Misinformation detection with learning from spatial-temporal propagation features and information content on Twitter

Anonymous ACL submission

Abstract

This study introduces the STU-User model, an innovative approach to detecting misinformation on social media, combining user behaviors with spatial-temporal analysis. The model incorporates an advanced neural network integrating spatial-temporal units (STU) with enhanced long short-term memory (LSTM) structures. Central to its design is the use of Bert-embeddings to analyze the patterns of users' historical interactions and connections in the network and incorporate them with a similarity measure to enhance accuracy. This analysis is then combined with the spatial-temporal aspects of the message propagation structure. We found that the STU-User model surpasses the performances of existing methods based on tests using public Twitter datasets. Theoretical and practical implications for policy-making and regulating social media in the misinformation era are discussed.

1 Introduction

With the advent and growing popularity of social media platforms and mobile devices, the ease of disseminating misinformation on these digital channels has increased considerably. The widespread circulation of this misinformation can cause public distress and lead to adverse effects on individuals, underscoring the urgency for automated detection mechanisms (Allcott and Gentzkow, 2017; Jin et al., 2017). Historically, the majority n of research in this domain has concentrated on text mining methodologies, employing supervised models grounded in feature engineering or leveraging advanced deep learning frameworks (Castillo et al., 2011; Ma et al., 2017; Song et al., 2019). In recent years, an emerging line of research has focused its focus on the spatial structure of message propagation, offering a novel and promising avenue for tackling misinformation (Liu and Wu, 2018; Yuan et al., 2019).

However, these innovative methods predominantly focus on spatial aspects, frequently neglecting crucial temporal dimensions intertwined with the spatial dynamics in message propagation (Zhang et al., 2021). This oversight presents a gap in fully understanding and countering the spread of misinformation. Furthermore, the role of individual users in the propagation of misinformation, often inadvertently, cannot be overlooked (Castillo et al., 2011; Yang et al., 2012). The absence of immediate fact-checking or corroborative information makes users prone to be guided by their preexisting beliefs and opinions (Lewandowsky et al., 2017; Vosoughi et al., 2018). In this paper, we assume that people's predispositions can be inferred from their historical social media activities. For example, users and their connected friends who typically engage in conservative content would be more likely to share misinformation that supports conservative views than liberal information (Anspach and Carlson, 2020).

Our study aims to develop a comprehensive and integrated model for misinformation detection. By incorporating features that encapsulate the spatial-temporal propagation and the content of historical user behaviors on social media platforms, we proposed the STU-User model. This model aims to improve proficiency in identifying and curtailing the proliferation of misinformation across social media networks. This investigation enriches the theoretical framework surrounding misinformation detection and delivers actionable strategies for its pragmatic deployment in diverse contexts.

2 Related work

Misinformation detection has become an imperative task in maintaining the integrity of online discourse. Current research in this field indicates a dichotomy in the methodologies used to identify and curtail the spread of misinformation. One

approach emphasizes the analysis of the propagation structure of the information (Jin et al., 2013; Sampson et al., 2016), while the other focuses on examining user behavior and the content of the information itself (Qazvinian et al., 2011; Popat, 2017). These methodologies gain heightened importance in early detection scenarios, especially due to the absence of fact-checkers or experts.

2.1 Propagation Structure in Misinformation Detection

Research on propagation structures aims to understand how information spreads within networks, with a focus on distinguishing the patterns of misinformation from factual content (Shu et al., 2017; Lazer et al., 2018). Vosoughi et al. (2018) found that false information propagates more widely and deeply than true information, especially in political contexts. Zhou and Zafarani (2020) identified unique spatial propagation patterns for rumors, indicating misinformation. Wu et al. (2019) developed a model using propagation paths to differentiate true from false news, outperforming traditional content-based methods through machine learning on propagation features.

Additionally, Gupta et al. (2022) explored how bots manipulate propagation structures to amplify misinformation, a phenomenon also noted by Addawood et al. (2019) in their study on bot behavior. Current research mainly focuses on spatial aspects, often neglecting temporal dimensions. Future research should integrate both spatial and temporal elements for a more comprehensive analysis.

2.2 User Behavior and Information Content in Misinformation Detection

The second crucial approach to detect misinformation involves a detailed examination of user behavior and the intrinsic content of the information. Tacchini et al. (2017) found that user engagement metrics like likes and shares can indicate content authenticity, as interaction patterns differ between true and false information. Similarly, Castelo et al. (2019) identified linguistic elements such as sentiment and complexity as critical indicators of misinformation.

Ciampaglia et al. (2018) developed a hybrid model combining user-centric features (e.g., account age and activity level) with textual features (e.g., subjectivity and evidence usage) to improve misinformation detection. Conroy et al. (2015) supported a multifaceted approach incorporating user credibility, content style, and emotional resonance.

Shu et al. (2018) and Ghosh et al. (2023) highlighted the importance of user profile data, noting that habitual misinformation spreaders often lack verifiable details and engage with similar accounts. Analyzing user behavior and historical messages within their information-sharing network is thus crucial for assessing content authenticity on social media platforms.

2.3 Comparative Analysis and Integrative Approaches

Researchers have recently explored integrating the structural dynamics of misinformation propagation with user behavior to address rampant misinformation on social media (Monti et al., 2019). Hangloo and Arora (2021) compared these two prevalent methods, suggesting that while each has strengths, a unified approach merging propagation structure analysis with user behavior and content scrutiny could be more effective. Zhou et al. (2020) supported this hypothesis, demonstrating that a multimodal methodology combining propagation and content attributes outperforms approaches using singular data types. Huang et al. (2023) also proposed integrating spatial and temporal information of message propagation to improve rumor detection accuracy, emphasizing the importance of including user behavior information.

In summary, misinformation detection research employs diverse methods. Propagation structure studies provide an overview of information spread, while examining user behavior and content focuses on personal interactions and text features. The trend is moving toward holistic models that combine these perspectives, indicating a shift toward more sophisticated and accurate misinformation detection methods (Yuan et al., 2019; Hu et al., 2022).

In this study, we present an innovative deep learning-based model STU-User, for misinformation detection that eliminates the necessity for external expert validation or fact-checking. This model merges the spatio-temporal aspects of misinformation spread with a novel embedding of users' historical interactions and behaviors within the information network. This integrative approach significantly improves the model's proficiency in precisely detecting misinformation, demonstrating the effectiveness of combining various methodologies in the evolution of misinformation detection.

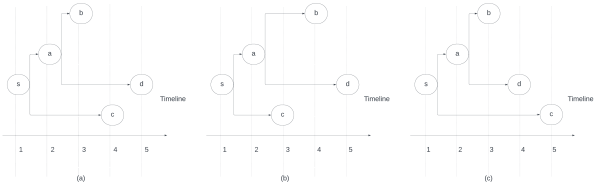

Figure 1: Three propagations share the same spatial structure,: $s \to a \to b$, $s \to a \to d$, $s \to c$, but in temporal structure, they are different: (a) $s \to a \to b \to c \to d$, (b) $s \to a \to c \to b \to d$, (c) $s \to a \to b \to d \to c$

3 Problem Statement

Let $P = \{P_1, P_2, \ldots, P_{|P|}\}$ denotes a collection of message propagation instances, where each propagation instance P_i is defined as $P_i := \{s_i, r_{i,1}, r_{i,2}, \ldots, r_{i,|P_i|-1}\}$. Here, s_i represents the initial user who posted the source message, and each r_i refers to corresponding users who engaged with the post (through retweets or replies). For every user $r_{i,j}$ responding to a post, their friend networks and historical post data are extracted for analysis. It is posited that these responsive messages are ordered chronologically, implying that the post time of message $r_{i,j'}$ precedes that of $r_{i,j''}$ if $j' < j''$, with the source message s_i being the earliest in the sequence for P_i. Although represented sequentially, these messages are interconnected through their retweet or reply relationships.

Figure 1. shows that although some propagations share the identical spatial structure in dissemination, they may still differ in temporal structure. In this task, each message propagation P_i is associated with a categorical label from the set C, which comprises four classifications: true information, misinformation, debunking information, and unverified information. The primary objective of this paper is to frame the task of misinformation detection as a supervised classification problem, endeavoring to develop a classifier f that maps from P to C.

4 Methodology

4.1 User content features

4.1.1 Internal and external impact

In the realm of misinformation detection, distinguishing between internal and external impacts is essential for a comprehensive analytical framework (Zhou et al., 2022). Internal impact refers to the individual and psychological aspects of information processing, influenced by personal beliefs, prior knowledge, and emotional engagement with content. These internal factors significantly affect an individual's tendency to accept and spread misinformation and serve as critical indicators for identifying vulnerability to deceptive narratives (Lazer et al., 2018). On the contrary, the external impact is anchored in the social and relational dimensions of information dissemination. It encompasses the effects of communal norms, the structure of social networks, and the frequency and nature of user interactions. The influence of these external elements is pivotal in developing information cascades, where perceived credibility and content dissemination are often more shaped by social dynamics than by factual accuracy. Grasping the nuances of these external dynamics is vital for a deeper insight into the mechanisms of misinformation spread across networks (Lazer et al., 2018; Zhou and Zafarani, 2020).

A holistic approach to misinformation detection, which incorporates both internal and external impacts, allows for a more nuanced understanding of the intrinsic qualities of content and the wider context of its spread. This dual analysis is instrumental in developing sophisticated detection algorithms capable of differentiating between inadvertent sharing of misinformation and intentional disinformation efforts. In addition, plain embeddings of user historical text content with language models would introduce irrelevant dimensions to the misinformation detection task, while a solid comparison analysis between the content and the source message is more efficient. Therefore, this section focuses on formulating a method to craft user content features for our deep learning model, aimed at the similarity measure between the user content and the source message.

4.1.2 Feature modeling

Suppose we identify a user u within a propagation chain where the veracity of the source information

s is yet to be confirmed. Given a source information content s created at time t, the user u has i historical posts $(p_1, p_2, p_3, \ldots, p_i)$, an associated attitude vector $(a_1, a_2, a_3, \ldots, a_i)$, and a posting time vector $(t_1, t_2, \ldots, t_i)$. The internal impact can be quantified as

$$I_{\text{internal}} = \sum_i S(E(s), a_i E(p_i)) \times \tau_{t-t_i}$$

where $E(s)$ represents the embedding feature vectors of the source message s, extractable via deep learning models employing transformer architectures, renowned for their proficiency in text semantics and various NLP (Natural Language Processing) tasks, such as machine translation and sentiment analysis. $S(E(s), E(n))$ is a similarity measure of the embedding features between m and n, typically employing cosine similarity. τ_{t-t_i} is a function that quantifies the temporal decay in impact from historical posts to current misinformation. The associated attitude vector $(a_1, a_2, a_3, \ldots, a_i)$, indicative of user behavior such as commenting, likes, or dislikes on the historical posts $(p_1, p_2, p_3, \ldots, p_i)$, is scaled within $[-1, 1]$. If the user did not interact with the historical post i, a_j will be set to 0.

Given misinformation content m created at time t, and a user u with a friends list of j: $(f_1, f_2, f_3, \ldots, f_j)$, each friend has i historical posts $(p_{j_1}, p_{j_2}, p_{j_3}, \ldots, p_{j_i})$ and an associated attitude vector $(a_{j_1}, a_{j_2}, a_{j_3}, \ldots, a_{j_i})$, emanating from user u, with a posting time vector $(t_{j_1}, t_{j_2}, t_{j_3}, \ldots, t_{j_i})$. Filtering out the posts older than 7 weeks from the time of the misinformation occurrence [1], the external impact can be measured as

$$I_{\text{external}} = \sum_j \sum_i (S(E(s), a_{j_i} E(p_{j_i})) \times \tau_{t-t_{j_i}}$$

Hence, the vector representing all impact features I_{all} will be

$$I_{\text{all}} = I_{\text{internal}} + I_{\text{external}}$$

which will characterize the user content features inputted into deep learning networks.

4.2 Propagation features

4.2.1 spatial-temporal unit (STU)

The Spatial-Temporal Unit (STU) is developed to model message propagation from a comprehensive spatial-temporal perspective. Unlike approaches that treat spatial and temporal structures separately, the STU integrates these aspects to form a cohesive model. In alignment with the principles of recurrent neural networks, the STU conceptualizes message propagation as a chronologically ordered sequence, applying an STU to each individual message within this sequence. Each STU, sharing parameters across the model, comprises three distinct components: a spatial capturer, a temporal capturer, and an integrator. These components collectively function to assimilate spatial-temporal information for each message.

In the context of a specific message propagation chain $P : P_i := \{s_i, r_{i,1}, r_{i,2}, \ldots, r_{i,|P_i|-1}\}$, where s represents the originating message and each r denotes a responsive message (either a retweet or reply), it is postulated that all messages within P adhere to a chronological sequence, thereby defining the temporal framework. For analytical simplicity, s is subsequently referred to as r_0. During data preprocessing, each message r_t within the range $\{0, 1, \ldots, |P| - 1\}$ is represented by a summative embedding of its constituent words, denoted as x_t. These embeddings are derived from the **vector of all impact features** I_{all} sourced from user content characteristics. The spatial configuration of P is theoretically modeled as a tree structure $T_i = \langle Pi, E_i \rangle$, in which E symbolizes a set of directed edges, indicative of retweet or reply connections between messages. For instance, if message r_t is a response to message $r_{t'}$, a directed edge $(r_t, r_{t'})$ is established, provided $0 < t < t < |P| - 1$.

As depicted in Figure 2, the STU architecture encompasses h_{r_t}, the hidden representation of the message propagation P up to the appearance of message r_t; $h_{r_{t'}}$, a temporary hidden state representing the propagation up to message r_t; and o_t, the outcome of classification based on h_{r_t}.

Adhering to the temporal order, messages within P sequentially enter the STU, constituting a chain of $|P|$ units. To ensure uniformity in all STUs, both the antecedent message and the progenitor of the source message r_0 are assigned a null message $\emptyset$; that is, $r_{-1} := \emptyset$ and $p(r_0) := \emptyset$. The term $p(r_t)$ is used to denote the progenitor of message r_t. Corresponding hidden states for these null messages are initialized as $\mathbf{0}$, implying $h_{r_{-1}} = h_{p(r_0)} = 0$.

[1] we choose 7 weeks because among the dataset we used, 7 weeks would involve 80% of the historical data of all users

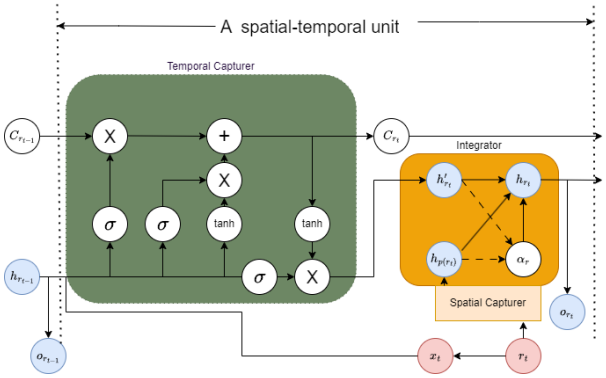

Figure 2: The architecture of an STU, which consists of three components, including spatial capturer, temporal capturer and integrator.

4.2.2 Spatial capturer

In the context of the current message r_t, where $0 \leq t \leq |P| - 1$, the function of the spatial capturer is to aggregate the hidden representation $h_{p(r_t)}$ from the parent message $p(r_t)$. This aggregation is crucial for encapsulating the spatial characteristics inherent in message propagation. For instance, as depicted in Fig. 1a, during the computation of hidden representations for messages a and c within the Spatial-Temporal Unit (STU), the spatial capturer retrieves the hidden representation of message s, denoted as r_0. Similarly, the computation of hidden representations for messages b and d involves the spatial capturer that obtains the hidden representation from message a. It is pertinent to note that the hidden representation $h_{p(r_t)}$ for any given message in the sequence is pre-computed, attributable to the antecedent positioning of $p(r_t)$ relative to r_t in the message propagation chain P.

4.2.3 Temporal capturer

In the context of the given message r_t, the temporal capturer is specifically designed to extract the temporal characteristics inherent in message propagation P. Given that the temporal aspect of the propagation is conceptualized as a sequential arrangement of messages, our temporal capturer employs a long-short-term memory (LSTM) model to effectively interpret this sequence, thereby acquiring the temporal attributes pertinent to the message propagation. In this model, C_{r_t} symbolizes the cell state at the specific time step t. The input to the temporal capturer comprises the message representation x_t of the current node r_t along with the hidden representation $h_{r_{t-1}}$ of the preceding message r_{t-1}. The resultant output of this process is a temporary hidden representation h'_{r_t}, which is subsequently integrated with the hidden representation $h_{p(r_t)}$ of the parent message r_t, thus encapsulating both spatial and temporal dynamics in the message propagation process.

4.2.4 Integrator

In our STU-User model, the outputs derived from both the spatial and temporal capturers are integrated using the softmax function to facilitate the prediction of the final output. This process leads to the acquisition of the temporary hidden representation h'_{r_t} and the hidden representation $h_{p(r_t)}$, which encapsulate the spatial-temporal dynamics of message propagation P up to the appearance of the current message r_t. To effectively amalgamate these distinct representations, we introduce an integrator that employs a self-attention mechanism, thus generating a unified hidden representation h_{r_t}.This self-attention mechanism is structured as a two-layer perceptron. It computes the attention coefficients for h'_{r_t} and $h_{p(r_t)}$ according to the subsequent formula:

$$\alpha_h = \text{softmax}\left(\frac{a \times \tanh(Wh)}{\sum_{h' \in \{h'_{r_t}, h_{p(r_t)}\}} a \times \tanh(Wh')}\right)$$

4.2.5 Output and model learning

In our model, the softmax function is utilized to ascertain the classification of message propagation P as an output:

$$o_t(P) = \text{softmax}(Vh_{r_t} + b)$$

where V and b represent the learnable weights and bias in the output layer, respectively. The implementation of $o_t(P)$ allows the Spatial-Temporal Unit (STU) to discern the classification results based on a subset of the information from message propagation P, specifically up to the juncture

of the current message r_t. This approach does not require the entirety of P's information, thus facilitating the early detection of misinformation, as will be demonstrated in our experimental analysis.

Let $P = \{P_1; P_2; \ldots; P_{|P|}\}$ denote a set of message propagations, each P_k being associated with a categorical label from set C, which is subdivided into three more detailed categories: true information, misinformation, and unverified information. We represent this class label for each $P_k \in P$ using a 4-dimensional one-hot vector y_k. The model's learning is guided by the cross-entropy loss function, augmented with an L_2 regularization term:

$$L = -\sum_{k=1}^{|P|} y_k^T \ln o_{|P_k|-1}(P_k) + \frac{1}{2}\|\Theta\|_2^2$$

where Θ signifies the complete set of parameters within the deep learning framework.

5 Experiment

5.1 Data

Our empirical investigation utilized two widely recognized Twitter datasets, Twitter15 and Twitter16, which are widely used in the field of misinformation detection research. The Twitter15 dataset encompasses 1490 instances of tweet propagation, while Twitter16 comprises 818 such instances. For both datasets, we aggregated the historical posts of users and their connections from the week preceding the dissemination of the information. This data collection was facilitated through the Twitter API[2]. Each dataset is annotated with four distinct labels: 'fake', 'true', 'unverified', and 'debunking of fake', where the latter denotes narratives that identify specific news stories as falsified. In our analysis, these labels are reclassified as "misinformation", "true information", "unverified information", and "debunking information", respectively.

5.2 Setup

In the experimental setup, we allocated 10% of the instances randomly as the development set, and divided the remaining data into training and testing sets in a 3:1 ratio for both Twitter15 and Twitter16 datasets. The dimensionality of the word-embedding vector was established at 300, while the dimension for the hidden state h was set at 100. For the similarity function S, we opted for the cosine similarity measure. Considering that user posts on Twitter often comprise fewer than 512 words, a pre-trained Sentence-RoBERTa model, an adaptation of RoBERTa through a Siamese network, was utilized to extract the embedding vectors for these posts. In instances where misinformation content exceeded 512 words, the Longformer model was employed to acquire the respective embedding vectors. The time decay function was computed using $\tau_{t-t_i} = \exp((1 - (t - t_i))$ which is inspired by (Woźniak et al., 1995).

Regarding the attitude vector, if a user did not engage (through comments, likes, or dislikes) with a historical post i, the corresponding a_i was set to 0. If a user expressed likes, a_i was assigned a value of 1, and it was set to -1 for dislikes. In all other cases, a_i was determined by the output of the Vader algorithm (Hutto and Gilbert, 2014), inputting the user's comments.

For modeling propagation features, we adhered to the optimization settings reported in the literature for all comparative methods. Our methodology was implemented using PyTorch, with parameter optimization executed via the Adam algorithm. The initial learning rate was set at 0.005, subject to gradual reduction during the training phase with a batch size of 64 for the training set. Optimal parameter settings were identified based on performance metrics observed in the validation set, constituting 10% of the total dataset. The loss function L was designated as the primary optimization objective.

Our comparative analysis included state-of-the-art baseline models such as SVM-TK, GRU-RNN, TD-RvNN, PPC, and GLAN. The performance evaluation of all models was conducted based on accuracy and F_1 scores for each category.

- **RFC:** a random forest classifier that utilizes user, linguistic and structure characteristics(Zhao et al., 2015).

- **SVM-TK:** Support Vector Machine with Tree Kernel integrates sophisticated tree-based kernel functions into the traditional SVM framework to classify misinformation, exploiting structural similarities within data to enhance detection accuracy(Ma et al., 2017).

- **GRU-RNN:** Gated Recurrent Unit - Recurrent Neural Network is designed for classification of misinformation by capturing sequential dependencies in data through its memory-based architecture, which allows for nuanced

[2]https://dev.twitter.com/rest/public

detection of patterns indicative of false information(Ma et al., 2016).

- **TD-RvNN:** Top-Down Recursive Neural Network leverages a tree-structured approach to analyze the propagation patterns of information, enabling the discernment of authentic news from misinformation effectively(Ma et al., 2018).

- **PPC:** Propagation Path Classification model strategically analyzes the paths of information dissemination across social networks, utilizing structural and temporal features to distinguish between authentic and misleading content in misinformation detection tasks(Liu and Wu, 2018).

- **GLAN:** Graph Learning-Attention Network model employs a novel graph-based learning mechanism combined with attention networks to effectively classify misinformation by capturing complex relational patterns and dependencies in data(Yuan et al., 2019).

5.3 Ablation study

In order to determine the relative importance of every module of the STU-User, we perform a series of ablation studies over the different parts of the model.

- **w/o User content feature:** Replacing the embeddings vector I_{all} by the **SOTA** word2vec method to represent x_t.

- **w/o Spatial feature:** Removing the spatial capturer component in STU, that is, set $h_{r_t} = h'_{r_t}$.

- **w/o Temporal feature:** Removing the temporal capturer component in STU, that is, set $h_{r_t} = h_{p(r_t)}$

6 Results

Tables 1 and 2 present the comparative performance outcomes of our proposed STU-User model against baseline models in the domain of misinformation detection on the Twitter15 and Twitter16 datasets, respectively. In instances where the STU-User model demonstrates significantly superior numerical results compared to the baselines, these figures are highlighted in boldface for each respective column. Notably, our model exhibits accuracy of 91.4% and 91.0% on the two Twitter datasets, respectively, underscoring its adaptability to varied tweet-based datasets. Furthermore, the exceptional performance of the STU-User model, leveraging both spatio-temporal and user content features, underscores its efficacy in addressing the challenges of misinformation detection on social media platforms.

Table 1: misinformation classification results of different classifiers in Twitter15
("M": Misinformation;"T":True information; "D": Debunking information; "U": Unverified information;)

CLASSIFIER	ACC	M F_1	T F_1	D F_1	U F_1
RFC	0.565	0.422	0.810	0.401	0.543
SVM-TK	0.667	0.669	0.619	0.772	0.645
GRU-RNN	0.641	0.634	0.684	0.688	0.571
TD-RVNN	0.723	0.758	0.682	0.821	0.654
PPC	0.842	0.875	0.811	0.818	0.790
GLAN	0.905	0.917	0.924	0.852	0.927
STU-USER	**0.914**	**0.923**	0.918	**0.881**	0.929

Table 2: misinformation classification results of different classifiers in Twitter16

CLASSIFIER	ACC	M F_1	T F_1	D F_1	U F_1
RFC	0.585	0.415	0.752	0.547	0.563
SVM-TK	0.662	0.623	0.643	0.783	0.655
GRU-RNN	0.633	0.715	0.617	0.577	0.527
TD-RVNN	0.737	0.743	0.662	0.835	0.708
PPC	0.863	0.898	0.820	0.843	0.837
GLAN	0.902	0.869	0.921	0.847	0.968
STU-USER	**0.910**	**0.892**	0.915	**0.876**	0.935

Table 3: Results of the ablation study

METHOD	ACC	M F_1	T F_1	D F_1	U F_1
TWITTER15					
STU-USER	0.914	0.923	0.918	0.881	0.929
W/O USER	0.836	0.812	0.823	0.811	0.792
W/O SPATIAL	0.883	0.894	0.837	0.866	0.824
W/O TEMPORAL	0.867	0.836	0.834	0.864	0.825
TWITTER16					
STU-USER	0.910	0.892	0.915	0.876	0.935
W/O USER	0.836	0.815	0.842	0.812	0.806
W/O SPATIAL	0.863	0.843	0.833	0.852	0.836
W/O TEMPORAL	0.874	0.826	0.835	0.812	0.829

Table 3 delineates the results of our ablation study. It is observed that any variant of the model, when deprived of specific components, demonstrates a reduction in both accuracy and F_1 scores

compared the complete STU-User model. This finding shows evidences that the holistic amalgamation of the three critical components—user content features, spatial propagation structure, and temporal propagation structure—is essential for the enhanced efficacy of the STU-User model in detecting misinformation.

7 Discussion

Traditional machine-learning approaches to misinformation detection involve training supervised models using statistical features derived from content, user characteristics, and message propagation. These methods require extensive preprocessing and feature engineering, which can be time-consuming and labor-intensive (Jin et al., 2013; Ma et al., 2017). Moreover, some features may be unavailable, inadequate, or impossible to extract. As shown in Tables 1 and 2, models relying on manually crafted features (e.g., RFC, SVM-TK) perform poorly, highlighting their limitations in generalizing relevant features.

Recent advancements in deep neural networks offer a promising alternative, addressing the shortcomings of traditional methods (Song et al., 2019; Huang et al., 2019). The GRU-RNN model, for example, captures temporal dependencies crucial for understanding misinformation spread. The PPC model combines structural and temporal elements to analyze content propagation across social networks. Tables 1 and 2 indicate that both GRU and PPC models outperform traditional classifiers, with PPC being more effective due to its integration of CNN and RNN to capture user feature variations. However, existing deep learning methods often model user features and temporal/spatial structures separately, lacking a unified approach.

The proposed STU-User model addresses this gap, showing superior accuracy and F_1 scores for misinformation and debunking categories, as demonstrated in Tables 1 and 2. GLAN also performs well for True and Unverified information categories, integrating graph-based learning with attention mechanisms to analyze relational patterns. The STU-User model's focus on embedding user content based on similarity with historical posts is particularly effective for identifying misinformation related to specific themes like political or health rumors (Zhou et al., 2020). Our analysis, especially in Table 3, emphasizes the complementary role of user content, spatial propagation, and temporal structure in misinformation detection. Removing user content features significantly impacts performance, underscoring the importance of encoding semantic similarity and the general context of misinformation.

Future work aims to enhance the STU-User model with additional data types, such as user profiles and geographic locations, and evaluate its effectiveness in early detection scenarios. We also encourage integrating multimodal data to further refine misinformation detection strategies.

8 Conclusion

In this study, we introduced the STU-User model to tackle the issue of rampant misinformation on social media. Unlike traditional detection methods, our approach integrates spatial-temporal and content-based features within message propagation networks. The Spatial-Temporal Unit (STU) in our neural network architecture, enhanced with parameter-sharing, modified LSTM structures, and a self-attention mechanism, has proven particularly effective. Additionally, the use of Bert-embedded similarity measures between user beliefs and source content offers a novel perspective on misinformation detection by emphasizing the importance of user and network interactions.

Empirical evaluations on publicly available Twitter datasets show that our model significantly outperforms existing benchmarks in misinformation detection. These findings highlight the efficacy of our integrated approach, which combines spatial-temporal propagation and content similarity features, thus enhancing the model's performance. This research not only enriches the academic discourse on misinformation detection methodologies but also holds considerable practical potential. It opens new avenues for algorithmic improvements and fact-checking initiatives, providing valuable tools for policymakers, social media platforms, and the broader community engaged in combating misinformation. Ultimately, this study provides both a comprehensive theoretical framework and actionable insights, paving the way for more effective and timely interventions against misinformation in the digital age.

9 Limitations

The STU-User model, while innovative, has several limitations that point to gaps in current research and areas for future improvement. One significant

research gap is the limited exploration of user behavior dynamics over extended periods. The model primarily focuses on short-term interactions and immediate user responses, overlooking how misinformation might affect users' long-term behavior and beliefs. This approach potentially misses deeper insights into the persistence and evolution of misinformation in social networks. Future research should aim to integrate longitudinal user data to capture these long-term effects, providing a more comprehensive understanding of misinformation dynamics.

Additionally, the granularity and variety of user profile data used in the current study are somewhat limited. The STU-User model employs basic user interaction metrics and Bert-embedded similarity measures, which may not fully capture the complexity of user behavior and influence. Essential user profile details, such as demographic information, political affiliations, and past exposure to misinformation, are not considered. Incorporating these details could significantly enhance the model's predictive power. Future research should address this by including more detailed and diverse user profile data, enabling a richer analysis of how different user characteristics influence the spread of misinformation.

Furthermore, the reliance on publicly available Twitter datasets limits the model's applicability to other social media platforms with different user demographics and interaction patterns. Propagation structures and user behaviors on platforms like Facebook, Instagram, or TikTok may differ significantly from those on Twitter, potentially affecting the model's performance and generalizability. Future studies should test the STU-User model across various social media platforms to validate its effectiveness and adaptability. Additionally, expanding the model to include multimodal data, such as images and videos, would provide a more holistic approach to misinformation detection, reflecting the diverse forms of content shared online.

In summary, while the STU-User model advances the field of misinformation detection by integrating spatial-temporal and content-based features, it has notable limitations. Addressing these gaps through future research, such as incorporating longitudinal user data, more detailed user profiles, and testing across various social media platforms, will be crucial. These improvements could enhance the model's comprehensiveness and effectiveness, providing more robust tools for combating misinformation in the digital age.

References

Aseel Addawood, Adam Badawy, Kristina Lerman, and Emilio Ferrara. 2019. Linguistic cues to deception: Identifying political trolls on social media. In *Proceedings of the international AAAI conference on web and social media*, volume 13, pages 15–25.

Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–236.

Nicolas M Anspach and Taylor N Carlson. 2020. What to believe? social media commentary and belief in misinformation. *Political Behavior*, 42(3):697–718.

Sonia Castelo, Thais Almeida, Anas Elghafari, Aécio Santos, Kien Pham, Eduardo Nakamura, and Juliana Freire. 2019. A topic-agnostic approach for identifying fake news pages. In *Companion proceedings of the 2019 World Wide Web conference*, pages 975–980.

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

Giovanni Luca Ciampaglia, Azadeh Nematzadeh, Filippo Menczer, and Alessandro Flammini. 2018. How algorithmic popularity bias hinders or promotes quality. *Scientific reports*, 8(1):15951.

Nadia K Conroy, Victoria L Rubin, and Yimin Chen. 2015. Automatic deception detection: Methods for finding fake news. *Proceedings of the association for information science and technology*, 52(1):1–4.

Shreya Ghosh and Prasenjit Mitra. 2023. How early can we detect? detecting misinformation on social media using user profiling and network characteristics. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 174–189. Springer.

Ankur Gupta, Neeraj Kumar, Purnendu Prabhat, Rajesh Gupta, Sudeep Tanwar, Gulshan Sharma, Pitshou N Bokoro, and Ravi Sharma. 2022. Combating fake news: Stakeholder interventions and potential solutions. *Ieee Access*, 10:78268–78289.

Sakshini Hangloo and Bhavna Arora. 2021. Fake news detection tools and methods–a review. *arXiv preprint arXiv:2112.11185*.

LinMei Hu, SiQi Wei, Ziwang Zhao, and Bin Wu. 2022. Deep learning for fake news detection: A comprehensive survey. *AI Open*.

Qi Huang, Chuan Zhou, Jia Wu, Luchen Liu, and Bin Wang. 2023. Deep spatial–temporal structure learning for rumor detection on twitter. *Neural Computing and Applications*, 35(18):12995–13005.

Qi Huang, Chuan Zhou, Jia Wu, Mingwen Wang, and Bin Wang. 2019. Deep structure learning for rumor detection on twitter. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

Fang Jin, Edward Dougherty, Parang Saraf, Yang Cao, and Naren Ramakrishnan. 2013. Epidemiological modeling of news and rumors on twitter. In *Proceedings of the 7th workshop on social network mining and analysis*, pages 1–9.

Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, Yu Wang, and Jiebo Luo. 2017. Detection and analysis of 2016 us presidential election related rumors on twitter. In *Social, Cultural, and Behavioral Modeling: 10th International Conference, SBP-BRiMS 2017, Washington, DC, USA, July 5-8, 2017, Proceedings 10*, pages 14–24. Springer.

David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. *Science*, 359(6380):1094–1096.

Stephan Lewandowsky, Ullrich KH Ecker, and John Cook. 2017. Beyond misinformation: Understanding and coping with the "post-truth" era. *Journal of applied research in memory and cognition*, 6(4):353–369.

Yang Liu and Yi-Fang Wu. 2018. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

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.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2017. Detect rumors in microblog posts using propagation structure via kernel learning. Association for Computational Linguistics.

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

Federico Monti, Fabrizio Frasca, Davide Eynard, Damon Mannion, and Michael M Bronstein. 2019. Fake news detection on social media using geometric deep learning. *arXiv preprint arXiv:1902.06673*.

Kashyap Popat. 2017. Assessing the credibility of claims on the web. In *Proceedings of the 26th international conference on world wide web companion*, pages 735–739.

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*, pages 1589–1599.

Justin Sampson, Fred Morstatter, Liang Wu, and Huan Liu. 2016. Leveraging the implicit structure within social media for emergent rumor detection. In *Proceedings of the 25th ACM international on conference on information and knowledge management*, pages 2377–2382.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36.

Kai Shu, Suhang Wang, and Huan Liu. 2018. Understanding user profiles on social media for fake news detection. In *2018 IEEE conference on multimedia information processing and retrieval (MIPR)*, pages 430–435. IEEE.

Changhe Song, Cheng Yang, Huimin Chen, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. 2019. Ced: credible early detection of social media rumors. *IEEE Transactions on Knowledge and Data Engineering*, 33(8):3035–3047.

Eugenio Tacchini, Gabriele Ballarin, Marco L Della Vedova, Stefano Moret, and Luca De Alfaro. 2017. Some like it hoax: Automated fake news detection in social networks. *arXiv preprint arXiv:1704.07506*.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *science*, 359(6380):1146–1151.

Piotr Woźniak, Edward Gorzelańczyk, and Janusz Murakowski. 1995. Two components of long-term memory. *Acta neurobiologiae experimentalis*, 55(4):301–305.

Liang Wu, Fred Morstatter, Kathleen M Carley, and Huan Liu. 2019. Misinformation in social media: definition, manipulation, and detection. *ACM SIGKDD explorations newsletter*, 21(2):80–90.

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

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

Qiang Zhang, Jonathan Cook, and Emine Yilmaz. 2021. Detecting and forecasting misinformation via temporal and geometric propagation patterns. In *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part II 43*, pages 455–462. Springer.

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

Xinyi Zhou, Kai Shu, Vir V Phoha, Huan Liu, and Reza Zafarani. 2022. "this is fake! shared it by mistake": Assessing the intent of fake news spreaders. In *Proceedings of the ACM Web Conference 2022*, pages 3685–3694.

Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. : Similarity-aware multi-modal fake news detection. In *Pacific-Asia Conference on knowledge discovery and data mining*, pages 354–367. Springer.

Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40.

A Example Appendix

This is an appendix.