IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 42, No. 3, June, 2025, pp. 1553-1564
Journal homepage: http://iieta.org/journals/ts

Detecting Fake News on Social Media via Multimodal Semantic Understanding and Enhanced Transformer Architectures

Meiling Xu, Feng Li*, Zhigang Miao, Zhuhua Han, Lei Wang, Gong Wang

School of Information and Artificial Intelligence, Hebei Finance University, Baoding 071066, China

Corresponding Author Email: chnlifeng2008@163.com

https://doi.org/10.18280/ts.420327

Received: 21 October 2024
Revised: 19 April 2025
Accepted: 30 April 2025
Available online: 30 June 2025

Keywords:
fake news detection, multimodal learning, social media, transformer model, text-image semantic matching, cross-modal attention

ABSTRACT

With the rapid development of social media, multimodal news combining text and images has become a primary vehicle for the spread of misinformation due to its strong dissemination power and broad audience reach. Existing fake news detection methods overly rely on textual analysis, making it difficult to effectively capture the deep semantic relationships between visual and textual modalities. Moreover, traditional neural networks often suffer from limited robustness and weak noise resistance when handling heterogeneous multimodal data. Although multimodal learning has recently been introduced into this field, challenges such as insufficient text-image semantic matching and a lack of architectural innovation persist. To address these issues, this study proposes a detection framework that integrates multimodal semantic understanding with an enhanced Transformer architecture. Specifically, a cross-modal attention mechanism is constructed to strengthen the interactive representation of visual and textual features through semantic alignment. Additionally, a hierarchical Transformer structure with a dynamic gating mechanism is designed to optimize the multimodal information fusion strategy, significantly improving the model's adaptability to complex social media scenarios. Experiments conducted on public datasets demonstrate that the proposed method achieves a noticeable improvement in detection accuracy compared to traditional models, while also exhibiting superior recall and noise robustness. This research offers a more efficient technological pathway for fake news detection on social media and promotes the broader application of multimodal learning in the domain of information content security.

1. INTRODUCTION

With the rapid development of the internet and social media, the way people access and share information has undergone profound changes [1-4]. However, this change has also brought about the problem of the proliferation of fake news, especially on social media platforms, where news reports combining images and text are more likely to spread widely and mislead [5, 6]. Fake news not only harms public opinion [7], but can also cause social panic and turmoil [8]. Therefore, how to effectively identify and detect fake news on social media has become an important issue that needs to be addressed.

The detection of fake news not only helps to maintain the healthy development of the information ecosystem but also enhances the public's trust in media information [9-11]. Currently, most fake news detection methods mainly rely on text analysis [12, 13]. However, the diverse forms of news on social media and the combination of images and text in news content make it difficult to meet the needs using only text analysis [14-17]. Cui and Yang [18] mention that multimodal learning methods can comprehensively utilize both image and text information, thereby improving the accuracy and robustness of fake news detection. Therefore, researching the application of multimodal learning in detecting fake news in social media images has important theoretical and practical significance.

Although some studies have attempted to apply multimodal learning to fake news detection, several issues still exist. For example, Tuerhong et al. [19] did not fully consider the deep semantic relationships between different modalities in the image-text matching process, leading to lower detection accuracy. Additionally, most existing methods rely on traditional neural network models, neglecting the potential advantages of the Transformer architecture in multimodal semantic understanding. Specifically, Ang and Lim [20] pointed out that existing methods perform poorly when handling large-scale heterogeneous data and are easily influenced by data noise.

This paper is divided into two parts: first, the image-text matching of social media news based on multimodal semantic understanding, and second, fake news detection in social media images based on an improved Transformer. By introducing a multimodal semantic understanding model, we hope to better capture the associations between images and text, improving the accuracy of image-text matching. Based on this, the improved Transformer model will further enhance the ability to detect fake news in social media images. This research not only provides new technical means for fake news detection but also offers strong support for the promotion of

multimodal learning in real-world applications, with significant research value and application prospects.

2. IMAGE-TEXT MATCHING OF SOCIAL MEDIA NEWS BASED ON MULTIMODAL SEMANTIC UNDERSTANDING

On social media platforms, news content is typically presented in a combination of images and text, with the images and text complementing each other to convey information. In the spread of fake news, images and text often work together, and fake information may be more deceptive and misleading through the combination of images and text. Therefore, relying solely on text analysis or image recognition to detect fake news often fails to achieve ideal results. To address this, this paper conducts research on image-text matching for social media news based on multimodal semantic understanding, aiming to effectively combine image and text information and fully explore the semantic correlations between them, thereby improving the accuracy of fake news detection. Image-text matching research helps to understand the interaction and semantic consistency between images and text in news reports, and by comparing the semantics of the images and text in the news content, it determines whether there are contradictions or misleading information, thus assessing the authenticity of the news.

The model constructed is based on an end-to-end image-text matching architecture, with a focus on improving fake news recognition by extracting key samples and analyzing their semantic features. To effectively extract the key semantics of the samples, this research adopts a sample diversion strategy based on Gaussian Mixture Models, filtering the image-text data to retain the core semantic information of each sample. The training of the model is divided into two stages. In the first stage, the model predicts the negative impact of key samples and evaluates their influence on fake news detection. In the second stage, the image-text matching model is formally trained by combining key semantics and negative impact. This phased training method enables the model to gradually optimize its image-text matching ability and improve the accuracy of fake news recognition.

In the specific implementation of image-text matching, this paper introduces a Siamese network structure to ensure that the image-text matching learning process is more stable and consistent. In each training step, the Siamese network X and X' models share parameters, ensuring the consistency of the structure and internal parameters of the two models, which is crucial for handling the semantic associations in image-text information. During the training process, the X' model is first pre-trained using the original image-text data to compute the matching loss function and predict the negative impact of each sample, i.e., the weights. These weights reflect the importance of the sample in fake news detection, further helping the model focus on key samples for learning. In the second stage, the X model retrains using the original image-text data, key samples, and negative impact weights, ultimately obtaining the optimized image-text matching model.

2.1 Image-text feature extraction and similarity measurement

In the process of feature extraction, this paper uses the pre-trained model CLIP as the baseline model. CLIP consists of two independent Transformer encoders, which are used to process visual and textual information, respectively. The visual encoder $d(\cdot)$ uses the pre-trained Vision Transformer encoder *ViTB*/32, which adds a layer of normalization before output to enhance the stability and accuracy of feature extraction. The text encoder $h(\cdot)$ consists of 12 layers of Transformer structures with a width of 512, equipped with 8 attention heads, to capture the semantic details in the text. Both encoders were pre-trained on the large-scale *WIT* dataset, ensuring their strong feature extraction capability. In specific applications, the visual encoder $h(\cdot)$ extracts high-dimensional visual features from the image, while the text encoder $h(\cdot)$ extracts high-dimensional textual features from the text. These features will be used in subsequent steps to calculate the similarity between the image and text. Let $F = (U, S)$ represent a social media news image-text dataset, and the similarity calculation between modalities can be represented as follows:

$$T\left(U_u, S_k\right) = \frac{d\left(U_u\right) \cdot h\left(S_k\right)}{\left\|d\left(U_u\right)\right\| \times \left\|h\left(S_k\right)\right\|} \tag{1}$$

This paper uses cross-entropy loss to measure the matching degree between image-text features. During training, the image and text are first input into the visual encoder and text encoder, respectively, to obtain the corresponding feature vectors. Then, the cosine similarity between the image and text feature vectors is calculated to evaluate the proximity of the image-text pair in the semantic space. Suppose the image or text is represented by a_u and b_u, with a batch size denoted by v, then we have:

$$\begin{aligned} ZR\left(a_u, b_u\right) &= -\log\left(\frac{\exp\left(T\left(a_u, b_u\right)\right)}{\sum_{k=1}^{v} \exp\left(T\left(a_u, b_u\right)\right)}\right) \\ LOSS_{ZR} &= \frac{1}{v}\sum_{u=1}^{v}\left(ZR\left(U_u, S_u\right) + ZR\left(S_u, U_u\right)\right) \end{aligned} \tag{2}$$

2.2 Key semantic retention

In the task of social media image fake news detection, the image-text matching model faces complex challenges, including the interference of noise samples and hard samples. Noise samples are those where the semantics between the image and text are completely inconsistent, while hard samples refer to those where there is partial inconsistency in the semantic information between the image and text. When these noise and hard samples are included during the training process, the model may overfit these erroneous samples, leading to deviations in the prediction of the actual image-text correspondence. According to the learning characteristics of deep neural networks (DNN), the model tends to first learn and memorize the obvious correspondences in simple samples, while for complex hard samples and noise samples, it may produce erroneous memories, thus affecting the overall model's performance. In image-text matching tasks, erroneous memories will make matching between some image-text pairs more difficult, especially when these erroneous samples have a high similarity to other image-text samples, making the model more susceptible to interference and resulting in incorrect fake news judgments. Therefore, to prevent the negative impact of hard and noise samples on the training process, retaining the key semantic information of simple

samples helps the model focus better on those image-text pairs with semantic consistency and strong discriminative power, thus effectively improving the accuracy of fake news detection.

The model identifies simple and hard samples by calculating the loss function value of the sample. Since DNNs tend to learn simple samples first, the loss value of simple samples is usually low, while the loss value of hard and noise samples is high. Based on this characteristic, this paper uses a Gaussian Mixture Model (GMM) to fit the loss distribution of the training dataset and uses the difference in sample loss to identify which samples are simple. This step provides the foundation for subsequent semantic retention. By analyzing the loss distribution, the model can accurately filter out those simple samples with clear semantics and low loss, which will be used as the core of key semantics. Let the mixture coefficient be represented by x_j, and the probability density of the j-th component be represented by $\psi(c|\phi_j)$, then we have:

$$o(c \mid \varphi) = \sum_{j=1}^{J} x_j \psi(c \mid \varphi_j) \tag{3}$$

Assume that the component with the smaller mean in the two-component Gaussian Mixture Model is represented by ϕ_j. The probability that the u-th image-text pair is correctly matched is the posterior probability o_u, and the calculation process is as follows:

$$o_u = o(\varphi_j \mid c_u) = \frac{o(\varphi_j) o(c_u \mid \varphi_j)}{o(c_u)} \tag{4}$$

Next, the model uses the selected simple sample set F_z to construct the key semantic library of the training set. In this process, the model matches each training sample pair (U_u, S_u) with the image and text that have the highest semantic similarity. This similarity measurement is performed by calculating the visual similarity T between the image U_u and all images in F_z, selecting the image $U^{U}{}_{u}$ that is most similar to U_u, and pairing the corresponding text $S^{U}{}_{u}$ with the original text S_u. In this way, the model can use U_u and $S^{U}{}_{u}$ as key semantic pairs to enhance the visual modal semantic information of the original image U_u. This process ensures the semantic consistency between the visual information and the corresponding text information for each sample, thereby improving the model's ability to understand the visual content. The threshold is represented by π. The selection process is as follows:

$$F_z = \{(U_k, S_k) \mid o_k \geq \pi\} \tag{5}$$

In the semantic retention of the text modality, the model also calculates the similarity T_s between the text S_u and all texts in F_z, and selects the text $U^{S}{}_{u}$ that is most similar to T_s. In this way, the textual information between the original text S_u and $U^{S}{}_{u}$ can be strengthened, ensuring a more accurate semantic understanding of the text modality. By mutually enhancing the text and image modalities, the model can better learn the semantic association between image and text, which is crucial for fake news detection. This is because fake news on social media often exhibits semantic mismatches, and the model needs to be able to recognize and correct such mismatches. Through this bidirectional enhancement, the model can not only better understand the independent semantics of images and texts, but also find deep connections between them, thus improving the accuracy of matching.

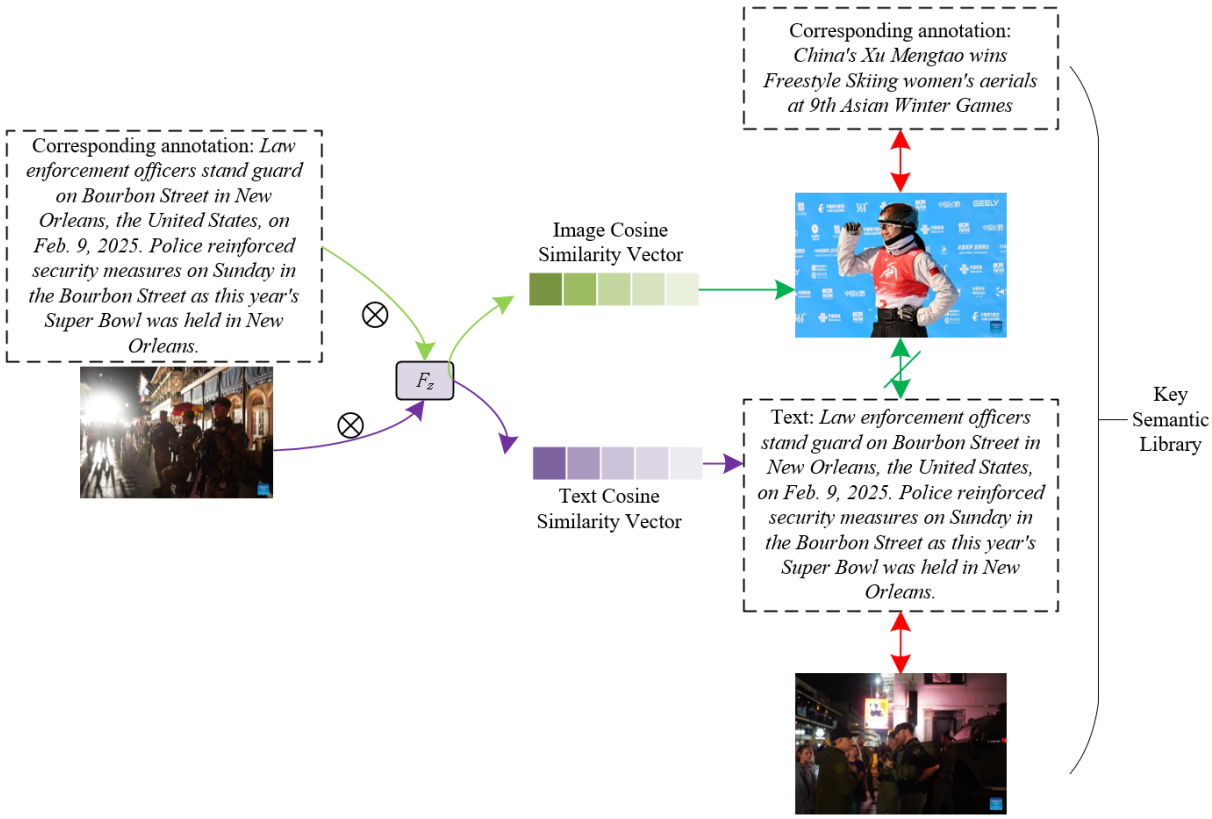


Figure 1. Schematic diagram of the construction process of the key semantic library

Figure 1 shows the schematic diagram of the construction process of the key semantic library. Finally, the constructed key semantic library L contains the key image-text pairs selected from the simple sample set. These pairs are used to enhance the visual and textual semantic information of the original samples during training. Let $L= \{(U^U_u, S^U_u), (U^S_u, S^S_u)\}$ represent the enhanced semantic information on the visual and text modalities for each training sample pair, and this information helps the model better capture the subtle differences between images and texts. Through this key semantic retention strategy, the model not only improves the robustness of fake news detection but also enhances the deep understanding of social media news content. Especially on social media platforms, the authenticity of news content often involves complex semantic changes, and the construction of the key semantic library effectively enhances the model's ability to handle complex image-text matching tasks, thereby helping to more accurately identify fake news.

2.3 Model training method

The optimization of the constructed model is divided into two stages, aiming to reasonably evaluate the impact of sample pairs on the model and adjust accordingly, thereby improving the accuracy of fake news detection. The goal of the first stage is to predict the negative impact of each sample pair on the model's performance and calculate the influence factor of each sample. This stage focuses on identifying which image-text pairs interfere with the model's training, especially those hard samples and noise samples related to fake news. The goal of the second stage is to optimize the main model based on the influence factors calculated in the first stage. In this stage, the model will adjust the training process with weighted samples, assigning higher weights to samples with smaller impacts to help the model better learn the semantic information from key samples.

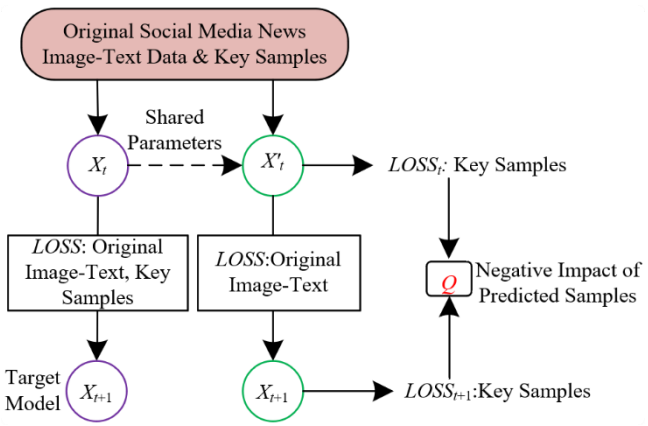


Figure 2. Model training strategy diagram

The main reason for predicting the negative impact of samples during model training lies in the memory characteristics of DNNs. In the learning process, DNNs tend to prioritize learning the obvious patterns in simple samples, while for hard samples, they are prone to learning incorrect correspondences. These erroneous learning results not only directly affect the model's prediction ability for hard samples, but may also negatively impact other samples that should have been correctly predicted. Specifically, when the model during training matches noise samples with the most similar simple samples in the key semantic library, these simple samples should play a positive role in the training. However, if the model learns incorrect matching relationships, it will reduce the loss of these simple samples, which in turn increases the loss of these simple samples, thereby affecting the model's prediction of other correct samples. Figure 2 shows a schematic diagram of the model training strategy.

The basic idea of training is to introduce an independent copy model X', which shares all parameters with the main model X but remains independent during updates. The core purpose of this design is to predict the negative impact each training sample has on the model by calculating the difference in loss values of the simple sample set L_y before and after training for both model X and X'. The samples in the simple sample set L_y generally have obvious semantic alignment and are more sensitive to fluctuations in model performance. Therefore, by analyzing the loss value changes of these samples, the model can indirectly assess whether there is interference from noise or difficult samples during training. Specifically, the loss value for image-text matching is represented by o_j, the loss value for text-image matching is represented by w_j, and X'_t represents the model before the t-th round of training. The loss for key samples before training is denoted as o^t_j and w^t_j, and after updating the parameters, the model X'_{t+1} is obtained. The loss for key samples after training is represented as o^{t+1}_j and w^{t+1}_j. For the training sample (U_j, S_j) corresponding to the key samples (U^U_j, S^U_j) and (U^S_j, S^S_j), the loss calculation process between them is as follows:

$$\begin{aligned} o_j &= ZR\left(U^U_j, S^U_j\right) + ZR\left(U^S_j, S^S_j\right) \\ w_j &= CE\left(S^U_j, U^U_j\right) + ZR\left(S^S_j, U^S_j\right) \end{aligned} \tag{6}$$

Next, by calculating the impact of the training sample (U_j, S_j) on the model performance, the loss change caused by this sample on the model can be quantified. In this process, the loss value change of the key samples (U^U_j, S^U_j) and (U^S_j, S^S_j) directly reflects the impact of the training sample (U_j, S_j) on the model performance. Specifically, if (U_j, S_j) is a difficult-to-predict noise sample, it may lead to a negative impact on the predictions of the key samples (U^U_j, S^U_j) and (U^S_j, S^S_j), causing their loss values to increase. In contrast, if (U_j, S_j) is an easily predictable simple sample, its impact should be positive, causing the loss values of the key samples to decrease. Specifically, the model performance change caused by (U_j, S_j) can be characterized by the change in the key sample loss values as follows:

$$e_j = \frac{1}{2}\left(\frac{o^t_j}{o^{t+1}_j} + \frac{w^t_j}{w^{t+1}_j}\right) \tag{7}$$

When $e_j<1$, it indicates that the loss value of the key sample has increased; otherwise, when $e_j \geq 1$, it indicates that the sample has no negative impact. During the model training process, this paper quantifies the negative impact of each sample pair (U_j, S_j) on the model performance by calculating the influence factor q_j. The influence factor q_j reflects the extent of the change in the loss values of the key samples caused by the training sample (U_j, S_j), thereby predicting the potential negative impact of this sample on the model. Specifically, the influence factor q_j evaluates the interference with the model learning process caused by the loss change induced by the training sample (U_j, S_j). If the q_j value is large, it indicates that the sample may severely affect the model's

prediction performance, and the model will take corresponding measures to reduce its weight; if the q_j value is small, it means the sample has little impact on the model's training and can continue to participate in training.

$$q_j = \begin{cases} \tanh\left(e_j\right), e_j < 1 \\ 1 \quad\quad\quad , Others \end{cases} \tag{8}$$

The principle of the retraining module is first reflected in retraining the main model X and introducing a weighted cross-entropy loss $LOSS_{QZD}$ to prevent training samples from causing performance degradation in the model. The influence factor q_j of each training sample reflects the degree to which the sample affects the model's performance, and based on these influence factors, weights are assigned to the samples during training. Those samples with larger negative impacts will have lower weights during updates, thereby reducing their interference with the model. The calculation process for $LOSS_{QZD}$ is:

$$LOSS_{QZR} = \frac{1}{v}\sum_{j=1}^{v} q_j \left(ZR\left(U_j, S_j\right) + ZR\left(S_j, U_j\right)\right) \tag{9}$$

Furthermore, since the image-text correspondences in difficult and noisy samples are often unreliable, relying solely on these samples for learning may lead the model to learn incorrect semantic relationships. To enhance the model's understanding of these unreliable semantic relationships, this paper designs an auxiliary learning mechanism using the cross-entropy loss of key samples, $LOSS_L$. Key samples are those selected from the training set that have high semantic matching with the original samples, and they provide stable and correct semantic alignment in both visual and textual modalities. The calculation process for $LOSS_L$ is:

$$LOSS_L = LOSS_{ZR}\left(U_j^U, S_j^U\right) + LOSS_{ZR}\left(U_j^S, S_j^S\right) \tag{10}$$

The total loss function is as follows:

$$LOSS = LOSS_{QZR} + LOSS_L \tag{11}$$

3. FAKE NEWS DETECTION IN SOCIAL MEDIA IMAGES BASED ON IMPROVED TRANSFORMER

After completing the image-text matching for social media news based on multimodal semantic understanding, this paper further conducts fake news detection in social media images based on the results of image-text matching. A fake news detection model for social media images based on an improved Transformer is constructed in this paper, which fully utilizes the results of the multimodal semantic understanding of image-text matching. By combining the propagation path of social media news and the semantic relationship between images and text, fake news detection is effectively performed. First, the multimodal image-text matching model is used to evaluate the semantic consistency between social media images and texts. The model will learn and analyze the potential relationships between images and texts to determine whether they can jointly express a consistent news theme or event. If the match between the image and text is low, and there is semantic inconsistency, the model will regard it as a possible fake news. In addition, the model will also combine the visual features of the social media image with keywords, sentiment, and contextual information in the news text to comprehensively determine the credibility of the news. When the image-text matching results show significant inconsistency, the model will further perform fake news detection using the improved Transformer architecture to determine whether the news is false or misleading.

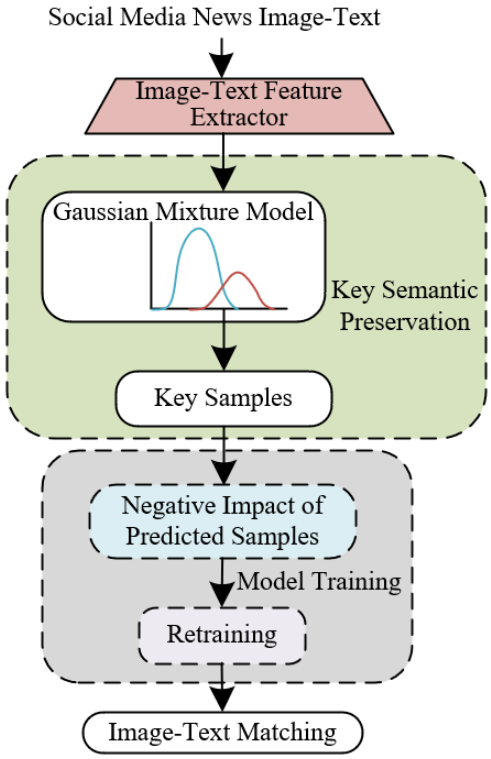


Figure 3. Overall structure diagram of the social media image fake news detection model

Specifically, the model uses a social media news-level attention module to learn the implicit semantic relationships between social media news. This module captures key semantic features in the news content through the improved Transformer structure, and uses the self-attention mechanism to weight the semantic importance of different social media news. At the same time, a structural embedding method is introduced to enhance the model's ability to learn spatial relationships when processing the propagation tree structure of social media news, so that the model not only understands the content of the news but also understands the position of the news in the propagation network, thereby more accurately evaluating the influence of each news piece in the propagation process. In the path-level attention module, the model further strengthens its handling of the social media news propagation path representation. Considering that each news propagation path has a different impact on fake news detection, the path-level attention mechanism aggregates the weights of each propagation path so that the model can focus on the most informative paths, thus more accurately capturing clues related to the spread of fake news. On this basis, the fake news classification module uses a classifier to predict the label of the entire social media news event. By learning and analyzing the final representation of the propagation tree structure, the authenticity of the news is ultimately determined. Figure 3 shows the overall structure diagram of the social media image fake news detection model.

3.1 Problem description

The specific problem description for fake news detection in social media images based on the improved Transformer is as follows: Given a fake news event R, which contains a series of social media news $S_1,S_2,...,S_{|R|}$ arranged in chronological order. Here, S1 represents the source social media news of event R, and the rest $\{S_2,S_3,...,S_{|R|}\}$ are related replies and retweets. These social media news events will be classified into four possible category labels: non-fake news, fake news verified as false, fake news verified as true, and fake news that cannot be verified. In a real social media environment, event R will form a propagation tree structure through multiple propagation paths O_u, where each path consists of a series of social media news nodes with reply relationships. The goal of the fake news detection task can be defined as a classifier d, which maps the fake news R to its corresponding category label B:

$$d : R \rightarrow B \tag{12}$$

3.2 Structural embedding

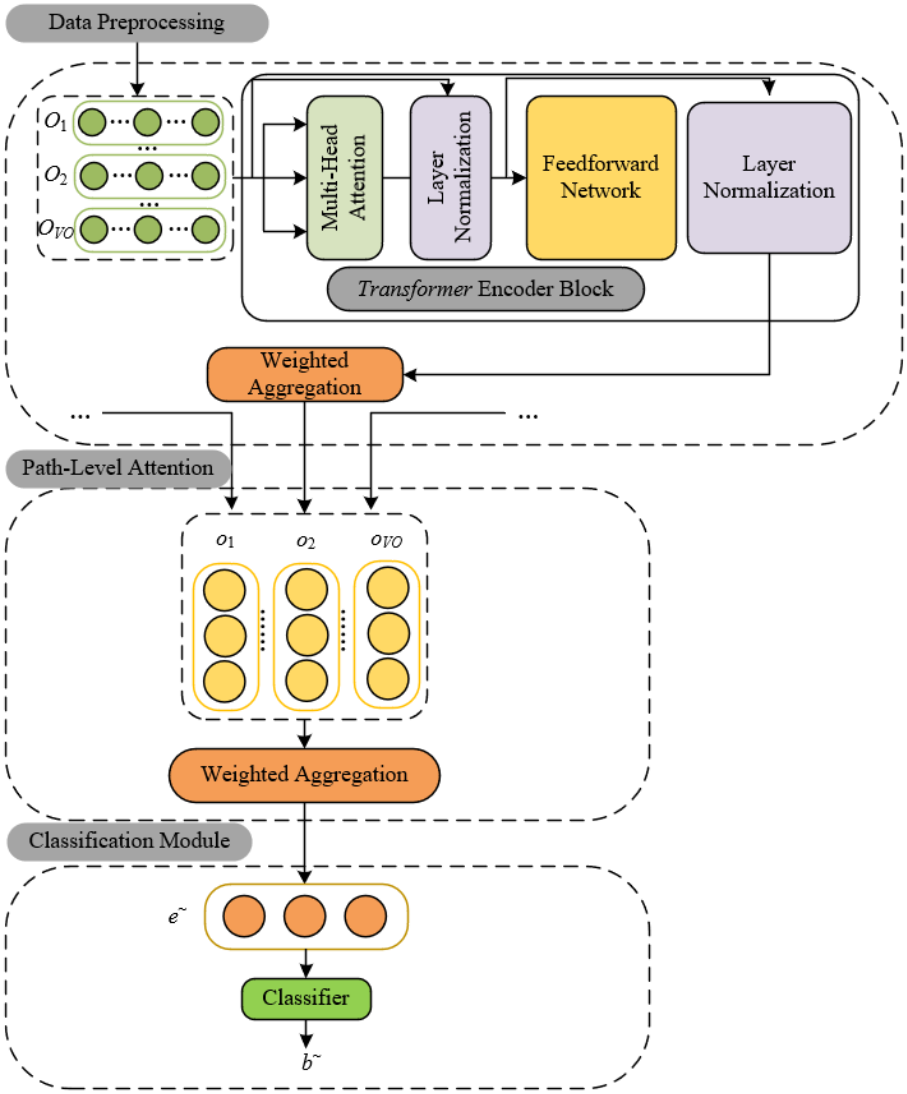


Figure 4. Structure diagram of the fake news social media news propagation tree model

Traditional Transformer models perform well when handling sequential data, but they often struggle with tree-structured data, as they cannot naturally capture the hierarchical relationships between nodes and the structural information in the propagation paths. Existing hierarchical embedding methods introduce horizontal and vertical position information for each node in a path, allowing the Transformer to understand the hierarchical relationships between nodes in the propagation tree. In a fake news social media news propagation tree, there is a flow of information from the root node to the leaf nodes. Therefore, to handle the tree-structured social media news propagation data of fake news events, the proposed hierarchical positional embedding can be applied to the fake news detection task, enabling the model to learn the horizontal and vertical position information of the fake news social media news propagation structure in space. Figure 4 illustrates the structure diagram of the fake news social media news propagation tree model. In this way, the model not only captures the semantic information of individual news content but also identifies the position of each news node in the propagation process and its relative relationships, thus providing richer information for fake news detection. By combining the propagation paths of social media news, the model can better understand the direction of news propagation, thereby improving the accuracy of fake news identification. Specifically, assume that the structural embedding representation of a social media news node S_{uk} in the propagation path O_u is denoted as t_{uk}. The trainable vertical

position embedding matrix R^n and horizontal position embedding matrix R^g contain the a-th and b-th row vectors, denoted as $r^n{}_a$ and $r^g{}_{bk}$, respectively. The concatenation operation is represented by ⊗, and the maximum depth of the $|O|$ propagation paths is denoted as g_o. The vertical position information of S_{uk} in the propagation path is encoded top-down by $a=|N^u{}_k|$, where the set of nodes passed through in the process of information flowing from the root node S_{u1} to node S_{uk} is represented as $N^u{}_k=\{S^u{}_k|1\leq j\leq k\}$. The number of times $S^u{}_k$ has appeared in the current propagation path is represented by $b=|G^u{}_k|$, which indicates the horizontal position information in the propagation tree structure. The set of paths where S_{uk} appears, including its previous paths, is represented as $G^u{}_k=\{S_{uk}|1\leq j\leq k\ ANDS_{jk}=S_{uk}\}$. For each social media news node in the path, the corresponding structural embedding representation is calculated.

$$t_{uk}=r_a^n\otimes r_b^g, a=\left|N_k^u\right|\ AND\ b=\left|G_k^u\right| \tag{13}$$

3.3 Path oversampling

The number of social media news propagation paths may vary significantly. Some propagation paths may be relatively short and lack sufficient user engagement and feedback, which can result in insufficient information in these paths and affect the accurate identification of fake news. To address this issue, this paper proposes a path oversampling method to ensure that each fake news event receives sufficient user feedback. Path oversampling resamples the original variable-length propagation path sequence to generate a fixed-length propagation path sequence, so that each news event can receive relatively balanced feedback information.

Specifically, for paths longer than the fixed length V_O, the model only uses the first V_O social media news propagation paths to ensure that each propagation path sequence can balance computational complexity and information effectiveness within the fixed-length constraint. During the path oversampling process, each social media news node S_{uk} has a corresponding structural embedding representation t_{jk}, which is further extended during path resampling. For a path O_j sampled from the original path O_u, the corresponding social media news node t_{jk} will inherit the structural embedding representation from S_{uk}, i.e., $t_{jk}=t_{uk}$. This mechanism ensures that the resampled path retains the same structural and semantic information as the original path.

In this way, the model does not lose the effective use of structural embeddings in the social media news propagation paths during resampling, thereby fully leveraging the advantages of multimodal semantic understanding. In the fake news detection task, this mechanism helps the model to perform a more in-depth analysis and comparison of image and text information, enabling the model to comprehensively consider the correlation between the image-text information and the propagation paths, improving the accuracy and robustness of fake news detection.

3.4 Social media news-level attention

In order to improve the model's understanding of social media news content by assigning different importance to different news nodes in the propagation path, this paper introduces a social media news-level attention process in the model. First, the social media news text S_{uk} is represented as a sequence of words $S_{uk}=\{Q_{uk1},Q_{uk2},...Q_{uk|Suk|}\}$, where each word is embedded into a fixed-dimensional vector q_{ukj}. Based on this, the maximum pooling method is used to extract the text representation s_{uk} of each social media news S_{uk}.

$$s_{uk}=\mathrm{MaxPooling}\left(\left\{q_{uk1},q_{uk2},...,q_{uk|S_{uk}|}\right\}\right) \tag{14}$$

To further enhance the model's understanding of the tree-like propagation structure, this paper adds structural embeddings t_{uk} to the text representation s_{uk}, strengthening the model's understanding of the spatial information of each news node in the propagation path. This process not only captures the position of each news node in the propagation path but also helps the model better understand the hierarchical relationship of news in the propagation tree and the mutual influence of upstream and downstream, providing structural support for subsequent multimodal understanding.

$$s_{uk}=s_{uk}+t_{uk} \tag{15}$$

Next, when processing the propagation path O_u, this paper considers the temporal dependencies of each news node, meaning each social media news is influenced not only by its immediate parent node but also potentially by earlier news, especially the root node's information. To learn these implicit temporal relationships, the model applies the Transformer encoder block to the sequence of social media news text representations along the path O_u. This process, through the V_u-layer Transformer encoder, can effectively capture complex dependencies between news nodes and enhance the model's understanding of temporal order information, further improving the model's performance in handling social media fake news. The encoder block in the Transformer structure is given by the following equation:

$$\left\{s_{u1}^{'},s_{s2}^{'},...,s_{u|O_u|}^{'}\right\}=TRANS\left(\left\{s_{u1},s_{u2},...,s_{u|O_u|}\right\}\right) \tag{16}$$

To handle the varying importance of different social media news in the propagation path representation, this paper proposes an attention mechanism to measure the importance of each news node in the path. The context vector s_{uk} for each news node is calculated using the following equation, and the weight of the news node is computed using the *LeakyReLU* activation function.

$$z_{uk}^s=x_s^S\cdot\mathrm{LeakyReLU}\left(Q_s s_{uk}^{'}\right) \tag{17}$$

Based on these weights, further normalization is performed using the following equation to obtain the attention weight $\alpha^s{}_{uk}$ for each news node.

$$\alpha_{uk}^s=\frac{\exp\left(z_{uk}^s\right)}{\sum_{j=1}^{|O_u|}\exp\left(z_{uk}^s\right)} \tag{18}$$

Finally, the weighted sum of all social media news representations is calculated using the following equation to obtain the representation of the entire propagation path o_u.

$$o_u=\sum_{k=1}^{|O_u|}\alpha_{uk}^s s_{uk}^{'} \tag{19}$$

3.5 Path-level attention

In the model, the path-level attention mechanism is designed to assign different importance to different social media news propagation paths, in order to effectively capture the most crucial path information for fake news detection. In this process, the representation o_u of each social media news propagation path O_u may contain multiple different news contents, and these contents have different impacts on the entire social media news propagation tree. Therefore, the model needs to use the path-level attention mechanism to identify which paths are more important for determining the authenticity of the news propagation tree. Specifically, the model introduces a path-level context vector $z^{O}{}_{u}$ to calculate the contextual information of each propagation path, which can reflect the importance of path O_u to the structure of the entire social media news propagation tree. The calculation of the path-level context vector is processed using the *LeakyReLU* activation function, and the weight vector x_o and the weight matrix Q_o are further used to optimize the evaluation of the path importance.

$$z_u^o = x_o^S \cdot \mathrm{LeakyReLU}\left(Q_o o_u\right) \tag{20}$$

Next, the model processes the path-level impact through the normalized importance weight $\alpha^{O}{}_{u}$, thus accurately weighting each path's contribution to the social media news propagation tree. In this process, the model ensures that the weight of each path reflects its relative importance in determining fake news. Finally, the weighted representations of all paths are summed to obtain the final representation $e^{\sim}$ of the social media news propagation tree, which integrates the key information from each path and helps the model make more accurate image-text matching and fake news detection.

$$\alpha_u^o = \frac{\exp\left(z_u^o\right)}{\sum_{j=1}^{V_o} \exp\left(z_u^o\right)} \tag{21}$$

$$\tilde{e} = \sum_{u=1}^{V_o} \alpha_u^o o_u \tag{22}$$

3.6 Fake news classification module

In this paper, the core task of the fake news classification module is to convert the representation $e^{\sim}$ obtained from the social media news propagation tree into the predicted label $b^{\sim}$ for fake news through a feed-forward neural network and a *softmax* layer. Specifically, the model receives the output $e^{\sim}$ from the improved Transformer model, which is the representation of the social media news propagation tree after passing through multiple layers of Transformer encoding, to perform fake news detection. In this step, the feed-forward neural network is used to further process the propagation tree representation and map it to a category label space. The *softmax* layer calculates the probability for each category and outputs the predicted label for fake news.

$$\tilde{b} = \mathrm{softmax}\left(Q_e \tilde{e} + y_e\right) \tag{23}$$

To measure the similarity between the model's predicted labels and the actual labels, this paper uses cross-entropy as the classification loss function. The cross-entropy calculation formula ensures that the model can optimize based on the difference between predicted and actual labels, minimizing the loss function, and continuously improving the accuracy of predictions during training. The smaller the cross-entropy, the closer the model's predicted labels are to the actual labels. At the same time, to prevent overfitting, this paper introduces an L2 regularization term into the loss function. L2 regularization penalizes all model parameters, encouraging the model to learn smoother parameters, thus improving the model's generalization ability. The regularization weight is controlled by the hyperparameter η, balancing the trade-off between classification error and model complexity.

$$\begin{aligned} &LOSS\left(b, \tilde{b}\right) = \\ &-\left[b \log \tilde{b} + (1-b) \log\left(1-\tilde{b}\right)\right] + \eta \|\varphi\|_2^2 \end{aligned} \tag{24}$$

4. EXPERIMENTAL RESULTS AND ANALYSIS

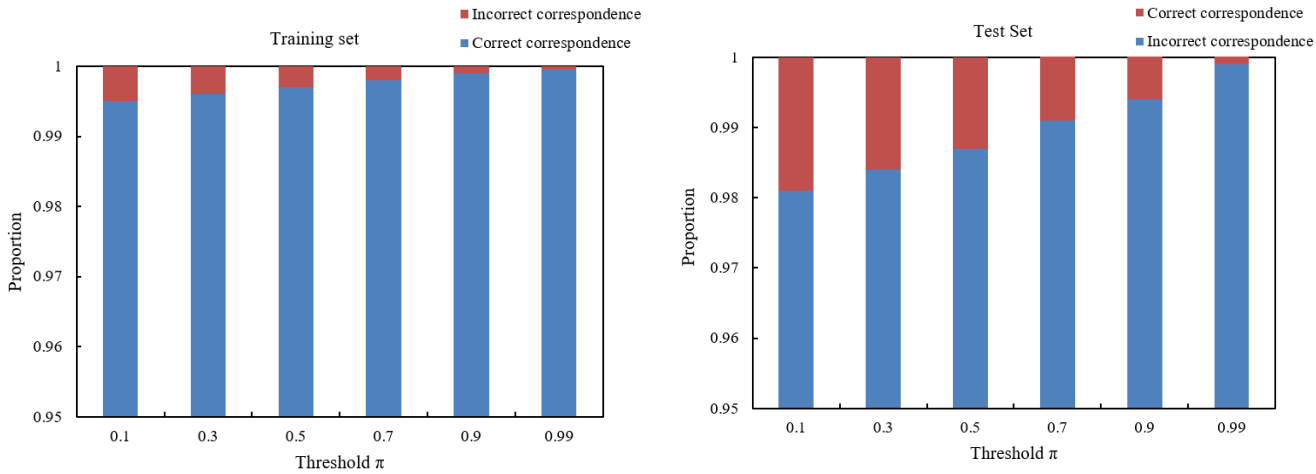
From the experimental results, the model proposed in this paper performs excellently in the social media news image-text matching task, especially in the two directions of "Image-to-Text Matching" and "Text-to-Image Matching." Specifically, the model achieves an R@1 of 81.2 in the "Image-to-Text Matching" task, significantly higher than other models, such as Self-Attention (78.9) and Stacked Cross Attention (77.3), demonstrating its strong ability in image and text matching. In the "Text-to-Image Matching" task, the model also achieves an R@1 of 67.9, clearly outperforming Perceptual Digital Quotient (55.6) and TMK+PDQF (61.2), indicating its efficiency in matching images and text. In addition, the model also outperforms most other models in the R@5 and R@10 metrics, especially in the text-to-image matching task, where it reaches R@5 and R@10 values of 93.5 and 97.5, demonstrating high accuracy and robustness (Table 1).

Table 1. Comparison results of different models for social media news image-text matching

Algorithm	Image-to-Text Matching			Text-to-Image Matching		
	R@1	*R*@5	*R*@10	*R*@1	*R*@5	*R*@10
Perceptual Digital Quotient	68.9	92.5	96.8	55.6	85.6	92.3
TMK+PDQF	76.5	94.5	97.8	61.2	88.9	94.5
TweetGage	76.4	95.3	97.5	61.3	88.5	94.6
AttentionRanker	77.9	94.2	97.6	62.8	91.2	94.2
Position Focused Attention Network	75.2	94.6	97.5	60.6	88.4	94.8
Stacked Cross Attention	77.3	94.8	97.4	61.8	91.3	94.6
Self-Attention	78.9	94.2	97.6	64.5	91.6	97.8
The Proposed Model	81.2	95.6	97.5	67.9	93.5	97.5

Table 2. Ablation experiments of the proposed social media news image-text matching model

Sample Set	Siamese Network Structure	Key Semantic Retention	Image-to-Text Matching			Text-to-Image Matching		
			R@1	*R*@5	*R*@10	*R*@1	*R*@5	*R*@10
Image Verification Corpus	√	√	86.5	96.5	97.5	71.5	91.2	94.2
	√		84.5	96.4	97.6	72.6	92.6	94.6
		√	84.2	96.3	97.2	72.4	92.4	94.7
			84.9	94.5	97.3	65.9	87.9	92.6
Weibo Dataset	√	√	78.9	96.8	97.5	72.8	92.6	94.5
	√		78.5	94.5	96.5	63.2	84.6	92.3
		√	78.5	94.2	96.4	61.8	84.2	92.4
			75.6	94.8	96.5	58.9	84.3	91.5
Top News Dataset	√	√	82.6	92.3	95.8	67.8	88.9	93.5
	√		77.9	94.2	97.2	58.2	81.5	87.6
		√	77.5	92.6	95.6	58.3	81.2	87.2
			65.9	88.6	92.3	51.9	77.9	88.6



1) Training set 2) Test set

Figure 5. Comparison of simple sample set composition at different thresholds

From the ablation experiment results in Table 2, the proposed social media news image-text matching model demonstrates stable performance and strong advantages under different experimental setups. By comparing the different combinations of Siamese network structure and key semantic retention mechanism, it can be observed that models incorporating both show performance improvements across all three datasets. For example, on the Image Verification Corpus dataset, after adding the Siamese network structure and key semantic retention, the model achieves the highest R@1, R@5, and R@10 values in both image-to-text and text-to-image matching tasks, reaching 86.5, 96.5, 97.5 and 71.5, 91.2, 94.2, respectively. This is significantly better compared to other combinations where some factors are removed, such as using only the Siamese network or only retaining key semantics. A similar trend is observed on the Weibo Dataset and Top News Dataset, particularly on the Weibo Dataset, where the joint use of the Siamese network and key semantic retention increases the R@1 value to 78.9, compared to other combinations, such as using only the Siamese network (78.5) or only retaining key semantics (78.5), demonstrating its strong fusion effect.

From Figure 5, it can be seen that with different thresholds on the training and test sets, the changes in correct and incorrect correspondences show a clear trend. In the training set, as the threshold gradually increases from 0.1 to 0.99, the proportion of correct correspondences increases from 0.995 to 0.9995, indicating that as the threshold increases, the model's matching accuracy on the training data improves continuously. Meanwhile, the proportion of incorrect correspondences decreases from 0.005 to 0.0005, showing that the model gradually reduces the probability of incorrect matching. Correspondingly, on the test set, the proportion of incorrect correspondences increases with the threshold, from 0.981 to 0.999, reflecting that the model begins to make more incorrect matches on the test data. At the same time, the proportion of correct correspondences decreases from 0.019 to 0.004, suggesting that as the threshold increases, the model's generalization ability weakens, leading to a reduction in the number of correctly matched samples. Based on these results, it can be concluded that the setting of the threshold plays an important role in the performance and generalization ability of multimodal learning models. In the training set, a lower threshold means the model accepts more samples, leading to higher accuracy, but in the test set, a higher threshold, while improving matching accuracy on the training set, leads to an increase in incorrect correspondences on the test set. This indicates that, in the task of social media image fake news detection, to optimize the model's performance, it is necessary to balance the threshold setting during both training and testing. A threshold that is too high may lead to overfitting, thus affecting the model's accuracy on unknown data.

From the experimental results shown in Table 3, the model proposed in this paper performs very well in the task of social media image fake news detection, especially in terms of accuracy (*Acc.*) and the F1 score for different types of fake news detection. For example, the overall accuracy of the proposed model reaches 0.912, far surpassing other models such as *RoBERTa* (0.579) and *BERT* (0.425), demonstrating

its high efficiency in detecting image-based fake news. Moreover, the model also excels in detecting different categories of fake news, particularly in the "Unverifiable Fake News" category, where the F1 score of the proposed model is 0.945, much higher than other models such as *FLAVA* (0.915) and *Reformer* (0.925). For the "Non-Fake News" and "Verified Fake News" categories, the F1 scores of the proposed model are 0.885 and 0.854, respectively, also higher than most comparison models, especially in the "Verified Fake News" category, where the performance shows a significant advantage over other models.

Table 3. Experimental results for social media image fake news detection

Model	*Acc.*	Non-Fake News	Verified Fake News	Verified True Fake News	Unverifiable Fake News
		F_1	F_1	F_1	F_1
BERT	0.425	0.385	0.265	0.625	0.335
XLNet	0.458	0.651	0.385	0.421	0.412
RoBERTa	0.579	0.746	0.425	0.539	0.556
ALBERT	0.312	0.712	0.086	0.425	0.036
Vision Transformer	0.568	0.746	0.412	0.569	0.512
Conformer	0.621	0.623	0.725	0.562	0.528
Reformer	0.889	0.845	0.826	0.925	0.879
Linformer	0.652	0.635	0.612	0.779	0.642
BEiT	0.725	0.651	0.768	0.826	0.715
DALL-E	0.768	0.758	0.765	0.845	0.756
CLIP	0.758	0.712	0.745	0.736	0.729
FLAVA	0.856	0.812	0.826	0.915	0.879
OPT	0.879	0.836	0.859	0.926	0.856
The Proposed Model	0.912	0.885	0.854	0.945	0.912

Table 4. Ablation experiment results for social media image fake news detection model

Model	*Acc.*	Non-Fake News	Verified Fake News	Verified True Fake News	Unverifiable Fake News
		F_1	F_1	F_1	F_1
The Proposed Model without the attention mechanism	0.912	0.879	0.912	0.923	0.879
The Proposed Model	0.888	0.885	0.889	0.925	0.865

From the ablation experiment results in Table 4, the performance differences between models with and without the attention mechanism in social media image fake news detection are quite noticeable. In the model without the attention mechanism, the overall accuracy (*Acc.*) is 0.912, which is quite good, especially in the "Verified True Fake News" category, where the F1 score is 0.923, significantly higher than other categories. However, in the model with the attention mechanism, there is an improvement in several metrics, although the overall accuracy drops to 0.888. The F1 scores for "Non-Fake News" and "Verified Fake News" increase to 0.885 and 0.889, respectively, and the F1 score for "Unverifiable Fake News" also increases to 0.865. In comparison, the addition of the attention mechanism seems to improve the model's ability to identify certain categories, even though the overall accuracy slightly decreases. These results suggest that the inclusion of the attention mechanism contributes to the performance improvement of the proposed social media image fake news detection model. Specifically, the attention mechanism helps to better capture the deep associations between images and text, especially when dealing with "Non-Fake News" and "Verified Fake News," which may be due to the attention mechanism's ability to focus more accurately on the parts critical for determining fake news.

From the experimental data in Figure 6, it can be seen that the number of words in the social media news text has a certain impact on the model's performance. In Dataset 1, the accuracy remains stable between 0.855 and 0.905 as the number of words increases from 10 to 100, showing little fluctuation, indicating that an increase in text length does not significantly improve the model's performance. In contrast, Dataset 2 shows a more noticeable change in accuracy, increasing from 0.755 (10 words) to 0.822 (100 words), presenting a gradual upward trend, though the change is relatively small. In comparison, Dataset 3 demonstrates the most significant improvement, with accuracy increasing from 0.935 to 0.96, showing that as the number of words increases, the model's performance improves significantly, especially when the text length exceeds 70 words, where the accuracy becomes stable, further indicating that longer texts help improve the model's ability to detect fake news.

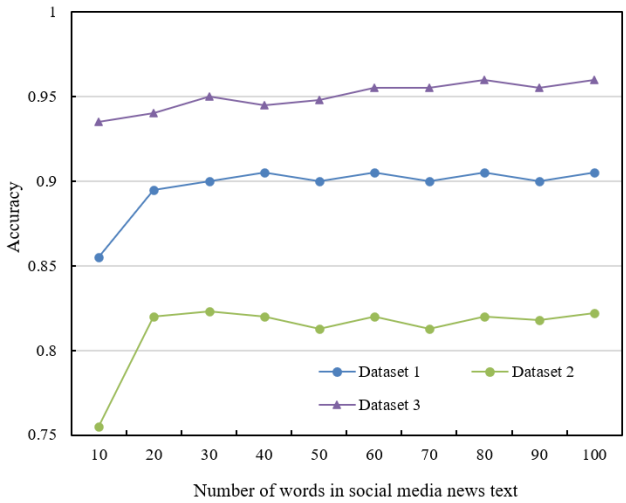


Figure 6. The impact of text length in social media news on experimental performance

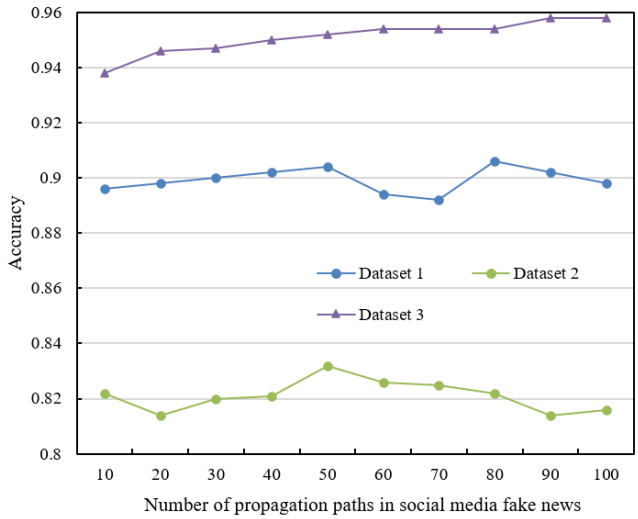


Figure 7. The impact of the number of propagation paths in social media news on experimental performance

From the experimental data in Figure 7, it can be seen that the number of propagation paths for social media fake news affects the model's accuracy to varying degrees. In Dataset 1, as the number of propagation paths increases, the accuracy fluctuates somewhat but remains stable between 0.896 and 0.906, showing a relatively steady performance. Even when the number of paths reaches 100, the accuracy only slightly decreases (0.898), indicating that the number of propagation paths has a limited impact on the model's performance for this dataset. In Dataset 2, the accuracy changes little across different path numbers, ranging from 0.822 to 0.816, indicating that the increase in propagation paths has no significant effect on this dataset. In contrast, Dataset 3 shows a more noticeable improvement, with accuracy increasing from 0.938 to 0.958, particularly when the number of propagation paths exceeds 70, where the accuracy stabilizes. This shows that more propagation paths can help the model better capture the propagation characteristics of fake news, improving its detection ability.

These results indicate that the impact of the number of propagation paths for social media fake news on the model depends on the characteristics of the dataset. In Dataset 3, the increase in propagation paths significantly improved the model's performance, likely because the propagation network in this dataset is more complex, and more paths provide the model with richer semantic information, allowing for more accurate identification of fake news. In contrast, in Datasets 1 and 2, the effect of the number of propagation paths on the model's performance is weak, possibly because these datasets have fewer propagation paths or redundant information, and thus cannot fully utilize the additional path information. Therefore, the number of propagation paths for social media fake news does affect the detection effect, but the magnitude of the effect depends on the characteristics of the dataset and the complexity of the propagation network.

5. CONCLUSION

This study mainly focused on the detection of fake news in social media images, using a multimodal semantic understanding-based image-text matching technology and an improved Transformer model. The research first enhanced the correlation between images and text through a multimodal semantic understanding model, effectively improving the accuracy of image-text matching. On this basis, the improved Transformer model further enhanced the ability to detect fake news in social media images, especially when dealing with fake news with complex and diverse features, showing significant advantages. Through a series of experiments, the proposed model performed excellently in terms of accuracy and F1 scores in different datasets, especially in key tasks for fake news identification, demonstrating strong robustness and accuracy.

However, despite the excellent results achieved by the proposed method, there are still some limitations. First, the model's detection ability is relatively weak for shorter texts or datasets with fewer propagation paths, indicating the model's limited adaptability in situations with small or simple amounts of information. Second, although multimodal learning enhances the ability of image-text matching, how to more efficiently fuse image and text information, avoiding redundancy or noise, remains an issue to be solved. Moreover, only specific model architectures and datasets were used in this study, so the model's generalization ability and its detection effectiveness for other types of fake news need further verification. Future research can explore the following directions: on the one hand, more diverse multimodal fusion techniques, such as self-attention mechanisms or image-text cross-modal alignment methods, can be introduced to further improve the fusion efficiency and performance of the model. On the other hand, future research can expand to more complex and uncertain social media data environments, combining more diverse data sources for fake news detection, and exploring how to optimize the model on different types of social media platforms. Additionally, further optimization of the model's real-time processing capabilities to improve its applicability in large-scale social media data is also an important direction for future research.

ACKNOWLEDGMENT

This paper was funded by Science Research Project of Hebei Education Department (Grant No.: QN2025018).

REFERENCES

[1] Ganesh, M., Raghunathan, S., Rajendran, C. (2013). Distribution and equitable sharing of value from information sharing within serial supply chains. IEEE Transactions on Engineering Management, 61(2): 225-236. https://doi.org/10.1109/TEM.2013.2271534

[2] Rauh, J., Banerjee, P.K., Olbrich, E., Jost, J., Bertschinger, N. (2017). On extractable shared information. Entropy, 19(7): 328. https://doi.org/10.3390/e19070328

[3] Creane, A. (2007). Productivity information in vertical sharing agreements. International Journal of Industrial Organization, 25(4): 821-841. https://doi.org/10.1016/j.ijindorg.2006.08.003

[4] Dyadic, G.B. (2014). An empirical investigation of extensible information sharing in supply chains. Information Resources Management Journal, 27(4): 1-22. https://doi.org/10.4018/irmj.2014100101

[5] Harrison, A. (2018). The effects of media capabilities on the rationalization of online consumer fraud. Journal of the Association for Information Systems, 19(5): 408-440.
[6] Oelrich, S., Siebold, N. (2024). Media framing in Wirecard's fraud scandal: Facts, failures, and spying fraudster fantasies. Critical Perspectives on Accounting, 100: 102755. https://doi.org/10.1016/j.cpa.2024.102755
[7] Zenone, M., Snyder, J. (2019). Fraud in medical crowdfunding: A typology of publicized cases and policy recommendations. Policy & Internet, 11(2): 215-234. https://doi.org/10.1002/poi3.188
[8] Ingold, P.V., Langer, M. (2021). Resume=Resume? The effects of blockchain, social media, and classical resumes on resume fraud and applicant reactions to resumes. Computers in Human Behavior, 114: 106573. https://doi.org/10.1016/j.chb.2020.106573
[9] D'ulizia, A., Caschera, M.C., Ferri, F., Grifoni, P. (2021). Fake news detection: A survey of evaluation datasets. PeerJ Computer Science, 7: e518. https://doi.org/10.7717/peerj-cs.518
[10] Parte, S.A., Ratmele, A., Dhanare, R. (2023). An efficient and accurate detection of fake news using capsule transient auto encoder. ACM Transactions on Asian and Low-Resource Language Information Processing, 22(6): 1-22. https://doi.org/10.1145/3589184
[11] Kausar, N., AliKhan, A., Sattar, M. (2022). Towards better representation learning using hybrid deep learning model for fake news detection. Social Network Analysis and Mining, 12(1): 165. https://doi.org/10.1007/s13278-022-00986-6
[12] Zrnec, A., Poženel, M., Lavbič, D. (2022). Users' ability to perceive misinformation: An information quality assessment approach. Information Processing & Management, 59(1): 102739. https://doi.org/10.1016/j.ipm.2021.102739
[13] Jwa, H., Oh, D., Park, K., Kang, J.M., Lim, H. (2019). exbake: Automatic fake news detection model based on bidirectional encoder representations from transformers (bert). Applied Sciences, 9(19): 4062. https://doi.org/10.3390/app9194062
[14] Smits, T., Warner, B., Fyfe, P., Lee, B.C.G. (2025). A fully-searchable multimodal dataset of the illustrated London news, 1842–1890. Journal of Open Humanities Data, 11(1): 1-13. https://doi.org/10.5334/johd.284
[15] Ramisa, A., Yan, F., Moreno-Noguer, F., Mikolajczyk, K. (2017). Breakingnews: Article annotation by image and text processing. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(5): 1072-1085. https://doi.org/10.1109/TPAMI.2017.2721945
[16] Marsh, E.E., Domas White, M. (2003). A taxonomy of relationships between images and text. Journal of Documentation, 59(6): 647-672. https://doi.org/10.1108/00220410310506303
[17] Rafeeq, A. (2024). Infographics as a storytelling tool in UAE newspapers: A comparative study of Gulf News and Al Bayan. Cogent Arts & Humanities, 11(1): 2406090. https://doi.org/10.1080/23311983.2024.2406090
[18] Cui, X., Yang, L. (2022). Fake news detection in social media based on multi-modal multi-task learning. International Journal of Advanced Computer Science and Applications, 13(7): 912-918. https://doi.org/10.14569/IJACSA.2022.01307106
[19] Tuerhong, G., Dai, X., Tian, L., Wushouer, M. (2024). An end-to-end image-text matching approach considering semantic uncertainty. Neurocomputing, 607: 128386. https://doi.org/10.1016/j.neucom.2024.128386
[20] Ang, G., Lim, E.P. (2023). Learning and understanding user interface semantics from heterogeneous networks with multimodal and positional attributes. ACM Transactions on Interactive Intelligent Systems, 13(3): 1-31. https://doi.org/10.1145/3578522