Stance-Conditioned Modeling for Rumor Verification

Gibson Nkhata, Susan Gauch
Department of Electrical Engineering and Computer Science
University of Arkansas
Fayetteville, AR 72701, USA
e-mails: {gnkhata, sgauch}@uark.edu

***Abstract*—The rapid spread of misinformation on social media platforms has heightened the need for effective rumor verification models. Traditional approaches primarily rely on textual content and transformer-based embeddings, but they often fail to incorporate conversational dynamics and stance evolution, limiting their effectiveness. We present a stance-conditioned rumor verification model that integrates Bidirectional Encoder Representations from Transformers (BERT) based source post embeddings, reply post embedding aggregation, and Bidirectional Long Short Term Memory (BiLSTM) encoding of stance labels to enhance rumor classification. By explicitly modeling stance progression and leveraging aggregated stance-conditioned reply embeddings, our approach captures critical discourse patterns that influence rumor veracity. Experiments on competitive benchmark tasks demonstrate that our model outperforms state-of-the-art baselines in Macro-F1 and accuracy, achieving superior performance across multiple datasets. Ablation studies confirm the effectiveness of each constituent model component, with early rumor detection analysis showcasing our model's ability to detect misinformation faster and more accurately than competing methods. Overall, this work presents a novel stance-conditioned approach to rumor verification that effectively captures conversational context and discourse interactions, providing a more robust and interpretable framework for combating online misinformation.**

***Keywords*-Rumor verification; stance-conditioned modeling; social media misinformation; embedding aggregation.**

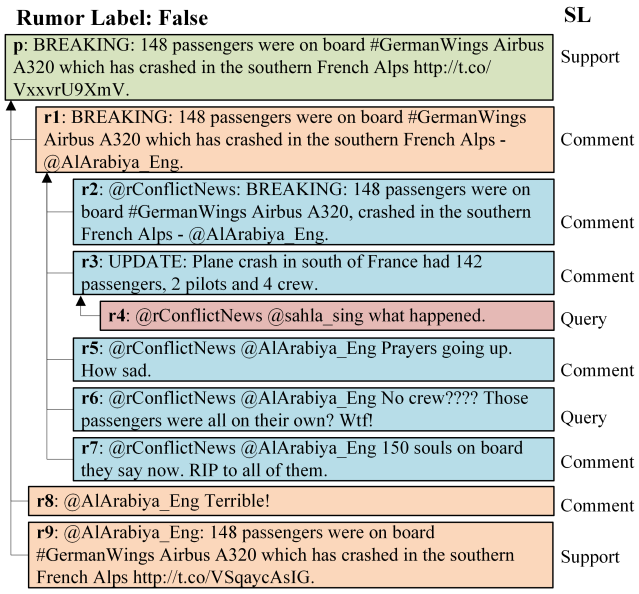

Figure 1. A sample thread C with a false veracity label. SL stands for Stance Labels.

I. Introduction

The exponential rise of social media platforms such as Twitter (rebranded as X) and Reddit has fueled the rapid spread of misinformation and rumors [1][2], making rumor verification a critical challenge. Traditional approaches primarily rely on transformer-based language models, such as Bidirectional Encoder Representations from Transformers (BERT) and Generative Pre-trained Transformer (GPT) [3][4] to analyze textual representations of posts. However, these methods often truncate conversational threads due to sequence length constraints and overlook valuable discourse signals, such as stance labels, that reflect user perspectives on rumors.

This work presents an enhanced rumor verification framework that effectively integrates the structure and stance dynamics of online discussions. Our approach builds upon prior work by incorporating stance labels as additional input features, embedded using a Bidirectional Long Short-Term Memory (BiLSTM) [5] network. Specifically, we extract source post embeddings using BERT [6] and concatenate them with stance-conditioned reply embeddings, where stance labels are sequentially modeled based on their temporal order in the conversation thread. The resulting feature representations are processed through a unified feed-forward layer for final classification.

Unlike prior studies that primarily rely on direct textual features [7][8], our model explicitly encodes stance signals, allowing it to capture the argumentative structure within rumor propagation. By preserving the full conversational context and avoiding truncation, our model offers a more holistic understanding of rumor veracity. Empirical results on benchmark rumor datasets demonstrate that our method significantly improves performance distinguishing rumor veracity classes, setting a new benchmark for rumor detection systems.

Furthermore, this study explores the task of early rumor detection, that focuses on identifying and assessing the veracity of emerging rumors in real-time as they propagate online. By detecting rumors at an early stage, this approach aims to mitigate the rapid spread of misinformation, enabling timely interventions and fact-checking before false narratives gain widespread traction. Figure 1 presents a sample discourse, showcasing how stances evolve. We leverage the evolution to model the temporal ordering of stance annotations with BiLSTM. The major contributions of this work are outlined as follows:

- **Novel rumor verification framework**: Presents a methodology that integrates BERT-based post embeddings and BiLSTM-based stance encoding to enhance rumor verification in conversational threads.
- **Avoiding sequence truncation**: Unlike prior approaches that truncate long conversation threads due to BERT's sequence length constraints, our model effectively aggregates embeddings without discarding crucial discourse information.
- **Leveraging stance labels**: Incorporates stance labels as an additional input feature, embedding them using a BiLSTM to capture the sequential stance evolution within a thread.
- **Early rumor detection**: Evaluates the model's ability to detect rumors at an early stage of the conversation, highlighting its real-world applicability for misinformation mitigation.

The rest of this paper unfolds as follows. Section II reviews the existing literature on rumor verification, and Section III delves into a comprehensive description of our approach. Section IV demonstrates experiments and provides a discussion of results. Finally, Section V presents the conclusion.

II. Related work

The proliferation of misinformation on social media has led to extensive research in rumor verification. Early studies primarily focused on content-based analysis, utilizing textual features and user metadata to assess veracity. Nonetheless, these approaches often overlooked the dynamic nature of conversations and the valuable insights provided by user stances within discussion threads.

Recently, Yang et al. [9] introduced a weakly supervised propagation model that leverages multiple instance learning for joint rumor verification and stance detection. This approach models the diffusion of claims through bottom-up and top-down trees, capturing the propagation structure of rumors. The model requires only bag-level labels concerning a claim's veracity, reducing the need for extensive labeled data. Experiments demonstrated promising performance in both claim-level rumor detection and post-level stance classification. Furthermore, Mai et al. [10] introduces a graph attention mechanism to effectively capture and process interactions within a conversational thread.

Jami et al. [11] conducted a comprehensive literature review on rumor stance classification in online social networks. They highlighted the importance of user viewpoints in predicting rumor veracity and discussed various approaches, datasets, and challenges in the field. The study emphasized the need for models that effectively utilize user stances to improve rumor verification systems.

Moreover, Khandelwal [12] explored a multi-task learning framework that jointly predicts rumor stance and veracity. By fine-tuning the Longformer model, the study addressed the limitations of sequence length in traditional transformer models, allowing for the processing of longer conversational threads without truncation. This approach underscored the benefits of handling extended contexts in rumor verification tasks.

Despite these advancements, challenges remain in effectively modeling the temporal dynamics of conversations and fully leveraging stance information. In addition, most of these models still suffer from sequence length limitations (since they rely on pretrained language models), often truncating crucial replies within a discourse. Our framework aims to address these gaps by proposing a sequential stance aggregation mechanism that accounts for the temporal ordering of replies and embedding stance labels using a BiLSTM [5] network, preserving the chronological order of replies. This method seeks to capture the evolution of discussions more effectively, providing a comprehensive understanding of rumor propagation and verification.

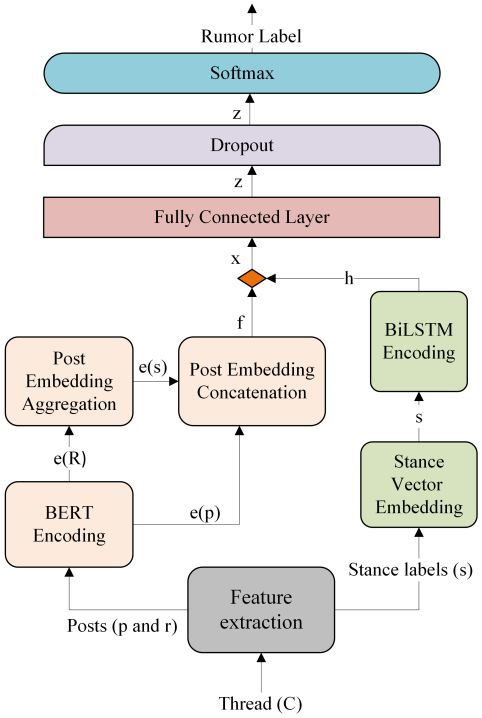

Figure 2. The model framework. $e(R)$, $e(p)$, and $e(s)$ represent e_R, e_p and e_s, respectively, in the main text.

III. Methodology

Our model consists of three main components: 1) **Post embedding representation**: BERT extracts contextual embeddings for the source and reply posts. 2) **Stance-aware sequence encoding**: A BiLSTM encodes the sequence of stance labels in temporal order. 3) **Unified feed-forward layer**: The post embeddings and stance representations are concatenated and fed into a classifier. Figure 2 illustrates our methodology.

A. Task Formulation

Given a conversational thread C (see Figure 1) consisting of a source post p and a set of reply posts $R = \{r_1, r_2, ..., r_n\}$, where n is the total number of reply posts, the goal of rumor verification is to classify the source post p into one of three categories: $y_c \in \{\text{true rumor}, \text{false rumor}, \text{unverified rumor}\}$. Each post (both p and r_i) is associated with a stance label s_i, where $s_i \in \{\text{support}, \text{deny}, \text{query}, \text{comment}\}$.

B. Post Embedding Representation and Aggregation

Each post x_i (both p and r_i) is tokenized and passed through a pre-trained BERT model. The mean-pooled hidden states are used as the post embedding:

$$\mathbf{e}_i = \text{BERT}(x_i) = \frac{1}{T}\sum_{t=1}^{T} h_t \tag{1}$$

where h_t represents the hidden state at position t of a given post, and T is the sequence length.

The source post embedding is:

$$\mathbf{e}_p = \text{BERT}(p) \tag{2}$$

The reply post embeddings are:

$$\mathbf{E}_R = \{\mathbf{e}_{r_1}, \mathbf{e}_{r_2}, ..., \mathbf{e}_{r_N}\} \tag{3}$$

To preserve stance information, we aggregate reply embeddings based on stance labels:

$$\mathbf{e}_s = \sum_{r_i \in R_s} \mathbf{e}_{r_i} \tag{4}$$

where R_s represents the set of replies with stance s. After aggregating embeddings for all four stances, these vectors are concatenated with the embedding of the source post to create a composite feature vector:

$$f = [e_p; e_s; e_d; e_q; e_c], \tag{5}$$

where subscripts (p, s, d, q, c) represent (source, support, deny, query, comment). Aggregating embeddings by stance allows the model to capture the distribution of opinions within a conversation thread. This method emphasizes the collective influence of each stance category, providing nuanced insights into the overall sentiment and credibility of the information. Prior research has highlighted the importance of analyzing specific stances, such as denial and questioning, in rumor detection, as they play a crucial role in assessing veracity [13]. On the same note, embedding aggregation addresses the challenge of thread sequence truncation, a common limitation in large language model-based approaches. By aggregating embeddings in this manner, the model can better discern patterns indicative of true, false, or unverified rumors.

C. Stance Label Encoding Using BiLSTM

Recent studies have demonstrated the efficacy of BiLSTMs in stance detection tasks. For instance, Jia et al. [14] proposed an improved BiLSTM approach that integrates external commonsense knowledge and environmental information to enhance user stance detection. Their method effectively captures the temporal progression of user viewpoints, leading to improved detection performance. Deviating from their approaches, in this work, each reply's stance label is first embedded into a continuous vector space:

$$\mathbf{s}_i = \text{Embed}(s_i) \tag{6}$$

These embedded stance vectors are then processed chronologically through a BiLSTM network:

$$\overrightarrow{h}_i, \overleftarrow{h}_i = \text{BiLSTM}(\mathbf{s}_i) \tag{7}$$

The final stance representation is obtained from the last hidden states:

$$\mathbf{h}_S = [\overrightarrow{h}_N; \overleftarrow{h}_N] \tag{8}$$

The utilization of a BiLSTM for encoding stance labels offers three primary advantages in this study. First, **capturing sequential dependencies**: conversations on social media often exhibit temporal dynamics, where the stance of a reply can influence and be influenced by preceding and subsequent replies. A BiLSTM processes the sequence in both forward and backward directions, effectively capturing these dependencies. This bidirectional processing ensures that the context from both past and future replies is considered, leading to a more comprehensive understanding of the stance dynamics within a thread. Next, **handling variable-length sequences**: social media threads vary in length and complexity. BiLSTMs are adept at managing such variability, allowing the model to process each thread appropriately without the need for strict length constraints. Finally, **enhanced contextual representation**: by encoding stance labels through a BiLSTM, the model generates contextually enriched representations that encapsulate the interplay between different stances over the course of the conversation. This enriched representation aids in distinguishing subtle nuances in stance expressions, that is crucial for accurate rumor verification.

D. Classification Layer

The final input to the classifier is the concatenation of the source post embedding, aggregated reply embeddings, and stance representation:

$$\mathbf{x} = [f; \mathbf{h}_S] \tag{9}$$

The classification module comprises a fully connected layer that projects the high-dimensional representation $\mathbf{x}$ onto the output space corresponding to the rumor classes (True, False, Unverified):

$$\mathbf{z} = \text{Dropout}\left(\mathbf{W}\mathbf{x} + \mathbf{b}\right), \tag{10}$$

where $\mathbf{W} \in \mathbb{R}^{C \times D}$ is a learnable weight matrix responsible for transforming the hidden representation $\mathbf{x}$ into the output space of *C* classes, and $\mathbf{b} \in \mathbb{R}^{C}$ denotes the bias term. Here, *D* represents the dimensionality of $\mathbf{x}$, while the number of classes is given by *C* = 3. $Dropout$ is used for regularization. The raw output $\mathbf{z}$ is subsequently passed through a softmax activation function to derive class probabilities:

$$\hat{y}_i = \text{softmax}(\mathbf{z_i}) = \frac{\exp(z_i)}{\sum_{j=1}^{C}\exp(z_j)}, \quad i = 1, \ldots, C, \tag{11}$$

where $\hat{y}_i$ represents the predicted probability for class i.

E. Training and Optimization Objective

The training process involves computing the discrepancy between the predicted probabilities $\hat{y}$ and the true labels using the cross-entropy loss function:

$$\mathcal{L} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i), \tag{12}$$

where y_i corresponds to the one-hot encoded ground truth label. The objective function for rumor verification aims to minimize the classification loss $\mathcal{L}$. The overall optimization seeks to enhance the model's ability to accurately classify rumor veracity. The objective function is expressed in detail as:

$$\mathcal{J}(\mathbf{y}, \hat{\mathbf{y}}) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} y_{i,c} \log \hat{y}_{i,c}, \tag{13}$$

where N is the total number of rumor events in a training batch.

IV. Experiments

This section assesses the performance of our model in comparison to state-of-the-art (SOTA) baselines and conducts a comprehensive analysis to gain deeper insights into the model's effectiveness.

A. Datasets

Experiments are conducted on three widely used and publicly available challenging benchmark datasets: SemEval-2017 [15], RumorEval-2019 [2], and PHEME [16]. Among these, RumorEval-2019 and PHEME extend the SemEval-2017 task, that comprises 325 rumor-related events and 5,568 tweets collected from eight major breaking news events.

RumorEval-2019 extends SemEval-2017 by incorporating additional test data and new Reddit-based content while utilizing all SemEval-2017 rumor events for training. It consists of 446 rumor-related conversational threads and a total of 8,574 posts. The claims in both SemEval-2017 and RumorEval-2019 are annotated with three veracity labels: True, False, or Unverified. Each post within a thread is assigned a stance label: Support, Deny, Query, or Comment. Conversely, PHEME enhances RumorEval-2017 by incorporating additional rumor events and data from nine major breaking news stories on Twitter. It contains 2,402 conversational threads and 105,354 tweets. Unlike RumorEval-2019, the additional data in PHEME is annotated solely with rumor veracity labels.

For SemEval-2017 and RumorEval-2019, we adhere to the standard train/validation/test split as defined in the original publications. Conversely, since PHEME does not provide an official dataset split, a conventional evaluation protocol is adopted, that follows a leave-one-out k-fold validation strategy, where each event is used as a test set in turn. Table I provides a detailed summary of the dataset statistics.

B. Data Preprocessing

In addition to standard data preprocessing techniques, such as the removal of null entries, this work employs hashtag processing and text normalization following the methodology proposed by [17]. Furthermore, inspired by the approach of [18], all hyperlinks in the text are replaced with url and all *@user* mentions are substituted with $mention$, as these transformations have been shown to enhance model performance in the aforementioned studies.

C. Experimental settings

The uncased-BERT-base version is employed to generate word embeddings for both the claim p and its corresponding replies R within a thread C. An alternative pre-trained language model, RoBERTa [19], was also evaluated; still, this model exhibited inferior performance compared to BERT and was thus excluded. Extensive experimentation with different hyperparameter settings was performed to identify the optimal configuration. The training process is conducted with a batch size of 16 threads, and the BERT tokenizer is configured with a maximum sequence length of 128. Optimization is carried out using the Adam optimizer [20] with a learning rate of 0.001. Dropout rate was set to 0.35. BiLSTM embedding dimension is set to (18, 19, 20) for (SemEval-2017, PHEME, RumorEval-2019), corresponding to the average thread lengths in these datasets. The experiments were conducted on two Quadro RTX 8000 GPUs, each equipped with 48 GB of VRAM.

Since PHEME contains only partial stance annotations, the model was initially trained on the stance-labeled RumorEval-2019 and SemEval-2017, omitting the stance-based embedding aggregation and the stance label encoding using BiLSTM. Given that these datasets exhibit a strong bias toward the *Comment* stance, we employed SMOTE [21] oversampling technique to balance the stance distribution and enhance model generalization. However, SMOTE was not applied to the rumor verification task to ensure a fair comparison with baseline approaches. The best-performing model from this pretraining stage was subsequently utilized to predict stance labels for PHEME.

D. Evaluation Metrics, Baselines, and Results

Model performance is assessed using macro F1-score and accuracy, with the best-performing model—determined based on validation macro F1-score—selected for final evaluation. All hyperparameters were meticulously fine-tuned using the development dataset, and the reported results are averaged over ten experimental runs. The model is compared against the following rumor detection baselines:

1) **eventAI** [22]: Securing first place in the RumorEval-2019 competition task [23], eventAI leverages multidimensional information and employs an ensemble learning strategy to improve rumor verification performance.
2) **Longformer** [12]: Introduces a fine-tuned Longformer, that is a multi-task learning framework with the bottom part predicting rumor stance and the upper part classifying rumor veracity.

TABLE I
DETAILED STATISTICS OF THE DATASETS.

Dataset	#Threads	#Tweets	Stance Distribution				Rumor Veracity Labels		
			#Support	#Deny	#Query	#Comment	#True	#False	#Unverified
SemEval-17	325	5,568	1,004	415	468	3,685	145	74	106
RumorEval-19	446	8,574	1,184	606	608	6,176	185	138	123
PHEME	2,402	105,354	-	-	-	-	1,067	638	697

3) **Coupled Hierarchical Transformer (CHT)** [24]: This method partitions conversational threads into multiple groups based on their hierarchical structure. Each group is independently processed using BERT to extract contextual features, that are subsequently integrated through a Transformer network for rumor verification.
4) **Joint Rumor and Stance Model (JRSM)** [18]: This approach utilizes a graph transformer to encode input data and a partition filter network to explicitly model rumor-specific, stance-specific, and shared interactive features, that are used for joint rumor and stance classification.
5) **SAMGAT** [25]: This employs Graph Attention Networks (GATs) [26] to model contextual relationships between posts. Although originally designed for binary rumor classification on the PHEME dataset (excluding the *Unverified* class), we adapt and retrain the model for our experimental setting, extending to three-class classification task.

Table II provides a comparative analysis of the performance of the models. The findings demonstrate that our model significantly outperforms the best-competing system, as validated by McNemar's test with a p-value < 0.05. Furthermore, our results exhibit a standard deviation in the range of 0.006–0.02 across all three datasets over the 10 experimental runs, indicating robust and consistent performance.

E. Discussion and Evaluation

We analyze why our approach achieves superior performance in comparison to the listed baselines. *eventAI* leverages ensemble learning but primarily relies on multidimensional handcrafted features, that may not generalize well across datasets. While *Longformer* effectively handles long text sequences, its multi-task learning framework is limited in capturing the structural relationships between stance and rumor veracity. *CHT* processes conversational threads in disjoint hierarchical groups, that disrupts temporal dependencies. *JRSM* treats rumor and stance classification as two separate tasks, but it does not fully exploit the interplay between them. *SAMGAT* relies on GATs to model contextual relationships, but it was originally designed for binary rumor classification and struggles with multi-class settings.

Unlike models such as *CHT*, that process hierarchical groups separately, our aggregation strategy preserves stance distribution and reduces information loss, allowing for better contextual reasoning. Reply posts contain rich contextual signals that indicate how a rumor is perceived within a conversation thread. Simply analyzing the source post alone (as some baselines do) ignores these critical interactions. Our model aggregates reply post embeddings grouped by stance type, ensuring that stance-conditioned representations provide a holistic view of the conversation. Aggregation, in emphasis, also mitigates the sequence length limitation of BERT by summarizing the impact of all replies in a stance-specific manner, preventing the loss of important context and allowing it to dynamically adapt to unseen data rather than relying on predefined feature extraction. Compared to SAMGAT, our model is more adaptable to three-class classification, as demonstrated by the substantial performance boost. Furthermore, while baselines implicitly incorporate stance, our model explicitly embeds and encodes stance labels. BiLSTM preserves the chronological order of stance evolution, that is critical for understanding how rumors develop over time and allowing it to capture stance progression and interactions.

Although only Twitter and Reddit data are used in our experiments, this work can be customized and extended to any social media platform actively engaging in fact-checking and where users participate in the subsequent conversations about a source claim. Therefore, our stance-conditioned modeling for rumor verification can also be generalized to Facebook, Instagram, Threads, etc. This will be incorporated into future work.

F. Ablation Study

To assess the contribution of each component, ablation experiments are conducted using the best-performing model on RumorEval-2017 and RumorEval-2019. The study involves systematically removing specific components and thus coming up with the following derivatives: 1) *-Replies*: Excludes reaction posts R, encoding only the source post p; 2) *-Emb agg*: Discards stance-conditioned embedding aggregation, instead encoding the entire rumor event as a single BERT embedding, constrained by the language model's maximum sequence length; 3) *-Stance-aware*: Omits the sequential modeling of stance labels using BiLSTM. The *Ours-whole* configuration represents the complete model.

Table III presents the results of the ablation study. *-Replies* leads to a significant drop in performance, indicating that contextual signals from replies are crucial for rumor verification, as previously highlighted. *-Emb agg* also results in lower performance. This highlights the importance of stance-conditioned embedding aggregation, that ensures that replies are grouped by stance type rather than processed as isolated inputs. Without aggregation, crucial stance patterns may be lost due to BERT's sequence length limitation, leading to incomplete contextual understanding. *-Stance-aware* furthermore negatively impacts

TABLE II
COMPARISON OF OUR RESULTS WITH BASELINE MODELS.

Model	SemEval-2017		RumorEval-2019		PHEME	
	Macro-F1	Acc	Macro-F1	Acc	Macro-F1	Acc
eventAI	0.618	0.629	0.577	0.591	0.342	0.357
Longformer	0.662	0.673	0.672	0.684	0.452	0.469
CHT	0.680	0.678	0.579	0.611	0.396	0.466
SAMGAT	0.702	0.709	0.542	0.562	0.409	0.418
JRSM	0.754	0.767	0.598	0.623	0.448	0.479
Ours	**0.774**	**0.781**	**0.636**	**0.648**	**0.641**	**0.643**

TABLE III
ABLATION STUDY RESULTS.

Model	RumorEval-2017		RumorEval-2019	
	Macro-F1	Acc	Macro-F1	Acc
-Replies	0.624	0.632	0.540	0.566
-Emb agg	0.642	0.649	0.552	0.579
-Stance-aware	0.647	0.651	0.548	0.564
Ours-whole	**0.774**	**0.781**	**0.636**	**0.648**

performance. This confirms that modeling stance evolution sequentially is beneficial, as stance shifts over time can indicate the credibility of a rumor. The *Ours-whole* configuration, that includes all modules, achieves the highest performance, validating the effectiveness of our stance-conditoned BiLSTM encoding and reply embedding aggregation.

G. Early Detection

Timely detection of rumors can mitigate their widespread dissemination. To assess early detection capabilities, we define detection checkpoints based on the elapsed time, spanning 24 hours, since the initial post. At each checkpoint, only replies accumulated up to that point are considered for model evaluation.

Figure 3 illustrates Macro-F1 and accuracy scores over time for early rumor detection on the SemEval-2017 dataset. Our model consistently outperforms all baselines throughout the 24-hour period, demonstrating superior effectiveness in detecting rumors early. While all models improve as more information becomes available, our model achieves significantly higher Macro-F1 scores early on, starting with an advantage at 4 hours and maintaining superior performance throughout. This suggests that our approach is more responsive to limited initial data, making it highly effective for early-stage rumor identification and particularly valuable in real-world misinformation scenarios where timely intervention is crucial.

H. Illustrative Example: Debunking a False Rumor

We discuss the modeling and debunking of a rumor event shown in Figure 1. The claim has a *False* veracity and a *Support* stance; our model accurately debunked it as *False* rumor. It can be observed from the diagram that more replies in the conversation thread contain *Comment* and *Query* stances. While *Comment* stance entail neutrality of users towards the claim,

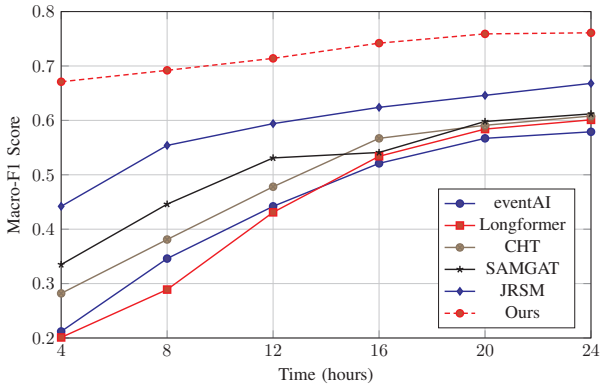

Figure 3. Early Rumor Detection Performance on SemEval-2017.

key insight is that some people who are exposed to a rumor, before deciding its veracity, will take the step of information inquiry to seek more information or express skepticism without specifically asserting whether it is false [13]. Moreover, three out of nine responses in the discourse are a repetition of the claim post, further intensifying doubt in the credibility of the source. These features enhance modeling stance progression and conversational dynamics, presenting cues for our model to discern signals that help in debunking a rumor.

V. CONCLUSION

This paper presents a novel stance-conditioned rumor verification model that integrates BERT-based source post embeddings and reply post embedding aggregation and BiLSTM encoding of stance labels, to enhance the detection of rumors in online discourse. Our findings highlight several key insights: the explicit incorporation of stance information significantly improves rumor verification, demonstrating that user reactions provide crucial contextual cues; processing stance sequences chronologically using BiLSTM preserves the natural evolution of discussions, leading to more context-aware representations; and leveraging stance-conditioned embedding aggregation mitigates the sequence length limitations of transformer-based models, ensuring a more comprehensive understanding of conversational dynamics. Early rumor detection analysis demonstrates that our model achieves faster and more accurate misinformation detection than competing methods, underscoring its practical utility in real-world misinformation detection.

While the model has shown success, its limitations include a heavy reliance on accurate stance annotations—which might not be consistently available—and training on datasets that may not fully represent real-world misinformation trends across diverse social media platforms. Additionally, the focus on textual content ignores the visual aspects (such as images, memes, and videos) that often accompany online rumors. Future work could reduce dependence on manually labeled data through weakly supervised and self-supervised learning, improve generalization via cross-platform adaptation, incorporate multi-modal data, and further explore extra structural dynamics like stance distribution, hierarchical level encoding, and attention mechanisms.

While the work has positive implications, ethical challenges and risks persist. False negatives and false positives could respectively suppress credible information or allow misinformation to spread, so human validation of predictions is recommended. The system's success could also enable misuse, such as censorship or targeting, requiring transparent deployment and strict ethical guidelines. Additionally, training data biases might lead to unfair outcomes; hence, evaluating and mitigating these biases is critical.

ACKNOWLEDGMENT

This work was supported by the National Science Foundation (NSF) under Award number OIA-1946391, Data Analytics that are Robust and Trusted (DART).

REFERENCES

[1] H. Gong, M. Zhang, Q. Liu, S. Wu, and L. Wang, "Breaking event rumor detection via stance-separated multi-agent debate", *arXiv preprint arXiv:2412.04859*, 2024.

[2] G. Gorrell *et al.*, "SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours", in *Proceedings of the 13th International Workshop on Semantic Evaluation*, J. May *et al.*, Eds., Minneapolis, Minnesota, USA: Association for Computational Linguistics, Jun. 2019, pp. 845–854. DOI: 10.18653/v1/S19-2147.

[3] A. Gupta, R. Kumar, and P. Sharma, "Rumor detection in online conversations using transformer-based language models", *Journal of Artificial Intelligence Research*, vol. 68, pp. 1023–1045, 2023.

[4] Y. Li, W. Zhang, and M. Chen, "Leveraging stance detection for improved rumor verification in social media", *ACM Transactions on Knowledge Discovery from Data*, vol. 16, no. 5, pp. 45–67, 2022.

[5] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures", *Neural Networks*, vol. 18, no. 5-6, pp. 602–610, 2005. DOI: 10.1016/j.neunet.2005.06.042.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *Bert: Pre-training of deep bidirectional transformers for language understanding*, 2019. arXiv: 1810.04805 [cs.CL].

[7] J. Khoo, D. Wu, and L. Tan, "Conversational context and rumor propagation: A neural approach", in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023, pp. 2123–2135.

[8] K. Shu, A. Sliva, and S. Wang, "Fake news and misinformation: A deep learning perspective", *IEEE Transactions on Computational Social Systems*, vol. 8, no. 3, pp. 556–573, 2021.

[9] R. Yang, J. Ma, H. Lin, and W. Gao, "A weakly supervised propagation model for rumor verification and stance detection with multiple instance learning", *arXiv preprint arXiv:2204.02626*, 2022.

[10] Q. Mai, S. Gauch, D. Adams, and M. Huang, *Sequence graph network for online debate analysis*, 2024. arXiv: 2406.18696 [cs.CL].

[11] S. Jami *et al.*, "Rumor stance classification in online social networks: The state-of-the-art, prospects, and future challenges", *arXiv preprint arXiv:2208.01721*, 2022.

[12] A. Khandelwal, "Fine-tune longformer for jointly predicting rumor stance and veracity", *arXiv preprint arXiv:2007.07803*, 2020.

[13] Z. Zhao, P. Resnick, and Q. Mei, "Enquiring minds: Early detection of rumors in social media from enquiry posts", in *Proceedings of the 24th international conference on world wide web*, 2015, pp. 1395–1405.

[14] P. Jia *et al.*, "An improved bilstm approach for user stance detection based on external commonsense knowledge and environment information", *Applied Sciences*, vol. 12, no. 21, p. 10 968, 2022.

[15] L. Derczynski *et al.*, "Semeval-2017 task 8: Rumoureval: Determining rumour veracity and support for rumours", *arXiv preprint arXiv:1704.05972*, 2017.

[16] A. Zubiaga, M. Liakata, and R. Procter, "Learning reporting dynamics during breaking news for rumour detection in social media", *arXiv preprint arXiv:1610.07363*, 2016.

[17] B. Han and T. Baldwin, "Lexical normalisation of short text messages: Makn sens a# twitter", in *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, 2011, pp. 368–378.

[18] N. Luo *et al.*, "Joint rumour and stance identification based on semantic and structural information in social networks", *Applied Intelligence*, vol. 54, no. 1, pp. 264–282, 2024.

[19] Y. Liu *et al.*, "Roberta: A robustly optimized bert pretraining approach", in *arXiv preprint arXiv:1907.11692*, 2019.

[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization", *International Conference on Learning Representations (ICLR)*, 2015. arXiv: 1412.6980 [cs.LG].

[21] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique", *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[22] Q. Li, Q. Zhang, and L. Si, "EventAI at SemEval-2019 task 7: Rumor detection on social media by exploiting content, user credibility and propagation information", in *Proceedings of the 13th International Workshop on Semantic Evaluation*, J. May *et al.*, Eds., Minneapolis, Minnesota, USA: Association for Computational Linguistics, Jun. 2019, pp. 855–859. DOI: 10.18653/v1/S19-2148.

[23] G. Gorrell *et al.*, "Rumoureval 2019: Determining rumour veracity and support for rumours", *arXiv preprint arXiv:1809.06683*, 2018.

[24] J. Yu, J. Jiang, L. M. S. Khoo, H. L. Chieu, and R. Xia, "Coupled hierarchical transformer for stance-aware rumor verification in social media conversations", Association for Computational Linguistics, 2020.

[25] Y. Li, Z. Chu, C. Jia, and B. Zu, "Samgat: Structure-aware multilevel graph attention networks for automatic rumor detection", *PeerJ Computer Science*, vol. 10, e2200, 2024.

[26] P. Veličković *et al.*, "Graph attention networks", in *International Conference on Learning Representations (ICLR)*, 2018.