claim veracity), which can jointly infer rumor veracity and the unseen post-level stance labels. Our two novel tree-based stance aggregation mechanisms (top-down and bottom-up) achieve promising results for both rumor verification and stance detection tasks compared with state-of-the-art supervised and unsupervised models.

REFERENCES

[1] Emily Allaway and Kathleen McKeown. 2020. Zero-Shot Stance Detection: A Dataset and Model Using Generalized Topic Representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 8913–8931.

[2] Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance Detection with Bidirectional Conditional Encoding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 876–885.

[3] Gaurav Bhatt, Aman Sharma, Shivam Sharma, Ankush Nagpal, Balasubramanian Raman, and Ankush Mittal. 2018. Combining neural, statistical and external features for fake news stance identification. In *Companion Proceedings of the The Web Conference 2018*. 1353–1357.

[4] 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*, Vol. 34. 549–556.

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

[6] Zheqian Chen, Ben Gao, Huimin Zhang, Zhou Zhao, Haifeng Liu, and Deng Cai. 2017. User personalized satisfaction prediction via multiple instance deep learning. In *Proceedings of the 26th International Conference on World Wide Web*. 907–915.

[7] Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the Properties of Neural Machine Translation: Encoder–Decoder Approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*. 103–111.

[8] 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 (2011), 2493–2537.

[9] Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. SemEval-2017 Task 8: RumourEval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. 69–76.

[10] Leon Derczynski, Kalina Bontcheva, Michal Lukasik, Thierry Declerck, Arno Scharl, Georgi Georgiev, Petya Osenova, Toms Pariente Lobo, Anna Kolliakou, Robert Stewart, et al. 2015. Pheme: Computing veracity—the fourth challenge of big social data. In *Proceedings of the Extended Semantic Web Conference EU Project Networking session (ESCW-PN)*.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 4171–4186.

[12] Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. 1997. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence* 89, 1-2 (1997), 31–71.

[13] James Foulds and Eibe Frank. 2010. A review of multi-instance learning assumptions. *The knowledge engineering review* 25, 1 (2010), 1–25.

[14] Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. 2019. SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation*. 845–854.

[15] Zhiwei Jin, Juan Cao, Yongdong Zhang, and Jiebo Luo. 2016. News verification by exploiting conflicting social viewpoints in microblogs. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. AAAI Press, 2972–2978.

[16] 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*, Vol. 34. 8783–8790.

[17] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1412.6980

[18] Jonathan Kobbe, Ioana Hulpuș, and Heiner Stuckenschmidt. 2020. Unsupervised stance detection for arguments from consequences. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 50–60.

[19] Elena Kochkina, Maria Liakata, and Isabelle Augenstein. 2017. Turing at SemEval-2017 Task 8: Sequential Approach to Rumour Stance Classification with Branch-LSTM. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. 475–480.

[20] Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. 2018. All-in-one: Multi-task learning for rumour verification. In *Proceedings of the 26th International Conference on Computational Linguistics*. 3402–3413.

[21] Quanzhi Li, Qiong Zhang, Luo Si, and Yingchi Liu. 2019. Rumor Detection on Social Media: Datasets, Methods and Opportunities. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*. 66–75.

[22] Hongzhan Lin, Jing Ma, Mingfei Cheng, Zhiwei Yang, Liangliang Chen, and Guang Chen. 2021. Rumor Detection on Twitter with Claim-Guided Hierarchical Graph Attention Networks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 10035–10047.

[23] Yusan Lin, Maryam Moosaei, and Hao Yang. 2020. OutfitNet: Fashion outfit recommendation with attention-based multiple instance learning. In *Proceedings of The Web Conference 2020*. 77–87.

[24] 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*. 1867–1870.

[25] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[26] Yang Liu and Yi-Fang Brook Wu. 2018. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In *Thirty-second AAAI conference on artificial intelligence*.

[27] 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*. 505–514.

[28] Michal Lukasik, Trevor Cohn, and Kalina Bontcheva. 2015. Classifying Tweet Level Judgements of Rumours in Social Media. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 2590–2595.

[29] Michal Lukasik, PK Srijith, Duy Vu, Kalina Bontcheva, Arkaitz Zubiaga, and Trevor Cohn. 2016. Hawkes processes for continuous time sequence classification: an application to rumour stance classification in twitter. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 393–398.

[30] Jing Ma and Wei Gao. 2020. Debunking Rumors on Twitter with Tree Transformer. In *Proceedings of the 28th International Conference on Computational Linguistics*.

5455–5466.

[31] Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. In *IJCAI International Joint Conference on Artificial Intelligence*, Vol. 2016. 3818–3824.

[32] 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)*. 708–717.

[33] Jing Ma, Wei Gao, and Kam-Fai Wong. 2018. Detect rumor and stance jointly by neural multi-task learning. In *Companion proceedings of the the web conference 2018*. 585–593.

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

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

[36] Marcelo Mendoza, Barbara Poblete, and Carlos Castillo. 2010. Twitter under crisis: Can we trust what we RT?. In *Proceedings of the first workshop on social media analytics*. 71–79.

[37] Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 9–14.

[38] 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)*. 1532–1543.

[39] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 2227–2237.

[40] Vahed Qazvinian, Emily Rosengren, Dragomir Radev, and Qiaozhu Mei. 2011. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. 1589–1599.

[41] Nir Rosenfeld, Aron Szanto, and David C Parkes. 2020. A kernel of truth: Determining rumor veracity on twitter by diffusion pattern alone. In *Proceedings of The Web Conference 2020*. 1018–1028.

[42] Kai Sheng Tai, Richard Socher, and Christopher D Manning. 2015. Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 1556–1566.

[43] Wei Wang, Yue Ning, Huzefa Rangwala, and Naren Ramakrishnan. 2016. A multiple instance learning framework for identifying key sentences and detecting events. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. 509–518.

[44] Penghui Wei, Nan Xu, and Wenji Mao. 2019. Modeling Conversation Structure and Temporal Dynamics for Jointly Predicting Rumor Stance and Veracity. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 4787–4798.

[45] 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*. IEEE, 651–662.

[46] 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*. 1–7.

[47] Feng Yu, Qiang Liu, Shu Wu, Liang Wang, and Tieniu Tan. 2017. A convolutional approach for misinformation identification. In *Proceedings of IJCAI*. 3901–3907.

[48] Li Zeng, Kate Starbird, and Emma S Spiro. 2016. # unconfirmed: Classifying rumor stance in crisis-related social media messages. In *Tenth International AAAI Conference on Web and Social Media*.

[49] Qiang Zhang, Shangsong Liang, Aldo Lipani, Zhaochun Ren, and Emine Yilmaz. 2019. From Stances' Imbalance to Their HierarchicalRepresentation and Detection. In *The World Wide Web Conference*. 2323–2332.

[50] 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*. 1395–1405.

[51] Xiaodan Zhu, Parinaz Sobhani, and Hongyu Guo. 2015. Long short-term memory over recursive structures. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning-Volume 37*. 1604–1612.

[52] 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 (2018), 1–36.

[53] Arkaitz Zubiaga, Elena Kochkina, Maria Liakata, Rob Procter, and Michal Lukasik. 2016. Stance Classification in Rumours as a Sequential Task Exploiting the Tree Structure of Social Media Conversations. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. 2438–2448.

[54] Arkaitz Zubiaga, Elena Kochkina, Maria Liakata, Rob Procter, Michal Lukasik, Kalina Bontcheva, Trevor Cohn, and Isabelle Augenstein. 2018. Discourse-aware rumour stance classification in social media using sequential classifiers. *Information Processing & Management* 54, 2 (2018), 273–290.

[55] Arkaitz Zubiaga, Maria Liakata, Rob Procter, Kalina Bontcheva, and Peter Tolmie. 2015. Towards detecting rumours in social media. In *Workshops at the Twenty-Ninth AAAI conference on artificial intelligence*.

[56] Arkaitz Zubiaga, Maria Liakata, Rob Procter, G Wong Sak Hoi, and Peter Tolmie. 2016. PHEME rumour scheme dataset: journalism use case. *PHEME rumour scheme dataset: journalism use case* (2016).

[57] Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS one* 11, 3 (2016), e0150989.