A Multi-modal Rumor Detection Model Based on Temporal Graph Attention Network

Shiming Li

Nanjing University of Aeronautics and Astronautics, Nanjing, Jiangsu, 211106, China

1071572330@qq.com

Abstract. Objective: To address the issue of insufficient mining of structural and temporal sequence features of information dissemination in existing rumor detection methods, a multi-modal rumor detection model based on temporal graph attention is designed. Methods: For the text modality, a RoBERTa pre-trained model is used as the basis, and GAT and GRU modules are introduced to extract and fuse mixed features of text and dissemination structure. For the image modality, ViT is used to extract image features. Multi-modal features are fused through self-attention and cross-attention mechanisms to complete rumor detection. Results: The accuracy and F1 value of the proposed model on the Twitter dataset reach 91.1% and 91.4%, respectively, achieving the best performance in the comparative experiments. Limitations: The performance of the model on other datasets has not been tested. Conclusion: The proposed model can effectively improve the rumor detection effect of multi-modal posts on social media.

Keywords: Rumor detection, Multi-modal, Graph attention network.

1 Introduction

In recent years, with the wide and deep penetration of the Internet and the continuous maturation and development of 5G technology, the number of Internet users in China has shown an explosive growth trend. Along with the growth of Internet users, the popularity of social media users has also increased. On social media platforms with an explosion of information, users are more likely to be attracted by eye-catching, concise, easy-to-understand, and content-rich image information. Therefore, multimodal content that combines text and images conveys more abundant and intuitive information and is more likely to be shared by social media users. As a result, rumor spreaders on social media platforms often adopt a text and image integrated dissemination method to spread rumors, demonstrating multiple characteristics such as rapid spread speed, wide influence range, high cost and low efficiency of manual verification and refutation. Such rumors are particularly worthy of attention due to their high concealment and potential harm.

Currently, the mainstream methods in the field of rumor detection mainly focus on in-depth learning and comprehensive integration of various features to achieve effec-

H. Sun et al. (eds.), *Proceedings of the 2025 10th International Conference on Social Sciences and Economic Development (ICSSED 2025)*, Advances in Economics, Business and Management Research 332,
https://doi.org/10.2991/978-94-6463-734-2_99

tive rumor identification. However, these methods often tend to ignore key information such as the time dimension during the processing, resulting in limited detection effects. Therefore, it is necessary to design a reasonable feature extraction scheme based on the characteristics of rumor spread on social media.

This paper proposes a social media rumor detection algorithm that integrates multimodal features and the dynamic evolution of rumor spread. By leveraging pre-trained models such as RoBERTa, a multimodal rumor detection model based on a temporal graph attention network is designed. Through an attention-based feature fusion method, rumors can be detected efficiently and accurately. The main contributions of this paper are as follows:

(1) By using a temporal graph attention network to represent the dynamic evolution features of rumor spread structure and integrating text features, the rumor detection effect in the single text modality is improved;

(2) Based on the rumor detection model of the temporal graph attention network, the model is extended from the single text modality to multimodal data, and the detection stability of the model is enhanced by utilizing the features of image data.

2 Related Research

2.1 Research on Text Rumor Detection

Compared with traditional language analysis paths, deep learning methods show incomparable advantages, mainly due to the deep neural network architecture can independently extract deep semantic background knowledge and automatically acquire potential text feature vectors, thus significantly improving the efficiency of rumor detection. In 2016, Ma et al.[1] based on two classical Recurrent Neural Network (RNN) architectures: The Long Short-Term Memory network (LSTM) and Gate Recurrent Unit (GRU) designed the first deep learning-based fake news detection model. In performance, it surpasses those models that rely on the combination of traditional machine learning algorithms and complex feature engineering, showing more superior performance. In 2017, Wang et al.[2] developed a false news detection model using Convolutional Neural networks (CNNS) and a pre-trained Word2Vec algorithm to obtain the feature vector representation of text. In order to explore the long-term dependence between text sequences of social media information, Chen et al.[3] introduced the Attention Mechanism and integrated it with RNN, aiming to selectively extract temporal feature representations from user feedback sequences.

In the early stage of rumor detection, a core problem is how to use the limited data available in the early stage of rumor transmission to effectively identify rumors. Qian et al.[4] proposed an innovative method based on probabilistic generation model, which aims to learn and build a probability distribution model about users' response to rumors by collecting rumor texts and corresponding user response data. In the prediction phase, the model can be used to generate user feedback on articles to be classified so that rumors can be accurately identified in the early stages of rumor propagation. Given the low-cost nature of social media account creation and the manufacture of software-operated social media bots, rumor makers may take advantage of this to act as cyber watermen, spreading malicious statements to influence and guide the overall attitude and stance of the social media user community. To solve the above problems, Ma et al.[5-6] designed a rumor detection model of GAN architecture inspired by Generative Adversarial Network. Through the adversarial learning mechanism between generator and discriminator, the model can learn more robust text feature representation, thus enhancing the stability and reliability of the rumor detection model. With the development and maturity of large language models, Yang et al.[7] proposed the CICAN model, which can effectively extract hierarchical features of text content and construct heterogeneous graphs to integrate multi-source information through the group intelligent semantic feature learning module and the knowledge enhancement module based on ChatGPT, which significantly improves the performance of rumor detection.

2.2 Research on Rumor Detection of Fused Images

The core of rumor detection technology based on image features is to deeply explore the information correlation between rumor text and picture. This correlation includes the consistency information, the complementarity information, and the potential semantic contradiction or non-consistency information between text and image. By mining these complex information relationships, the model can identify rumors more accurately, thus improving the efficiency of rumor detection. In order to fully capture complex semantic patterns contained in picture information and have scalability in different application scenarios, and considering the superiority of deep neural networks in automatically capturing complex semantic patterns of text and pictures, Jin et al.[8] further proposed the first deep-learning based multimodal rumor detection method att-RNN in 2017, and pointed out that there is some implicit correlation between text and pictures. To capture this correlation effectively, they pioneered the design of a unique visual neuronal attention mechanism. The mechanism works by calculating the attention weights derived from text features and social features of social media users, and multiplying these weights with high-level image features to filter out the more important image features. Although they had early insight into the importance of mining the correlation information between text and image to improve the accuracy of the rumor detection model, the att-RNN model designed by them had limitations in cross-modal information fusion, which did not fully consider the information complementarity between the rumor text and image. Nor does it solve the problem of inconsistent representation space caused by the difference of learning algorithms. In order to enhance the generalization performance of the multi-modal false news detection model in response to newly emerging news events on social media, Wang et al.[9] proposed a multi-task rumor detection framework based on adversarial learning mechanism in 2018. The Enhanced Adversarial Multi-task Network (EANN). The model constructs independent representation modules for text information and image information respectively. With the help of adversarial learning strategy, the EANN model can learn the common feature representation of text and image information in all social media events, and remove the specific information of each event in the process. Although the EANN model takes into account the differences between text and image information representation space, it does not model the relational information between text and image. Peng et al.[10] proposed a multimodal rumor detection method (MRML) based on deep metric learning, which significantly improves the performance of rumor detection by capturing intra- and inter-modal feature relationships through metric learning and comparison learning.

2.3 Research on Neural Network Rumor Detection

Traditional machine learning techniques have made remarkable achievements in dealing with Euclidean spatial problems, but the results are often unsatisfactory when dealing with non-Euclidean spatial data. In recent years, the emerging graph neural network technology provides an effective solution to this problem. Graph neural networks usually model information in the form of graph structure and extract relevant features from it, which can efficiently capture the dependency relationship between different entities. Bian et al.[11] adopted a node update strategy combining top-down and bottom-up in order to fully capture the propagation characteristics and diffusion mode of rumors. On this basis, a rumor detection model based on Graph Convolutional Neural Networks (GCN) is proposed. The model can accurately capture the characteristics of rumor propagation and diffusion by integrating the characteristic information obtained by the two node updating methods. Ke et al.[12] proposed an innovative rumor detection model that combines the advantages of multi-head attention mechanisms with GCN. In order to capture information comprehensively, all source tweets, retweet behavior, and global associations between users are integrated into a heterogeneous graph to obtain rich heterogeneous information. On this basis, multi-head attention mechanism is used to dig deep into the semantic connotation of source tweets and further extract the key information. Hu et al.[13] proposed an innovative rumor detection method, which is based on multi-relation propagation tree and aims to comprehensively consider inter-layer dependency and intra-layer sibling node dependency in the process of information propagation. On this basis, the method can adaptively enhance the key nodes in the graph, thus achieving a significant improvement in rumor detection. Yang et al.[14] proposed an innovative rumor detection model, which combines the advantages of gated neural networks and GCN. In this model, a gating unit is cleverly designed to select and combine the initial feature vectors entering GCN and the feature vectors processed by GCN. In this way, a better and more informative representation of node features can be obtained. The graph neural network-based method focuses on efficient acquisition of text features, user behavior features, propagation path features and diffusion pattern features. In contrast, the traditional deep learning method has shortcomings in processing temporal expression between tweets, while this method has shortcomings in this aspect. Sun et al.[15] combines graph neural networks and adversarial training to model structural features in the user's propagation path to enhance robustness and achieve good results in text rumor detection. Huang et al.[16] proposed for the first time the use of Causal Graph Attention Network (CGAN) to handle the multimodal rumor detection task, which not only identifies rumors more accurately but also enhances the interpretability of the results by constructing causal graphs to capture potential causal relationships in multimodal data.

3 Model Design

In this paper, a multimodal rumor detection model based on time graph attention network is proposed, which adopts two-flow model architecture and uses RoBERTa and ViT pre-trained models to extract text and image feature vectors respectively. In the process of text feature extraction, graph attention network is introduced to help RoBERTa model better understand the rumor detection task, so as to obtain more quality text feature vectors. In order to further improve the fusion effect of text and image features, the cross-attention mechanism is used to dynamically adjust the attention weight. In this way, the fusion between text feature vectors and image feature vectors is promoted, and their semantic consistency and information integrity are improved. The model structure is shown in Fig. 1.

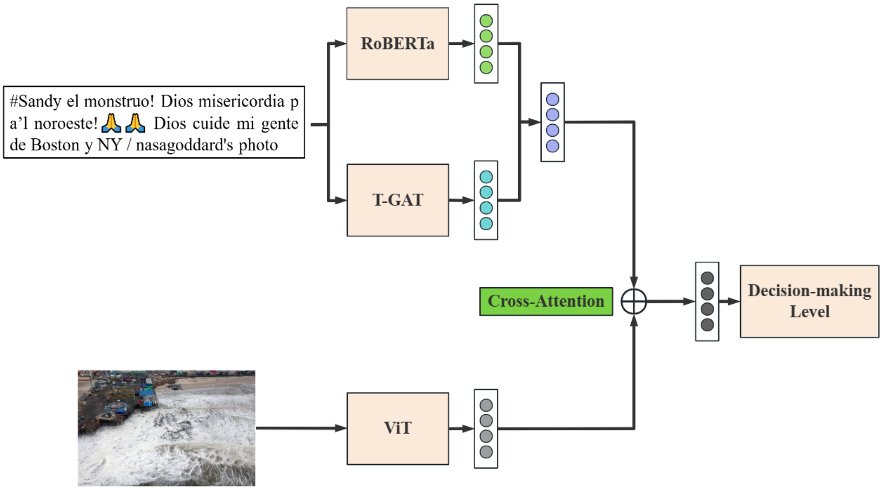

Fig. 1. Structure of multimodal rumor detection model based on time graph attention network.

3.1 Text Feature Extraction

(1) Figure attention network module

In the graph attention network module, the process of information aggregation depends on the integration of various sub-propagation structures and text features. The two major input components of this module are the specific content of the posts and the communication structure between the posts. Each node in the communication structure, such as forwarding, comments and other different types of posts, should have a different influence on the original post. Therefore, the use of graph attention network is to assign different weights to each node, while using multiple relational headers to regulate the flow of information from neighboring nodes, so as to show better performance than traditional recurrent neural networks in retaining structural information.

Specific to the input of the graph attention network model, it includes the adjacency matrix a_{ij} of the text content of the current post t_i and the propagation structure, which is used to represent the correlation between neighboring nodes i and j assign different weights to each node. Since information aggregation has not yet taken place in the initial phase, the hidden state h_1 is replaced by t_1 directly. Then, a new feature matrix h is constructed by combining the hidden feature vector h of each node with the hidden feature vector of the root node of layer $j-1$. For the definition of node t_1 feature aggregation, the mathematical representation is shown in formula (1) (2).

$$f_{nodei}^{l+1} = ||_{k=1}^{K} \sum_{j \in N_i} a_{ij}^{lk} W_k^l f_j^l \tag{1}$$

$$f_l = concat(f_l, f_{l-1}^{root}) \tag{2}$$

Where, $||_{k=1}^{K}$ represents the concatenation operation of vectors collected by different relational heads, and W_k^l is a trainable weight matrix. a_{ij} is the attention coefficient of the node i to j, the mathematical expression as shown in formula (3) (4).

$$a_{ij}^{lk} = \frac{exp\left(av_i^{lk}\right)}{\sum_{j \in N_i}(av_i^{lk})} \tag{3}$$

$$v_j = LeakyRelu\left(w_2 f_j^{l+1}\right) \tag{4}$$

Similarly, the feature aggregation of propagation feature nodes U_i also adopts the above method.

The representation of the neighbor aggregation of a post node is essentially a fusion of information from the same type of neighbor node into the node representation, and all that remains to be done is to learn the representation of the post node from all the pattern nodes. Different nodes contain type information, which requires the importance of distinguishing node types. Here, this study uses the attention mechanism to automatically learn the importance of nodes in different patterns, and finally uses the learned coefficients to carry out weighted fusion, as shown in formula (5).

$$g_j^i = w_2 tanh(w_1[h_i', f_i'] + b_1) + b_2 \tag{5}$$

Where, w_1, w_2 is trainable learning parameter, $tanh$ represents activation function, b_1, b_2 respectively represents bias data, h_i', f_i' respectively represents domain information.

(2) Gated cycle unit module

The Gated Cycle Unit (GRU) module is designed to further learn the features extracted at previous levels, thereby improving the effectiveness of rumor detection. In the process of rumors spread on social media, historical information always occupies a pivotal position, which is crucial for the model to capture the dynamic evolution pattern and law of rumor spread. GRU is based on RNN, RNN can only record the historical state of the last moment, and can not communicate the information of a longer structure. In order to enrich the representation of historical states, a time-attention mechanism is introduced in this study, so that multiple historical states can be processed in parallel as hidden states and input into the GRU. At the same time, in order to improve the performance of the model, a learning strategy based on the fusion of two loss functions is proposed.

In detail, this process first involves concatenating a series of historical states $M = [h_{t-w}, \cdots, h_{t-2}, h_{t-1}]$ and calculating them through a nonlinear activation function A_{tt}, as shown in formula (6), and calculating the attention weight of the concatenated vector accordingly. In this process, R and Q respectively represent the weight vector and matrix obtained through learning. Finally, the final hidden state is obtained by multiplying and multiplying the calculated attention weights M. The mathematical representation is shown in formula (7).

$$A_{tt} = r^T tanh(QM) \tag{6}$$

$$h_{t-1} = softmax\left(r^T tanh(QM)\right) * M \tag{7}$$

GRU has a reset door and an update door. The reset gate determines the extent to which information at the current time is used to update the hidden state, and the update gate determines the extent to which the hidden state at the current time is jointly determined by the hidden state at the previous time and the candidate hidden state at the current time. In order to obtain the characteristics of the propagation structure more accurately, this study improved on the basis of the traditional GRU. Due to the difference between the historical state and the current state, the difference of the communication structure of the two time periods also plays a crucial role in studying the evolution of the communication structure of the historical state. Therefore, in order to refine the difference between the historical state and the current state, this study will make $h_{t-1} - f_{t-1}$ as a part of the update gate, integrate the original update strategy in the GRU for the evaluation of the historical state. The specific calculation is shown in formulas (8) to (11).

$$z_t = sigmod(W_z f_{t-1}, h_{t-1} - f_{t-1}, U_z h_{t-1}) \tag{8}$$

$$R_t = sigmod(W_r f_{t-1}, U_r h_{t-1}) \tag{9}$$

$$f_{t-1} = tanh\left(W_h f_{t-1} + U_h(R_t h_{t-1})\right) \tag{10}$$

$$h_t = (1 - z_t)h_{t-1} + z_t f_{t-1} \tag{11}$$

The time graph-based attention method takes the discrete time subpropagation structure as the input, and the output is the feature vector for classification. The first is the T-step loop, which inputs each subpropagation structure into the GAT to extract the relevant features. Time attention is used to record the historical features of each sub-propagation structure, and GRU is used to record the current and historical states and generate the feature representation. Finally, after training, the minimization loss function is used to capture the features of the propagation structure changing over time through this model. Secondly, in order to enhance the learning ability of the propagation structure in each sub-time period, this study adds a local constraint loss function to the complete loss function, that is, the Euclidean-distance between the output of the previous moment and the output of the current moment is calculated, as shown in formula (12).

$$L_{time} = \frac{1}{N}\frac{1}{T}\sum_{i}^{T} D(H_t, H_{t-1}) \tag{12}$$

Where N represents the number of batch samples, T represents the number of sub-propagation structures, and D is a measure of distance between features. The loss function is learned from end to end. Samples of the same category are gathered together in the feature space, and samples of different categories are gathered into different feature Spaces. Cross-entropy is used as a local loss function to achieve accurate detection of rumors. Cross entropy is shown in equation (13).

$$L_c = -\sum_{i}^{N}\sum_{c}^{M} y_{ic}\, log(y'_{ic}) \tag{13}$$

Where y_i and y'_i are the lalel of the post and the probability that the prediction is positive, respectively, and M represents the number of categories. Finally, a set of samples is given, whose loss function is $L_c + \mu L_{time}$, where $\mu \in [0,1]$ is the hyperparameter of the equilibrium loss function.

3.2 Image Feature Extraction

In this paper, ViT (Vision Transformer) model is used to complete image feature extraction. After the image is fed into the model, the ViT first divides the image into several image blocks, and then converts each image block into a vector representation through an embedding layer. In this process, the input image must be split into fixed size (e.g. 16×16, 32×32) image blocks. Given an image $X \in R^{H\times W\times C}$($H$: height, W: width, C: number of channels), ViT needs to reshape X into a flat sequence of 2D image blocks:

$$v \in R^{N\times(P^2\cdot C)} \tag{14}$$

Where P^2 represents the resolution of the image block, N represents the number of image blocks, HW/P^2 namely. These vector representations are then fed into a multi-layer Transformer encoder for processing after linear mapping and positional embedding stitching (to preserve positional information between image blocks). In the Transformer encoder, the vector representation of each image block is self-attentional computed with the vector representation of other image blocks, enabling each image block to capture global information about the entire image. In the process of image processing, the self-attention mechanism plays a key role, which can capture and learn the internal relationship between the various blocks in the image. Meanwhile, the Feedforward Neural Network (FNN) is responsible for performing nonlinear transformation operations on the vector representation of each image block to further enrich its feature representation. Image embedding vectors are generated after passing through multiple Transformer encoders:

$$H_{image} \in R^{N\times H} \tag{15}$$

During the training process, the ViT model also adopts the strategy of pre-training and fine-tuning. Specifically, the ViT model is first pre-trained on a large-scale image data set, a process that enables it to learn features that are prevalent in images. These generic representations of image features provide a strong foundation for models that enable them to cope with a variety of image analysis tasks. The ViT model is then fine-tuned to optimize for specific rumor detection tasks. In this stage, the model will adjust its parameters according to the needs of the rumor detection task to better complete the task. This pre-training approach makes the ViT model have a strong generalization ability, even on small data sets can achieve better performance.

3.3 Feature Fusion

This paper mainly uses two kinds of attention mechanisms: self-attention mechanism and cross-attention mechanism. The self-attention mechanism is used to calculate the correlation between features in a single mode, and the cross-attention mechanism is used to calculate the correlation between features in the text mode and the image mode.

In the specific calculation process, for a given input sequence $X = [x_1, x_2, \cdots, x_m]$, each element is first transformed to a different dimension space by linear mapping, and then the set of Query, Key and Value vectors is obtained. Then, the similarity scores between each query vector and all key vectors are calculated, which reflect the relevance between them. Then, through Softmax normalization of these similarity scores, the weight distribution of each query vector relative to all value vectors is obtained. The final representation of each query vector is obtained by summing the corresponding value vectors weighted according to their weights.

Specifically, the Query, Key, Value matrix is generated first, and the similarity between each Query and each Key is calculated to obtain the set of weight coefficients. Secondly, the weight coefficients are normalized using Softmax function to ensure that the sum of weights is 1. Finally, the normalized weight coefficient is used to sum the corresponding values, and the output vector weighted by the attention mechanism is obtained. The steps are shown in formulas (16) to (18).

$$Sim(Query, Key) = Query \otimes Key^T \tag{16}$$

$$a_i = Softmax\big(Sim(Q, K)\big) \tag{17}$$

$$Attention(Q, K, V) = \sum a_i \otimes Value \tag{18}$$

The cross-attention mechanism is used to calculate the correlation between two different sequences. Given a text sequence $X = [x_1, x_2, \cdots, x_m]$ and an image sequence $Y = [y_1, y_2, \cdots, y_m]$, first, they are mapped to different dimensions by a linear transformation to obtain the query vector Q of the text sequence and the key vector $K$and value vector V of the image sequence. Then, the query vector Q of each text sequence is compared with all the key vectors K of the image sequence one by one, and the correlation between them is measured by calculating the similarity between them, and an attention score representing the strength of the respective association is generated.

By normalizing the attention scores, the weight distribution of query vectors for each text sequence and value vectors for all image sequences is obtained. Finally, multiply and sum the value vectors of the text sequence with the corresponding weights to obtain the final representation of the text image fusion sequence, as shown in formula (19).

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{19}$$

4 Experiment and Analysis

4.1 Experimental Data

In the experimental phase, this paper uses the common standard data set proposed by Peng et al.[10], which is a collection of news data from Twitter platform from 2012 to 2015. The dataset has been divided into the training set and the test set. The training set (devset) contains 6225 real posts and 9596 fake posts from 17 events. testset contains 1107 real posts and 1121 fake posts from 35 events. The events of the training set and the test set do not overlap, so the posts in the training set and the test set cover different events. Each piece of data in a data set is multimodal, meaning it includes text content, pictures or videos, and multiple social contexts. An example of multimodal data is shown in Fig. 2.

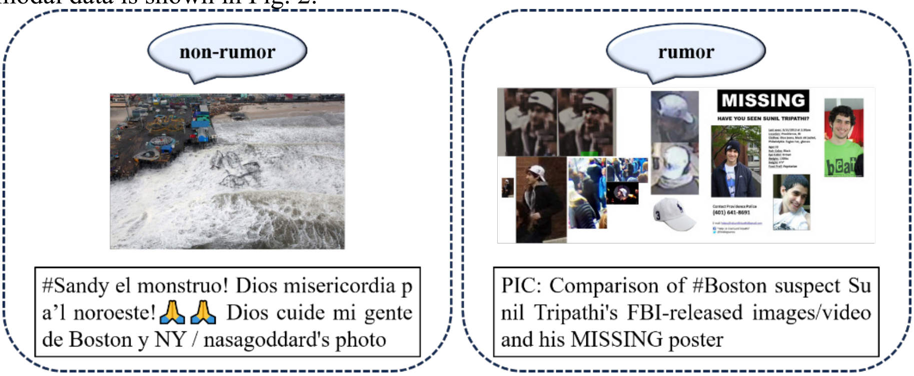

Fig. 2. Example of multimodal data.

4.2 Experimental Parameter Setting and Evaluation Index

In this paper, Python3.8 is selected as the programming environment and PyTorch, a deep learning framework, is adopted. At the same time, it integrates widely used deep learning components such as Torch-vision, Transformers library and scikit-learn to complete the construction of models and the visualization of experimental data results. All experiments were performed on a remote GPU cloud platform with a core configuration consisting of an Intel(R)Xeon(R)Gold 5218 CPU processor with 64GB of running memory, and two Tesla P100 GPU graphics cards with 16GB of video memory each.

In feature extraction, the RoBERTa model provided by Transformers library, specifically Roberta-base version, is used to process text data. Image feature extraction adopts the ViT model based on Transformers architecture to process the pictures attached to the post. Parameter Settings during model training are shown in Table 1.

Table 1. Model training parameters table.

Parameter	Parameter value
Batch size	128
Learning rate	3e-5
Preheating learning rate	0.1
The optimizer	AdamW
Regularization	0.1
Task loss function	Cross entropy loss

In this paper, accuracy, precision, recall and F1-score are used as evaluation indicators of the model. For binary classification problem, the specific calculation method is shown in formulas (20) to (23).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{20}$$

$$Precision = \frac{TP}{TP+FP} \tag{21}$$

$$Recall = \frac{TP}{TP+FN} \tag{22}$$

$$F_1 = \frac{2\times Precision\times Recall}{Precision+Recall} \tag{23}$$

4.3 Analysis of Comparative Test Results

In order to evaluate the effectiveness of the proposed social media rumor detection model based on graph attention network and multi-modal feature fusion, a series of comparison experiments were designed, including the comparison experiments of single text model, single image model and multi-modal feature fusion model. The baseline model of the specific comparison test is selected as follows:

(1) Single text model

DT-Rank: Uses decision trees to model the credibility of a post, and determines whether it is a rumor by analyzing the text features and propagation patterns of the post.

SVM-TS: Use manually extracted features to build time series, and analyze the feature changes of the text at different time points to determine whether it is a rumor.

Tanh-rnn: captures the time-dependent and semantic features of text sequences through RNN, uses hyperbolic tangent function (tanh) as the activation function to process sequence data, which has smooth transition characteristics, effectively alleviates the problem of gradient disappearance or gradient explosion, and realizes the rumor detection of text data on social media.

Bi-GCN: By processing the propagation characteristics and diffusion characteristics of rumors in the two propagation directions of top-down and bottom-up at the same time, it can capture the dynamic changes of rumors in a more comprehensive way to improve the accuracy of detection.

RvNN: A recursive neural network based on the rumor propagation tree structure captures the rumor propagation path and text characteristics on social media, and performs the final rumor detection by aggregating the information of all nodes in the tree.

RDGCN: Construct the graph structure of rumors based on GCN, study the regional propagation characteristics of rumors, and introduce Bernoulli-Poisson (BP) model to learn the regional propagation characteristics of rumors through unsupervised learning.

(2) Single image model

In the comparison experiment of rumor detection in single image modes, a pre-trained convolutional neural network was selected as the baseline model, including ResNet-50, VGG-19, DenseNet-161 and ViT model.

(3) Multi-modal model

att-RNN: The model processes text data through LSTM, extracts image features through VGG, combines social context features, introduces attention mechanism, and dynamically adjusts the weights of different modal features to achieve deep feature fusion.

EANN: The model extracts Text features through the text-CNN model and image features through the VGG-19 model, and introduces an event discriminator to remove event-specific features in the training process, thereby learning event-invariant features with more generalization ability and improving the model's generalization ability on new events.

MAVE: The model adopts an end-to-end training method, learns the shared representation of text and image through variational autoencoder, and then captures the correlation between the cross-modes, and only uses the text and image content of posts, which does not require additional social context information, and is more convenient in data acquisition.

SAFE: The model extracts text semantic features through CNN and image visual features through RNN. Firstly, the text and image features are independently rumor detected, and similarity is introduced. The text, image features and their similarity information are jointly learned for the final rumor detection.

The comparative test results are shown in Table 2. In general, the social media rumor detection model proposed in this paper based on the fusion of graph attention network and multi-modal features has better detection effect than the selected baseline model. The fusion model of integrated graphic mode is generally superior to the single mode in rumor detection. This result shows that there is significant complementarity between information from different modes in rumor detection tasks. Although the accuracy of the model that relies on the image mode alone is relatively low, the semantic information contained in the image mode can effectively supplement and strengthen the deficiencies of the text mode in the expression of semantic features, thus improving the overall detection effect. Compared with other multimodal fusion models, the rumor detection model based on graph attention network and multimodal feature fusion proposed in this paper has better performance, with an accuracy of 91.1% and an F1 value of 91.4%, showing better rumor detection results in experiments. This result suggests that the attention mechanism can more effectively fuse the relevant features of text data and image data.

Table 2. Comparative test results.

Mode	Model	Accuracy rate	Precision rate	Recall rate	F1
Text	DT-Rank	0.473	0.447	0.436	0.451
	SVM-TS	0.574	0.569	0.575	0.570
	Tanh-RNN	0.851	0.845	0.866	0.835
	Bi-GCN	0.867	0.858	0.841	0.872
	RvNN	0.737	0.662	0.743	0.708
	RDGCN	0.878	0.882	0.879	0.880
	Textual model	**0.897**	**0.883**	**0.879**	**0.889**
Graphics	ResNet-50	0.599	0.591	0.607	0.598
	VGG-19	0.583	0.535	0.587	0.568
	DenseNet-161	0.607	0.593	0.607	0.609
	ViT	**0.681**	**0.634**	**0.637**	**0.641**
Multimodal	att-RNN	0.774	0.773	0.789	0.783
	EANN	0.831	0.846	0.818	0.825
	MAVE	0.829	0.848	0.803	0.817
	SAFE	0.888	0.873	0.859	0.867
	Textual model	**0.911**	**0.916**	**0.917**	**0.914**

4.4 Analysis of Ablation Results

In order to verify the validity of the main modules of the multimodal rumor detection model proposed in this paper based on time graph attention network, the ablation study was further conducted in this paper, and the variant of the rumor detection model in which a module was removed was named as follows:

-time attention: Removes the time attention module. This variant inputs the hidden state of the last timestamp directly into the GRU module.

-GAT: The text rumor detection model of graph attention network module is removed, and this variant is extracted by traditional graph neural network (GNN) and input into GRU module.

-GRU: The text rumor detection model of the GRU module is removed. This variant directly performs rumor detection by concatenating features.

Each ablation experiment was repeated 5 times, and the test set was used for prediction. The ablation experiment results were shown in Fig. 3. According to the experimental results, it can be seen that the removal of any module will cause the overall performance of the model to a certain extent, which indicates that each module is necessary for the model proposed in this paper.

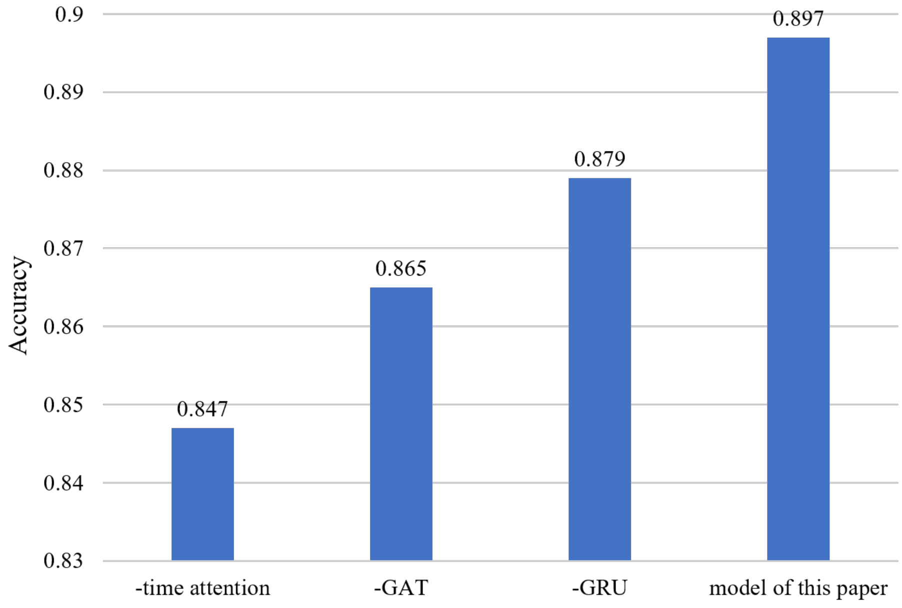

Fig. 3. Ablation results.

5 Conclusion

In this paper, a multimodal rumor detection model based on graph attention network is proposed, which focuses on two aspects: the aggregation of rumor propagation features and text features and the fusion of text features and image features. By introducing the graph attention network based on time and the multimodal fusion algorithm based on attention mechanism, the rumor detection effect of the model is improved to some extent.

This paper also has some shortcomings. Rumors are dynamic and constantly growing, so future research should pay more attention to early detection of rumors. Exploring the application of the rumor detection model with multi-modal feature fusion on time series data and analyzing the changing trend and evolution law of rumor propagation will help provide more accurate rumor detection methods.

References

1. Ma j, Gao W, Mitra P, et al. Detection Rumors from Microblogs with Recurrent Neural Networks[C]//Proceedings of the 25th International Joint Conference on Artificial Intelligence. 2016: 3818-3824.

2. Wang W Y. “Liar, Liar Pants on Fire”: A New Benchmark Dataset for Fake News Detection[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017: 422-426.
3. Chen T, Li X, Yin H, et al. Call attention to rumors: deep attention based recurrent neural networks for early rumor detection[C]//Proceedings of the 22nd Pacific-Asia Conference on Knowledge Discovery and Data Mining. 2018: 4052.
4. Qian F, Gong C, Sharma K, et al. Neural User Response Generator: Fake news detection with collective user intelligence[C]//Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence. 2018: 3834-3840.
5. Ma J, Gao W, Wong K F Detect rumors on twitter by promoting information campaigns with Generative Adversarial Learning[C]//Proceedings of the 28th international conference on World Wide Web. 2019: 3049-3055.
6. Ma J, Li J, Gao W, et al. Improving Rumor Detection by Promoting Information Campaigns with Transformer-based Generative Adversarial Learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2021: 1-1.
7. Yang C, Zhang P, Qiao W, et al. Rumor Detection on Social Media with Crowd Intelligence and ChatGPT-Assisted Networks[C]// Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023: 5705-5717.
8. Jin Z, Cao I, Guo H, et al. Multimodal fusion with recurrent neural networks for rumor detection on microblogs[C]//Proceedings of the 25th ACM International Conference on Multimedia. 2017: 795-816.
9. Wang Y, Ma F, Jin Z, et al. EANN: event adversarial neural networks for multi-modal fake News detection[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018: 849-857.
10. Peng L, Jian S, Li D, et al. MRML: Multimodal Rumor Detection by Deep Metric Learning[C]//IEEE International Conference on Acoustics, Speech and Signal Processing, 2023: 1-5.
11. Bian T, Xiao X Xu T, et al. Rumor detection on social media with bi-directional graph convolutional networks[C]//Proceedings of the AAAI conference on artificial intelligence. 2020, 34(01): 549-556.
12. Ke Z, Li Z, Zhou C, et al. Rumor detection on social media via fused semantic information and a propagation heterogeneous graph[J]. Symmetry, 2020, 12(11): 1806.
13. Hu D, Wei W, Zhou W, et al. A Multi-Relational Propagation Tree-Based Rumor Detection Method[J]. Journal of Computer Research and Development, 2021, 58(7): 1395-1411. in Chinese.
14. Yang Y, Wang L, Wang Y. Rumor Detection by Integrating Source Information and Gated Graph Neural Network[J]. Journal of Computer Research and Development, 2021, 58(7): 1412-1424. in Chinese.
15. Sun T, Qian Z, Dong S, et al. Rumor Detection on Social Media with Graph Adversarial Contrastive Learning[C]// Proceedings of the ACM Web Conference 2022. 2022: 2789-2797.
16. Huang J, Cao D, Lin D. Multimodal Rumor Detection with Causal Graph Attention Network[C]// 2023 IEEE 29th International Conference on Parallel and Distributed Systems (ICPADS). 2023: 17-21.