

Article

Multimodal Social Media Fake News Detection Based on 1D-CCNet Attention Mechanism

Yuhan Yan , Haiyan Fu * and Fan Wu 

School of Information Science and Technology, Hainan Normal University, Haikou 571158, China; 202212083900014@hainnu.edu.cn (Y.Y.); 202212083900013@hainnu.edu.cn (F.W.)

* Correspondence: 240058@hainnu.edu.cn

Abstract: Due to the explosive rise of multimodal content in online social communities, cross-modal learning is crucial for accurate fake news detection. However, current multimodal fake news detection techniques face challenges in extracting features from multiple modalities and fusing cross-modal information, failing to fully exploit the correlations and complementarities between different modalities. To address these issues, this paper proposes a fake news detection model based on a one-dimensional CCNet (1D-CCNet) attention mechanism, named BTCM. This method first utilizes BERT and BLIP-2 encoders to extract text and image features. Then, it employs the proposed 1D-CCNet attention mechanism module to process the input text and image sequences, enhancing the important aspects of the bimodal features. Meanwhile, this paper uses the pre-trained BLIP-2 model for object detection in images, generating image descriptions and augmenting text data to enhance the dataset. This operation aims to further strengthen the correlations between different modalities. Finally, this paper proposes a heterogeneous cross-feature fusion method (HCFFM) to integrate image and text features. Comparative experiments were conducted on three public datasets: Twitter, Weibo, and Gossipcop. The results show that the proposed model achieved excellent performance.

Keywords: fake news; multimodal; attention mechanism; feature fusion



Citation: Yan, Y.; Fu, H.; Wu, F. Multimodal Social Media Fake News Detection Based on 1D-CCNet Attention Mechanism. *Electronics* **2024**, *13*, 3700. <https://doi.org/10.3390/electronics13183700>

Academic Editors: Patrick Siarry and Junaid Rashid

Received: 26 July 2024

Revised: 10 September 2024

Accepted: 12 September 2024

Published: 18 September 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In recent years, the widespread adoption of the internet has revolutionized information dissemination, with social media becoming the primary channel through which people access information, express opinions, and exchange ideas. With the substantial increase in users of social platforms such as Twitter, Weibo, and Facebook, the amount of information on social media has also increased exponentially. However, users can easily post or share information without verifying its authenticity, which leads to the widespread dissemination of fake news in the form of text, images, videos, and audio. If left unchecked, the rapid spread of such fake news could have significant economic and social implications.

Early efforts in fake news detection primarily focused on analyzing text content. However, research approaches targeting multimodal features have arisen in response to the growing diversity of news content and the continuous development and use of deep learning. These methods employ deep neural networks to extract features from various modalities and integrate them for detecting fake news. Although multimodal fake news detection methods have demonstrated good performance, existing methods still have some drawbacks, such as the inability to fully exploit correlations between different modalities and adequately integrate multimodal features, among others. Figure 1 shows an example of fake news in the Twitter dataset, including text and images.

Researchers have begun to focus on multimodal methods for detecting fake news that involves a large amount of text and images. Despite significant progress in multimodal fake news detection, existing methods still face some issues. Firstly, when processing text data, current methods mainly focus on semantic features of the text while neglecting the

deep contextual and spatial features. Secondly, when processing graphical and textual information, most existing multimodal detection methods independently extract features of images and texts, while ignoring the correlation between different modalities. Uncovering the implicit consistency between different modalities is more effective in identifying fake news that does not align with the text and images. Finally, most current methods simply concatenate the obtained graphical and textual feature vectors for fusion. However, this approach is prone to introducing noise and generating excessive redundant information, potentially affecting the effectiveness of fake news detection.



Figure 1. These are examples of news from Twitter dataset, with the following text content: “Boston Bomber suspect is missing person Sunil Tripathi”, “The detail of these photos used to identify the Boston Marathon bombing suspect is bananas...”.

To address these issues, this study makes the following contributions:

- This article proposes a multimodal model for detecting fake news, called BTCM. This method utilizes BERT and BLIP-2 model encoders to extract text and image features separately and employs a 1D-CCNet attention mechanism module to compress these features, enhancing the key aspects within each modality. Meanwhile, relationship reinforcement learning is conducted through joint image descriptions. In the modality fusion section, this paper proposes the heterogeneous cross-feature fusion method (HCCFFM) to integrate image features and text features. This results in high-performance fake news detection.
- This paper proposes a 1D-CCNet attention mechanism. This method can simultaneously capture horizontal and vertical relationships between different positions at the word and pixel levels in images, modeling long-distance dependencies in text and image features. This enhances the ability to capture both global and local information. Additionally, this method can adaptively learn and adjust attention weights for different positions, capturing valuable contextual information.
- This paper proposes a multimodal heterogeneous cross-feature fusion method. This method combines text and image features through batch matrix multiplication to obtain a feature matrix, which is then fused using weighted methods for the diagonal, upper triangular, and lower triangular parts to produce the final feature vector for classification. This method enhances the most relevant parts of the feature matrix, fully integrates features from different modalities, mitigates the interference of redundant information, and reduces noise.

The organization of this paper is outlined as follows: Section 2 offers an extensive review of deep learning techniques used for detecting fake news. The architecture of the multimodal fake news detection technique based on the 1D-CCNet attention mechanism is discussed in Section 3. Section 4 explains the experiment in this article and analyzes the experiment results. Lastly, Section 5 concludes the article.

2. Related Works

Fake news is intentionally fabricated and verifiably false information [1]. These falsehoods often manifest in various forms, such as rumors, misinformation, and pseudoscience.

In the early stages, fake news was verified manually, requiring experts to use domain knowledge for assessment. This process was not only time-consuming and labour-intensive, but also lacked timeliness. Researchers started using classic machine learning methods like naive Bayes once machine learning emerged [2], along with decision trees [3] and support vector machines [4], for fake news classification. Subsequently, due to the development and application of deep learning, neural network-based techniques for detecting fake news have drawn more and more attention. Nowadays, detection methods have evolved from simple single-peak methods to more sophisticated multi-peak methods. This section presents an in-depth examination of the development of technologies for detecting fake news.

2.1. Single Modal Fake News Detection

Based on the single-modal approach, methods for detecting fake news can be classified into two types: text-based and image-based. The core of text-based methods for detecting fake news lies in constructing and selecting high-quality features to represent news text information. In news text data, commonly used features primarily include semantic and dissemination characteristics of the news. There are significant differences in the semantic style between real and fake news; thus, semantic features can be extracted from various aspects of the text. These aspects include paragraph count [5], the percentage of specific words [6], the number of symbols [7], writing style [8], and language style [9]. However, these methods only focus on certain aspects of the text, making it difficult to comprehensively address the complex characteristics of fake news.

To enhance the semantic features of news, researchers have integrated various features for analysis. Moreover, advancements in machine learning have prompted researchers to apply these methods to fake news detection. For instance, Pérez-Rosas et al. [10] utilized different linguistic features of text, including grammatical features, punctuation, and psychological traits, in conjunction with a linear support vector machine (SVM) classifier for news classification, achieving a classification accuracy of 74%. Qazvinian et al. [11] explored textual topic features and employed Bayesian networks as classifiers for fake news detection. Kwon et al. [12] performed a comprehensive analysis of rumor propagation patterns on Twitter. They focused on identifying the characteristics of fake news concerning rumor propagation time, structure, and language. They developed three classifiers based on decision trees, support vector machines, and random forests, which achieved notable results. However, this single-modal approach lacks comprehensiveness and flexibility in analyzing fake news content. Compared to traditional machine learning techniques, deep learning methods demonstrate significantly superior performance in single-modal fake news detection. Deep learning methods transform news text features into pixel matrices or tensors and continuously update these features during model training, exhibiting exceptional performance in fake news detection. For instance, Ma et al. [13] were the first to employ deep learning to solve problems related to detecting fake news, by representing sentence-level semantic characteristics with recurrent neural networks (RNNs) and hidden layer vectors. This method is effective in processing and retaining contextual information within the text, particularly for lengthy texts, owing to its internal cyclic structure. Yu et al. [14] employed convolutional neural networks (CNNs) in their model to investigate deeper interactions among fundamental features, automatically extracting valuable features from raw data and capturing spatial and local patterns, thus improving the performance of fake news detection. Taher et al. [15] examined the efficiency of CNNs and RNNs in fake news identification. They evaluated CNNs, RNNs, bidirectional LSTM networks, and unidirectional LSTM networks based on news text features.

Additionally, methods such as word embeddings, transformers, sentence transformers, and document embeddings are effective in capturing vocabulary relationships, managing complex contexts, and understanding the core content of the text. Verma et al. [16] proposed WELFake, a two-phase benchmark model for fake news detection that integrates word embedding (WE) with linguistic features and machine learning classification. Their model preprocesses data using linguistic features and then applies voting classification to

improve accuracy. The WELFake model achieved a 96.73% accuracy rate, outperforming existing models such as BERT and CNN by notable margins. On a different front, Low et al. [17] addressed the challenge of distinguishing between fake news and satire, an area not previously tackled with transformers. Their study applied a DistilBERT-based classifier and introduced non-standard tokenization schemes and varied pre-training strategies, resulting in a 5.2% improvement in F1 score and a 6.4% improvement in accuracy over existing methods. This approach demonstrated the potential of transformers in enhancing fake news detection and distinguishing it from benign satire. Truică et al. [18] suggested a model to handle the issue of detecting fake news that was based on a sentence transformer and BiLSTM, including single language and cross language tasks, and conducted transfer training on a German dataset. In contrast, Truică and Apostol [19] proposed a new method using document embedding to accurately label news articles as reliable or false by constructing multiple models. Their research showed that document encoding performs better in fake news detection compared to complex deep neural network models, and this method outperformed traditional methods on five large-scale news corpora, emphasizing the importance of document embedding in advancing the accuracy of detecting fake news. DANES, an ensemble architecture of deep neural networks intended for context-aware fake news detection, was introduced by Truică et al. [20] in 2024. This model combines a textual content-based text branch with a social context-based social branch to create a comprehensive network embedding. Their initial findings, assessed on datasets such as Twitter15 and BuzzFace, showed encouraging accuracy, outperforming state-of-the-art methods by successfully combining textual and social aspects. Even with a small amount of training data, DANES shows how merging different contextual sources can improve the identification of bogus news.

In addition to text-based detection methods, image-based detection methods are also essential for detecting fake information. P. Zhou et al. [21] proposed a dual-stream Faster R-CNN model that effectively utilizes rich tampering features in RGB and noise streams to detect diverse types of fake images. Z. Jin et al. [22] analyzed images in multimedia datasets of news and articles, examining visual and statistical image features to predict the accuracy of associated articles. P. Qi et al. [23] employed attention mechanisms combined with neural networks to dynamically integrate image frequency domain features and pixel domain features for news detection, achieving promising results.

2.2. Multimodal Fake News Detection

Despite the high accuracy of deep learning-based single-modal detection models for fake news detection, these models frequently require substantial amounts of training data, leading to high training costs and time consumption. Moreover, concentrating solely on a single modality may restrict the effectiveness of detection, as the benefits of multimodal information are not fully utilized. In contrast, multimodal learning, which concurrently leverages both text and image content for news verification, captures the relationships between different modalities, addresses both textual and visual features, and reduces training costs. Consequently, researchers have started investigating methods that leverage multimodal information to tackle the fake news detection challenge.

Various categories of information sources can be considered as distinct modalities. News frequently includes multiple modalities including text, images, videos, and audio. The essence of multimodal news detection technology lies in developing a model framework that can effectively learn and integrate diverse forms of information present in news. This typically involves employing distinct feature extractors for text and images, integrating the features obtained from these modalities, and subsequently inputting the combined features into a classifier for categorization. Jin et al. [24] were the first to use a multimodal model for fake news detection, integrating extracted features from news images, text, and social context through an attention mechanism. This approach comprehensively captures various aspects of news content, thereby enhancing the ability to identify fake news. The model can dynamically prioritize significant features while ignoring unimportant details

thanks to the attention technique. Khattar et al. [25] proposed a Multimodal Variational Autoencoder (MVAE) that learns shared representations across two modalities. After training, this model can be utilized for detecting fake news by reconstructing both modalities from the learned shared representations. Singhal et al. [26] used the BERT model to extract text features and the VGG19 model to obtain image features. This approach ensured that the extracted features were of high quality and expressive power, improving the robustness and generalization ability of their fake news detection model to adapt to various types of news content. Wang et al. [27] constructed an event-adversarial neural network (EANN) that uses a task-specific event discriminator to better utilize multimodal information in order to acquire feature representations of textual and visual content in articles. The network inputs multimodal information and outputs event categories, using the auxiliary task to better understand multimodal information and enhance fake news detection. Zhou et al. [28] proposed a Similarity-Aware Fake News Detection Model (SAFE) to detect mismatched images and text in news articles, assuming a strong correlation between images and text in news. The model first uses image2sentence to convert images into text representations, then uses TextCNN to extract features from both text and images. It calculates new similarities and loss functions through fully connected modality interaction layers for classification. Qian et al. [29] employed pre-trained models to learn representations of news text and images, then input these learned representations into a multimodal context attention network to integrate features from different modalities. R. K. Kaliyar et al. [30] proposed a deep convolutional neural network model (FNDNet) that includes pre-trained word embeddings called GloVe and a CNN with multiple hidden layers to extract discriminative features for fake news classification. Zhang et al. [31] integrated semantic, visual, and textual data into multimodal fake news detection by leveraging the complementarity between modalities. To address the issue of CNNs' insufficient extraction of high-quality visual features, Yang et al. [32] used transformers to extract features from text and images and distinguished news authenticity by calculating the similarity between these modality features. Kananian et al. [33] distinguished themselves by treating each text as a graph, using a graph convolutional network (GCN) as a text encoder, and employing ResNet as an image encoder for fake news detection.

Key research has made significant advances not only in terms of detecting fake news, but also in disaster management and misinformation control. Arachie et al. [34] developed an unsupervised learning method to detect large-scale hazard-related subevents in tweets, using a crisis management ontology and spectral clustering. Li et al. [35] introduced three structural immunization strategies and a heuristic approach for managing false information in signed social networks, highlighting the importance of negative relationships and achieving better performance. Most recently, Apostol et al. [36] proposed a distributed system integrating social media data and deep learning for real-time, geolocation-based disaster reporting, offering enhanced speed and accuracy. Truică et al. [37] introduced a deep learning-based fake news detection system that not only accurately detects fake news but also mitigates its spread through a real-time network-aware strategy. This study showcases how advanced deep learning techniques and network analysis methods can be effectively employed to address the issue of misinformation on social media. These studies have collectively improved real-time disaster response and misinformation control with innovative tools and techniques. Coban et al. [38] proposed AIN, a community-based network immune algorithm. AIN utilizes network information to detect harmful content disseminators and achieves immunity by generating partitions and their rankings. The experimental results show that AIN performs better in immune networks than existing solutions such as NetShield and SparseShield, with faster convergence speed and fewer iterations. On the other hand, Petrescu et al. [39] proposed Sparse Shield, aimed at addressing the issue of harmful content dissemination in social networks. Sparse Shield uses an innovative approach to push harmful content to the bottom of user information streams, thereby reducing its negative impact on the network rather than directly scrutinizing users. This method not only minimizes interference with user freedom to the greatest extent

possible, but also enhances the effectiveness of immunity. Sparse Shield performs more effectively than existing state-of-the-art technologies with regard to scalability and mitigation of harmful speech. These studies demonstrate the application and advantages of different methods in network immunity, and provide valuable references for future work. Both AIN and Sparse Shield have demonstrated significant performance improvements on empirical data, highlighting the potential for effective prevention and control of harmful content dissemination in social networking environments.

Although previous methods have demonstrated good performance, they typically handle different modalities separately, failing to fully utilize the correlation and complementarity between them. Additionally, these methods often use simple concatenation to fuse feature vectors from different modalities, which leads to excessive redundant information and reduces data utilization efficiency. To deal with these issues, this paper suggests a model framework based on the 1D-CCNet attention mechanism and a multimodal heterogeneous cross-fusion method. This method can fully utilize the correlation between different modalities and capture the spatial relationship between different positions in text at the word level and images at the pixel level to strengthen important parts of the modality. At the same time, it fully integrates the obtained features to avoid redundant information interference and achieve better performance.

3. Materials and Methods

This article proposes a multimodal fake news detection model utilizing the 1D-CCNet attention mechanism. The structure of the suggested multimodal model of detecting fake news is illustrated in Figure 2, and it comprises text and visual feature extractors, the 1D-CCNet attention mechanism, and a heterogeneous cross-feature fusion module. Detecting fake news involves a binary classification task. Given a dataset containing textual and visual information about fake news, this paper first retrieves image descriptions and concatenates them with the original text data to create new textual data. Subsequently, features are extracted from text and images, textual and visual context information is captured using the 1D-CCNet attention mechanism, and features from different modalities are fused using a multimodal heterogeneous cross-fusion method. The model labels data with 0 or 1, where 0 represents fake news, and 1 represents real news. This paper will provide a detailed explanation of each component module in the sections that follow.

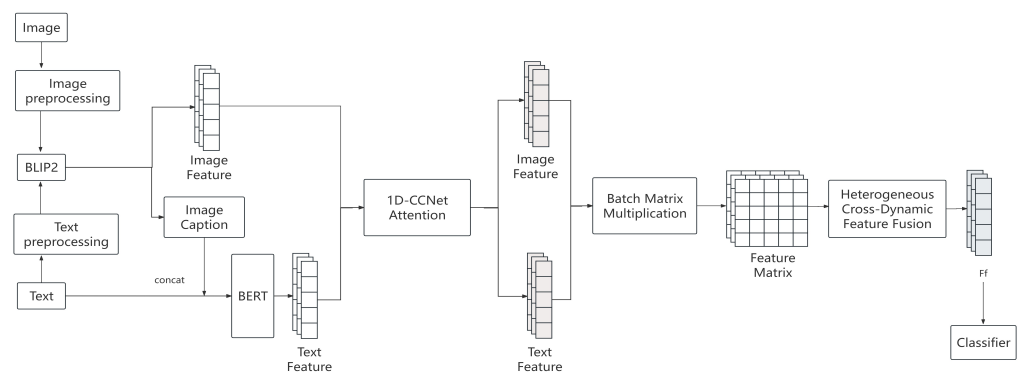


Figure 2. The structure diagram of the fake news detection model (BTCM) proposed in this paper.

3.1. Data Preprocessing

3.1.1. Text Data Preprocessing

To enhance the quality of news data, this paper performed preprocessing operations on textual and visual content before analyzing the news data. Initially, this paper preprocessed the news text. Text data preprocessing involved text standardization, cleaning, traditional Chinese conversion, and spell correction to produce the final news text, *T*. The primary purpose of this step is to filter out information noise in the text, thereby ensuring the quality of the final textual data.

3.1.2. Image Data Preprocessing

- **Image Preprocessing:**
This paper preprocessed images from news articles. Considering the small number of images with imbalanced aspect ratios and low resolutions in the dataset, these images were excluded. Subsequently, in order to improve the picture captioning model's performance, all images were resized to 224×224 pixels, resulting in the final news images.
- **Image Caption:**
To augment textual data and leverage the correlations between different modalities of text and image, this paper employed the language generation capability of the pre-trained BLIP-2 model to derive image descriptions, thereby transforming image data V from a visual to textual modality. Subsequently, these image descriptions T_i were concatenated with the original textual data T_o to generate the final textual data T for feature extraction. The formulas used are as follows:

$$T_i = \text{BLIP-2}(V) \quad (1)$$

$$T = \text{concat}(T_o, T_i) \quad (2)$$

Equations (1) and (2) describe the process of extracting image descriptions. Here, V represents the image data, T_i denotes the image description, T_o stands for the original text data, and T refers to the combined text data of the image description and the original text data.

3.2. Text Feature Extraction

Pre-trained language models are currently one of the most effective methods for extracting text features. These models are trained on large-scale data in an unsupervised manner, encompassing substantial grammatical knowledge and implicit information between sentences, thereby greatly enhancing the semantic representation capability of the models. Based on the Transformer encoder architecture, the BERT [40] model captures contextual relationships between texts through stacked self-attention layers, possessing strong language understanding capabilities. It undergoes pre-training via self-supervised learning, resulting in rich text representations. This paper employs the BERT pre-trained model as the primary component for extracting features from news texts. Firstly, the Bert-Tokenizer is used to tokenize and encode news text data. After tokenizing and adding [CLS] and [SEP] tokens, the textual data T is represented as a token list T_t , as shown in the following formula:

$$T_t = \text{Tokenzier}(T) \quad (3)$$

Next, the tokenized text data is further fed into the BERT model to extract semantic features, syntactic structures, contextual relationships, and other information, resulting in the final word feature vectors T_f of the text. This process effectively extracts both local features and global semantic information, providing richer and more comprehensive feature representations for subsequent text classification or other tasks. The formula is as follows:

$$T_f = \text{BERT}(T_t) \quad (4)$$

Equation (4) describes the process of extracting text features using the BERT model, where T_t is the token list and T_f is the text feature vector.

3.3. Image Feature Extraction

The content of images in news articles is crucial for detecting fake news. Typically, news articles with maliciously altered images or images that do not match the text content are likely fake. Therefore, it is crucial to fully utilize image features and the correlation between images and text. The use of vision-language pre-training models usually incurs high computational costs, and simply freezing the pre-trained parameters of vision and language models may result in difficulties aligning the visual and textual feature spaces.

To address this issue, this paper introduces the BLIP-2 [41] model to extract image features. BLIP-2 is a multimodal Transformer model composed of three main components: a visual encoder, a cross-modal Q-Former module, and a language generation module. A lightweight Querying Transformer (Q-Former) is introduced by the BLIP-2 model. It is pre-trained in two stages: multimodal learning from a frozen visual encoder is guided in the first stage, and multimodal learning from a frozen text encoder is guided in the second stage. This approach significantly improves training efficiency and reduces many training parameters, thereby saving computational costs.

This article uses the frozen image encoder module in the BLIP-2 model to obtain image pixel feature vectors V_f . The formula is as follows:

$$V = \{V_{class}, V_1, \dots, V_n\} \quad (5)$$

$$V_f = \text{BLIP-2}(V) \quad (6)$$

Equations (5) and (6) describe the process of extracting image features using the BLIP-2 model, where V represents the image data and V_f denotes the image feature vector.

3.4. 1D-CCNet Attention Mechanism

To fully capture the contextual information in text-image data and better process the spatial and global relationships, this paper proposes the 1D-CCNet attention mechanism. Compared to operating on sentence-level and image-level features, this paper operates on word-level features for text and pixel-level features for images.

This paper improves CCNet [42] by replacing the two-dimensional convolution layers with one-dimensional convolution layers and reducing the dimensions of all criss-cross attention components. This allows the model to directly inherit the hidden states of word-level and pixel-level features from the text and image pre-trained encoder. The operation process of this module is shown in Figure 3, presented with a feature map $H \in R^{W \times H}$. Firstly, two one-dimensional convolution layers with a kernel size of 1 are used in this method on H to produce feature maps Q and K , and $\{Q, K\} \in R^{W \times H}$. After obtaining Q and K , the attention map A , $A \in R^{(H+W-1) \times (W \times H)}$ is produced through an affinity fusion operation. We obtain a vector Q_u at each position u in Q . Meanwhile, we obtain a scalar for each position u in the two-dimensional feature map Ω_u by extracting the set of feature values from K within the same column or row as the u , where $\Omega_{i,u}$ is the i -th element of Ω_u . The following is the affinity fusion formula.:

$$g_{i,u} = Q_u \cdot \Omega_{i,u}^T \quad (7)$$

Here, $g_{i,u} \in G$ denotes the relationship between features Q_u and $\Omega_{i,u}$, which is $G \in R^{(H+W-1) \times (W \times H)}$, and that is then passed through a softmax layer to obtain the attention map A .

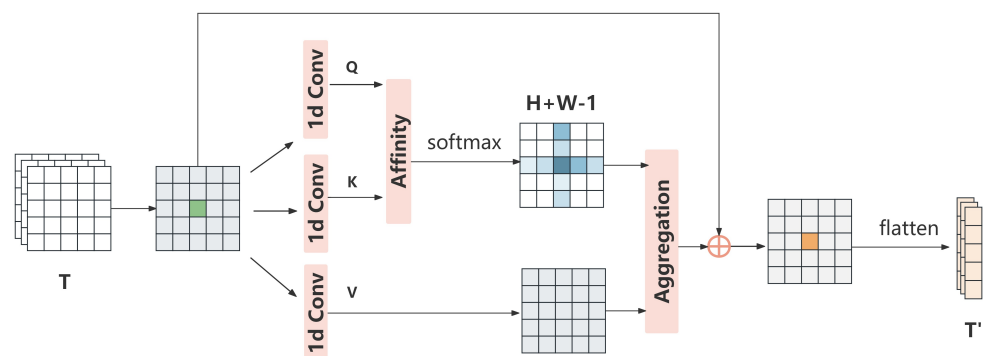


Figure 3. 1D-CCNet attention mechanism structure diagram.

Next, a one-dimensional convolution layer with a kernel size of 1 is used to process H , and generate $V \in R^{W \times H}$. For each position u , a set $V_u \in R^{(H+W-1)}$ is obtained, where V_u is the set of feature values within the identical column or row as the position u . The attention map A is aggregated with V_u , as described by the following formula:

$$H' = \sum_{i=0}^{W+H-1} A_{i,u} V_{i,u} + H \quad (8)$$

Here, H' denotes the new feature value at position u , $A_{i,u}$ is the scalar value of the attention map A at position (i,u) , and $V_{i,u}$ is the feature map's value at that position (i,u) .

Consequently, the text features vector T_f and the image features vector V_f obtained earlier are inputted into the 1D-CCNet attention mechanism module, where weights are computed for words and pixels to weight their vector representations. This captures the important relationships and contextual information across various positions in the input text sequence and visual features. The last text feature vector e_t and image feature vector e_v are obtained as follows:

$$e_t = \text{1D-CCNet}(T_f) \quad (9)$$

$$e_v = \text{1D-CCNet}(V_f) \quad (10)$$

Equations (9) and (10) describe the process of processing text and image feature vectors using the 1D-CCNet module, where e_t is the processed text feature vector and e_v is the processed image feature vector. These two vectors will be fused in subsequent steps.

3.5. Feature Fusion

When detecting fake news, textual content and visual content complement each other, so it is crucial to effectively integrate the feature information of these two modalities. Previous methods have often relied on simple connections between text and image information. Although the operation was simple, this approach often failed to fully utilize effective information from the data and was prone to introducing noise and redundant information. Therefore, it damaged the accuracy of detection and had a negative impact on model performance.

This article proposes a multimodal heterogeneous cross-fusion method to solve the problem, as shown in Figure 4. This method effectively utilizes the fusion of features from different modalities, while reducing the interference of redundant information and minimizing noise. Specifically, after obtaining the text feature vector and image feature vector, this article transposes the image feature vector while keeping the text feature vector unchanged, and then uses batch matrix multiplication to batch multiply the text feature vector and image feature vector to obtain the fused feature matrix. Then, the fused feature matrix is matrix multiplied with the identity mask matrix and the symmetry mask matrix, respectively. This operation divides the fused feature matrix into two parts, namely, the diagonal part and the upper and lower triangular parts. These two parts are flattened to obtain two eigenvectors, namely, diagonal eigenvectors M_1 and non diagonal eigenvectors M_2 . Finally, the weighted method is applied to concatenate these two vectors, resulting in the final feature vector F_f for classification. The formula is as follows:

$$M_1 = \text{Mask}_1(e_t \otimes e_v) \quad (11)$$

$$M_2 = \text{Mask}_2(e_t \otimes e_v) \quad (12)$$

$$F_f = \alpha M_1 + (1 - \alpha) M_2 \quad (13)$$

Equations (11)–(13) describe the process of fusing text features with image features. Here, e_t denotes the text features vector, and e_v denotes the image features vector. The mask matrix Mask_1 is used to obtain the main diagonal features, while Mask_2 is used to obtain the off-diagonal features. F_f represents the final fused feature vector.

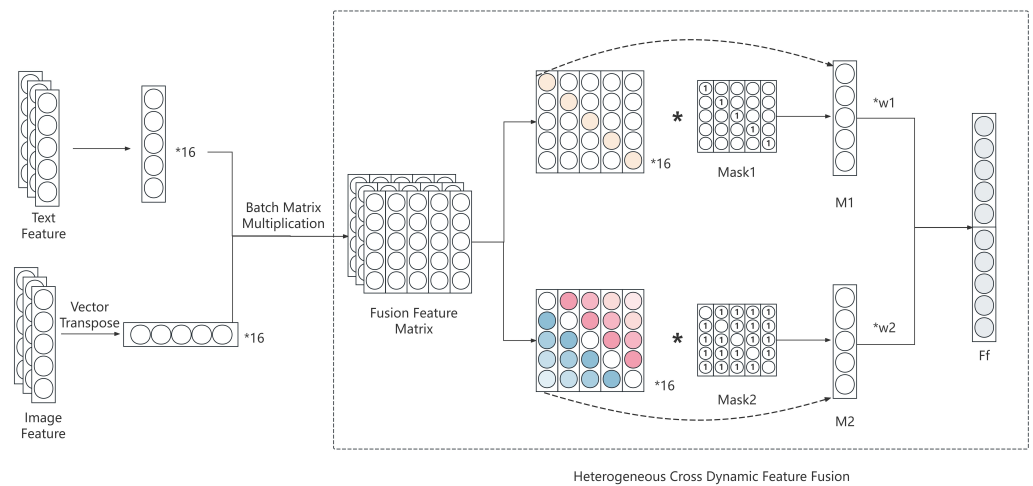


Figure 4. Multimodal heterogeneous cross-fusion structure diagram.

3.6. Output

The output layer comprises a linear transformation layer and a batch normalization layer, followed by a ReLU activation function layer. The fused feature vector F_f is linearly transformed to obtain F_f' . Then, a completely connected layer with a softmax activation function is utilized to output the probabilities of real and fake news. The formula is as follows:

$$y = \text{softmax}(w \times F_f' + b) \quad (14)$$

Here, y represents the predicted probabilities, w denotes the parameters of the fully connected layer, and b represents the bias term.

This paper utilizes the binary cross-entropy (BCE) loss function with logits. This method is simple yet powerful for handling binary classification tasks, improving numerical stability and computational efficiency, thereby aiding in achieving better optimization results. The formula is as follows:

$$\text{BCEWithLogitLoss}(y, \hat{y}) = \frac{1}{N} \sum_{i=1}^N [y_i \cdot \log(\sigma(\hat{y}_i)) + (1 - y_i) \cdot \log(1 - \sigma(\hat{y}_i))] \quad (15)$$

N is the number of samples, and $\sigma(x)$ is the sigmoid function used to convert logits into probability values ranging from 0 to 1. Its definition is given in the formula:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (16)$$

y_i is the true label of the i -th sample (0 or 1), and $\hat{y}_i$ is the model's output logits.

4. Results and Discussion

4.1. Dataset

The evaluation in this paper primarily uses the official English dataset Twitter [43], Gossipcop, and the Chinese dataset Weibo, as shown in Table 1. These three datasets all include text and images. The text section includes news content, tags, and other information. Among them, the tag includes two items, (0, 1), which, respectively, represent real news and fake news.

Table 1. Dataset statistics.

Statistic	Fake News	Real News	Total	Labels (F,R)	Text Length
Twitter	7021	5974	12,995	1, 0	50–150
Weibo	5974	4779	9528	1, 0	100–140
Gossipcop	3212	5078	8290	1, 0	100–700

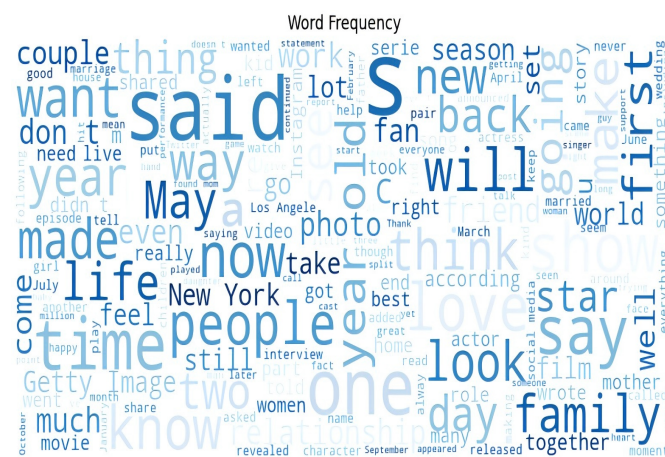
The Twitter dataset was sourced from the MediaEval2015 fake news dataset, which includes both a development set and a test set. It contains 17,000 tweets and associated images from the Twitter platform and is a commonly used multimodal dataset.

The Weibo dataset comprises fake news verified by the official Weibo fact-checking platform and real news validated by Xinhua News Agency, with both types of news, including text and images.

The Gossipcop [44] dataset is an annotated dataset for detecting fake news, containing news entries from the Gossipcop website. However, in order to classify the degree of the news from fake to real, the Gossipcop dataset provides the output as rating scores which are set on a scale of 1–10, where scores less than five are classified as fake, and scores higher than five are classified as real news. The dataset includes news headlines, body text, and true/fake labels, which are used to train and evaluate text classification models and fake news detection algorithms.

Figures 5 and 6 present a visual analysis of the Gossipcop dataset. To explore the features of the three datasets used in this study, we created word clouds that visually display the key terms and phrases within the datasets. These keywords helped in understanding the themes or topics covered in the datasets, revealing the language style of the news texts and providing insights into the textual characteristics of fake news. By examining the high-frequency words, we observed that the three datasets predominantly contain vocabulary related to life, time, and tasks. This suggests that the texts in these datasets are oriented towards descriptions of individuals and daily life.

Additionally, we analyzed the text length distribution across the three datasets, and explored length differences across different labels. The text length distributions varied among datasets, with Twitter and Weibo tending toward shorter texts, primarily around 100 words, while the Gossipcop dataset features longer texts, predominantly around 500 words. However, overall, the study focused on short-text processing. Differences in text length across different labels were relatively small, so they were relatively unhelpful for distinguishing between real and fake content.

**Figure 5.** Word cloud map of Gossipcop dataset.

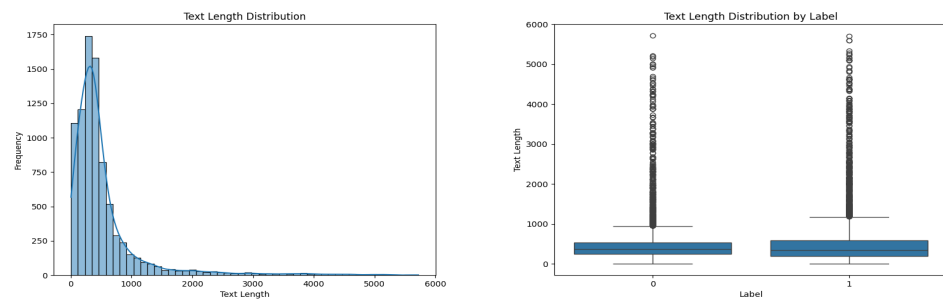


Figure 6. Text length distribution map (**left**) and comparison chart of different labels (**right**) of Gossipcop dataset. We conducted data visualization analysis on three datasets separately, but only the visualization results of the Gossipcop dataset are presented here.

Specifically, the Weibo dataset contains news texts and user attribute information, while Twitter only has news texts and propagation trees, and Gossipcop only has news texts and images. Therefore, for fairness, we only used news texts and image information from the three datasets in the experimental process. Table 2 shows the specific content of user attribute information for the Weibo dataset. This paper can more comprehensively evaluate the model’s generalizability and robustness by using datasets in different languages.

Table 2. Weibo dataset user attributes.

User Attribute	Value
Verified	false/true
Description	false/true
Gender	female/male
Messages	Above 0
Followers	Above 0
Location	Province Name
Time	Above 0
Friends	Above 0

The three datasets have a comparable number of real and fake news samples, avoiding issues of sample imbalance. This paper first performs basic preprocessing on these datasets to improve data quality, including filtering and adjusting the text and images.

4.2. Evaluating Indicator

This study uses accuracy, precision, recall, and F1-score as measures to assess the performance of the fake news detection model. These metrics were selected due to their comprehensive and balanced approach to assessing binary classification performance. Accuracy provides an overall measure of correct classifications, whereas precision ensures that identified fake news is indeed fake, thereby reducing false positives. Recall focuses on capturing as many actual fake news instances as possible, thus minimizing missed detections. The F1-score integrates precision and recall, offering a balanced perspective on model performance by addressing the trade-offs between the two. These metrics are widely recognized for their effectiveness in evaluating binary classification tasks [45]. Given the relatively balanced sample sizes of the two classes in our dataset, these indicators facilitate a more effective comparison of our method with the benchmark model. Thus, this paper employs these four metrics. The fundamental measures in classification are represented by True Positives (*TP*), True Negatives (*TN*), False Positives (*FP*) and False Negatives (*FN*). The formulas are as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (17)$$

$$Precision = \frac{TP}{TP + FP} \quad (18)$$

$$Recall = \frac{TP}{TP + FN} \quad (19)$$

$$F1\text{-score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (20)$$

4.3. Experimental Setup

This article uses the Python 3.9 programming language for the experiment. Table 3 shows the parameter values of all models. BERT is a pre-trained language model used in this article, with a vector dimension of 768; the BLIP-2 model is used for image feature extraction, with an input dimension of 224×224 for the image. We conducted experimental comparisons on commonly used optimizers and chose the AdamW optimization algorithm to update the model parameters. For the learning rate, we compared it within the range of 1×10^{-4} – 1×10^{-7} and discovered that a learning rate of 1×10^{-5} exhibited a model with more steady convergence. Based on the hardware available in this experiment and the experimental comparison results, we set the batch to 16. Since the results of this experiment tend to stabilize after 40 rounds, we set the epoch to 50.

Table 3. Hyperparameter configuration in this article.

Hyper Parameters	Parameter Values
Optimizer	AdamW
batch	16
Learning Rate	1×10^{-5}
epoch	50

4.4. Baseline Model

EANN: EANN includes a multimodal feature extractor and a fake news detector. To ensure experimental fairness, the event detector component was removed. The model integrates text and visual features to identify fake news, effectively combining features from different modalities to improve detection accuracy. By considering textual and visual information, EANN can analyze news content more comprehensively and assess its relevance to the task.

SAFE: SAFE extracts text features using LSTM and extracts image features using a pre-trained VGG model. By integrating the semantic content of both text and images, SAFE captures deep semantic connections in the news, enhancing the ability to identify false information and providing strong support for evaluating news authenticity.

SpotFake: SpotFake uses the BERT model to extract semantic features from text and the VGG model to extract visual features from images. These two features are concatenated to form a multimodal representation, which is then used directly for news classification. This approach captures semantic information from the news and considers the visual content in images, combining text and visual features to improve fake news detection's robustness and accuracy.

MVAE: MVAE uses BiLSTM and the VGG19 model to extract text and image features, respectively, and then concatenates these features. A variational autoencoder is then employed to learn the distribution of multimodal data and perform classification by reconstructing multimodal features. This method considers the semantic information from text and images, and efficiently combines features from both modalities, enhancing the model's ability to understand complex multimodal data and improving classification accuracy.

Att-RNN: Att-RNN combines attention mechanism and a recurrent neural network (RNN) to integrate text, visual, and user social features, merging image features into the collaborative representation of text and social context. In this experiment, the user social feature component was removed.

MCNN [46]: MCNN uses the BERT model and ResNet50 to extract text and image features, respectively, and aligns these features through BiGRU layers. The model organically combines text and image features using deep learning techniques, effectively capturing textual semantics and visual information and enhancing the ability of the model to understand cross-modal data. This multimodal fusion approach allows MCNN to perform exceptionally well with complex data, making it particularly suitable for tasks like fake news detection that require the integration of diverse information.

CAFE [47]: CAFE develops an ambiguity-aware multimodal fake news detection approach that dynamically integrates unimodal features and cross-modal correlations. By leveraging context-aware mechanisms, CAFE effectively addresses ambiguities in multimodal data, enhancing the model's ability to aggregate and interpret information from diverse sources.

4.5. Results

This section compares the proposed model with recent benchmark models based on four evaluation metrics, as shown in Table 4. The suggested model performed better than the benchmark models on a number of measures, according to the data. The model's performance is presented in this paper in terms of F1-score, accuracy, precision, and recall. On the Twitter fake news dataset and the Weibo fake news dataset, the proposed model achieved accuracy scores of 0.925 and 0.909, respectively, surpassing the accuracy of the benchmark models. Additionally, on the Twitter fake news dataset, the precision for detecting fake and real news reached 0.922 and 0.967, respectively, approximately 3% to 9% higher than the benchmark models. However, on the Weibo dataset, the accuracy and F1-score for fake news, as well as the recall for real news, were all lower than those of the MCNN model. The MCNN model employs a BiGRU layer to align text and image features. The bidirectional nature of BiGRU allows it to capture contextual information from both directions, thereby enhancing the interaction and integration between text and image. Additionally, the gated mechanism of GRU mitigates the vanishing gradient problem. However, the proposed method exhibits limitations in alignment across different modalities. Meanwhile, on the Gossipcop dataset, the accuracy and fake news precision were also slightly lower compared to the benchmark dataset, but our method outperformed the benchmark models with regard to the accuracy of real news and the F1-score of fake news.

Table 4. Comparison of results for models across three datasets.

Dataset	Model	Accuracy	Fake News			Real News		
			Precision	Recall	F1	Precision	Recall	F1
Twitter	Att-RNN	0.741	0.765	0.665	0.719	0.652	0.725	0.707
	EANN	0.710	0.815	0.692	0.754	0.875	0.901	0.890
	SAFE	0.763	0.671	0.503	0.582	0.637	0.772	0.703
	MVAE	0.742	0.812	0.722	0.761	0.719	0.787	0.770
	SpotFake	0.851	0.780	0.881	0.852	0.828	0.655	0.752
	MCNN	0.872	0.901	0.862	0.871	0.907	0.901	0.912
	Ours	0.925	0.922	0.924	0.922	0.927	0.925	0.926
Weibo	Att-RNN	0.753	0.675	0.740	0.755	0.677	0.702	0.770
	EANN	0.791	0.735	0.704	0.762	0.762	0.778	0.727
	SAFE	0.820	0.831	0.815	0.821	0.821	0.830	0.819
	MVAE	0.831	0.822	0.852	0.795	0.811	0.802	0.873
	SpotFake	0.851	0.863	0.891	0.853	0.850	0.831	0.827
	MCNN	0.850	0.882	0.870	0.905	0.871	0.881	0.870
	Ours	0.909	0.838	0.963	0.894	0.967	0.871	0.915
Gossipcop	Att-RNN	0.743	-	-	-	-	-	-
	EANN	0.856	0.702	0.518	0.594	0.841	0.881	0.875
	SAFE	0.838	0.758	0.558	0.643	0.857	0.937	0.895
	MVAE	0.815	-	-	-	-	-	-
	SpotFake	0.851	0.732	0.375	0.494	0.866	0.942	0.914
	CAFE	0.867	0.732	0.490	0.587	0.887	0.957	0.921
	Ours	0.857	0.696	0.592	0.639	0.893	0.930	0.912

With regard to time performance, we conducted experiments on three datasets with training sets consisting of 10,396, 7622, and 6632 pairs of data, respectively. The total estimated model parameter size was 1752.515 MB. Due to variations in text length and image size across different datasets, the training times varied accordingly. On average, the processing time per news pair was 0.023 s, resulting in an average training time of 2.5 to 3.9 min per epoch. With 50 epochs per experiment, the average total training time per experiment ranged from 125 to 195 min, which was less than 3.5 h overall. Compared to other similar multimodal models, the training time of our model fell within a reasonable range. Overall, while enhancing performance, our model maintained a reasonable training time, demonstrating commendable time efficiency.

This article conducted five experiments on each of these three datasets, and the average value of each evaluation metric was taken as the final experimental result. Table 5 shows the standard deviation of the five results for all evaluation indicators. It can be seen that for the Twitter dataset, the standard deviation of all rating indicators was less than 0.01, indicating that the results of our model on the Twitter were relatively stable, with small fluctuations, and can be trusted. However, on the other two datasets, the standard deviation of these indicators was larger, indicating that the experimental results were not as stable as the Twitter dataset, especially on the Gossipcop dataset, where fluctuations were greater. This indicates that the model's ability to handle longer texts still needs improvement. However, overall, these standard deviations were within a normal range, which indicates that the experimental results in this article are relatively stable and reliable.

Figure 7 shows the accuracy curves for the proposed method's training and validation on the three datasets, respectively. For the Twitter dataset, the BTCM model achieved accuracy of 0.984 and 0.925 on the training and validation samples, respectively. For the Weibo dataset, the BTCM model achieved accuracy of 0.985 and 0.909 on the training and validation samples, respectively. For the Gossipcop dataset, the BTCM model achieved accuracy of 0.966 and 0.857 on the training and validation samples.

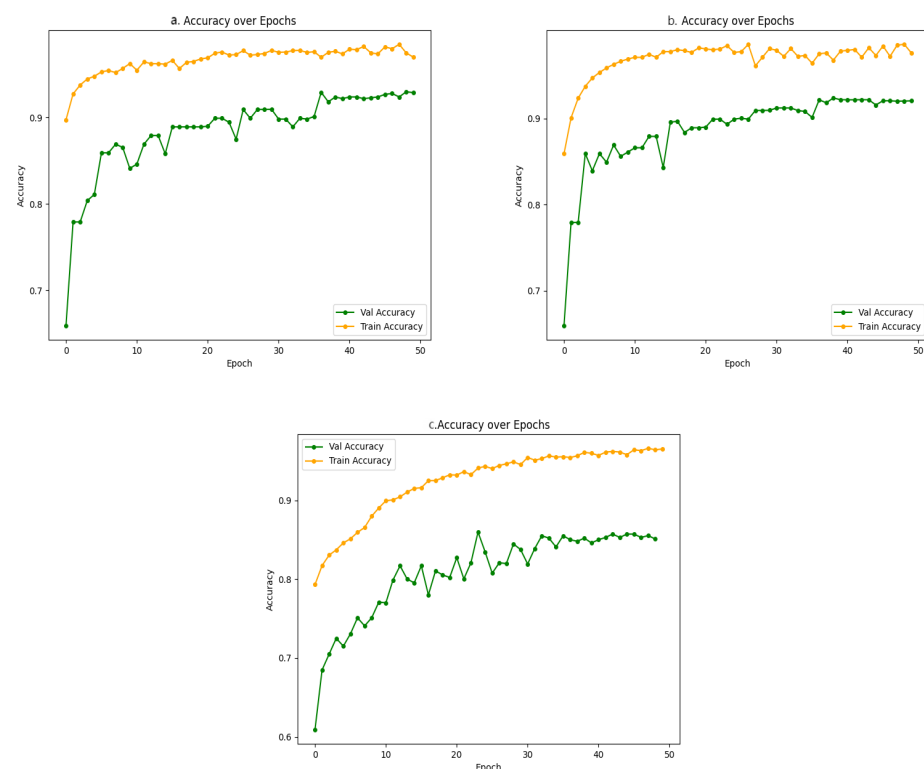
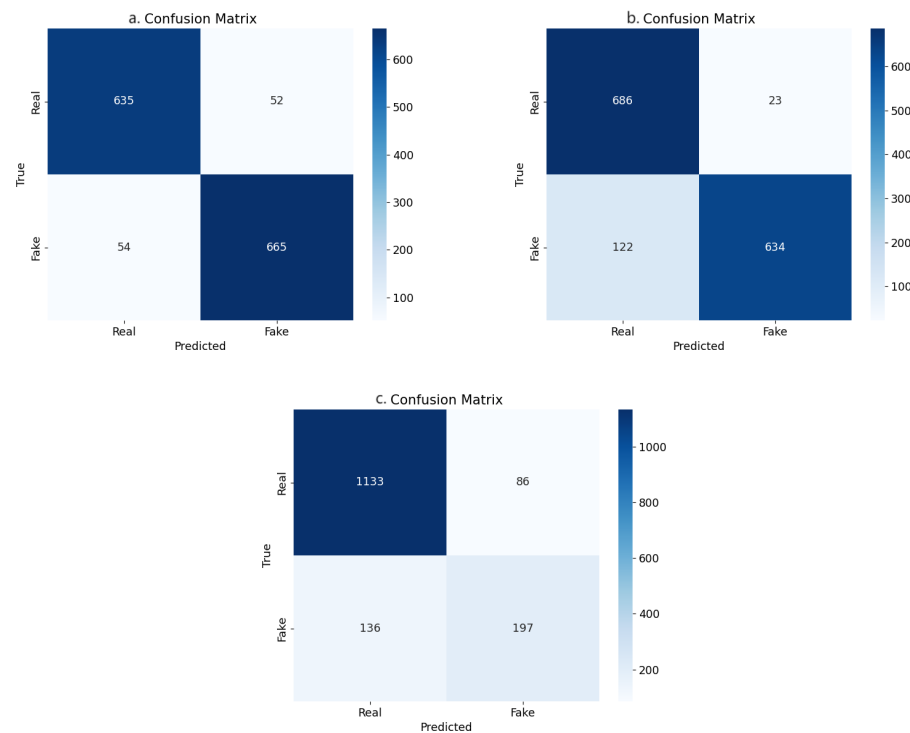


Figure 7. The variation curves of training and validation accuracy on Twitter (a), Weibo (b), and Gossipcop (c) datasets.

Table 5. Standard deviation of each indicator in the three datasets.

Dataset	Accuracy	Fake News			Real News		
		Precision	Recall	F1	Precision	Recall	F1
Twitter	0.0042	0.0029	0.0048	0.0043	0.0072	0.0040	0.0052
Weibo	0.0102	0.0156	0.0162	0.0161	0.0080	0.0076	0.0079
Gossipcop	0.0296	0.0105	0.0126	0.0119	0.0103	0.0112	0.0109

One approach for assessing how well classification models perform, especially in supervised learning tasks with classification issues, is the confusion matrix. Based on four important indicators, it shows a comparison between the model's projected and actual results: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). True Positives indicate the quantity of samples correctly identified as positive, True Negatives indicate the quantity of samples correctly identified as negative, False Positives indicate the quantity of samples incorrectly identified as positive, and False Negatives indicate the quantity of samples incorrectly identified as negative. Figure 8 displays the confusion matrices for results of the BTCM model on the Twitter (a), Weibo (b), and Gossipcop (c) datasets, respectively.

**Figure 8.** The confusion matrices for the Twitter (a), Weibo (b), and Gossipcop (c) datasets.

4.6. Ablation Experiment

To clarify the functionality of each module in the model, this study performs ablation experiments to verify the contributions of the DCCA (1D-CCNet attention mechanism) and HCFFM modules (heterogeneous cross-feature fusion method) to our method. Table 6 displays the results of the experiment. In the model, “BTCM-DC” refers to the experimental results with the DCCA module removed, “BTCM-HC” refers to the experimental results with the HCFFM module removed, and “BTCM” refers to the results of the complete model. Removing the DCCA module reduced accuracy by 4.3%, 2%, and 2% on the three datasets, respectively; removing the HCFFM module reduced accuracy by 3.3%, 1.8%, and 1.9% on the three datasets. The results of the experiment demonstrate that the DCCA and HCFFM

modules contribute to the fake news detection task, with a more pronounced effect on the Twitter dataset, thus improving the model's detection performance.

Table 6. BTCM results of ablation experiment.

Dataset	Model	Accuracy	F1
Twitter	BTCM-DC	0.882	0.873
	BTCM-HC	0.892	0.885
	BTCM	0.925	0.922
Weibo	BTCM-DC	0.889	0.860
	BTCM-HC	0.891	0.872
	BTCM	0.909	0.894
Gossipcop	BTCM-DC	0.837	0.388
	BTCM-HC	0.838	0.390
	BTCM	0.857	0.412

Figures 9 and 10 show the accuracy and loss curves of the BTCM compared with the BTCM-DC and BTCM-HC models on the three datasets, respectively. As can be shown, the BTCM model's accuracy is noticeably superior to the other two ablation models, and on the Twitter and Weibo datasets, the accuracy curve of the BTCM model fluctuates less, and the results in the middle and later stages are more stable. This indicates that the DCCA module for enhancing text feature extraction and the multimodal heterogeneous cross-fusion method proposed in this study improve the fake news detection model's performance.

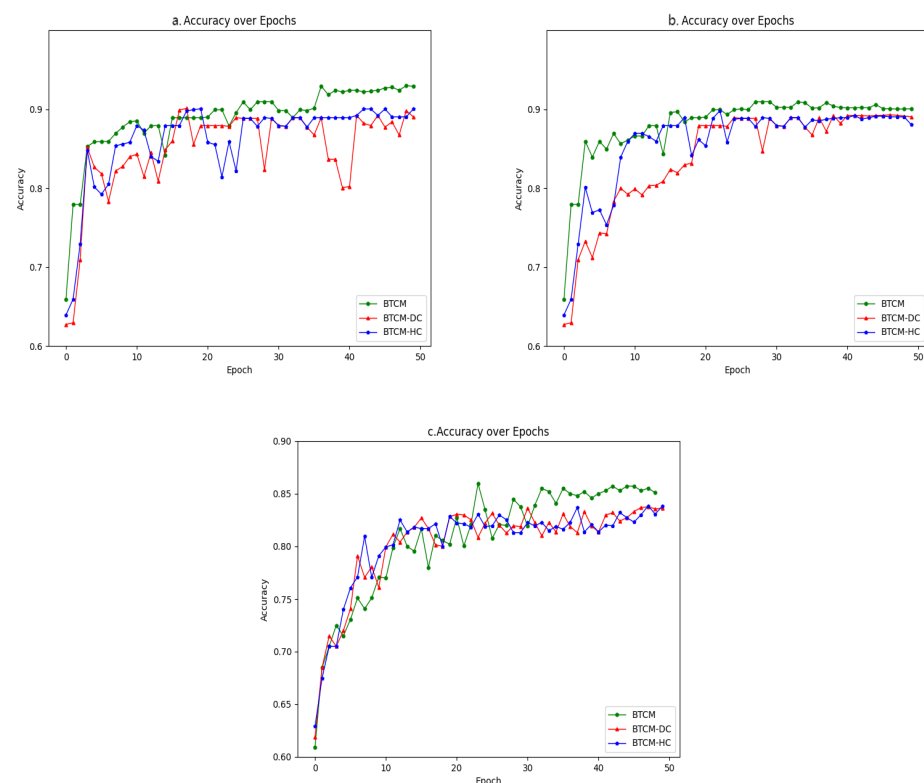


Figure 9. Comparison curves of accuracy of ablation model on Twitter (a), Weibo (b), and Gossipcop (c) datasets.

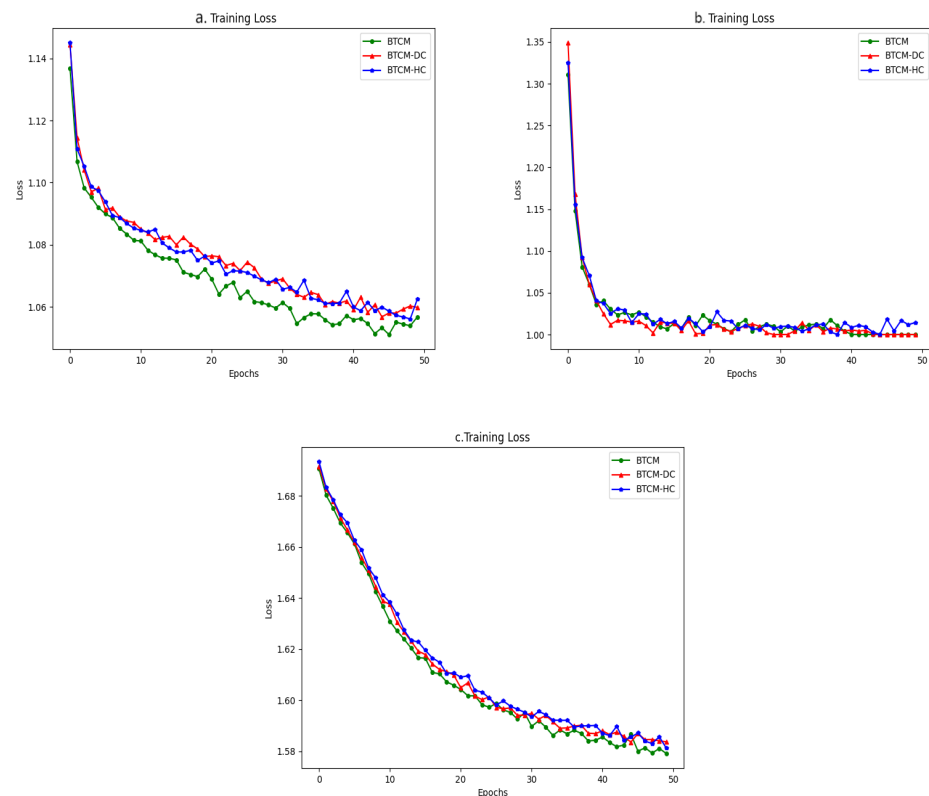


Figure 10. Comparison curves of loss of ablation model on Twitter (a), Weibo (b), and Gossipcop (c) datasets.

5. Conclusions

This article suggests a fake news detection model, BTCM, utilizing a 1D-CCNet attention mechanism. To some extent, this method solves the problems of insufficient feature extraction of news content and insufficient fusion of different modalities of information. It uses a BERT model to extract text features and a BLIP-2 encoder to extract image features. Moreover, the 1D-CCNet attention mechanism module is introduced to process input text and image sequences, enhancing the important aspect of bimodal features, effectively focusing on the most relevant parts between text and image, and being able to obtain contextual information of text and image spatial distance, improving the model's ability to capture global and local information from data. Meanwhile, the BLIP-2 pre-trained model is used for object detection in images, generating image descriptions and enhancing text data to enhance data diversity. This operation aims to more effectively enhance the correlation between different modalities. Finally, a multimodal heterogeneous cross-feature fusion method (HCFFM) is proposed to fuse image and text features, improving performance by reducing noise, interference from redundant information, and redundant information. According to the experimental results, our method performs better than the benchmark method on official public datasets from Twitter, Weibo, and Gossipcop, with accuracies of 0.925, 0.909, and 0.857, respectively. Especially on the Twitter dataset, the effect is significant, with higher accuracy, recall, and F1-score than the benchmark model for both fake and real news. Although this article has achieved certain results in multimodal fake news detection, the model proposed in this article still has certain limitations in feature extraction and modal fusion:

- (1) Although the model in this article tries to extract text and image features as much as possible, it cannot completely solve the problems of correlation and heterogeneity between different modalities, which may result in the model not fully utilizing the features of different modalities in the detection process.

- (2) Official public datasets such as Twitter, Weibo, and Gossipcop were used in the experiment and achieved good results. However, there may be limitations in certain other fields or datasets, and these results cannot be directly extended to other types of fake news or other social media platforms. Therefore, the applicability of the model in different datasets or domains needs further validation.
- (3) The main focus of this article is in the field of computer science. In the future, collaboration with disciplines such as journalism, psychology, sociology, and communication can be considered to gain a more thorough understanding of the mechanisms involved in processing, disseminating, and receiving news information, thereby improving the accuracy of identifying and predicting fake news.
- (4) Although detecting fake news is a crucial step, preventing its spread is equally important. To effectively address the spread of fake news, future work can explore the use of proactive defense strategies and community-driven approaches. Active defense strategies can prevent the spread of false news by monitoring and intervening in real-time. The community-driven approach improves the accuracy of information by mobilizing community members to identify and report fake news. It is recommended to explore in detail the practical applications of these strategies in future research, as well as how they complement detection models to form a comprehensive system for preventing and controlling fake news.

Author Contributions: Conceptualization, Y.Y. and F.W.; methodology, Y.Y. and F.W.; software, Y.Y. and F.W.; validation, H.F.; formal analysis, Y.Y.; investigation, Y.Y.; resources, Y.Y.; data curation, H.F.; writing—original draft preparation, Y.Y.; writing—review and editing, H.F.; visualization, Y.Y.; supervision, Y.Y. and H.F.; project administration, Y.Y. and H.F.; funding acquisition, H.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Hainan Provincial Natural Science Foundation Project (622RC675); National Natural Science Foundation of China (62262019).

Data Availability Statement: The data presented in this study are available in the public domain: <https://github.com/OwenLeng/rumor-detection-include-twitter15-twitter16data-> (accessed on 25 July 2024).

Acknowledgments: The authors would like to acknowledge and thank all reviewers for their constructive and helpful reviews.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Ruchansky, N.; Seo, S.; Liu, Y. Csi: A hybrid deep model for fake news detection. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, Singapore, 6–10 November 2017; pp. 797–806.
2. Rish, I. An empirical study of the naive Bayes classifier. In Proceedings of the IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, Seattle, WA, USA, 3 August 2001; pp. 41–46.
3. Safavian, S.R.; Landgrebe, D. A survey of decision tree classifier methodology. *IEEE Trans. Syst. Man Cybern.* **1991**, *21*, 660–674. [CrossRef]
4. Noble, W.S. What is a support vector machine? *Nat. Biotechnol.* **2006**, *24*, 1565–1567. [CrossRef] [PubMed]
5. Volkova, S.; Shaffer, K.; Jang, J.Y.; Hodas, N. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Vancouver, BC, Canada, 30 July–4 August 2017; pp. 647–653.
6. Potthast, M.; Kiesel, J.; Reinartz, K.; Bevendorff, J.; Stein, B. A stylometric inquiry into hyperpartisan and fake news. *arXiv* **2017**, arXiv:1702.05638.
7. Castillo, C.; Mendoza, M.; Poblete, B. Information credibility on twitter. In Proceedings of the 20th International Conference on World Wide Web, Hyderabad, India, 28 March–1 April 2011; pp. 675–684.
8. Chen, Y.; Conroy, N.J.; Rubin, V.L. Misleading online content: Recognizing clickbait as “false news”. In Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection, Seattle, WA, USA, 13–17 April 2015; pp. 15–19.
9. Feng, S.; Banerjee, R.; Choi, Y. Syntactic stylometry for deception detection. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Jeju, Republic of Korea, 8–14 July 2012; pp. 171–175.
10. Pérez-Rosas, V.; Kleinberg, B.; Lefevre, A.; Mihalcea, R. Automatic detection of fake news. *arXiv* **2017**, arXiv:1708.07104.

11. Qazvinian, V.; Rosengren, E.; Radev, D.; Mei, Q. Rumor has it: Identifying misinformation in microblogs. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, Edinburgh, UK, 27–29 July 2011; pp. 1589–1599.
12. Kwon, S.; Cha, M.; Jung, K.; Chen, W.; Wang, Y. Prominent features of rumor propagation in online social media. In Proceedings of the 2013 IEEE 13th International Conference on Data Mining, Dallas, TX, USA, 7–10 December 2013; IEEE: Piscataway, NJ, USA, 2013; pp. 1103–1108.
13. Ma, J.; Gao, W.; Mitra, P.; Kwon, S.; Jansen, B.J.; Wong, K.-F.; Cha, M. *Detecting Rumors from Microblogs with Recurrent Neural Networks*; AAAI Press: Washington, DC, USA, 2016.
14. Yu, F.; Liu, Q.; Wu, S.; Wang, L.; Tan, T. A Convolutional Approach for Misinformation Identification. In Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI), Melbourne, Australia, 19–25 August 2017; pp. 3901–3907.
15. Taher, Y.; Moussaoui, A.; Moussaoui, F. Automatic fake news detection based on deep learning, FasTtext and news title. *Int. J. Adv. Comput. Sci. Appl.* **2022**, *13*, 1–7. [\[CrossRef\]](#)
16. Verma, P.K.; Agrawal, P.; Amorim, I.; Prodan, R. WELFake: Word embedding over linguistic features for fake news detection. *IEEE Trans. Comput. Soc. Syst.* **2021**, *8*, 881–893. [\[CrossRef\]](#)
17. Low, J.F.; Fung, B.C.; Iqbal, F.; Huang, S.C. Distinguishing between fake news and satire with transformers. *Expert Syst. Appl.* **2022**, *187*, 115824. [\[CrossRef\]](#)
18. Truică, C.O.; Apostol, E.S.; Paschke, A. Awakened at CheckThat! 2022: Fake news detection using BiLSTM and sentence transformer. In Proceedings of the CEUR Workshop Proceedings, Bologna, Italy, 5–8 September 2022; Volume 3180, pp. 49–757.
19. Truică, C.O.; Apostol, E.S. It's all in the embedding! Fake news detection using document embeddings. *Mathematics* **2023**, *11*, 508. [\[CrossRef\]](#)
20. Truică, C.O.; Apostol, E.S.; Karras, P. DANES: Deep neural network ensemble architecture for social and textual context-aware fake news detection. *Knowl.-Based Syst.* **2024**, *294*, 111715. [\[CrossRef\]](#)
21. Zhou, P.; Han, X.; Morariu, V.I.; Davis, L.S. Learning rich features for image manipulation detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1053–1061.
22. Jin, Z.; Cao, J.; Zhang, Y.; Zhou, J.; Tian, Q. Novel visual and statistical image features for microblogs news verification. *IEEE Trans. Multimed.* **2016**, *19*, 598–608. [\[CrossRef\]](#)
23. Qi, P.; Cao, J.; Yang, T.; Guo, J.; Li, J. Exploiting multi-domain visual information for fake news detection. In Proceedings of the 2019 IEEE International Conference on Data Mining (ICDM), Beijing, China, 8–11 November 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 518–527.
24. Jin, Z.; Cao, J.; Guo, H.; Zhang, Y.; Luo, J. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In Proceedings of the 25th ACM International Conference on Multimedia, Mountain View, CA, USA, 23–27 October 2017; pp. 795–816.
25. Khattar, D.; Goud, J.S.; Gupta, M.; Varma, V. Mvae: Multimodal variational autoencoder for fake news detection. In Proceedings of the World Wide Web Conference, San Francisco, CA, USA, 13–17 May 2019; pp. 2915–2921.
26. Singhal, S.; Shah, R.R.; Chakraborty, T.; Kumaraguru, P.; Satoh, S. Spotfake: A multi-modal framework for fake news detection. In Proceedings of the 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM), Singapore, 19–21 September 2019; pp. 39–47.
27. Wang, Y.; Ma, F.; Jin, Z.; Yuan, Y.; Xun, G.; Jha, K.; Su, L.; Gao, J. 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, London, UK, 19–23 August 2018; pp. 849–857.
28. Zhou, X.; Wu, J.; Zafarani, R. Safe: Similarity-aware multi-modal fake news detection. *arXiv* **2020**, arXiv:2003.04981.
29. Qian, S.; Wang, J.; Hu, J.; Fang, Q.; Xu, C. Hierarchical multi-modal contextual attention network for fake news detection. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, 11–15 July 2021; pp. 153–162.
30. Kaliyar, R.K.; Goswami, A.; Narang, P.; Sinha, S. FNDNet—A deep convolutional neural network for fake news detection. *Cogn. Syst. Res.* **2020**, *61*, 32–44. [\[CrossRef\]](#)
31. Zhang, G.; Giachanou, A.; Rosso, P. SceneFND: Multimodal fake news detection by modelling scene context information. *J. Inf. Sci.* **2024**, *50*, 355–367. [\[CrossRef\]](#)
32. Yang, P.; Ma, J.; Liu, Y.; Liu, M. Multi-modal transformer for fake news detection. *Math. Biosci. Eng. MBE* **2023**, *20*, 14699–14717. [\[CrossRef\]](#) [\[PubMed\]](#)
33. Kananian, M.; Badiel, F.; Ghahramani, S. GRaMuFeN: Graph-based Multi-modal Fake News Detection in Social Media. *arXiv* **2023**, arXiv:2310.07668.
34. Arachie, C.; Gaur, M.; Anzaroot, S.; Groves, W.; Zhang, K.; Jaimes, A. Unsupervised detection of sub-events in large scale disasters. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 354–361.
35. Li, A.W.; Xu, X.K.; Fan, Y. Immunization strategies for false information spreading on signed social networks. *Chaos Solitons Fractals* **2022**, *162*, 112489. [\[CrossRef\]](#)
36. Apostol, E.-S.; Truică, C.-O.; Paschke, A. ContCommRTD: A distributed content-based misinformation-aware community detection system for real-time disaster reporting. *IEEE Trans. Knowl. Data Eng.* **2024**, 1–12. [\[CrossRef\]](#)

37. Truică, C.O.; Apostol, E.S.; Nicolescu, R.C.; Karras, P. MCWDST: A minimum-cost weighted directed spanning tree algorithm for real-time fake news mitigation in social media. *IEEE Access* **2023**, *11*, 125861–125873. [\[CrossRef\]](#)
38. Coban, Ö.; Truică, C.O.; Apostol, E.S. CONTAIN: A community-based algorithm for network immunization. *arXiv* **2023**, arXiv:2303.01934.
39. Petrescu, A.; Truică, C.O.; Apostol, E.S.; Karras, P. Sparse shield: Social network immunization vs. harmful speech. In Proceedings of the 30th ACM International Conference on Information & Knowledge Management, Gold Coast, QLD, Australia, 1–5 November 2021; pp. 1426–1436.
40. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
41. Li, J.; Li, D.; Savarese, S.; Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Proceedings of the International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023; pp. 19730–19742.
42. Huang, Z.; Wang, X.; Huang, L.; Huang, C.; Wei, Y.; Liu, W. Ccnet: Criss-cross attention for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 603–612.
43. Boididou, C.; Papadopoulos, S.; Zampoglou, M.; Apostolidis, L.; Papadopoulou, O.; Kompatsiaris, Y. Detection and visualization of misleading content on Twitter. *Int. J. Multimed. Inf. Retr.* **2018**, *7*, 71–86. [\[CrossRef\]](#)
44. Shu, K.; Mahudeswaran, D.; Wang, S.; Lee, D.; Liu, H. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big Data* **2020**, *8*, 171–188. [\[CrossRef\]](#) [\[PubMed\]](#)
45. Rainio, O.; Teuho, J.; Klén, R. Evaluation metrics and statistical tests for machine learning. *Sci. Rep.* **2024**, *14*, 6086. [\[CrossRef\]](#) [\[PubMed\]](#)
46. Zhang, Y.; Zhou, D.; Chen, S.; Gao, S.; Ma, Y. Single-image crowd counting via multi-column convolutional neural network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 589–597.
47. Chen, Y.; Li, D.; Zhang, P.; Sui, J.; Lv, Q.; Tun, L.; Shang, L. Cross-modal Ambiguity Learning for Multimodal Fake News Detection. In Proceedings of the ACM Web Conference 2022, Lyon, France, 25–29 April 2022; pp. 2897–2905.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.