Coupled Hierarchical Transformer for Stance-Aware Rumor Verification in Social Media Conversations

Jianfei Yu[1], Jing Jiang[2], Ling Min Serena Khoo[3], Hai Leong Chieu[3], and Rui Xia[1]
[1] School of Artificial Intelligence, Nanjing University of Science & Technology, China
[2] School of Information Systems, Singapore Management University, Singapore
[3] DSO National Laboratories, Singapore
{jfyu, rxia}@njust.edu.cn, jingjiang@smu.edu.sg,
{klingmin, chaileon}@dso.org.sg

Abstract

The prevalent use of social media enables rapid spread of rumors on a massive scale, which leads to the emerging need of automatic rumor verification (RV). A number of previous studies focus on leveraging stance classification to enhance RV with multi-task learning (MTL) methods. However, most of these methods failed to employ pre-trained contextualized embeddings such as BERT, and did not exploit inter-task dependencies by using predicted stance labels to improve the RV task. Therefore, in this paper, to extend BERT to obtain thread representations, we first propose a Hierarchical Transformer[1], which divides each long thread into shorter subthreads, and employs BERT to separately represent each subthread, followed by a global Transformer layer to encode all the subthreads. We further propose a Coupled Transformer Module to capture the inter-task interactions and a Post-Level Attention layer to use the predicted stance labels for RV, respectively. Experiments on two benchmark datasets show the superiority of our Coupled Hierarchical Transformer model over existing MTL approaches.

1 Background

Recent years have witnessed a profound revolution in social media, as many individuals gradually turn to different social platforms to share the latest news and voice personal opinions. Meanwhile, the flourish of social media also enables rapid dissemination of unverified information (i.e., rumors) on a massive scale, which may cause serious harm to our society (e.g., impacting presidential election decisions (Allcott and Gentzkow, 2017)). Since manually checking a sheer quantity of rumors on social media is naturally labor-intensive and time-consuming, it is crucial to develop an automatic rumor verification approach to mitigate their harmful effect.

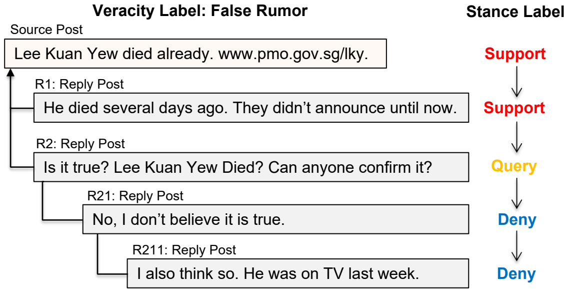

Veracity Label: False Rumor
Stance Label
Source Post
Lee Kuan Yew died already. www.pmo.gov.sg/lky.
Support
R1: Reply Post
He died several days ago. They didn't announce until now.
Support
R2: Reply Post
Is it true? Lee Kuan Yew Died? Can anyone confirm it?
Query
R21: Reply Post
No, I don't believe it is true.
Deny
R211: Reply Post
I also think so. He was on TV last week.
Deny

Figure 1: An example conversation thread with both rumor veracity label and stance labels. Each post has a stance label towards the claim in the source post, and the source claim was later identified as *false rumor*.

Rumor verification is typically defined as a task of determining whether the source claim in a conversation thread is *false rumor*, *true rumor*, or *unverified rumor* (Zubiaga et al., 2018a). In the literature, much work has been done for rumor verification (Liu et al., 2015; Ma et al., 2016; Ruchansky et al., 2017; Chen et al., 2018; Kochkina and Liakata, 2020). Among them, one appealing line of work focuses on exploiting stance signals to enhance rumor verification (Zubiaga et al., 2016), since it is observed that people's stances in reply posts usually provide important clues to rumor verification (e.g., in Fig. 1, if the source claim is denied or queried by most replies, it is highly probable that the source claim contains misinformation and is *false rumor*).

This line of work has attracted increasing attention in recent years. A number of multi-task learning (MTL) methods have been proposed to jointly perform stance classification (SC) and rumor verification (RV) over conversation threads, including Sequential LSTM-based methods (Li et al., 2019), Tree LSTM-based methods (Kumar and Carley,

[1] Note that the concept of hierarchy in this paper is different from that in Yang et al. (2016), as we use hierarchy to refer to a neural structure that first models the local interactions among posts within each subthread, followed by modeling the global interactions among all the posts in the whole thread.

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 1392–1401,
November 16–20, 2020. ©2020 Association for Computational Linguistics

2019), and Graph Convolutional Network-based methods (Wei et al., 2019). These MTL approaches are mainly constructed upon the MTL2 framework proposed in Kochkina et al. (2018), which aims to first learn shared representations with shared layers in the low level, followed by learning task-specific representations with separate stance-specific layers and rumor-specific layers in the high level.

Although these MTL approaches have shown the usefulness of stance signals to rumor verification, they still suffer from the following shortcomings: (1) The first obstacle lies in their single-task models for SC or RV, whose randomly initialized text encoders such as LSTM tend to overfit existing small annotated corpora. With the recent trend of pre-training, many pre-trained text encoders such as BERT have been shown to overcome the overfitting problem and achieve significant improvements in many NLP tasks (Devlin et al., 2019). However, unlike previous sentence-level tasks, our SC and RV tasks require the language understanding over conversation threads in social media. Since BERT is unable to process arbitrarily long sequences due to its maximum length constraint in the pre-training stage, it remains an open question how to extend BERT to our SC and RV tasks. (2) Another important limitation of previous studies lies in their multi-task learning framework. First, the MTL2 framework used in existing methods fails to explicitly model the inter-task interactions between the stance-specific and rumor-specific layers. Second, although it has been observed that people's stances in reply posts are crucial to rumor verification, the stance distributions predicted from stance-specific layers have not been utilized for rumor veracity prediction in the MTL2 framework.

To address the above two shortcomings, we explore the potential of BERT for stance-aware rumor verification, and propose a new multi-task learning model based on Transformer (Vaswani et al., 2017), named Coupled Hierarchical Transformer. Our main contributions can be summarized as follows:

- To extend BERT as our single-task model for SC and RV, we propose a Hierarchical Transformer architecture. Specifically, we first flatten all the posts in a conversation thread into a long sequence, and then decompose them evenly into multiple subthreads, each within the length constraint of BERT. Next, each subthread is encoded with BERT to capture the local interactions between posts within the subthread, and then a Transformer layer is stacked on top of all the subthreads to capture the global interactions between posts in the whole conversation thread.
- To tackle the limitations of the MTL2 framework, we first design a Coupled Transformer Module to capture the inter-task interactions between the stance-specific and the rumor-specific layers. Moreover, to utilize the stance distributions predicted for each post, we propose to concatenate them with its associated post representations, followed by a post-level attention mechanism to automatically learn the importance of each post for the final rumor verification task.

Evaluations on two benchmark datasets demonstrate the following: First, compared with existing single-task models, our Hierarchical Transformer brings consistent performance gains on Macro-F_1 for both SC and RV tasks. Second, our Coupled Hierarchical Transformer outperforms the state-of-the-art multi-task learning approach by 9.2% and 6.3% on Macro-F_1 for the two benchmarks, respectively.

2 Related Work

Stance Classification: Although stance classification has been well studied in different contexts such as online forums (Hasan and Ng, 2013; Lukasik et al., 2016; Ferreira and Vlachos, 2016; Mohammad et al., 2016), a recent trend is to study stance classification towards rumors in different social media platforms (Mendoza et al., 2010; Qazvinian et al., 2011). These studies can be roughly categorized into two groups. One line of work aims to design different features to capture the sequential property of conversation threads (Zubiaga et al., 2016; Aker et al., 2017; Pamungkas et al., 2018; Zubiaga et al., 2018b; Giasemidis et al., 2018). Another line of work attempts to apply recent deep learning models to automatically capture effective stance features (Kochkina et al., 2017; Veyseh et al., 2017). Our work extends the latter line of work by proposing a hierarchical Transformer based on the recent pre-trained BERT for this task. Moreover, we notice that our BERT-based hierarchical Transformer is similar to the model proposed in (Pappagari et al., 2019), but we want to point out that our model design in the input and output layers is specific to stance classification, which is different from their work.

Rumor Verification: Due to the negative impact of various rumors spreading on social media, rumor verification has attracted increasing attention in recent years. Existing approaches to single-task rumor verification generally belong to two groups. The first line of work focuses on either employing a myriad of hand-crafted features (Qazvinian et al., 2011; Yang et al., 2012; Kwon et al., 2013; Ma et al., 2015) including post contents, user profiles, information credibility features (Castillo et al., 2011), and propagation patterns, or resorting to various kinds of kernels to model the event propagation structure (Wu et al., 2015; Ma et al., 2017). The second line of work applies variants of several neural network models to automatically capture important features among all the propagated posts (Ma et al., 2016; Ruchansky et al., 2017; Chen et al., 2018). Different from these studies, the goal in this paper is to leverage stance classification to improve rumor verification with a multi-task learning architecture.

Stance-Aware Rumor Verification: The recent advance in rumor verification is to exploit stance information to enhance rumor verification with different multi-task learning approaches. Specifically, Ma et al. (2018a) and Kochkina et al. (2018) respectively proposed two multi-task learning architectures to jointly optimize stance classification and rumor verification based on two different variants of RNN, i.e., GRU and LSTM. More recently, Kumar and Carley (2019) proposed another multi-task LSTM model based on tree structures for stance-aware rumor verification. Our work bears the same intuition to these previous studies, and aims to explore the potential of the pre-trained BERT to this multi-task learning task.

3 Methodology

In this section, we first formulate the task of stance classification (SC) and rumor verification (RV). We then describe our single-task model for SC and RV, followed by introducing our multi-task learning framework for stance-aware rumor verification.

3.1 Task Formulation

Given a Twitter corpus, let us first use $\mathbb{D} = \{C_1, C_2, \ldots, C_{|\mathbb{D}|}\}$ to denote a set of conversation threads in the corpus. Each thread C_i is then assumed to consist of a post with the source claim S^0 and a sequence of reply posts sorted in chronological order, denoted by R^1, R^2, ... , R^N.

For the SC task, given an input thread C_i, we assume that each post (including a source post and reply posts) in the thread is annotated with a stance label towards the source claim, namely *support*, *deny*, *query*, and *comment*. Formally, let $\mathbf{s}$ = (s^0, s^1, ..., s^N) denote the sequence of stance labels, and the goal of SC is to learn a sequence classification function g: $S^0, R^1, \ldots, R^N \rightarrow s^0, s^1, \ldots, s^N$.

For the RV task, we assume that each input thread C_i is associated with a rumor label y_i, which belongs to one of the three classes, namely *false rumor*, *true rumor*, and *unverified rumor*. The goal of RV is to learn a classification function f: $C_i \rightarrow y_i$.

3.2 Hierarchical Transformer for Stance Classification and Rumor Verification

In this subsection, we present our proposed Hierarchical Transformer, which is a single-task learning framework encompassing the tasks of SC and RV. Fig. 2 illustrates the overview of our model, which mainly consists of four modules, including input thread transformation, local context encoding, global context encoding, and output layers.

Motivation: Although BERT has been widely adopted in various NLP tasks (Devlin et al., 2019), its application to our SC and RV tasks is not trivial. First, most previous studies employed BERT to obtain token-level representations for sentence or paragraph understanding, while our SC and RV tasks primarily require sentence-level representations for conversation thread understanding. Second, due to the maximum length constraint during the pre-training stage, BERT cannot be directly applied to encode arbitrarily long sequences, e.g., conversation threads in our tasks. Although truncating the input sequences is a feasible solution, it will inevitably ignore many posts that might be crucial for rumor verification.

Our main idea to address the limitations above is to divide the long sequence of a thread into shorter sequences, each within the length constraint of BERT, and to use a hierarchical model to capture the global interactions at the top layer.

Input Thread Transformation: First, to obtain post-level representations, we insert two special tokens, i.e., [CLS] and [SEP], to the beginning and the end of each post, where the [CLS] token is intended to represent the semantic meaning of the post following it. We then sort the transformed posts in each thread C_i in chronological order, followed by flattening them into a long sequence. Second, to eliminate the maximum length constraint,

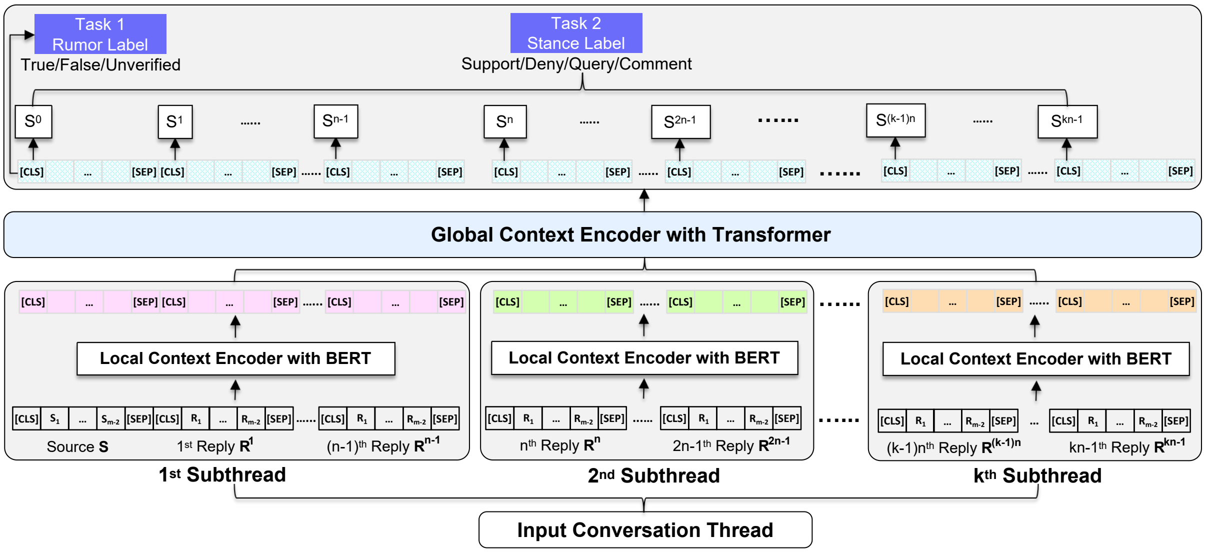


Figure 2: Our Single-Task Model (Hierarchical Transformer) for Stance Classification and Rumor Verification.

we propose to decompose the flattened sequence into multiple subthreads, so that each subthread has the same number of posts, and the sequence length of each subthread satisfies the length constraint.

Formally, let $C_i = (S^0, R^1, \ldots, R^N)$ denote the flattened thread, where S^0 is the source post, and R^j refers to the j-th reply post. As shown in the bottom of Fig. 2, we assume that C_i is decomposed into k subthreads, each subthread consists of n consecutive posts, and each post consists of m tokens[2]. For the j-th post in the thread C_i, let us use $\mathbf{P}_j = (\mathbf{x}^j_{\text{CLS}}, \mathbf{x}^j_1, \ldots, \mathbf{x}^j_{m-2}, \mathbf{x}^j_{\text{SEP}})$ to denote its input representations, where each token $\mathbf{x}$ is represented by summing up its word embeddings, segment embeddings and position embeddings. For the l-th subthread in C_i, we use $\mathbf{B}_l = (\mathbf{P}_{l0}, \mathbf{P}_{l1}, \ldots, \mathbf{P}_{l(n-1)})$ to refer to it.

Local Context Encoding (LCE): Next, we employ the pre-trained BERT to separately process the k subthreads to capture the local interactions between adjacent posts within each subthread:

$$\mathbf{h}_l = \text{BERT}(\mathbf{B}_l), \quad l = 1, 2, \ldots, k \tag{1}$$

where $\mathbf{h}_l \in \mathbb{R}^{nm \times d}$ is the hidden representation generated for the l-th subthread.

Global Context Encoding (GCE): To further capture the global interactions between all the posts in the whole conversation thread, we propose to first concatenate the hidden representations of each subthread: $\mathbf{h} = \mathbf{h}_1 \oplus \mathbf{h}_2 \oplus \ldots \oplus \mathbf{h}_k$. We then feed $\mathbf{h}$ to a standard Transformer layer as follows:

$$\widetilde{\mathbf{h}} = \text{LN}(\mathbf{h} + \text{MH-ATT}(\mathbf{h})), \tag{2}$$

$$\mathbf{H} = \text{LN}(\widetilde{\mathbf{h}} + \text{FFN}(\widetilde{\mathbf{h}})), \tag{3}$$

where MH-ATT and FFN respectively refer to the multi-head self-attention and the feed forward network (Vaswani et al., 2017), and LN refers to layer normalization (Ba et al., 2016).

Output Layers: Based on the global hidden representation $\mathbf{H}$, we further stack the output layers to make predictions for SC and RV, respectively. Specifically, for the SC task, we treat the hidden state of the j-th [CLS] token as the representation for the j-th post, followed by adding a softmax layer to classify its stance towards the source claim:

$$p(s^j \mid \mathbf{H}^j_{\text{CLS}}) = \text{softmax}(\mathbf{W}_s^\top \mathbf{H}^j_{\text{CLS}} + \mathbf{b}_s), \tag{4}$$

where $\mathbf{W}_s \in \mathbb{R}^{d \times 4}$ and $\mathbf{b}_s \in \mathbb{R}^4$ are learnable parameters. Moreover, for the RV task, we add a softmax layer over the last hidden state of the first [CLS] token for rumor veracity prediction:

$$p(y \mid \mathbf{H}^0_{\text{CLS}}) = \text{softmax}(\mathbf{W}_r^\top \mathbf{H}^0_{\text{CLS}} + \mathbf{b}_r), \tag{5}$$

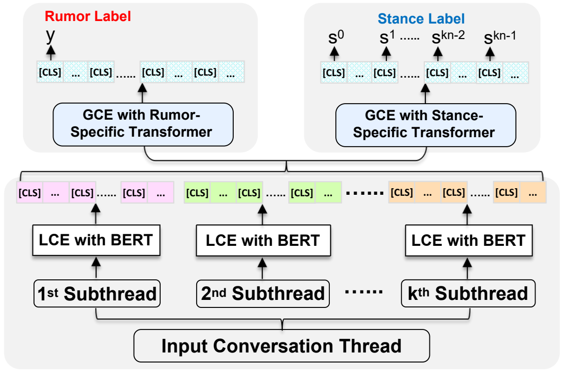
where $\mathbf{W}_r \in \mathbb{R}^{d \times 3}$ and $\mathbf{b}_r \in \mathbb{R}^3$ are weight and bias parameters.

3.3 Coupled Hierarchical Transformer for Stance-Aware Rumor Verification

Based on the above single-task model (i.e., Hierarchical Transformer), we describe our proposed multi-task learning (MTL) framework for stance-aware rumor verification in this subsection.

Baseline MTL Framework: To exploit the stance signals for rumor verification, a widely used MTL

[2] Note that for parallel computing, each post is padded or truncated to have the same number of tokens, i.e., m, and each subthread is padded to have the same number of posts, i.e., n.



Rumor Label
y
[CLS] ... [CLS] ...... [CLS] ... [CLS] ...
GCE with Rumor-Specific Transformer
Stance Label
s0 s1 ...... skn-2 skn-1
GCE with Stance-Specific Transformer
LCE with BERT
LCE with BERT
LCE with BERT
1st Subthread
2nd Subthread
......
kth Subthread
Input Conversation Thread

Figure 3: Baseline Multi-Task Learning Framework (MTL2) for Stance-Aware Rumor Verification.

framework is the MTL2 model proposed in Kochkina et al. (2018), which assumes that the SC and RV tasks share the low-level neural layers but the high-level layers are specific to each task. As illustrated in Fig. 3, to adapt our Hierarchical Transformer to this MTL2 framework, we propose to share the input and LCE modules between SC and RV, followed by employing separate GCE and output modules for these two tasks, respectively.

Motivation: However, as mentioned before, this baseline MTL framework has two major limitations. First, it fails to consider the inter-task interaction. Since the GCE module in SC is supervised to capture salient stance-specific features such as *no doubt*, *agree* and *fake news*, these features can be leveraged to guide the GCE module in RV to capture those important rumor-specific features closely related to stance features. Moreover, since both stance-specific and rumor-specific features are intuitively crucial to RV, it is necessary to effectively integrate them. Second, it ignores the sequential stance labels predicted from the output module in SC. Actually, the predicted stance distributions for each post can capture the temporal evolution of public stances towards the source claim, which may reflect indicative clues for veracity prediction.

Coupled Transformer Module: To model inter-task interactions, we devise a Coupled Transformer Module with two coupled components in Fig. 4: a stance-specific Transformer and a cross-task Transformer.

Concretely, we first employ a standard Transformer layer (i.e., Eqn (2) and Eqn (3)) to obtain stance-specific representations $\mathbf{P}$ in the right channel. Next, to learn the inter-task interactions in the left channel, we design a multi-head stance-aware attention mechanism (MH-SATT) by treating $\mathbf{P}$ as queries, and $\mathbf{h}$ as keys and values, which essentially leverages stance-specific features in $\mathbf{P}$ to guide our model to pay more attention to stance-aware rumor-specific features. Specifically, the i-th head of MH-SATT is defined as follows:

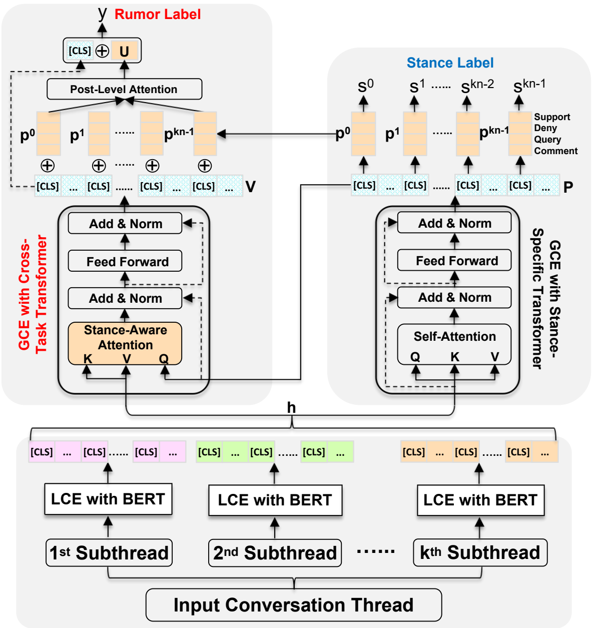


Figure 4: Our Multi-Task Learning Framework (Coupled Hierarchical Transformer) for Stance-Aware Rumor Verification.

$$\mathrm{SATT}_i(\mathbf{P},\mathbf{h}) = \mathrm{softmax}(\frac{[\mathbf{W_q P}]^\top [\mathbf{W_k h}]}{\sqrt{d/z}})[\mathbf{W_v h}]^\top, \quad (6)$$

where $\{\mathbf{W_q}, \mathbf{W_k}, \mathbf{W_v}\} \in \mathbb{R}^{d/z\times d}$ are parameters, and z is the number of heads.

Moreover, to integrate stance-specific and rumor-specific features, we propose to add a layer norm together with a residual connection as follows:

$$\widetilde{\mathbf{V}} = \mathrm{LN}(\mathbf{P} + \text{MH-SATT}(\mathbf{P},\mathbf{h})). \quad (7)$$

Finally, we add a feed-forward network and a layer normalization to get the rumor-stance hybrid representations $\mathbf{V}$:

$$\mathbf{V} = \mathrm{LN}(\widetilde{\mathbf{V}} + \mathrm{FFN}(\widetilde{\mathbf{V}})). \quad (8)$$

Post-Level Attention with Stance Labels: To address the second limitation, we propose to concatenate each post's stance distribution and its corresponding hidden representation, followed by a post-level attention layer to automatically learn the importance of each post.

Specifically, as shown in Fig. 4, we first use Eqn (4) to predict the stance distribution of the j-th post in the right channel, denoted by $\mathbf{p}^j$. We then treat the hybrid representation of the j-th [CLS] token (i.e., $\mathbf{V}^j_{\mathrm{CLS}}$) as the representation of the j-th post, and concatenate it with $\mathbf{p}^j$, followed by feeding them to a post-level attention layer to obtain the stance label-aware thread representation $\mathbf{U}$:

			Stance Labels				Rumor Veracity Labels		
Dataset	#Threads	#Tweets	*#Support*	*#Deny*	*#Query*	*#Comment*	*#True*	*#False*	*#Unverified*
SemEval-17	325	5,568	1,004	415	464	3,685	145	74	106
PHEME	2,402	105,354			-		1,067	638	697

Table 1: Basic statistics of the SemEval-2017 dataset and the PHEME dataset.

$$u_j = \mathbf{v}^\top \tanh\Big(\mathbf{W}_h(\mathbf{V}^j_{\text{CLS}} \oplus \mathbf{p}_j)\Big), \tag{9}$$

$$\alpha_j = \frac{\exp(u_j)}{\sum_{l=1}^{N}\exp(u_l)}, \tag{10}$$

$$\mathbf{U} = \sum_{j=1}^{N} \alpha_j(\mathbf{V}^j_{\text{CLS}} \oplus \mathbf{p}_j). \tag{11}$$

Output Layers: Finally, since $\mathbf{V}^0_{\text{CLS}}$ and $\mathbf{U}$ can be considered as the token-level thread representation and the post-level thread representation respectively, we propose to concatenate them to predict the veracity label of the source claim:

$$p(y \mid \mathbf{V}^0_{\text{CLS}}, \mathbf{U}) = \text{softmax}\Big(\mathbf{W}^\top(\mathbf{V}^0_{\text{CLS}} \oplus \mathbf{U}) + \mathbf{b}\Big), \tag{12}$$

where $\mathbf{W} \in \mathbb{R}^{(2d+4)\times 3}$ and $\mathbf{b} \in \mathbb{R}^3$ are weight and bias terms.

Model Training: To optimize all the parameters in our Coupled Hierarchical Transformer, we adopt the alternating optimization strategy to minimize the following objective function, which is a combination of the cross-entropy loss of the two tasks:

$$\begin{aligned}\mathcal{J} = -\Big(&\frac{1}{M}\sum_{i=1}^{M}\log p(y_i \mid \mathbf{V}^0_{\text{CLS}}, \mathbf{U}) \\ &+ \frac{1}{M'}\sum_{k=1}^{M'}\sum_{j=1}^{N}\log p(s^j \mid \mathbf{P}^j_{\text{CLS}})\Big),\end{aligned} \tag{13}$$

where M and M' refer to the number of samples for the tasks of RV and SC, respectively.

4 Experiments

In this section, we first evaluate our single-task model on both stance classification (SC) and rumor verification (RV), followed by evaluating our multi-task learning model on RV. Finally, we perform further analysis to provide deeper insights into our proposed multi-task learning model.

4.1 Experiment Setting

Dataset: To demonstrate the effectiveness of our proposed approaches, we carry out experiments on two benchmark datasets, i.e., **SemEval-2017** and **PHEME**. Table 1 shows the basic statistics of the two datasets.

Specifically, **SemEval-2017** is a widely used dataset from SemEval-2017 Challenge Task 8, which contains 325 Twitter conversation threads discussing rumors (Derczynski et al., 2017). The dataset has been split into training, development, and test sets, where the former two sets are related to eight events and the test set covers two additional events. Since each thread is annotated with a rumor veracity label and each post in the thread is annotated with its stance towards the source claim, this dataset is used for evaluating both SC and RV tasks in this work.

PHEME is a well known dataset for RV, which contains 2402 Twitter conversation threads discussing nine events. For fair comparison with existing approaches, we perform cross-validation experiments based on leave-one-event-out settings: for each fold, all the threads related to one event are used for testing, and all the threads related to the other eight events are used for training. Following previous studies (Kochkina et al., 2018; Wei et al., 2019), **PHEME** is only used for evaluating the performance of RV.

Since the class distribution of the two datasets are imbalanced, we employ Macro-F_1 as the main evaluation metric and accuracy as the secondary evaluation metric for both tasks.

Parameter Settings: Our models are based on the pre-trained uncased *BERT$_{base}$* model (Devlin et al., 2019), where the number of BERT layers is 12 and the number of attention heads is z = 12. Moreover, for both Hierarchical Transformer and Coupled Hierarchical Transformer, we set the learning rate as 5e-5, and the dropout rate as 0.1. Due to memory limitation, for each conversation thread, the number of subthreads is set to k = 6, and the maximum input length of each subthread is set as 512. For each subthread, the number of posts is set to n = 17, and the number of tokens in each post is fixed to m = 30. Moreover, the batch size is respectively set as 4 and 2 for Hierarchical Transformer and Coupled Hierarchical Transformer, respectively. We implement all the models based on PyTorch with a *24GB NVIDIA TITAN RTX* GPU.

	Single Stance Type Evaluation				Overall Evaluation	
Method	Support-F_1	Deny-F_1	Query-F_1	Comment-F_1	Macro-F_1	Accuracy
SVM (Pamungkas et al., 2018)	0.410	0.000	0.580	**0.880**	0.470	0.795
BranchLSTM (Kochkina et al., 2018)	0.403	0.000	0.462	0.873	0.434	0.784
Temporal ATT (Veyseh et al., 2017)	-	-	-	-	0.482	**0.820**
Conversational-GCN (Wei et al., 2019)	0.311	0.194	**0.646**	0.847	0.499	0.751
Hierarchical Transformer (Ours)	**0.421**	**0.255**	0.520	0.841	**0.509**	0.763

Table 2: Results of stance classification on the SemEval-2017 dataset.

		SemEval-2017 Dataset		PHEME Dataset	
Setting	Method	Macro-F_1	Accuracy	Macro-F_1	Accuracy
Single-Task	BranchLSTM (Kochkina et al., 2018)	0.491	0.500	0.259	0.314
	TD-RvNN (Ma et al., 2018b)	0.509	0.536	0.264	0.341
	Hierarchical GCN-RNN (Wei et al., 2019)	0.540	0.536	0.317	0.356
	HiTPLAN (Khoo et al., 2020)	0.581	0.571	0.361	0.438
	Hierarchical Transformer (Ours)	**0.592**	**0.607**	**0.372**	**0.441**
Multi-Task	BranchLSTM+NileTMRG (Kochkina et al., 2018)	0.539	0.570	0.297	0.360
	MTL2 (Veracity+Stance) (Kochkina et al., 2018)	0.558	0.571	0.318	0.357
	Hierarchical PSV (Wei et al., 2019)	0.588	0.643	0.333	0.361
	MTL2-Hierarchical Transformer (Ours)	0.657	0.643	0.375	0.454
	Coupled Hierarchical Transformer (Ours)	**0.680**†	**0.678**†	**0.396**†	**0.466**†

Table 3: Results of rumor veracity prediction. Single-Task indicates that stance labels are not used during the training stage. † indicates that our Coupled Hieararchical Transformer model is significantly better than the best compared system with p-value $<$ 0.05 based on McNemar's significance test.

4.2 Main Results

4.2.1 Evaluation on Single-Task Models

In this subsection, we compare our proposed Hierarchical Transformer with existing single-task models for SC and RV, respectively.

Stance Classification (SC): We first consider the following competitive approaches that focus on SC only: (1) *SVM* is a baseline method that feeds conversation-based and affective-based features to linear SVM (Pamungkas et al., 2018); (2) *BranchLSTM* is an LSTM-based architecture designed by Kochkina et al. (2018), which focuses on modeling the sequential branches in each thread; (3) *Temporal ATT* is an attention-based model proposed by Veyseh et al. (2017), which treats each post's adjacent posts in a conversation timeline as its local context, followed by employing attention mechanism over the local context to learn the importance of each adjacent post; (4) *Conversational GCN* is the state-of-the-art approach recently proposed by Wei et al. (2019), which leverages graph convolutional network to model the relations between posts in each thread.

We report the SC results in Table 2. First, it is clear to observe that our Hierarchical Transformer model performs much better than all the compared systems on Macro-F_1. Second, compared with previous approaches, our model shows its strong capability of detecting posts belonging to the *support* and *deny* stances. This is crucial for veracity prediction, because the *support* and *deny* stances usually provide important clues to identify the *true* and *false* rumors respectively (see Fig. 5). All these observations demonstrate the general effectiveness of our Hierarchical Transformer model.

Rumor Verification (RV): We then consider several competitive systems that focus on RV only: (1) *RvNN* is a recursive neural network model based on top-down tree structure, which is proposed by Ma et al. (2018b); (2) *Hierarchical GCN-RNN* is a variant of *Conversational GCN* for veracity prediction; (3) *PLAN* is the state-of-the-art approach recently proposed by Khoo et al. (2020), which uses a randomly initialized Transformer to encode each conversation thread.

We report the RV results of compared systems on **SemEval-2017** and **PHEME** in the top part of Table 3. First, compared with earlier methods for RV, we observe that our Hierarchical Transformer model gains significant improvements, outperforming *Hierarchical GCN-RNN* by 5.2 and 5.5 absolute percentage points on Macro-F_1 for the two datasets, respectively. Second, even compared with the recent state-of-the-art model *PLAN*, our model

Methods (TFM: Transformer)	Macro-F_1	Accuracy
Hierarchical TFM	**0.372**	**0.441**
- Truncating Input & Removing Global TFM	0.354	0.409
Coupled Hierarchical TFM	**0.396**	**0.466**
- Removing Post-Level Attention	0.385	0.430
- Replacing Cross-Task TFM with TFM	0.390	0.456

Table 4: Ablation study on the PHEME dataset.

can still bring moderate performance gains on the two datasets. Since *PLAN* is based on randomly initialized Transformer whereas our model is based on pre-trained Transformer (i.e., BERT), this shows the usefulness of employing pre-trained models for RV, which agrees with our first motivation.

4.2.2 Evaluation on Multi-Task Models

In this subsection, we evaluate the effectiveness of our Coupled Hierarchical Transformer model, and consider several multi-task learning frameworks for stance-aware rumor verification: (1) *BranchLSTM+NileTMRG* is a pipeline approach, which first trains a *BranchLSTM* model for SC, followed by a SVM classifier for RV (Kochkina et al., 2018); (2) *MTL2* is the MTL framework proposed in (Kochkina et al., 2018), which shares a single LSTM channel but uses two separate output layers for SC and RV, respectively; (3) *Hierarchical PSV* is a hierarchical model proposed by (Wei et al., 2019), which first learns content and stance features via *Conversational-GCN*, followed by exploiting temporal evolution for RV via *Stance-Aware RNN*; (4) *MTL2-Hierarchical Transformer* is our adapted MTL2 model which is introduced in Section 3.3.

In the bottom part of Table 3, we can first find that all the multi-task learning models achieve better performance than their corresponding single-task baselines across the two datasets, which verifies the usefulness of stance signals for RV. Second, among all the multi-task learning approaches, it is clear to observe that our Coupled Hierarchical Transformer model consistently achieves the best results on both **SemEval-2017** and **PHEME**, which outperforms the second best method by 2.3 and 2.1 absolute percentage points on Macro-F_1 for the two datasets, respectively. These observations show the superiority of our proposed model over previous multi-task learning methods for stance-aware rumor verification.

4.3 Ablation Study

To examine the impact of each key component in our single-task and multi-task approaches, we further perform ablation study in this subsection.

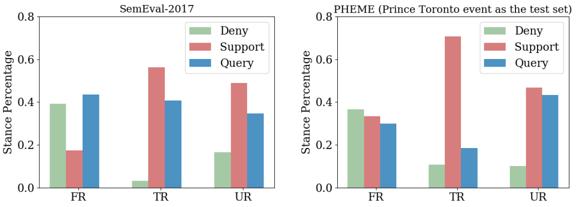


Figure 5: Correlation between predicted stance classes (y-axis) and predicted rumor labels (x-axis) from our Coupled Hierarchical Transformer on test sets of our two datasets.

As shown in Table 4, for our proposed Hierarchical Transformer, we can see that if we directly apply BERT to our RV task (i.e., truncating the input thread and removing the global Transformer layer), the performance will drop significantly. This is in line with our first motivation, and also demonstrates the effectiveness of our proposed model.

Moreover, for our multi-task learning framework (i.e., Coupled Hierarchical Transformer), the post-level attention layer shows its indispensable role because of the significant performance drop after removal. Meanwhile, replacing our cross-task Transformer with the standard Transformer will lead to moderate performance drop in both datasets, which also suggests its importance to our full model.

4.4 Correlation Between Predicted Stance Labels and Veracity Labels

To better understand the usefulness of stance signals to veracity prediction in our Coupled Hierarchical Transformer, we first analyze the correlation between predicted stance classes and predicted veracity labels on our two datasets. Since the *comment* stance is not crucial for rumor verification, we focus on the other three stance classes, i.e., *deny*, *query*, and *support*.

As shown in Fig. 5, we can clearly see that *true rumor* is more closely associated with the *support* stance, whereas *false rumor* is generally dominated by the other two stances *deny* and *query*. This suggests that our multi-task learning model has implicitly learnt that the stance signal can provide important clues to rumor verification.

Case Study: To provide deeper insights into our Coupled Hierarchical Transformer, we carefully choose one representative sample from our test set, and show the stance and veracity prediction results as well as the attention weights of each post learnt in the post-level attention layer. Due to space limitation, we only show five posts with the top-5 attention weights in the thread.

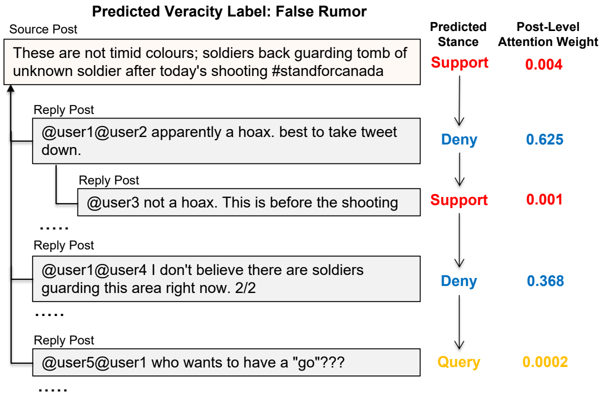

Predicted Veracity Label: False Rumor
Source Post
These are not timid colours; soldiers back guarding tomb of unknown soldier after today's shooting #standforcanada
Predicted Stance
Post-Level Attention Weight
Support
0.004
Reply Post
@user1@user2 apparently a hoax. best to take tweet down.
Deny
0.625
Reply Post
@user3 not a hoax. This is before the shooting
Support
0.001
.....
Reply Post
@user1@user4 I don't believe there are soldiers guarding this area right now. 2/2
Deny
0.368
.....
Reply Post
@user5@user1 who wants to have a "go"???
Query
0.0002
.....

Figure 6: Stance classes and rumor labels predicted by Coupled Hierarchical Transformer on a test sample in PHEME dataset.

In Fig. 6, we can see that although the source claim is supported by some replies, our model learns to pay much higher attention weights to the two posts with *deny* stance while primarily ignoring the other posts, which may help our model correctly predict its veracity label as *false rumor*.

5 Conclusion

In this paper, we first examined the limitations of existing approaches to stance classification (SC) and rumor verification (RV). To tackle these limitations, we first proposed a single-task model (i.e., Hierarchical Transformer) for SC and RV, followed by designing a multi-task learning framework with a Coupled Transformer module to capture inter-task interactions and a Post-Level Attention Layer to use stance distributions for the RV task. Experiments on two benchmarks show the effectiveness of our single-task and multi-task learning methods.

Acknowledgments

We would like to thank three anonymous reviewers for their valuable comments. This research is supported by DSO grant DSOCL18009, the Natural Science Foundation of China (No. 61672288, 62076133, and 62006117), and the Natural Science Foundation of Jiangsu Province for Young Scholars (SBK2020040749) and Distinguished Young Scholars (SBK2020010154).

References

Ahmet Aker, Leon Derczynski, and Kalina Bontcheva. 2017. Simple open stance classification for rumour analysis. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 31–39.

Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–36.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of WWW*.

Tong Chen, Xue Li, Hongzhi Yin, and Jun Zhang. 2018. Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection. In *Proceedings of PAKDD*.

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 SemEval*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*.

William Ferreira and Andreas Vlachos. 2016. Emergent: a novel data-set for stance classification. In *Proceedings of NAACL*.

Georgios Giasemidis, Nikolaos Kaplis, Ioannis Agrafiotis, and Jason Nurse. 2018. A semi-supervised approach to message stance classification. *IEEE TKDE*, 32(1):1–11.

Kazi Saidul Hasan and Vincent Ng. 2013. Stance classification of ideological debates: Data, models, features, and constraints. In *Proceedings of IJCNLP*.

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 AAAI*.

Elena Kochkina and Maria Liakata. 2020. Estimating predictive uncertainty for rumour verification models. In *Proceedings of ACL*.

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 SemEval*.

Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. 2018. All-in-one: Multi-task learning for rumour verification. In *Proceedings of COLING*.

Sumeet Kumar and Kathleen Carley. 2019. Tree LSTMs with convolution units to predict stance and rumor veracity in social media conversations. In *Proceedings of ACL*.

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

Quanzhi Li, Qiong Zhang, and Luo Si. 2019. Rumor detection by exploiting user credibility information, attention and multi-task learning. In *Proceedings of ACL*.

Xiaomo Liu, Armineh Nourbakhsh, Quanzhi Li, Rui Fang, and Sameena Shah. 2015. Real-time rumor debunking on twitter. In *Proceedings of CIKM*.

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 ACL*.

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 *Proceedings of IJCAI*.

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 CIKM*.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2017. Detect rumors in microblog posts using propagation structure via kernel learning. In *Proceedings of ACL*.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2018a. Detect rumor and stance jointly by neural multi-task learning. In *Companion Proceedings of the The Web Conference 2018*.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2018b. Rumor detection on twitter with tree-structured recursive neural networks. In *Proceedings of ACL*.

Marcelo Mendoza, Barbara Poblete, and Carlos Castillo. 2010. Twitter under crisis: can we trust what we rt? In *Proceedings of the 3rd Workshop on Social Network Mining and Analysis, SNAKDD 2009, Paris, France, June 28, 2009*.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of SemEval*.

EW Pamungkas, V Basile, and V Patti. 2018. Stance classification for rumour analysis in twitter: Exploiting affective information and conversation structure. In *2nd International Workshop on Rumours and Deception in Social Media (RDSM 2018)*, volume 2482, pages 1–7.

Raghavendra Pappagari, Piotr Zelasko, Jesús Villalba, Yishay Carmiel, and Najim Dehak. 2019. Hierarchical transformers for long document classification. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*.

Vahed Qazvinian, Emily Rosengren, Dragomir R Radev, and Qiaozhu Mei. 2011. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of EMNLP*.

Natali Ruchansky, Sungyong Seo, and Yan Liu. 2017. Csi: A hybrid deep model for fake news detection. In *Proceedings of CIKM*.

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 *Proceedings of NIPS*.

Amir Pouran Ben Veyseh, Javid Ebrahimi, Dejing Dou, and Daniel Lowd. 2017. A temporal attentional model for rumor stance classification. In *Proceedings of CIKM*.

Penghui Wei, Nan Xu, and Wenji Mao. 2019. Modeling conversation structure and temporal dynamics for jointly predicting rumor stance and veracity. In *Proceedings of EMNLP*.

Ke Wu, Song Yang, and Kenny Q. Zhu. 2015. False rumors detection on sina weibo by propagation structures. In *Proceedings of ICDE*.

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*, MDS '12, pages 13:1–13:7.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of NAACL*.

Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018a. Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys (CSUR)*, 51(2):32.

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

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):e0150989.