MAMFND: Multimodal Attention Mechanism for Enhanced Fake News Detection on Social Media

Mei Yang[1*], Yin Xie[2]

College of Big Data Statistics, Guizhou University of Finance and Economics, Guiyang 550025, China
*Corresponding Author

Received: 01 January 2025/ Revised: 10 January 2025/ Accepted: 16 January 2025/ Published: 31-01-2025

Abstract— *In response to the growing prevalence of multimodal false information on social media platforms, traditional single-modal models and basic feature concatenation approaches in multimodal models exhibit limitations in effectively detecting fake news. Therefore, this paper presents a multimodal approach for detecting fake news, integrating a multimodal attention mechanism known as MAMFND (Multimodal Attention Mechanism for Fake News Detection). Initially, we utilize pretrained BERT (Bidirectional Encoder Representations from Transformers) and Swin Transformer (Swin Transformer: Hierarchical Vision Transformer using Shifted Windows) models to extract features from text and images, respectively. Subsequently, we introduce a fusion strategy based on attention mechanisms to integrate textual and visual features. To better capture the intrinsic relationships between text and images, we also input the textual features into a BiLSTM (Bi-directional Long Short-Term Memory) model for temporal sequence modeling, followed by an additional attention-based fusion with visual features. Finally, we extract information from the two rounds of feature fusion and input it into a fake news detection model for classification. Experimental results demonstrate that, on the Weibo and CCF competition datasets, the MAMFND model achieved average accuracy improvements of approximately 9.4% and 5.6%, respectively, compared to baseline models.*

Keywords— Fake News Detection, Multimodal Features Fusion, Multimodal Attention Mechanism, Deep learning.

I. INTRODUCTION

Social media serves as a vital daily information source for the public, thereby making fake news detection a crucial task for enhancing the credibility of disseminated information. The continuous proliferation of fake news on social media poses a significant societal concern, underscoring the pressing need for effective detection methods. The spread of unverified and deceptive information on social media, commonly referred to as online fake news, is characterized by falsehood, exaggeration, provocation, and malicious intent. This pervasive issue extends beyond the mere disruption of social order, also resulting in considerable harm to individuals, businesses, and governments.

The onset of the COVID-19 pandemic witnessed an alarming surge in the spread of fake news across social media platforms. Amid this crisis, individuals and entities spread unsubstantiated claims, such as the efficacy of specific drugs in curing COVID-19. These claims prompted many to procure and use these drugs without scientific validation, sometimes leading to delays in seeking proper medical treatment. Moreover, false assertions circulated, suggesting that the novel coronavirus could be transmitted through the air or that consuming highly alcoholic disinfectants could prevent infection. These misleading narratives not only lacked a factual and scientific basis but also posed significant threats to public safety.

However, existing methods that rely on manual labor or single modalities cannot effectively detect fake news in multimodal data on social media. In the digital age, the sheer volume of information has reached an unprecedented scale, rendering accurate fake news detection through manual identification and single-modal methods increasingly unfeasible. Consequently, there is an urgent imperative to develop more effective multimodal models to combat the proliferation of misinformation in this complex landscape

Many studies have made significant progress in the field of fake news detection using multimodal approaches that combine text and images. For example, Yaqing et al.[1] employed the Text-CNN (Convolutional Neural Networks for Sentence Classification) model to extract text features and the VGG19 (Visual Geometry Group Network 19) model to extract image features, followed by a simple concatenation of these multimodal characteristics applied to fake news detection. Similarly, Qipeng et al.[2] enhanced the model's semantic modeling by recognizing visual entities and text. They utilized the ER-NIE (Enhanced Representation through Knowledge Integration) model to model text features and combined the VGG19 model for image feature extraction, along with incorporating recognized entities and text for semantic enrichment. Likewise, Mengjia et al.[3] utilized BiLSTM models and the VGG19 model to separately extract text and image features, and then concatenated these multimodal features for fake news detection. In addition, the MVAE[4] model introduced a variational autoencoder, learning latent representations of multimodal features by feeding concatenated features into an autoencoder. The MSRD[5] model also considered textual information embedded in images, and after concatenating features from both modalities, reparameterized the multimodal representation by sampling random variables. However, these studies largely relied on the simple concatenation of multimodal features and did not fully exploit text features to understand image features, or they unilaterally applied image-based text attention mechanisms, thereby failing to harness the full potential of text features.

In response to the aforementioned challenges in fake news detection within the realm of social media, this paper introduces an approach that incorporates a multimodal attention mechanism. This approach is designed to address the complexities arising from the vast and diverse nature of online content. To achieve this, we initially employ pretrained BERT models for the extraction of textual features and utilize pretrained Swin Transformer models for the extraction of image features. This strategy leverages the strengths of these pretrained models and incorporates multimodal attention mechanisms into image feature extraction, thereby compensating for the limitations associated with training data scarcity and producing more effective representations for fake news detection.

Furthermore, we incorporate attention mechanism fusion both before and after the integration of text features into a Bidirectional Long Short-Term Memory (BiLSTM) model. The former facilitates the amalgamation of latent information from both images and text, while the latter enhances the model's capacity to capture the nuanced temporal aspects of fake news. This is particularly important because the BiLSTM model excels in modeling the sequential nature of textual data. Ultimately, the resultant multimodal features are channeled into a dedicated fake news detection model for classification.

In this study, we utilized BERT and Swin Transformer models for feature extraction from different modalities. This approach not only enables a more precise capture of features representing text and images but also harnesses dual Transformer encodings, thereby enhancing the model's overall performance. Additionally, we introduced an attention mechanism to effectively fuse the underlying semantic relationships between text and images, mitigating information loss that might occur with simple feature concatenation. Furthermore, by introducing text features into the BiLSTM model and applying an additional fusion step, we further strengthen the modeling of the intrinsic semantics of fake news.

The primary innovative aspects of this research are presented below:

1.Utilizing pretrained BERT and Swin Transformer models for feature extraction, thereby enhancing the efficiency of the process and enabling the seamless integration of features derived from diverse modalities.

2.Introducing an attention mechanism to amalgamate original text features with those extracted by the BiLSTM model, thereby enabling the model to more effectively uncover and exploit the inherent relationships between raw text and features.

3.The resulting multimodal features comprise a combination of features extracted by the BERT and Swin Transformer models, as well as multimodal attention-driven fusion features, which collectively facilitate the enhanced preservation and utilization of information gleaned by the feature extractors.

II. RELATED WORK

Fake news detection is a task that focuses on analyzing multimedia data, including text, images, and videos, to determine the authenticity and credibility of information, thereby distinguishing between real and fake news. Current research in fake news detection can be categorized into two main approaches: single-modal and multimodal.

2.1 Single Modal Approaches:

Single-modal fake news detection methods rely solely on text data to discern fake news. These methods typically employ natural language processing techniques, including sentiment analysis, entity recognition, and keyword extraction.

For example, Liu et al.[6] used the hidden layers of convolutional neural networks for fake news detection, while Qi et al.[7] fused frequency domain and spatial domain information into a CNN (Convolutional Neural Network) model. Song Yurong et al.[8] considered events and their relationships and proposed a fake news detection model based on graph convolutional networks. Latif et al.[9] identified key features of text through principal component analysis and then utilized BiLSTM for fake news identification. Roshan et al.[10] leveraged text feature extraction techniques such as TF-IDF (Term Frequency-Inverse Document Frequency) and hash vectorizers. Zhang et al.[11] identified connections by establishing an internal relationship model between circular content and online communities sharing the same topic. The NER-SA[12] combined natural language processing and named entity recognition to identify fake information. Yuan et al.[13] presented an interpretable fake news analysis method based on stance information. The EBGCN[14] model adaptively controlled message passing based on prior beliefs within the observation graph, replacing fixed edge weights in the propagation graph. Song et al. [26] introduced a method for detecting fake news based on a dynamic propagation graph to capture missing dynamic propagation information in static networks and classify fake news. Truică et al.[15] proposed a novel document embedding method for fake news detection. Li Yuechen et al.[16] combined the BERT model with the RCNN model for fake news detection, while Zhou Lina et al.[17] constructed convolutional neural networks for food safety fake news detection, considering variations across different domains.

These studies provide valuable insights and techniques that contribute to the broader understanding of the challenges and advancements in the domain of false news detection and information verification. Although these single-modal models are research significant and perform well in testing, they fail to fully utilize information from the image modality in fake news, thus presenting limitations.

2.2 Multimodal Approaches:

With the increasing incorporation of image content in fake news, the use of multimodal techniques for fake news detection has gained momentum. Numerous studies have employed various methods and models to extract features from the image modality and combine them with diverse text features for fake news detection. For example, Cao Juan et al. [31] used an LSTM model for text features extraction and VGG19 for visual characteristic extraction, followed by the fusion of information from these two modalities for fake news detection. Meng Jie et al.[18] thoroughly considered the relationship between text and image features, employing attention mechanisms for both inter-modal and intra-modal fusion. The MFAN model, as presented in[19], notably addresses the crucial factors of complementarity and alignment among various modalities. By effectively integrating images, vision, and social commentary, it achieves a superior level of performance. The Att-MFNN model[20] extracted sentiment features from text, fused them with text features extracted by the BERT model and image features, and used them for fake news detection. Liu et al.[21] integrated image description information into the text to bridge the semantic gap between text and images. The MSRD model[5] fully considered textual information embedded in images, concatenated features from both modalities, and reparameterized the multimodal representation through sampling random variables. The EANN[1] model encoded text and image features separately using the Text-CNN model and VGG19 model and then concatenated the obtained features for fake news detection. The MCAN model[22] designed multiple co-attention modules to combine frequency domain and spatial domain information in image features for fake news detection. These multimodal models, which take image information into account, have performed remarkably well and outperformed single-modal models.

However, these approaches also have certain limitations. Most of them simply concatenate features extracted from the two modalities without deeply considering the underlying connections between text and images. Additionally, in the extraction of image features, most models use convolutional neural networks (CNNs) without considering the advantages of extracting image features on the basis of the Transformer framework. Recently, the Swin Transformer model has demonstrated outstanding performance in image feature extraction and image classification, particularly achieving remarkable results in the field of multimodal classification. Therefore, this paper summarizes the aforementioned experiences, overcomes some of the limitations in previous studies, and proposes a multimodal attention mechanism fusion-based fake news detection method.

III. THE PROPOSED MODEL

In this paper, we propose a multimodal fake news detection model that integrates textual and visual information. The model utilizes the BERT model to extract textual features and the Swin Transformer model to extract image features. An attention mechanism is employed to fuse features from both modalities. The entire model comprises three main components (as illustrated in Fig.1): a multimodal feature extractor, a multimodal attention-based feature fusion module, and a multimodal fake news classifier.

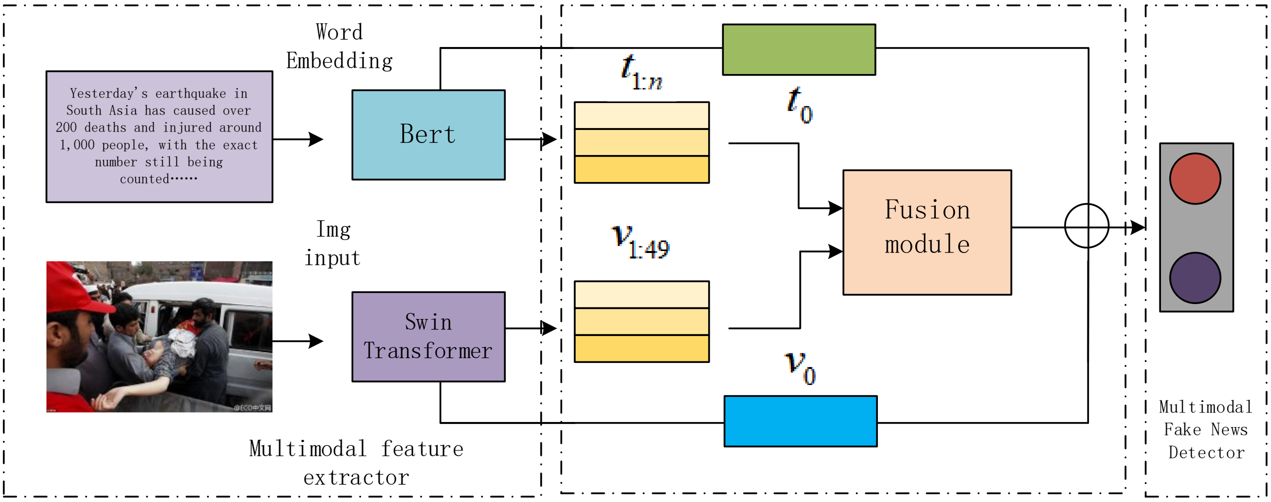
Word Embedding
Yesterday's earthquake in South Asia has caused over 200 deaths and injured around 1,000 people, with the exact number still being counted……
Bert
Img input
Swin Transformer
Multimodal feature extractor
$t_{1:n}$
t_0
$v_{1:49}$
v_0
Fusion module
Multimodal Fake News Detector

FIGURE 1: Overall procedure of proposed model

First, the multimodal feature extractor employs pretrained BERT and Swin Transformer models to extract features from textual and image data, respectively. These pretrained models facilitate enhanced and expedited understanding of both textual and visual information. Second, the multimodal attention-based feature fusion module is responsible for merging the features extracted from both modalities. During this process, the textual features are input into a BiLSTM model to capture temporal relationships among the features, thereby enhancing the fusion of latent connections between different modalities. Finally, the multimodal fake news classifier receives the processed features and performs fake news classification. The specific operations are described as follows.

3.1 Multimodal Feature Extractors:

The extractors for multimodal features comprise both text and image feature extractors. To better capture key semantic information within text and address issues such as polysemy and parallel training, this study adopts the BERT model, based on the Transformer framework, as the text feature extractor. BERT is a bidirectional pre-trained language model based on the Transformer architecture. Its core concept involves leveraging a substantial amount of unlabeled text data for pre-training to acquire rich language representations. Unlike previous unidirectional pre-training methods, BERT achieves robust natural language modeling through innovative techniques, including bidirectional masking, the Transformer architecture, unsupervised pre-training, Masked Language Modeling (MLM) tasks, and Next Sentence Prediction (NSP) tasks.

Initially, the Bert model converts input text into corresponding vector representations. The i -th word in the text can be represented as $e_i \in R^k$, where k has a default dimension for feature extraction, typically set at 768. Consequently, text of length n can be transformed into the following representation:

$$e_{0:n} = [e_0, e_1, \dots, e_n] \tag{1}$$

Where e_0 is the special symbol $[CLS]$ in front of each text, $e_{1:n}$ stands for the feature vector corresponding to the text.

After undergoing training, the pre-trained Bert model can be referred to as f_{bert}. After converting the text to vector form and inputting it into f_{bert}, the model can reconstruct each token of word and calculate it as follows:

$$t_{0:n} = f_{bert}(e_{0:n}) \tag{2}$$

Where $t_{0:n}$ is the output of $e_{0:n}$ after passing through the model.

The presence of fake news in images often includes important details that are misleading. Thus, it becomes crucial to extract such features from the images. This has a significant impact on the effectiveness of feature fusion in models. Numerous experimental results have demonstrated that the Swin Transformer model has reached remarkable performance within the domain of computer vision. Developed by Microsoft Research, this model is a combination of traditional CNNs and Transformer architectures, utilizing the latter's powerful modeling able to catch long-distance relationships of dependencies in images. Furthermore, the use of pre-trained models can aid in understanding the information present in images.

In this study, both the Bert model for text feature extraction and the Swin Transformer model for image feature extraction were chosen as Transformer encoders, which facilitate compatibility in subsequent feature fusion. Compared to traditional ViT (Vision Transformer) models, the Swin Transformer model offers higher computational efficiency and better image representation capabilities. It retains the sensitivity of CNNs to local features while also possessing the strong modeling ability of Transformers. (as illustrated in Fig.2): Firstly, the image is partitioned into 56 patches using the patch partition function of the model. Each patch then undergoes linear transformation followed by attention mechanisms calculated through conventional and sliding windows. Finally, the dimensionality is reduced and the number of features is decreased through fusion and merge layers, resulting in the final output of feature representations.

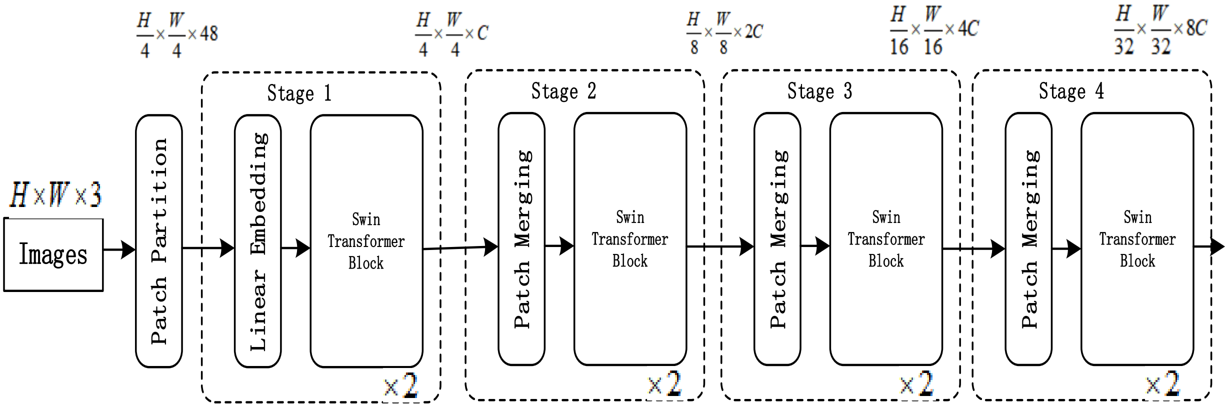

FIGURE 2: Overall procedure of Swin Transformer

The Swin Transformer differs from the traditional Transformer framework in that it employs regular window W-MSA and sliding window SW-MSA for attention mechanism computation. (as illustrated in Fig.3): windows are connected by MLP layers with activation functions, and residual connections are employed between modules.

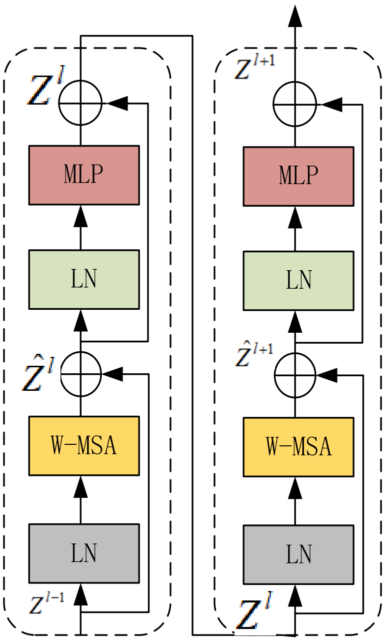

FIGURE 3: The blocks of Swin Transformer

The Swin Transformer model employs the local receptive field of convolutional neural networks (CNNs) for attention calculation by computing the relationships between each patch within its local window. The formula for calculating the window-related quantity is specified as follows:

$$\hat{Z}^l = W - MSA(LN(Z^{l-1})) + Z^{l-1}$$
$$Z^l = MLP(LN(\hat{Z}^l)) + \hat{Z}^l$$

$$\hat{Z}^{l+1} = SW - MSA(LN(Z^l)) + Z^l$$
$$Z^{l+1} = MLP(LN(\hat{Z}^{l+1})) + \hat{Z}^{l+1} \quad (3)$$

The initial representation of an image is denoted as v, and after undergoing the effect of the Swin Transformer model, it is represented as f_{swin}. The feature extractor transforms the image into the following representation:

$$v_0, v_{1:49} = f_{swin}(v) \quad (4)$$

Where, $v_0 \in R^k$ represents the feature vector employed for classification tasks after passing through the fully connected layer of the model, and $v_{1:49} \in R^k$ represents the image feature matrix extracted by the model.

3.2 Multimodal Attention Mechanism based Multimodal Feature Fusion:

The purpose of the multimodal attention mechanism based multimodal feature fusion module is to fuse the features extracted by the feature extractor and explore the potential connections between im-age and text features. In order to achieve deep inter-action between text information and visual information, this paper calculates the similarity scores between text feature vectors and image feature vectors. By using the similarity weights to reconstruct the text features, the final multimodal feature representation matrix is obtained by concatenating the two attention mechanisms' fused feature matrices. (as illustrated in Fig.4):

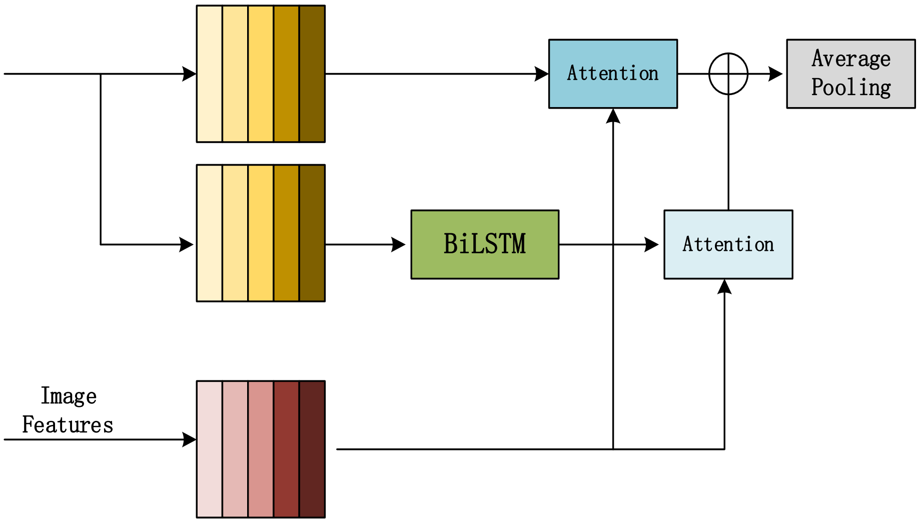

FIGURE 4: Overall Fusion Framework

The multimodal attention mechanism based multimodal feature fusion module proposed in this pa-per is divided into two parts: the first part aims to fuse unprocessed text features based on image attention, with the purpose of mining better expressions of text features through information from images; the second part involves inputting text features into a BiLSTM model for time series modeling and then fusing them again based on image attention. This helps to uncover potential temporal relationships between word-level features modeled by Bert and thus better capture text feature matrices through attention mechanisms. The specific operations are as follows:

If the text features are input into a BiLSTM model, they can be represented as f_{lstm}. The following mathematical formula can be used:

$$tm = f_{lstm}(t_{1:n}) \quad (5)$$

Where, $tm \in R^{n \times m}$ represents the feature matrix obtained after passing through the BiLSTM model, and$m$ represents the dimensionality.

The formula for fusing the attention mechanism between the textual and modality features is as follows:

$$Attention(Q, K, V) = softmax(QK^T/\sqrt{d})V \quad (6)$$

Where, Q, K, and V represent the query matrix, key matrix, and value matrix respectively. d is a scaling coefficient used to prevent the denominator from becoming too large and is typically the dimensionality of the vectors.

To fuse the attention mechanism across modalities, the following representation can be used:

$$\alpha = Attention(t_{1:n}W_{q1}, v_{1:49}W_{k1}, v_{1:49}W_{v1}) \tag{7}$$

$$\beta = Attention(tmW_{q2}, v_{1:49}W_{k2}, v_{1:49}W_{v2}) \tag{8}$$

Where, $\alpha, \beta \in R^{n\times k}$ represents the fused features after fusing modalities, and W_{q1}, W_{k1}, W_{v1}, W_{q2}, W_{k2}, W_{v2} is the weight matrix. To obtain a feature vector that can represent the fused features, average pooling is used for extraction:

$$\gamma = AvgPool(\alpha \oplus \beta) \tag{9}$$

Where, $\gamma \in R^{1\times 2k}$ represents the fused feature matrix after pooling.

The final multimodal feature representation of the model is:

$$s = t_0 \oplus v_0 \oplus \underline{\gamma} \tag{10}$$

Where, t_0, v_0 represents the features extracted by the model feature extractor, and $\underline{\gamma}$ is a modified version of γ.

3.3 Multimodal Fake News Detector:

A multimodal fake news detector takes the final representation of multiple modalities as input, and uses fully connected layers and normalization layers to classify fake news into true and false fake news. The normalization formula is as follows:

$$y = \sigma(ws + b) \tag{11}$$

Here, w is the weight matrix, b is the bias, σ is the normalization function, and y is the predicted probability of being fake. The cross-entropy function is used as the objective function in this study, with fake news assigned a value of 0 and non-fake news assigned a value of 1. The formula is as follows:

$$L = -\sum_{i=1}^{m}[y_i \, log\, p_i + (1 - y_i)\, log(1 - p_i)] \tag{12}$$

IV. EXPERIMENT

4.1 Experimental Setup:

In this study, all textual data were tokenized using the Chinese tokenizer provided by BERT. Images underwent scaling and normalization procedures to ensure a uniform size of 224×224 pixels. The experimental setup employed a Linux operating system, a Tesla P100-PCIE-16GB GPU, and CUDA version 12.0. The model parameters were set as follows: a text sentence length with an average value of 120 based on the dataset, a batch size of 12, a learning rate of 0.01, ReLU as the activation function, and reconstruction of the text feature with a dimensionality of 768 after modeling. The BiLSTM model comprises 128 neurons. For image characterization extraction, the input size was set to 224×224×3, and the neural network architecture preceding the image feature extractor mirrored that of the Swin Transformer. Finally, the model outputs both extracted image features and fully connected classification features.

4.2 Dataset:

The experimental datasets used in this study consist of two publicly available social media datasets, as presented in Table 1. The Weibo dataset was obtained from the EANN[1] dataset and subsequently cleaned and reorganized for use in this study. The CCF competition dataset was collected from the “Internet Fake News Detection during the Epidemic” competition organized by the China Computer Federation (CCF). This dataset encompasses eight domains: health, economy, technology, entertainment, society, military, politics, and education. The data processing methods applied to both datasets are similar to those used for the microblog dataset.

TABLE 1
STATISTICS OF THE DATASET

	Weibo	CCF competition
Non-fake news	3642	7500
Fake news	4203	7500
Total	7845	15000

4.3 Baseline Model and Evaluation Metrics:

To evaluate the effectiveness of the model, several common multimodal models were selected for comparison. The specific configurations of the models are as follows:

EANN[1] model: The EANN is an end-to-end adversarial neural network, Composed of three elements: a feature-fetching part, a rumor-detecting part, and an event-differentiating part. In this study, the event discriminator module was removed for comparison purposes.

MENG[3] model: The MENG proposed by Meng et al., is a rumor detection model based on adversarial neural networks, divided into cross-modal feature extractors, rumor detectors, and event discriminators. Similarly, the event discriminator module was removed for comparison purposes.

att-RNN[23] model: This model fuses features processed by recurrent neural networks and convolutional neural networks. In this study, the social context module was removed for comparison purposes.

MMDF[18] model: The MMDF model aims to fully leverage the temporal nature of features modeled by recurrent neural networks and to extract features from corresponding convolutional neural networks. The model performs deep integration of these features.

PTCA[24] model: The model aims to perform modality-specific feature fusion using cross-attention mechanisms. Pre-trained models are employed by both the text feature extraction component and the visual characteristic extraction component.

In this study, the evaluation metrics used were confusion matrices, consist of accuracy, precision, recall, and F1-score. The outcomes obtained from the model were used for evaluation. The predicted results were classified into four categories: true positive (fake news correctly identified as fake), false negative (fake news incorrectly identified as non-fake), false positive (non-fake news incorrectly identified as fake), and true negative (non-fake news correctly identified as non-fake).

4.4 Experimental Results and Analysis:

Table 2 presents the performance of the MAM-FND model alongside baseline models on both the Weibo and CCF competition datasets. The table highlights the best performance achieved by each model type, indicated in bold. The accuracy and F1 score comparison graphs are depicted in Fig.5 and Fig.6.

TABLE 2
COMPARISON OF MODEL PERFORMANCE

Dataset	Model	Accuracy	Non-fake news			Fake news		
			Precision	Recall	F1	Precision	Recall	F1
Weibo	EANN	0.819	0.778	0.858	0.816	0.858	0.777	0.816
	MENG	0.773	0.758	0.771	0.764	0.787	0.775	0.781
	att-RNN	0.779	0.742	0.824	0.781	0.821	0.738	0.777
	MMDF	0.718	0.689	0.746	0.716	0.749	0.693	0.720
	PTCA	0.850	0.782	0.950	0.858	0.943	0.758	0.841
	MAMFND	**0.882**	**0.837**	**0.934**	0.883	**0.932**	**0.834**	**0.881**
CCF	EANN	0.941	0.949	0.933	0.941	0.934	0.950	0.942
	MENG	0.885	0.880	0.891	0.885	0.890	0.878	0.884
	att-RNN	0.904	0.916	0.889	0.902	0.892	0.918	0.905
	MMDF	0.900	0.900	0.895	0.897	0.895	0.901	0.898
	PTCA	0.970	0.977	0.962	0.969	0.963	0.977	0.970
	MAMFND	**0.977**	0.972	**0.982**	**0.977**	**0.982**	**0.971**	**0.977**

The outcome suggests that the PTCA model and our proposed model significantly outperform other multimodal models in detecting fake news. The attention mechanism-based multimodal feature fusion is clearly superior to simple feature concatenation. Furthermore, attention mechanism fusion for both image and text features can better explore potential connections between different modalities. However, the accuracy of the MAMFND model is lower than that of some feature concatenation models. This suggests that using a Transformer-based feature extractor can better integrate attention mechanisms across modalities and enhance the accuracy of fake news detection.

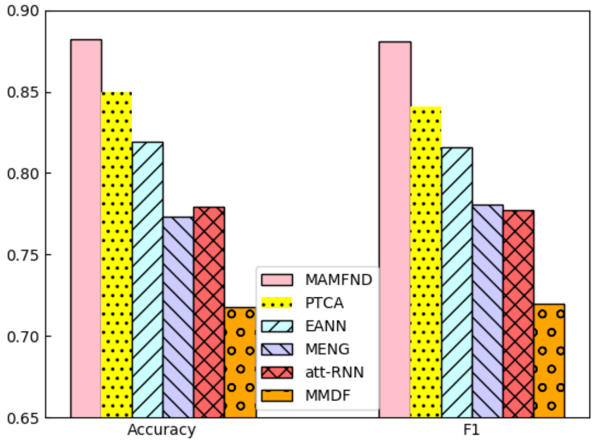

FIGURE 5: Results on Weibo

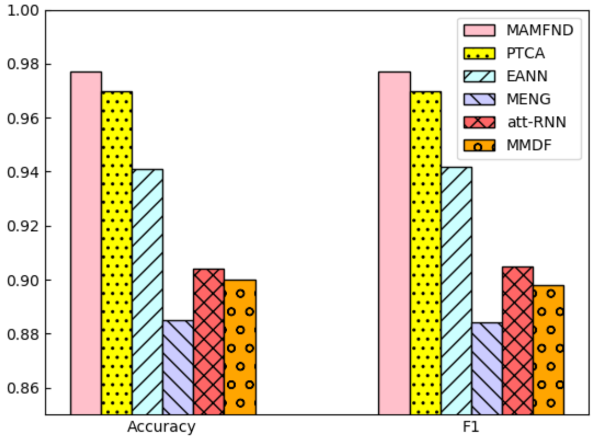

FIGURE 6: Results on CCF

The experimental results presented in Fig.5 and Fig.6, and Table 2 consistently indicate that the proposed model exhibits superior performance compared to the baseline models across both datasets. In the Weibo dataset, the proposed model outperforms the EANN by about 7.1%, MENG by 10.9%, att-RNN by 10.2%, MMDF by 16.2%, and PTCA by 3% in terms of accuracy. In the CCF competition dataset, the accuracy of the proposed model is approximately 3.6%, 8.7%, 7.3%, 7.7%, and 0.7% higher than those of the EANN, MENG, att-RNN, MMDF, and PTCA models, respectively. These results further identify the effectiveness of the model and indicate the advantageous of multimodal feature attention mechanism fusion.

4.5 Multimodal Feature Visualization:

To better demonstrate the effectiveness of the MAMFND model, we perform dimensionality reduction visualization on the multimodal final feature representations of relevant models. Fig.7 shows the visualization results of the models using the t-SNE algorithm on two datasets, where red and blue represent fake news and non-fake news, respectively. Each subfigure represents: (a) Visualization of the multimodal final feature representations of the PTCA model on the Weibo dataset. (b) Visualization of the multimodal final feature representations of the MAMFND model on the Weibo dataset. (c) Visualization of the multimodal final feature representations of the Att-RNN model on the CCF dataset. (d) Visualization of the multimodal final feature representations of the MAMFND model on the CCF dataset.

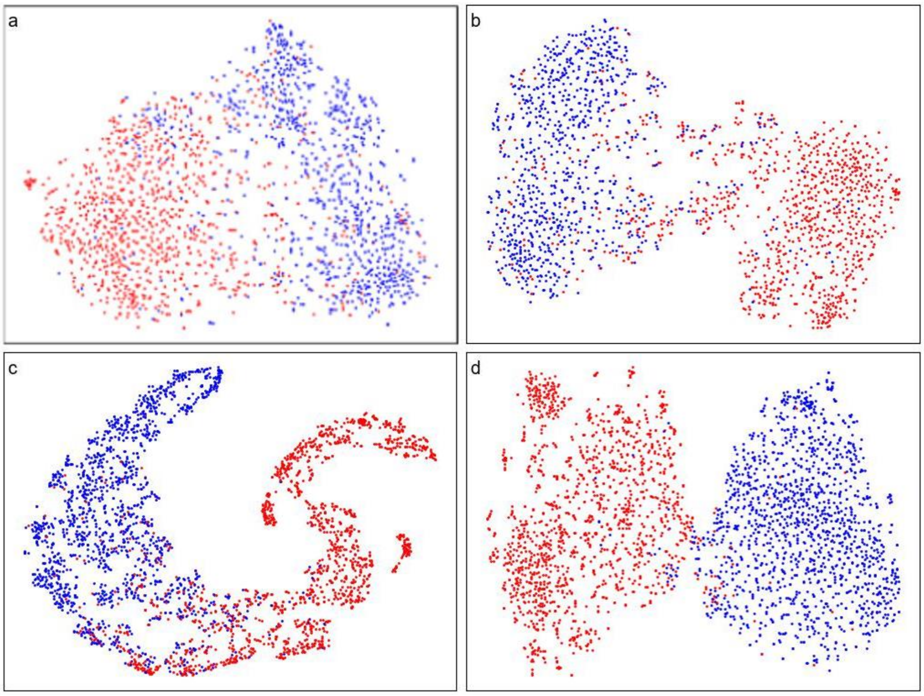

FIGURE 7: Visualization of the multimodal final representations of relevant models on the dataset using t-SNE algorithm

From Fig.7, it can be observed that the reference model has made numerous misclassifications, while the proposed model is able to better distinguish between the red and blue regions, resulting in clearer separation and more concentrated clustering. This demonstrates that the proposed model outperforms the reference model and further supports the claim that attention mechanism fusion can better explore potential connections among modalities.

4.6 Ablation Experiment:

To gain a clear understanding of the role of each module in the model, an ablation experiment was conducted on the Weibo dataset, and the experimental results are presented in Table 3. In this context, "w/o lstm" represents the removal of attention mechanisms in the time series modeled by the LSTM. "w/o att" represents the deletion of the attention mechanism fusion module, where image and text features are connected through concatenation instead. "w/o img" stands for the removal of the image feature extraction module, retaining only the text feature extraction module. "w/o text" stands for the removal of the text feature extraction module, retaining only the image feature extraction module. "MAMFND" represents retaining all modules of the model, and the visualization diagram of the ablation experiment is shown in Fig.8.

TABLE 3
RESULTS OF THE ABLATION EXPERIMENT

Dataset	Model	Accuracy	Precision	Recall	F1
Weibo	w/o lstm	0.870	0.929	0.814	0.867
	w/o att	0.861	0.933	0.791	0.856
	w/o img	0.818	0.839	0.807	0.823
	w/o text	0.807	0.808	0.826	0.817
	MAMFND	**0.882**	**0.932**	**0.834**	**0.881**

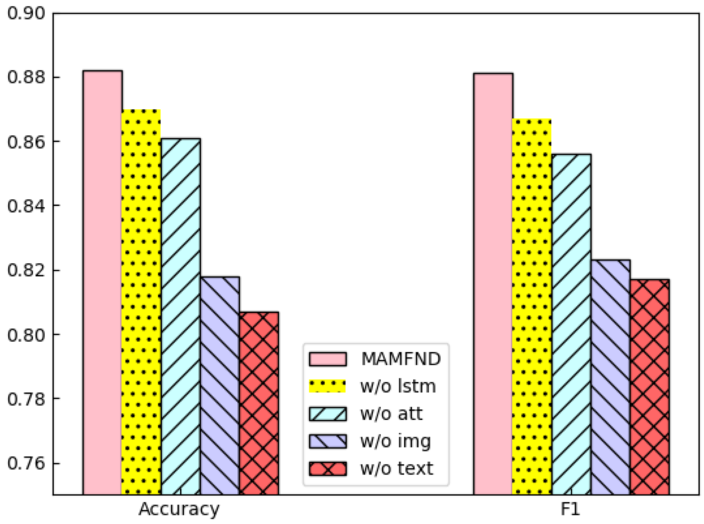

FIGURE 8: Visualization of the results of the ablation experiment

The results of the ablation experiment show that when the BiLSTM model's attention mechanism is not fused, the accuracy decreases by approximately 1.2%, demonstrating that fusing the attention mechanism with time series on text features is capable of enhancing the model's fake news detection capability. When the attention mechanism between text and image features is removed, the accuracy of our proposed model decreases by approximately 2%. This indicates that fusing attention mechanisms between text and image features can better explore potential correlations between modalities and yields better results compared to simple feature concatenation. Compared to unimodal detection of fake news, the accuracy of multimodal models decreases by approximately 7%. This demonstrates that multimodal approaches contain more information than unimodal ones and are more effective in improving fake news detection, thereby making multimodal approaches of great significance.

V. CONCLUSION

The widespread dissemination of fake news can have serious negative impacts on society. Compared to pure text-based fake news, multimodal fake news containing both text and images is more likely to have a greater impact, as people are more easily attracted by images and tend to overlook the fake news embedded within. This can lead to misunderstandings about certain events or issues among the public, potentially causing panic, chaos, and even violent conflicts. Therefore, it is necessary to detect fake news using multimodal technology. Within this paper, we introduce a multimodal feature attention mechanism fusion model for detecting fake news. The model employs a feature extractor based on the Transformer framework to extract image and text information. It captures the interaction between modalities through attention mechanisms and further mines the potential correlations between image and text features by modeling time series on text features to enhance the role of attention mechanisms. The experimental results on relevant datasets show that the text model outperforms the baseline model in all aspects. Relevant data indicate that fake news not only includes text and images but also user comments after browsing, which contain authenticity information about fake news. Therefore, in future work, we will combine text, images, and comments to verify the authenticity of fake news.

ACKNOWLEDGEMENTS

This research was funded by the Guizhou Provincial Science and Technology Fund (Qian Kehe Basic-ZK [2021] General 337), supported by the Fund of the State Key Laboratory of Public Big Data, Guizhou University (No. PBD2023-35).

REFERENCES

[1] Y. Wang *et al.*, "Eann: Event adversarial neural networks for multi-modal fake news detection," in *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*, 2018, pp. 849-857.

[2] P. Yuan, J. Cao, and Q. Sheng, "Semantics-enhanced multi-modal fake news detection," *J. Comput. Res. Dev,* vol. 58, p. 1456, 2021.

[3] M. Jiana, W. Xiaopei, L. Ting, L. Shuang, and Z. Di, "Cross-modal rumor detection based on adversarial neural network," *Data Analysis and Knowledge Discovery,* vol. 6, no. 12, pp. 32-42, 2023.

[4] D. Khattar, J. S. Goud, M. Gupta, and V. Varma, "Mvae: Multimodal variational autoencoder for fake news detection," in *The world wide web conference*, 2019, pp. 2915-2921.

[5] Liu Jinshuo, Feng Kuo, J. Z. Pan, Deng Juan, and W. Lina., "MSRD: Multi-Modal Web Rumor Detection Method," *Computer Research and Development,* vol. 57, no. 11, pp. 2328-2336, 2020.

[6] Z. Liu, Z. Wei, and R. Zhang, "Rumor detection based on convolutional neural network," *Journal of Computer Applications,* vol. 37, no. 11, p. 3053, 2017.

[7] P. Qi, J. Cao, T. Yang, J. Guo, and J. Li, "Exploiting multi-domain visual information for fake news detection," in *2019 IEEE international conference on data mining (ICDM)*, 2019: IEEE, pp. 518-527.

[8] W. Xin-yan, S. Yu-rong, and S. Bo, "Sina Microblog Rumor Detection Method Based on Weighted-graph Convolutional Network," *Journal of Chinese Computer Systems,* vol. 42, no. 08, pp. 1780-1786, 2021.

[9] A. A. Ali, S. Latif, S. A. Ghauri, O.-Y. Song, A. A. Abbasi, and A. J. Malik, "Linguistic Features and Bi-LSTM for Identification of Fake News," *Electronics,* vol. 12, no. 13, p. 2942, 2023.

[10] R. Roshan, I. A. Bhacho, and S. Zai, "Comparative Analysis of TF–IDF and Hashing Vectorizer for Fake News Detection in Sindhi: A Machine Learning and Deep Learning Approach," *Engineering Proceedings,* vol. 46, no. 1, p. 5, 2023.

[11] Y. Zhang, W. Song, Y. H. Koura, and Y. Su, "Social bots and information propagation in social networks: Simulating cooperative and competitive interaction dynamics," *Systems,* vol. 11, no. 4, p. 210, 2023.

[12] C.-M. Tsai, "Stylometric fake news detection based on natural language processing using named entity recognition: In-domain and cross-domain analysis," *Electronics,* vol. 12, no. 17, p. 3676, 2023.

[13] L. Yuan, H. Shen, L. Shi, N. Cheng, and H. Jiang, "An explainable fake news analysis method with stance information," *Electronics,* vol. 12, no. 15, p. 3367, 2023.

[14] C. Yuan, Q. Ma, W. Zhou, J. Han, and S. Hu, "Jointly embedding the local and global relations of heterogeneous graph for rumor detection," in *2019 IEEE international conference on data mining (ICDM),* 2019: IEEE, pp. 796-805.

[15] L. Wei, D. Hu, W. Zhou, Z. Yue, and S. Hu, "Towards propagation uncertainty: Edge-enhanced bayesian graph convolutional networks for rumor detection," *arXiv preprint arXiv:2107.11934,* 2021.

[16] C. Song, Y. Teng, Y. Zhu, S. Wei, and B. Wu, "Dynamic graph neural network for fake news detection," *Neurocomputing,* vol. 505, pp. 362-374, 2022.

[17] Y. Li, L. Qian, and J. Ma, "Early detection of micro blog rumors based on BERT-RCNN model," *Information studies: Theory & Application,* pp. 173-177, 2021.

[18] J. Alghamdi, Y. Lin, and S. Luo, "Towards COVID-19 fake news detection using transformer-based models," *Knowledge-Based Systems,* vol. 274, p. 110642, 2023.

[19] Z. Jin, J. Cao, H. Guo, Y. Zhang, and J. Luo, "Multimodal fusion with recurrent neural networks for rumor detection on microblogs," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 795-816.

[20] J. Zheng, X. Zhang, S. Guo, Q. Wang, W. Zang, and Y. Zhang, "MFAN: Multi-modal Feature-enhanced Attention Networks for Rumor Detection," in *IJCAI*, 2022, vol. 2022, pp. 2413-2419.

[21] S. Xiong, G. Zhang, V. Batra, L. Xi, L. Shi, and L. Liu, "TRIMOON: Two-Round Inconsistency-based Multi-modal fusion Network for fake news detection," *Information fusion,* vol. 93, pp. 150-158, 2023.

[22] P. Liu, W. Qian, D. Xu, B. Ren, and J. Cao, "Multi-modal fake news detection via bridging the gap between Modals," *Entropy,* vol. 25, no. 4, p. 614, 2023.

[23] H. Xiao-xia, G. Tuerhong, M. Wushouer, and W. Song, "Weibo rumors continuous detection model combining BERT word embedding and BiLSTM," *Journal of Northeast Normal University(Natural Science Edition),* vol. 55, no. 01, pp. 65-71, 2023, doi: 10.16163/j.cnki.dslkxb20210905001.

[24] D. Abudureyimu, M. Bo, Y. Yating, and W. Lei, "Attention based multi-feature fusion neural network for fake news detection," *Journal of Xiamen University(Natural Science),* vol. 61, no. 04, pp. 608-616, 2022.