

Article

Vae-Clip: Unveiling Deception through Cross-Modal Models and Multi-Feature Integration in Multi-Modal Fake News Detection

Yufeng Zhou ¹, Aiping Pang ^{1,2,*} and Guang Yu ³¹ Electrical Engineering College, Guizhou University, Guiyang 550025, China; z874234138@163.com² Key Laboratory of “Internet+” Collaborative Intelligent Manufacturing in Guizhou Province, Guiyang 550025, China³ School of Management, Harbin Institute of Technology, Harbin 150001, China; yug@hit.edu.cn

* Correspondence: appang@gzu.edu.cn

Abstract: With the development of internet technology, fake news has become a multi-modal collection. The current news detection methods cannot fully extract semantic information between modalities and ignore the rumor properties of fake news, making it difficult to achieve good results. To address the problem of the accurate identification of multi-modal fake news, we propose the Vae-Clip multi-modal fake news detection model. The model uses the Clip pre-trained model to jointly extract semantic features of image and text information using text information as the supervisory signal, solving the problem of semantic interaction across modalities. Moreover, considering the rumor attributes of fake news, we propose to fuse semantic features with rumor style features using multi-feature fusion to improve the generalization performance of the model. We use a variational autoencoder to extract rumor style features and combine semantic features and rumor features using an attention mechanism to detect fake news. Numerous experiments were conducted on four datasets primarily composed of Weibo and Twitter, and the results show that the proposed model can accurately identify fake news and is suitable for news detection in complex scenarios, with the highest accuracy reaching 96.3%.

Keywords: multi-modal; fake news detection; variational autoencoder; rumor style feature



Citation: Zhou, Y.; Pang, A.; Yu, G.

Vae-Clip: Unveiling Deception through Cross-Modal Models and Multi-Feature Integration in Multi-Modal Fake News Detection. *Electronics* **2024**, *13*, 2958. <https://doi.org/10.3390/electronics13152958>

Academic Editor: Chiman Kwan

Received: 25 June 2024

Revised: 21 July 2024

Accepted: 23 July 2024

Published: 26 July 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

The rapid advancement of internet technology, combined with the widespread use of social media due to its convenience, affordability, and speed, has resulted in 4.95 billion people being online out of a global population of 7.95 billion [1]. Social networking platforms have become the primary source for accessing information, with millions of users continuously browsing and sharing content. However, this convenience and speed also make social media an ideal environment for the spread of fake news. Today, to increase the credibility of fake news, it spreads in the form of a combination of text and images [2]. Compared to real news, fake news has the characteristics of being novel, spreading quickly, having a deep impact, and reaching a broad audience [3]. Fake news misleads readers with fabricated content, not only distorting the message but also inducing negative emotions among readers, sometimes even manipulating large public events. Conceptualizing truth or accuracy is at the core of all human efforts, and the spread of fake news fundamentally challenges this core, undermining people’s trust in social media [4]. Only by effectively identifying false information can better response strategies be developed. Therefore, detecting fake news plays a crucial role in creating a healthy social networking platform environment.

The earliest fake news detection was performed manually by experts. Although this method had high accuracy, it was inefficient, time-consuming, and difficult to detect fake news in a timely manner. As the number of internet users increased, this method was gradually phased out. Since manual inspection could not effectively solve the problem of news

detection, automatic detection technology emerged [5]. Early researchers manually added a series of labels such as “user behavior”, “user information”, and “news dissemination time” and used them in machine learning to detect the authenticity of news [6–9]. However, these labels showed randomness in fake news detection and were not utilized in all news detection models. With the maturation of deep learning models, using deep learning models for fake news detection has become the mainstream method in this field. Mining news content for authenticity detection has attracted extensive attention from researchers. Many researchers have used semantic information extracted from news copy to detect authenticity [10,11]. However, with the continuous development of network technology, visual information began to inundate social media platforms, and more and more news began to spread in the form of copy combined with images. Fake news also updated accordingly, and in order to attract readers’ attention, not only was the text fabricated but the images were also processed [12,13]. The use of image information to detect news has proven that images also play a role in detecting fake news. Today’s fake news is a collection of multiple modes, and the information between different modes has complementary effects. Using a single-mode information extraction detection model cannot meet the needs of this collection. At the same time, visual information can provide readers with a more direct visual impact and make it easier for readers to be convinced of the news content. Therefore, only by combining and considering the information content of the two modalities can proper authenticity detection for news be conducted. This paper introduces a multi-modal fake news detection model based on CLIP, where fusion features are obtained by standardizing and weighting the similarity between text and image features [14]. Knowledge distillation is a technique used to extract and transfer insights from one single-modal feature extraction model to another, enhancing the relationship between various modalities [15]. A cross-modal fuzzy assessment mechanism has been established to assign weights to text and image information [16]. Additionally, Ref. [17] employs a similarity measurement module to evaluate the alignment between images and text for fake news detection. It introduces and assesses four types of image-text similarities: textual similarity, semantic similarity, contextual similarity, and post-training similarity [18]. The assessment reveals that enhancing the similarity between text and images contributes to the detection of fake news.

Although previous research has made significant progress in news detection and outperformed traditional deep learning models in performance, they still cannot cope with the following challenges:

The challenge of cross-modal information feature extraction: News spread on social media today encounters a semantic gap between image and text modalities. Despite conveying identical semantic information, text and images exhibit significant feature differences due to variations in expression. In the processing of textual and visual features, the latent semantic correlations between them have not been thoroughly explored. This makes it challenging to obtain effective and generalizable features, leading to suboptimal outcomes in fake news detection.

The challenge of rumor feature extraction: A result in Science shows that fake news has higher “novelty” than real news. In this paper, “novelty” is defined as “rumor-style features”. For the text information of fake news, its content is intertwined with rumor-style features, which is a unique attribute of fake news. Existing methods ignore this attribute, resulting in poor model generalization and detection results.

To address the problem of accurately identifying multi-modal fake news and obtaining features that can accurately detect news, we propose a news detection model based on Vae-Clip. The model consists of five modules: a cross-modal semantic feature extraction module, a similarity measurement module, a rumor feature extraction module, a feature fusion module, and a news detection module. Among them, the cross-modal semantic feature extraction module extracts the features of image and text information based on text information as the supervisory signal. The similarity measurement module measures the similarity of image and text features. The rumor feature extraction module is used

to separate and extract rumor-style features from the text content of news. Finally, the image-text semantic features and the text's rumor-style features are fused in the feature fusion module. Numerous experiments conducted on four datasets benchmarked on Weibo and Twitter demonstrate that our proposed method achieves better detection results than existing methods.

Our main contributions are summarized as follows:

A Vae-Clip-based news detection model is introduced to leverage the correlation between cross-modal information for the joint extraction of features from both images and text. By using text data as a supervisory signal, the model extracts image features that represent the intrinsic qualities of the images, transcending specific categories. This method promotes semantic interaction between image and text features.

The use of the cosine similarity theorem to encourage similar information in image-text features to be closer in the embedding space. The obtained similarity scores of image-text semantic features serve as one of the criteria for news detection.

The utilization of a variational autoencoder to extract rumor features belonging to fake news, based on the rumor attribute of fake news, enhancing the model's generalization performance and improving the detection effectiveness of news.

The structure of the paper is as follows: Section 2 reviews relevant work in this field. Section 3 presents the problem description and introduces and derives the model proposed herein. Section 4 showcases extensive experimental results and analysis. Conclusions drawn from the experiments are summarized in Section 5.

2. Related Work

This section provides a brief review of the most pertinent methods in fake news detection. Fake news is defined as news that can be verified as false [19]. The objective of news detection is to assess the authenticity of news, typically framed as a binary classification problem (true or false) in deep learning models [19]. The primary challenge lies in accurately classifying news based on its features, as illustrated in Figure 1. Consequently, we divide fake news detection tasks into three categories based on the features used: machine learning-based detection, single-modal feature detection, and multi-modal feature detection. The following sections will explore these three categories in detail.

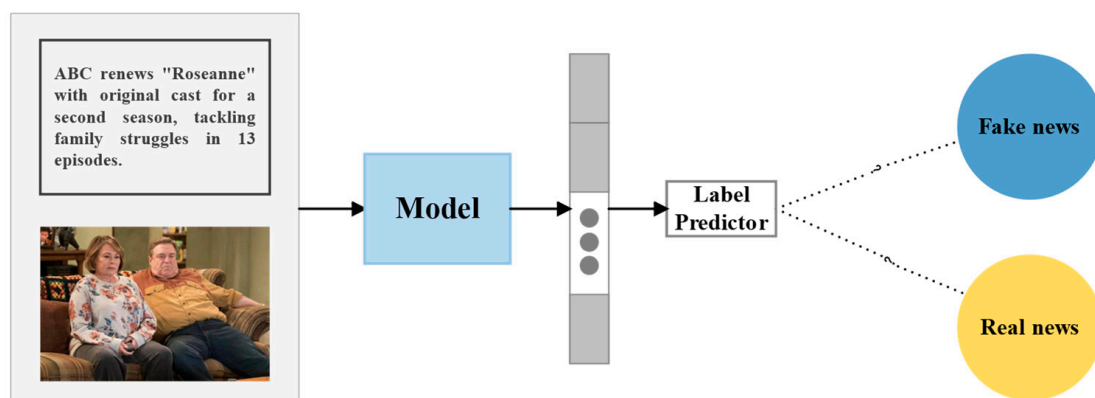


Figure 1. Problem overview diagram.

2.1. Machine Learning-Based Detection

Early news detection tasks involved setting multiple manual labels based on the user's social environment. These labels were used for feature extraction, followed by the use of support vector machine classifiers, decision tree classification, and random forest algorithms for fake news detection [6–9]. The Naive Bayes model tracks users' usage of specific words in news articles to assess the likelihood of the news being true or false [6]. The study introduces a time window for tracking rumor spread and evaluates rumors based on user features during different periods [7]. Features are developed from the users' language

information and text complexity, and a linear support vector machine is used to verify the authenticity of news [8]. Statistical data from article words are utilized to create word features, which are subsequently analyzed using a logistic regression model to identify false information [20]. Additionally, the study considers news credibility, user engagement, and the number of comments to extract features, employing multiple regression models to predict fake news [21]. These methods not only consume a lot of manpower but also have very strict requirements for datasets and have poor universality, making it difficult to effectively solve the problem of news detection.

2.2. Single-Modal Fake News Detection

With the continuous development of deep neural network models, researchers have applied them to the field of news detection. Single-modal detection model: Researchers utilized deep neural networks to process information from a single modality for news detection. Features are extracted from text content for news detection [10,11,22]. For rumor detection, researchers use a recursive neural network to model rumor text as variable-length time series, capturing rich semantic information [10]. An attention mechanism is employed to combine text features with temporal features, assigning different weights to specific words for more precise rumor detection [11]. The study integrated news text content with user information and features from post comments for a thorough analysis [23–28]. When extracting features from tweet text, a graph network is constructed by incorporating user features to visualize the interactions among users [24]. This method captures the correlation between tweets and user interactions, enhancing the prediction of tweet authenticity. By applying a shared attention mechanism to both news text and its comments, the system generates features that aid in detecting news authenticity [26]. Additionally, news text, headlines, keywords, and user comments are simultaneously analyzed and organized into a series of subgraphs [25]. Through a hierarchical path-aware kernelized graph attention network, it filters out information conducive to the detection of news authenticity. A fine-grained reasoning framework is established for post content, post comments, and users associated with the post [27]. By integrating human information processing models and prior knowledge, the accuracy and interpretability of fake news detection are enhanced. Leveraging historical data, they established an evidence extraction model that infers news authenticity by combining features derived from text content and evidence [29,30]. Researchers also extracted features from image content for news detection [12,13]. They converted visual information into pixel domain features and frequency domain features, which are then combined to assess the authenticity of news [12]. Additionally, they extracted features from multiple pre-trained models and combined them to judge news authenticity [13]. These methods only focused on information from a single modality, ignoring the fact that content between modalities is complementary and can enhance the detection effect of news. Moreover, nowadays, news is a collection of multiple modalities, and it is not enough to judge the authenticity of news based on single-modality content alone.

2.3. Multimodal Fake News Detection

Multimodal News Detection Task: Researchers utilized deep neural networks to analyze information from various modalities for news detection. They proposed the Bert-Vgg19 model, which leverages the advanced Bert-Vgg19 pre-trained model to extract both image and text features [31]. An end-to-end online rumor detection framework was introduced, based on a recurrent neural network with attention mechanisms [32]. In this framework, visual features are combined with text features and social context features, collectively employed to train a LSTM network. Additionally, a fine-grained fusion model was introduced, using scaled attention to integrate textual words and images [22]. The model captures representative information of each modality at different levels and achieves the fusion of modalities at the same level and across different levels through mixers, thus establishing strong connections between modalities [33]. A similarity measurement module assesses

text and image features [17]. Researchers obtained features related to event relationships based on news event features that demonstrate high generalization performance [34]. They proposed a BERT-based TTEC detection model that uses contrastive learning to leverage historical events for learning more effective multi-modal features for news detection [35]. In order to minimize the disparity between modalities and optimize the utilization of multi-modal information, researchers introduced a mechanism for cross-modal consistency learning [15,16,36–38]. Researchers extracted image pattern information, image semantic information, and textual information from the collected news [36]. These inputs are then processed by separate coarse classifiers, which use different perspectives to determine the authenticity of the news articles. This approach utilizes a task-based attention mechanism to assign weights to different modalities, with cross-modal fuzziness acting as a measure of the differences between them [16]. When the cross-modal fuzziness is weak, it employs single-modal information for detecting fake news. In instances of strong cross-modal fuzziness, it integrates textual and visual features for fake news detection using the outer product matrix of text and image. This approach can be regarded as employing a mechanism based on cross-modal fuzziness to weigh modal information. Researchers employed knowledge distillation to transfer insights from one single-modal feature extraction model to another, enhancing the correlation between modalities [15]. They selectively performed feature extraction across modalities [39,40]. A model was developed that extracts information relevant to the target modality from another source modality, while preserving the unique characteristics of the target modality [40]. Researchers utilized entity information to retrieve and extract data from both images and text, deeply exploring the semantic connections between textual and visual information [39]. While previous methods performed well in the task of detecting fake news, there is still room for improvement in achieving genuine cross-modal semantic interaction. Obtaining effective multi-modal features poses challenges in the face of complex news detection scenarios. Additionally, many news detection methods have not fully considered the unique attributes of fake news, impacting the model's generalization capability. Table 1 summarizes the distinguishing features utilized by the cited methods.

Table 1. Features used in news detection by various methods.

Machine Learning-Based Detection		Features of Utilization		
Method	Particular Words	User Language Characteristics	Word Count of the Article	News Comment Count
Naive Bayes model [7]	✓			
Vector machine [8]		✓		
Logistic regression model [20]		✓	✓	
Multiple regression model [21]		✓		✓
Single-modal detection		Features of utilization		
method	image features		text features	
[10,11,24–30]				✓
[12,13]	✓			
Multimodal fake news detection		methods for combining image and text features		
method	feature fusion		minimize modality differences	
[22,31–35]	✓			
[15,16,36–40]				✓

In summary, despite significant progress in multi-modal fake news detection models, they have not completely addressed the challenges of “rumor feature challenges” and “cross-modal feature extraction challenges”. To tackle these issues, we introduce a new network model (Vae-Clip) that collaboratively completes five modules. It not only captures potential semantic information between images and text but also delves into mining the rumor feature attributes associated with fake news. The model accurately identifies fake

news across multiple complex datasets and demonstrates robust performance in various challenging scenarios.

3. Methodology

To achieve more comprehensive fake news detection, this article proposes the Vae-Clip fake news detection model (see Figure 2), which aims to learn rumor feature representation and content feature representation for fake news detection. The model consists of five modules, including a cross-modal semantic feature extraction module (1), a rumor feature extraction module (2), a similarity measurement module (3), a feature fusion module (4), and a news detection module (5). Since online news contains different forms of information, such as text mixed with images, the model first embeds multi-modal information and then passes the obtained text and image information into the cross-modal semantic feature extraction module to learn semantic representation and calculate the semantic features of the text and image using the similarity measurement module. Additionally, the model uses the rumor feature extraction module to extract rumor features from the embedded text information. In the multi-modal feature fusion module, the semantic features and rumor features are fused. Finally, the obtained final features are used for authenticity detection in the news detection module.

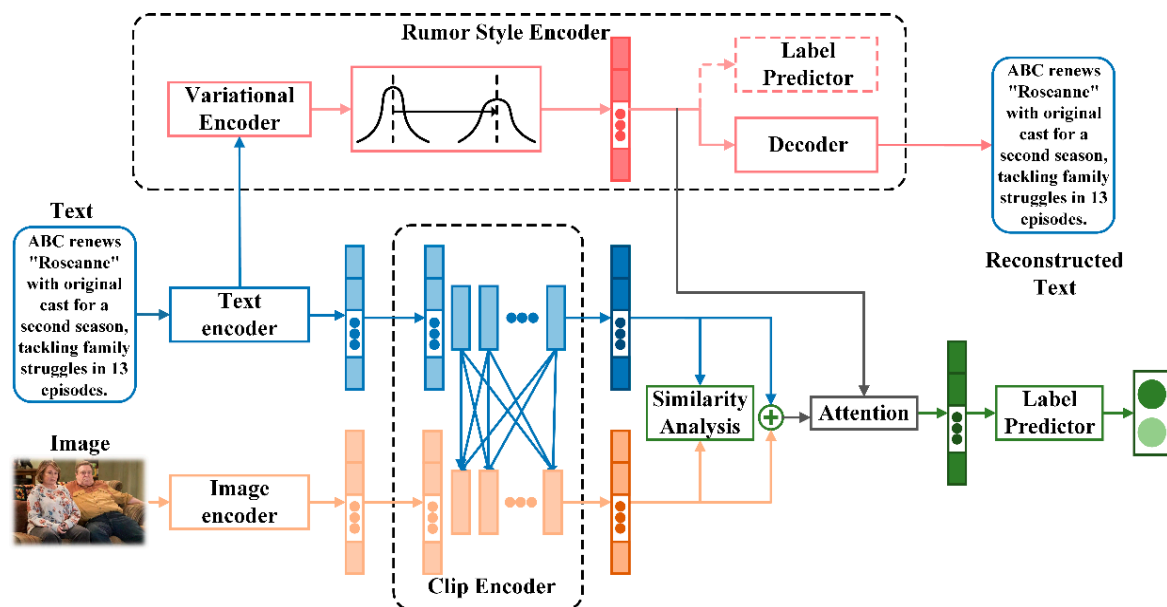


Figure 2. The structure of the Vae-Clip model.

3.1. Cross-Modal Feature Extraction Module

The text encoder E_x^T is a transformer [41] that adds [SOS] and [EOS] markers to text sequences and treats the highest layer activation of the transformer at the [EOS] marker as the text feature representation. After layer normalization, this representation is linearly projected into a multi-modal embedding space. Additionally, the encoder employs a masked self-attention mechanism to preserve its ability to utilize pre-trained language models for initialization.

The visual encoder E_x^I utilized is the ViT (Vision Transformer) model [42]. It is a visual image-encoding model based on attention mechanisms. Unlike traditional convolutional neural networks, the ViT model uses a transformer architecture to extract feature representations of images. The model splits the input image into multiple blocks, reshapes each image block into a one-dimensional vector, and then passes these vectors to the transformer to learn feature representations through a self-attention mechanism.

The image content X_{img} and original the text content X_{text} are respectively encoded by the text encoder E_x^T and visual encoder E_x^I to obtain embedded text feature representation

f_{text} and image feature representation f_{img} . These two features contain all the semantic information of the image and text, but they have weak correlations, and a semantic gap exists between them. Therefore, we use the Clip pre-training model [43] to embed the two features f_{text} and f_{img} into the same embedding space. As shown in Figure 3, cosine similarity is used to measure the similarity between images and text, making similar images and text closer in this space. In the embedding space, using f_{text} as a supervisory signal, a contrastive search is performed on the image-text features to extract f_{img} and f_{text} with semantic relevance. The text feature representation f_{Clip-T} and visual feature f_{Clip-I} representation are then concatenated to form a multi-modal feature representation, as shown in Equation (1).

$$f_m = f_{Clip-T} \oplus f_{Clip-I} \quad (1)$$

where f_m is the multi-modal feature representation, which is the output of the multi-modal feature extractor. Within the scope of this study, the multi-modal feature extractor is represented as $G_f(M; \theta_f)$, where M is the input to the multi-modal feature extractor, and θ_f represents the parameters to be learned.

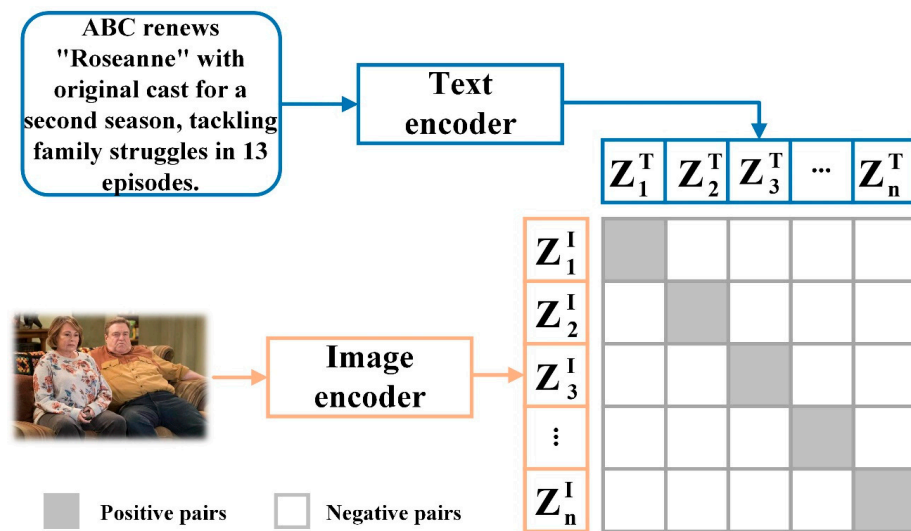


Figure 3. Cross-modal semantic feature extraction module framework.

Previous work has focused on extracting coarse and fine features within a single modality. The introduction of the Clip model connects the feature extraction of different modalities, allowing for the detection of more subtle semantic clues and providing a more detailed analysis of news semantics. This addresses the issue of semantic interaction across modalities in the field of news detection.

3.2. Similarity Measure Module

After obtaining multi-modal features f_{Clip-T} and f_{Clip-I} through the cross-modal feature extraction module, we establish a similarity measurement module that incorporates the measured values of the similarity between image and text features into the loss function of the fake news detection task. The objective is to guide the model in learning the intricate cross-modal correlations between image and text modalities through the optimization process. This initiative aims to stimulate the presentation of more closely aligned similarity information in the shared space of image-text features, providing the model with a more comprehensive and profound understanding and expression of cross-modal semantics [44]. The following text provides a detailed description of the similarity measurement module.

In this module, the distance between the visual and textual features is detected based on the cosine similarity theorem, as shown in Equation (2).

$$s = \frac{f_{Clip-T} \cdot f_{Clip-I}}{\|f_{Clip-T}\| \times \|f_{Clip-I}\|} \quad (2)$$

where f_{Clip-T} and f_{Clip-I} are the semantic features of text and image, s is the calculated similarity score, and its value ranges from -1 to 1 ; a higher value indicates that the textual and visual features are closer in the shared space, suggesting more similarity in information. After obtaining the similarity score, a softmax layer is added to the module to characterize the strength of the similarity and map it to $[0, 1]$, as shown in Equation (3).

$$p^s = \text{sigmoid}(s) \quad (3)$$

where sigmoid is the core function of the softmax layer, reflecting the strength of the similarity. We used the similarity of image-text features as one of the criteria for news detection, incorporating it into the loss function of the model to enhance the similarity of image-text features. Equations (4) and (5) are used to characterize the relationship of the loss function.

$$\mathcal{L}_s(\theta_f, \theta_d) = -\mathbb{E}_{(m,y) \sim (M,Y_d)} [(1-y) \log(1-p^s) + y \log p^s] \quad (4)$$

$$(\hat{\theta}_f, \hat{\theta}_d) = \underset{\theta_f, \theta_d}{\operatorname{argmin}} \mathcal{L}_s(\theta_f, \theta_d) \quad (5)$$

where θ_f is the parameter of the multi-modal feature extractor, Y_d is the label indicating the authenticity of the news, and θ_d is all the parameters of the model. We seek to minimize the detection loss function $\mathcal{L}_s(\theta_f, \theta_d)$ by finding the optimal parameters $\hat{\theta}_f$ and $\hat{\theta}_d$.

3.3. Rumor Feature Extraction Module

In this paper, the multi-modal feature extraction is supervised by text information, so it is necessary to analyze the rumor style of the text. This module is designed to explore the characteristics of fake news and extract its style features. Its network is represented by the part in Figure 2 and Table 2. A variational autoencoder (VAE [45]) is used as the model for extracting text rumor style features. This is because the model can modify the feature distribution in the latent space without changing the semantic content, and obtain the feature distribution for fake news attributes, that is, the rumor style feature f_{style} . In the entire model framework, a multi-layer perceptron (MLP) is used to extract rumor information from the embedded features [46], as shown in Equation (6).

$$[\mu_s, \log \sigma_s^2] = E_x^s(f_{text}, \theta_{E_x^s}) = \text{MLP}_{style}(f_{text}) \quad (6)$$

where μ and σ are the mean and standard deviation of the rumor information distribution, E_x^s is the rumor information encoder, and $\theta_{E_x^s}$ is the parameter of the rumor information encoder. Then, the latent distribution f_{style} is extracted from the style latent variable x_s [46], as shown in Equation (7).

$$x_s \sim \mathcal{N}(\mu_s, \sigma_s^2 I) \quad (7)$$

To ensure that the semantic information of the text is not altered when extracting rumor information, this article reconstructs the text using a decoder. Cross-entropy loss is used for prediction, and the difference between two probability distributions is measured using KL divergence. Minimizing the KL divergence here means optimizing the parameters

of the probability distribution (theta and phi) to make them very similar to the target distribution (a normal distribution) [45], as shown in Equation (8):

$$\mathcal{L}_{rec_t} = -\mathbb{E}_{qE(x_s|x)}[\log p(x|x_z)] + \lambda_{kl} KL(qE(x_s|x) || p(x_s)) \quad (8)$$

where λ_{kl} is the parameter that balances the reconstruction loss and the KL term, $p(x_s)$ is the prior distribution with standard normal distribution $\mathcal{N}(0, I)$, $qE(x_s|x)$ is the distribution of $\mathcal{N}(\mu_s, \sigma_s^2 I)$, and $\mathcal{L}_{rec_t}$ is the reconstruction loss of the feature.

Table 2. Network structure of the rumor feature extraction module.

Module	Structure
Encoder	A Conv1d, 1 inchannel, 64 outchannels, 3 kernelsize, 2 stride
	A Conv1d, 64 inchannels, 128 outchannels, 3 kernelsize, 2 stride
	A Conv1d, 128 inchannels, 256 outchannels, 3 kernelsize, 2 stride
	A Conv1d, 256 inchannels, 512 outchannels, 3 kernelsize, 2 stride
Decoder	Four Normalization layer
	A Conv1dTranspose1d, 512 inchannels, 256 outchannels, 3 kernelsize, 2 stride
	A Conv1dTranspose1d, 256 inchannels, 128 outchannels, 3 kernelsize, 2 stride
	A Conv1dTranspose1d, 128 inchannels, 64 outchannels, 3 kernelsize, 2 stride
	A Conv1dTranspose1d, 64 inchannels, 1 outchannel, 3 kernelsize, 2 stride
	Four Normalization layer

In order to ensure that the extracted rumor style features are attributes of fake news, a label predictor is set up in the rumor style feature extraction module to make a true or false judgment on the news. The predictor is formulated as Equation (9):

$$y_x^s = P_x^s(\mu_s; \theta_{P_x^s}) \quad (9)$$

$\theta_{P_x^s}$ is the parameter of the label predictor, and y_x^s is the output. The loss function of the predictor is defined as Equation (10):

$$\mathcal{L}_{X_s}(\theta_{E_x^s}, \theta_{P_x^s}) = -\mathbb{E}_{(p,y) \sim (P,Y_d)}[y \log(y_x^s) + (1-y) \log(1-y_x^s)] \quad (10)$$

where $\mathcal{L}_{X_s}$ is the label prediction loss. This trains the rumor feature extractor by combining the label prediction loss $\mathcal{L}_{X_s}$ and the feature reconstruction loss $\mathcal{L}_{rec_t}$, which is represented by Equations (11) and (12):

$$\mathcal{L}_{E_x^s} = \mathcal{L}_{rec_t} + \mathcal{L}_{X_s} \quad (11)$$

$$(\hat{\theta}_{E_x^s}, \hat{\theta}_{P_x^s}) = \arg \min_{\theta_{E_x^s}, \theta_{P_x^s}} (\mathcal{L}_{E_x^s}, \mathcal{L}_{P_x^s}) \quad (12)$$

where $\mathcal{L}_{E_x^s}$ represents the total loss of the rumor feature extraction module. In the course of this investigation, the optimal parameter $\theta_{E_x^s}$ for the rumor feature extractor is derived through the minimization of the total loss of the rumor feature extraction module.

3.4. Feature Fusion Module

Through the cross-modal feature extraction module and the rumor feature extraction module, we obtain the multi-modal feature f_m and the rumor feature f_{style} . In the feature fusion module, the attention mechanism is used to assign weights to these two features to highlight the most valuable information. Then, the features are fused according to the weights to obtain the final feature for news discrimination. The specific formulas are Equations (13)–(15):

$$u^c = U^T(W^c f_i^c + b^c) \quad (13)$$

$$\alpha_i = \frac{\exp(u^c)}{\sum_i \exp(u^c)} \quad (14)$$

$$R_f = \sum_i \alpha_i f_i^c \quad (15)$$

where f_c^i is the set of features f_m and f_{style} . u^c is the output of a fully connected layer, whose input is set f_c^i . α_i is the weight calculated for each feature, and R_f is the final feature obtained after fusion.

3.5. News Detection Module

The news detection module, as shown in Figure 2, interfaces with the feature fusion module and uses the final feature R_f for news discrimination. This feature is a high-dimensional vector. To map the final feature to the true/false label of the news, the news detection module first uses a fully connected layer, an activation function layer, and Dropout regularization to process this high-dimensional vector and reduce it to a two-dimensional vector. The softmax function is used to achieve mapping between features and labels, predicting the probability of true/false news and achieving accurate binary classification tasks. The specific formula is as follows (Equation (16)):

$$p^c = \text{softmax}(W_p \cdot R_f + B_p) \quad (16)$$

where p^c is the predicted probability of the news being true or false. Y_d is the actual label indicating whether the news is true or false, and $\mathcal{L}_c$ is the cross-entropy loss between p^c and Y_d . The formula is shown as follows (Equation (17)):

$$\mathcal{L}_c(\theta_t, \theta_f, \theta_d) = -\mathbb{E}_{(m,y) \sim (M,Y_d)} [(1-y) \log(1-p^c) + y \log p^c] \quad (17)$$

where $\mathcal{L}_c$ represents the predicted tag loss, Y_d represents the tag set of the news, and θ_t represents the parameters of the news detector.

In the news detection module, the similarity score between the image and text features is also used as one of the judgment criteria. Therefore, the similarity measurement score is included as a part of the total loss for model training, resulting in the following formula (Equation (18)):

$$\mathcal{L}_{final} = \mathcal{L}_s + \mathcal{L}_c \quad (18)$$

where $\mathcal{L}_s$ is the score of similarity measurement and $\mathcal{L}_c$ is the loss of label prediction.

In this article, the optimal parameters of the overall model are obtained by minimizing the total loss, which is formulated as Equation (19).

$$(\hat{\theta}_t, \hat{\theta}_f, \hat{\theta}_d) = \arg \min_{\theta_t, \theta_f, \theta_d} \mathcal{L}(\theta_t, \theta_f, \theta_d) \quad (19)$$

The main training method used in this article is to minimize the cross-entropy loss for discriminating news articles and obtain the optimal model parameters. Algorithm 1 describes the training algorithm of the model. The first step is to learn the optimal parameters of the rumor information encoder to extract rumor features. Then, during the news detection phase, the process is repeated to obtain the optimal model parameters by minimizing $\mathcal{L}_s + \mathcal{L}_c$.

Algorithm 1: Model Training

Input: Data X_{text} and X_{img}
Output: Learned parameters $\theta_{E_x^s}, \theta_{P_x^s}, \theta_f, \theta_d, \theta_t$

```

1  // Rumor feature extraction;
2  for each batch sampled from  $X_{text}$  do
3      a) compute loss  $\mathcal{L}_{x_s}(\theta_{E_x^s}, \theta_{P_x^s})$ ;
4      b) Take a gradient step for  $\mathcal{L}_{x_s}$ ;
5  end
6  repeat
7      // Fake news detection;
8      for each batch sampled from  $\{X_{text}, X_{img}\}$  do
9          (a) compute loss  $\mathcal{L}_s(\theta_f, \theta_d)$ ;
10         (b) Compute loss  $\mathcal{L}_c(\theta_t, \theta_f, \theta_d)$ ;
11         (c) Take a gradient step for  $\mathcal{L}_s + \mathcal{L}_c$ ;
12     end
13 until convergence;
```

4. Experiment

In this section, first an introduction and description of the datasets used are provided, followed by experimental results of the Vae-Clip model on four datasets. Additionally, a comparative experiment was conducted to compare the Vae-Clip model with the listed baseline methods. Finally, an ablation experiment was set up to analyze the performance of each module. The comparative experiment demonstrated that the Vae-Clip outperforms other models in news detection tasks, which is Problem 1. The ablation experiment proved that all five modules in the model contribute to its effectiveness in news detection, which is Problem 2.

4.1. Experimental Settings

In this section, the article will provide a detailed description of the dataset used in the experiment, as well as the model configuration of the Vae-Clip model on the dataset.

Below are the implementation details of the Vae-Clip model. Firstly, text information was embedded by a text encoder to obtain a 512-dimensional vector. A fully connected layer was added in the rumor feature extraction module to output a 16-dimensional vector. After being embedded through a visual encoder, the original image information becomes a vector, which was then fed into the cross-modal feature extraction module together with the text embedding feature. Through feature extraction using the Clip pre-training model, 512-dimensional vectors were obtained for both the image and text features. In the similarity measurement module, a 512-dimensional space was established for sharing weights between image and text features. Furthermore, to prevent overfitting during training, ReLU layers and Dropout layers with a forgetting rate of 0.3 were added to the model. In order to obtain optimal model parameters, the article used the Adam optimizer for optimization. Detailed experimental parameters are provided in Table 3, and the detailed configuration of the Variational Autoencoder is given in Table 2.

The dataset was divided into training, validation, and testing sets in a ratio of 7:1:2. The data was shuffled and grabbed in batches of 32. The model learning rate was set to 10^{-3} , and the model was trained for 100 rounds. The optimal model parameters from the testing set were used for model testing.

Table 3. Model experiment parameters.

Cross-modal Feature Extraction Module				
Text input features	Image input features	Similarity measurement space	Semantic feature output dimension	ReLU layer dropout rate
512	512	512	256	0.3
Rumor feature extraction module				
Text input features	Number of convolutional layers	Number of deconvolutional layers	Rumor feature output	Number of normalization layers
512	4	4	16	4
Model training parameters				
Batchsize	Epoch	Learning rate	Optimizer	Training loss function
64	100	0.001	Adam	Cross-entropy loss

4.1.1. Datasets

To comprehensively evaluate the effectiveness of the model, we selected four publicly recognized datasets for fake news detection tasks, including two Chinese datasets and two English datasets, showed in Table 4. Specifically, these datasets mainly came from the Twitter and Weibo platforms, which are the most popular social media platforms for users abroad and in China, respectively, and the news on these platforms is widely viewed and disseminated. The remaining datasets contained real data from the GossipCop and PolitiFact websites. Therefore, we believe these four datasets can fully represent social news from different linguistic and cultural backgrounds.

Table 4. Statistical data of datasets.

	Weibo1	Weibo2	Twitter	FakeNewsNet
fake_news	3737	3055	7221	5705
real_news	3986	2609	5882	5330
account	7723	5664	13,103	11,035

Weibo1: The dataset proposed consists of 7723 news articles, including 3737 fake news and 3986 real news. These news articles were all combinations of images and text, and the real news came from authoritative news sources in China, such as the Xinhua News Agency. The fake news was collected from May 2016 to January 2017 and verified by the Weibo official rumor-refuting system [3].

Weibo2: The dataset was processed and news articles with low-quality images and those that could not have images downloaded were removed, resulting in a dataset with a total of 5664 news articles. Among them, there were 2609 real news and 3055 fake news. For fake news, news segments judged by the Weibo Community Management Center as misinformation were selected, while for real news, news segments collected during the same period as fake news were selected, and all these real news segments were verified by Weibo's suspicious news segment verification platform [47].

Twitter: The dataset is a dataset from Twitter, used to verify the use of multimedia tasks and to detect false multimedia content on social media. This paper selected news articles combining English text with images, with a total of 13,103 news articles, including 5882 real news and 7221 fake news [48].

FakeNewsNet: The dataset is real data from the GossipCop and PolitiFact websites, which contain news content with professional journalist and expert annotation labels, as well as social context information. After screening out low-quality and invalid images, this paper obtained a total of 11,035 news articles with text and images, including 5330 real news and 5705 fake news [49,50].

4.1.2. Evaluation Metrics

In the experiments of this chapter, we used accuracy, AUC curves, precision, recall, and F1 scores as performance evaluation metrics for news detection, with Formulas (18) to (21). In this section, we will demonstrate the effectiveness of the Vae-Clip model on four datasets.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (20)$$

$$Precision = \frac{TP}{TP + FP} \quad (21)$$

$$Recall = \frac{TP}{TP + FN} \quad (22)$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (23)$$

TP represents true positive, which is when the model predicts a label of 1 for real news. TN represents true negative, which is when the model predicts a label of 0 for fake news. FP represents false positive, which is when the model predicts a label of 0 for real news. FN represents false negative, which is when the model predicts a label of 1 for fake news.

4.2. Experimental Results

Table 5 presents the experimental results of the Vae-Clip model on four datasets, which are divided into evaluation metrics for true news and false news in the table. The table provides a more detailed evaluation of the model's performance in classifying true and false news. The experimental results show that the accuracy of Vae-Clip on all four datasets is above 87%, with the highest being 96.3%. This validates the effectiveness of the model in news detection and shows stable results on large and complex datasets.

Table 5. Experimental results of Vae-Clip on four datasets.

Dataset	Accuracy	AUC	Fake News			Real News		
			Precision	Recall	F1	Precision	Recall	F1
Weibo1	0.963	0.990	0.941	0.986	0.963	0.986	0.942	0.964
Weibo2	0.938	0.972	0.939	0.946	0.942	0.936	0.928	0.932
Twitter	0.877	0.877	0.869	0.828	0.973	0.895	0.960	0.764
FakeNewsNet	0.877	0.879	0.949	0.776	0.854	0.833	0.964	0.894

We focused on the binary classification task of detecting the authenticity of news and therefore used the AUC (area under the ROC curve) to evaluate the classification performance of the model. The AUC value ranges from 0.5 to 1, with larger values indicating better performance, stronger ability to distinguish between positive and negative samples, and better classification performance. Figure 4 shows the ROC curves of the model on the four datasets, which is a curve with the false positive rate (FPR) as the x-axis and the true positive rate (TPR) as the y-axis. The AUC value is reflected in the area under the ROC curve. From Figure 4, it can be seen that the Vae-Clip model has a high score for AUC, indicating that the model is effective for the binary classification task.

With the key challenge being the extraction of features that can correctly identify news, we aimed to achieve accurate news authenticity detection. Therefore, we used the t-SNE technique to visualize the features extracted by the Vae-Clip model [51]. As shown in Figure 5, Vae-Clip clearly distinguishes between true and false news on the four datasets and achieves a good clustering effect on the true and false labels. This also proves that the features extracted by the model can correctly classify true and false news. Therefore, the model proposed in this paper is capable of achieving accurate and high-precision detection

of true and false news, even in the face of challenges such as rumor feature extraction and cross-modal semantic feature extraction.

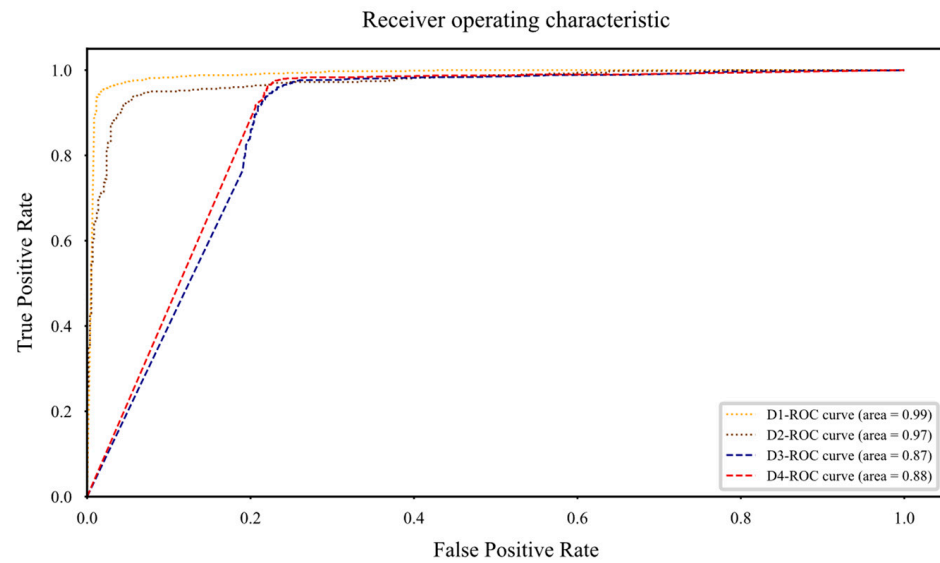


Figure 4. ROC curves of the Vae-Clip model on four datasets.

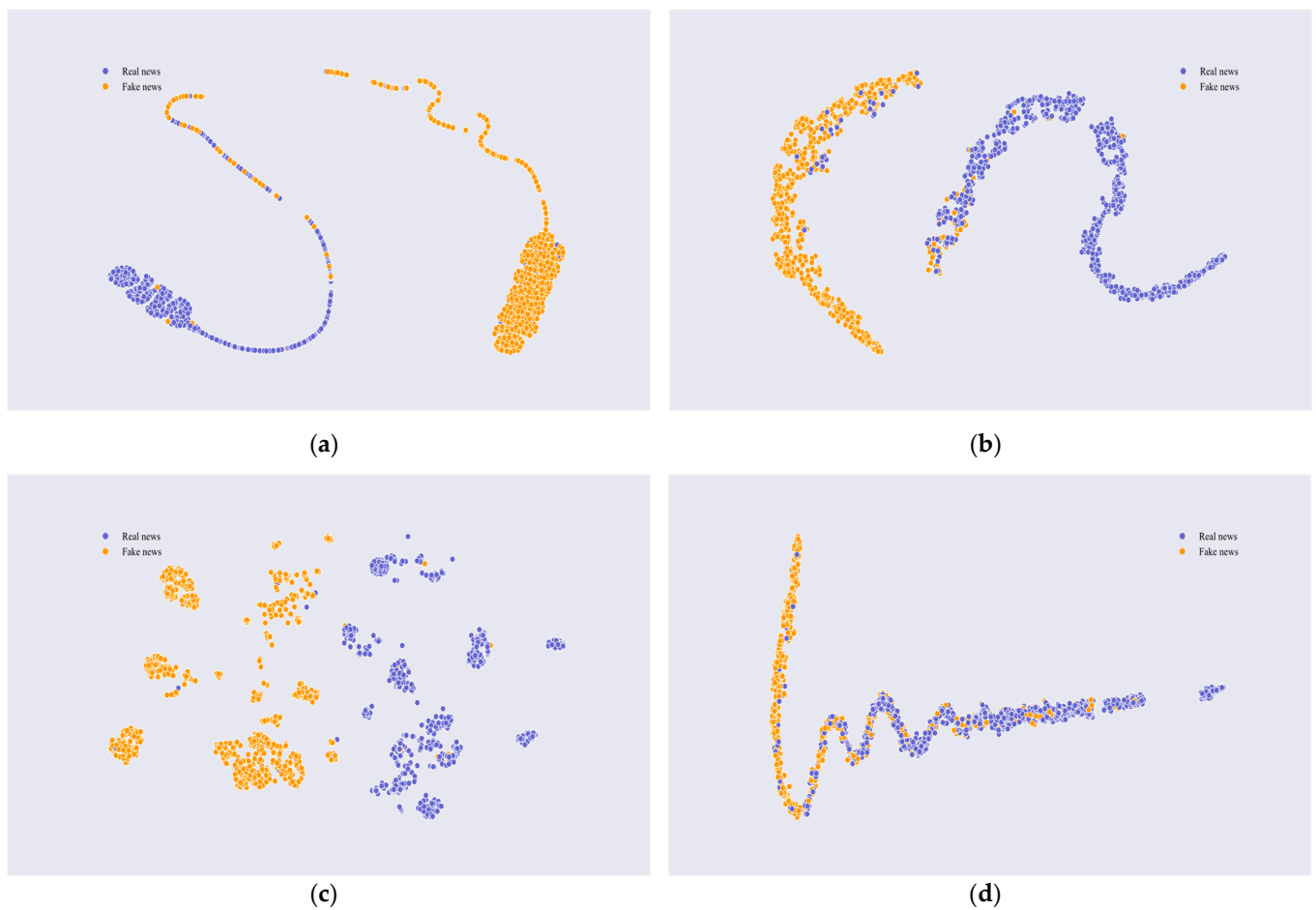


Figure 5. Visualization of features extracted by the Vae-Clip model on four datasets using t-SNE. (a) Feature representation on the Weibo1 dataset, (b) feature representation on the Weibo2 dataset, (c) feature representation on the Twitter dataset, and (d) feature representation on the FakeNewsNet dataset.

4.3. Comparisons

In this section, to assess the advancements of the Vae-Clip model, we conducted a comparative evaluation against other methods across four task scenarios. The superior performance of our model illustrates its effectiveness in overcoming the challenges of cross-modal and rumor feature extraction, leading to more accurate fake news detection.

4.3.1. Baselines

1. BMR: [36] Utilizing image pattern information, image semantic information, and textual information, inputting them into individual coarse classifiers and leveraging diverse perspectives for the detection of authenticity in news.

2. CAFE: [16] Establishing a model that employs a cross-modal fuzziness mechanism to assign weights to image-text information, thus detecting fake news based on these fused features.

3. CMC: [15] Using distillation to transfer knowledge from one single-modal feature extraction model to another, thereby improving the correlation between different modalities.

4. FND-CLIP: [14] A multi-modal fake news detection model based on CLIP is proposed, obtaining fused features through the standardization and weighted similarity of text and image features.

5. EANN (event adversarial neural network): [52] Aims to improve the detection of multi-modal news by using an event domain discriminator. Multiple granularity convolutional layers are used for feature extraction in text feature extraction, which is then fused with image features. For the sake of comparison in the experiment, we removed the event domain discriminator of this model.

6. att-RNN: [15] This model is a multi-modal fake news detection framework that fuses text, visual, and social contextual features using an attention mechanism.

7. SpotFake: This model [53] extracts image and text features using Bert and Vgg-19 and concatenates the two features to predict news.

8. MKN (multi-modal knowledge-aware event memory network): [54] This model is an event-level multi-modal fake news detection framework that uses visual information and external knowledge to assist in fake news detection tasks. To ensure fairness in the experiment, we removed external knowledge components and event memory networks.

4.3.2. Comparison Result

This section aims to compare the performance of Vae-Clip and baseline models on various metrics using four datasets. The performance evaluation metrics used in this study include detection accuracy, AUC, precision, recall, and the F1 score.

Table 6 presents the experimental comparison results of Vae-Clip and the baseline models on the Chinese dataset. As shown in Table 6, the proposed Vae-Clip model outperforms the baseline methods in all five evaluation metrics, including accuracy, AUC, precision, recall, and F1 score, which answers the first research question. Specifically, the accuracy of the Vae-Clip model reached 96.3% on the Weibo1 dataset, which is 3.1% higher than the highest accuracy achieved by the baseline method. On the Weibo2 dataset, the Vae-Clip model achieved a 2.3% improvement over the highest accuracy of the baseline method. In the Chinese dataset, the text content contains rich semantic information and various exaggerated writing techniques to attract readers. The features extracted by the Vae-Clip model contained the inherent characteristics of “rumor features” in fake news, which led to better performance than other baseline models.

Table 7 presents the experimental comparison results of the Vae-Clip model and baseline models on the English dataset. From Table 7, it can be seen that the Vae-Clip model proposed in this article outperforms the baseline method in all five evaluation metrics: accuracy, AUC, precision, recall, and F1 score, thus answering the first question. The accuracy of the Vae-Clip model on the Twitter dataset reached 87.7%, which is 1.9% higher than the highest accuracy achieved by the baseline method. On the FakeNewsNet dataset, the Vae-Clip model improved by 0.9% compared to the highest accuracy of the

baseline method. Overall, the performance on the Twitter and FakeNewsNet datasets are worse than that on the Weibo1 and Weibo2 datasets. The reason for this is that many texts in the Twitter and FakeNewsNet datasets were too simple and lacked diversity due to their irregular writing, which reduces the effectiveness of the model. Weibo1 and Weibo2 datasets contained richer textual information, allowing the model to extract more semantic information close to news, resulting in excellent performance on the Weibo1 and Weibo2 datasets. The BMR, CMC, and FND-CLIP models were similarly affected by this influence. While pursuing cross-modal consistency between textual and image features, they depend on the collaborative effect of both modalities. When the semantic features in textual content are not rich enough, obtaining universally expressive features becomes challenging, thus hindering the achievement of better detection results. The Vae-Clip model relies on both textual and rumor features. Although its performance is affected, it achieves better detection results compared to other baseline methods.

Table 6. Experimental results of different methods on the Weibo1 and Weibo2 datasets, with the performance of our model highlighted in bold.

Dataset	Method	Accuracy	AUC	Fake News			Real News		
				Precision	Recall	F1	Precision	Recall	F1
Weibo1	BMR	0.904	0.934	0.932	0.863	0.897	0.880	0.942	0.910
	CAFE	0.824	0.907	0.819	0.817	0.818	0.829	0.831	0.830
	CMC	0.892	0.956	0.917	0.853	0.884	0.871	0.927	0.898
	FND-CLIP	0.887	0.956	0.861	0.914	0.887	0.914	0.861	0.887
	EANN	0.800	0.891	0.815	0.773	0.793	0.792	0.830	0.811
	att-RNN	0.932	0.954	0.940	0.920	0.930	0.937	0.940	0.938
	SpotFake	0.922	0.953	0.937	0.935	0.936	0.932	0.918	0.925
	MKN	0.834	0.919	0.805	0.867	0.835	0.866	0.803	0.833
	ours	0.963	0.990	0.941	0.986	0.963	0.986	0.942	0.964
Weibo2	BMR	0.915	0.942	0.928	0.915	0.921	0.902	0.916	0.909
	CAFE	0.805	0.869	0.815	0.826	0.821	0.792	0.780	0.787
	CMC	0.885	0.902	0.871	0.923	0.897	0.903	0.840	0.871
	FND-CLIP	0.907	0.924	0.898	0.933	0.916	0.918	0.876	0.897
	EANN	0.846	0.930	0.823	0.910	0.864	0.879	0.770	0.821
	att-RNN	0.881	0.941	0.878	0.906	0.892	0.886	0.852	0.869
	SpotFake	0.876	0.928	0.901	0.865	0.883	0.849	0.888	0.868
	MKN	0.856	0.939	0.883	0.845	0.863	0.827	0.868	0.847
	ours	0.938	0.972	0.939	0.946	0.942	0.936	0.928	0.932

In general, models BMR, CMC, FND-CLIP, and SpotFake exhibit stable performance across four experimental scenarios. SpotFake, leveraging powerful pre-trained models Vgg-19 and Bert, extracts features that closely capture the semantic information of news, emphasizing the significance of semantic information in news. In contrast to other models, BMR, CMC, and FND-CLIP aim to enhance inter-modal correlations through task-based attention mechanisms, obtaining cross-modal expressive features beneficial for news detection. Experimental results demonstrate that enhancing inter-modal correlations can improve the model's performance in detecting news in multi-modal scenarios.

By comparing the experimental results, this paper answers question one, which essentially asks whether the proposed model can address the “rumor feature extraction challenge” and “cross-modal semantic information extraction challenge.” The Vae-Clip model is capable of extracting “rumor features” and cross-modal semantic information, making it suitable for large datasets and achieving better detection results than the other methods.

Table 7. Experimental results of different methods on the Twitter and FakeNewsNet datasets, with the performance of the proposed model in bold.

Dataset	Method	Accuracy	AUC	Fake News			Real News		
				Precision	Recall	F1	Precision	Recall	F1
Twitter	BMR	0.842	0.835	0.786	0.947	0.859	0.930	0.733	0.820
	CAFE	0.807	0.779	0.831	0.788	0.809	0.785	0.828	0.806
	CMC	0.834	0.858	0.778	0.942	0.852	0.923	0.722	0.810
	FND-CLIP	0.858	0.827	0.807	0.948	0.872	0.934	0.765	0.841
	EANN	0.741	0.729	0.854	0.619	0.718	0.670	0.880	0.760
	att-RNN	0.841	0.831	0.831	0.768	0.798	0.779	0.773	0.776
	SpotFake	0.683	0.792	0.793	0.660	0.720	0.635	0.800	0.708
	MKN	0.877	0.869	0.828	0.973	0.895	0.960	0.764	0.851
	ours	0.868	0.876	0.940	0.765	0.843	0.825	0.957	0.886
FakeNewsNet	BMR	0.785	0.820	0.883	0.618	0.727	0.738	0.929	0.822
	CAFE	0.860	0.835	0.931	0.801	0.861	0.799	0.930	0.856
	CMC	0.855	0.873	0.937	0.738	0.826	0.809	0.956	0.877
	FND-CLIP	0.809	0.828	0.880	0.681	0.768	0.769	0.920	0.838
	EANN	0.784	0.819	0.879	0.619	0.726	0.738	0.927	0.822
	att-RNN	0.853	0.861	0.942	0.728	0.821	0.803	0.961	0.875
	SpotFake	0.816	0.835	0.885	0.683	0.771	0.771	0.923	0.840
	MKN	0.877	0.879	0.949	0.776	0.854	0.833	0.964	0.894
	ours	0.842	0.835	0.786	0.947	0.859	0.930	0.733	0.820

4.4. Ablation Analysis

In this section, we evaluated the specific contribution and importance of each module within the overall model by assessing various performance metrics across four datasets. We also validated the effectiveness of eliminating cross-modal differences and incorporating rumor features in multi-modal fake news detection. Therefore, we started with the most basic model configuration and gradually added each module, evaluating their individual and combined impacts on the model's performance. This approach allowed us to identify which components were critical for enhancing the accuracy and robustness of fake news detection, thereby providing a comprehensive understanding of the model's structure and functionality.

Effectiveness of Model Combining Image-Text Information: To assess the model's ability to effectively integrate image-text information, we compared the experimental results of Vae-Clip-F with those of Vae-Clip-V and Vae-Clip-T. The results indicate improved performance in detecting multi-modal news, suggesting that the model can effectively leverage image-text information for more accurate fake news detection. This demonstrates that the model can extract and fuse features from different types of data, thereby enhancing overall detection performance. Such fusion provides more comprehensive information, helping to identify subtle differences in fake news.

Effectiveness of Rumor Features in Fake News Detection: To assess the potential contribution of rumor features to fake news detection, we compared the experimental results of Vae-Clip-R with those of Vae-Clip-F and Vae-Clip-S. The results from Table 8 clearly show that incorporating rumor features improves the model's news detection performance compared to the model without such features. Since rumor features represent unique attributes of fake news, their inclusion helps the model achieve more accurate news detection. This indicates that rumor features can serve as important supplementary information, enhancing the model's ability to recognize fake news. These features may include the linguistic characteristics of fake news, aiding the model in better understanding and capturing the essence of fake news.

Table 8. Ablation experiments of the Vae-Clip model.

Dataset	Method	Accuracy	AUC	Fake News			Real News		
				Precision	Recall	F1	Precision	Recall	F1
Weibo1	Vae-Clip-V	0.755	0.832	0.716	0.820	0.764	0.804	0.694	0.745
	Vae-Clip-T	0.833	0.904	0.839	0.810	0.824	0.827	0.855	0.841
	Vae-Clip-F	0.909	0.954	0.919	0.891	0.905	0.901	0.926	0.913
	Vae-Clip-R	0.942	0.982	0.922	0.962	0.941	0.963	0.923	0.943
	Vae-Clip-S	0.933	0.963	0.936	0.925	0.930	0.931	0.941	0.936
	Vae-Clip	0.963	0.990	0.941	0.986	0.963	0.986	0.942	0.964
Weibo2	Vae-Clip-V	0.772	0.846	0.773	0.818	0.795	0.771	0.719	0.744
	Vae-Clip-T	0.896	0.910	0.900	0.894	0.897	0.879	0.898	0.889
	Vae-Clip-F	0.902	0.935	0.915	0.901	0.908	0.886	0.902	0.894
	Vae-Clip-R	0.925	0.947	0.932	0.911	0.922	0.918	0.938	0.928
	Vae-Clip-S	0.915	0.958	0.904	0.942	0.923	0.929	0.882	0.905
	Vae-Clip	0.938	0.972	0.939	0.946	0.942	0.936	0.928	0.932
Twitter	Vae-Clip-V	0.788	0.831	0.763	0.721	0.741	0.726	0.666	0.787
	Vae-Clip-T	0.823	0.844	0.789	0.922	0.849	0.886	0.708	0.787
	Vae-Clip-F	0.831	0.867	0.780	0.952	0.858	0.925	0.690	0.791
	Vae-Clip-R	0.845	0.861	0.798	0.952	0.868	0.928	0.721	0.811
	Vae-Clip-S	0.841	0.850	0.793	0.955	0.866	0.931	0.708	0.804
	Vae-Clip	0.877	0.869	0.828	0.973	0.895	0.960	0.764	0.851
FakeNewsNet	Vae-Clip-V	0.774	0.825	0.886	0.710	0.788	0.779	0.934	0.850
	Vae-Clip-T	0.846	0.849	0.923	0.729	0.814	0.801	0.948	0.868
	Vae-Clip-F	0.856	0.860	0.933	0.742	0.827	0.810	0.954	0.876
	Vae-Clip-R	0.864	0.878	0.947	0.749	0.836	0.816	0.963	0.884
	Vae-Clip-S	0.868	0.869	0.936	0.769	0.844	0.827	0.954	0.886
	Vae-Clip	0.877	0.879	0.949	0.776	0.854	0.833	0.964	0.894

Vae-Clip-V is a model for extracting image information using the cross-modal feature extraction module. Vae-Clip-T is a model for extracting textual information using the cross-modal feature extraction module. Vae-Clip-F is a model that combines the image and textual features extracted by the cross-modal feature extraction module. Vae-Clip-R is a model that combines semantic features and rumor features. Vae-Clip-S is a model that incorporates a similarity measurement metric into the loss function for news detection in the Vae-Clip-F model. Vae-Clip is the complete model that integrates rumor features based on Vae-Clip-S.

Effectiveness of Similarity Measurement in Enhancing Cross-Modal Correlation for Fake News Detection: This study endeavors to improve cross-modal correlation by introducing a similarity measurement module, aiming to bring image-text representations closer in a shared space. To assess the module's effectiveness, we compared the experimental results of model Vae-Clip-S with those of Vae-Clip-F, as well as Vae-Clip and Vae-Clip-R. The experimental outcomes in Table 8 clearly indicate that the model's fake news detection performance has universally improved with the incorporation of the similarity measurement module. This further substantiates that enhancing the correlation between modalities enables the model to achieve more precise news detection. Through this approach, the model can more effectively integrate image-text information, establishing stronger connections between multi-modal data. This not only enhances the model's detection capabilities but also highlights the importance and potential of cross-modal information fusion in fake news detection.

To offer a more comprehensive depiction of the ablation experiments, we have represented the experimental results in Figure 6. It is clear from the graph that as the modules are added, the model's performance steadily improves across various metrics.

The visualization in this paper is based on the best-performing dataset, Weibo1, with similar results observed in other datasets. As shown in Figure 7, the blue dots represent each piece of true news, while the yellow dots represent false news. As the model improves, the clustering of blue and yellow dots becomes more distinct, indicating a clearer separation between true and false news. This demonstrates that as the model is continuously refined,

the extracted features for news verification become more effective in distinguishing between true and false news.

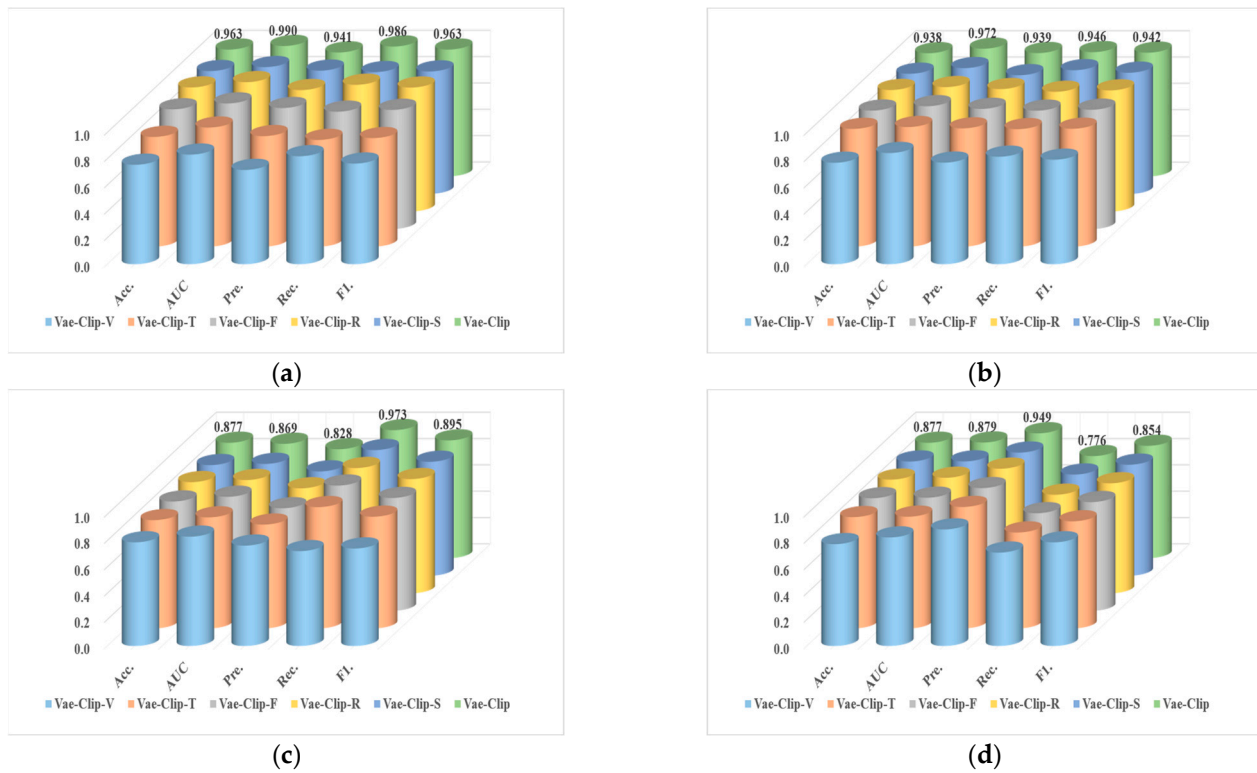


Figure 6. Results of ablation experiments. (a–d) Represent the experimental results of Vae-Clip-V, Vae-Clip-T, Vae-Clip-F, Vae-Clip-R, Vae-Clip-S, and Vae-Clip on datasets Weibo1, Weibo2, Twitter, and FakeNewsNet, respectively.

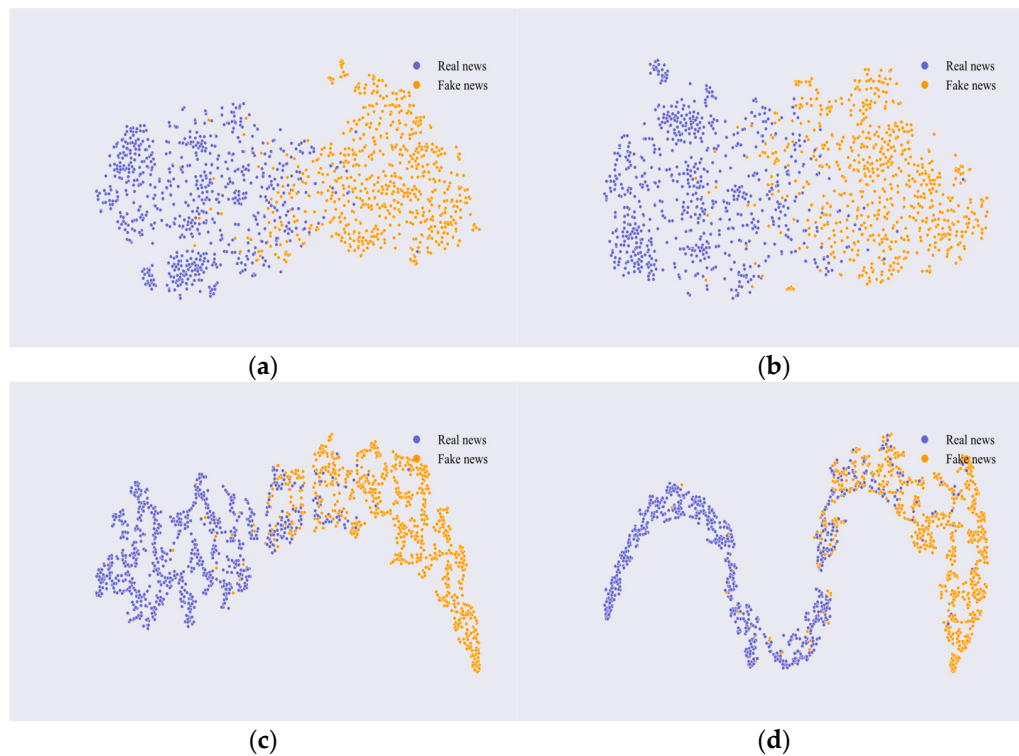


Figure 7. Cont.



Figure 7. Displays the visualization of learned feature representations using t-SNE on the Weibo1 dataset. Panels (a–f) represent the feature representations extracted by Vae-Clip-V, Vae-Clip-T, Vae-Clip-F, Vae-Clip-R, Vae-Clip-S, and Vae-Clip, respectively.

5. Conclusions

Existing methods fall short in integrating cross-modal semantic information and extracting rumor features, leading to subpar performance in rumor detection. To address this issue, this study introduces the Vae-Clip model, which significantly enhances rumor detection accuracy through two core modules: cross-modal feature extraction and rumor feature extraction. In four different experimental settings, this model achieved a peak accuracy of 96.3%. It utilizes textual information as a supervisory signal to extract image-text features and maps these features into a shared feature space, thus achieving more precise cross-modal semantic alignment. Moreover, the model employs a variational autoencoder to extract rumor features, effectively capturing the unique expressions of fake news. These features enable the Vae-Clip model to exhibit superior accuracy and robustness in fake news detection across complex scenarios, particularly in terms of precise image-text alignment and in-depth analysis of fake news. Therefore, the application of the Vae-Clip model will greatly enhance the capability of social media platform review systems to detect fake news, thereby reducing the spread of false information. However, this study has some limitations, including a limited dataset size and insufficient computational resources, which may impact the model's generalization capability and optimization performance. The limited dataset size could lead to diminished performance when the model encounters unseen data. Additionally, the scarcity of computational resources restricts the complexity of model training and parameter tuning, potentially leading to suboptimal optimization. Future research could address these issues by expanding the dataset size, integrating more diverse data sources, and exploring more efficient computational methods to further improve the model's performance and robustness. Based on the high accuracy of the Vae-Clip model, future research can also delve into the dynamic rules of information dissemination in complex networks and develop more precise prediction methods and control schemes.

Author Contributions: Y.Z.: Conceptualization, Methodology, Investigation, Writing—original draft, Data curation. A.P.: Methodology, Supervision, Validation, Resources, Writing—review and editing. G.Y.: Methodology, Conceptualization, Supervision, Data curation, Validation, Resources, Writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the National Natural Science Foundation of China (72074060), the Guizhou Provincial Postgraduate Research Fund [YJSKYJJ [2021]010], and the Department of Education of Guizhou Province, QianJiaoJi, China [2022]043.

Data Availability Statement: All the datasets used in this study are available in the referenced articles.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

- Hölig, S.; Behre, J.; Schulz, W. *Reuters Institute Digital News Report 2022: Ergebnisse für Deutschland*; Verlag ans-Bredow-Institut: Hamburg, Germany, 2022.
- Rastogi, S.; Bansal, D. A review on fake news detection 3T's: Typology, time of detection, taxonomies. *Int. J. Inf. Secur.* **2022**, *22*, 177–212. [[CrossRef](#)] [[PubMed](#)]
- Capuano, N.; Fenza, G.; Loia, V.; Nota, F.D. Content-Based Fake News Detection with Machine and Deep Learning: A Systematic Review. *Neurocomputing* **2023**, *530*, 91–103. [[CrossRef](#)]
- Vosoughi, S.; Roy, D.; Aral, S. The spread of true and false news online. *Science* **2018**, *359*, 1146–1151. [[CrossRef](#)] [[PubMed](#)]
- Zhou, X.; Zafarani, R.; Shu, K.; Liu, H. Fake News. In Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, Melbourne, Australia, 11–15 February 2019; pp. 836–837.
- Granik, M.; Mesyura, V. Fake news detection using naive Bayes classifier. In Proceedings of the 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), Kyiv, Ukraine, 29 May–2 June 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 900–903.
- Khan, J.Y.; Khondaker, M.T.I.; Afroz, S.; Uddin, G.; Iqbal, A. A benchmark study of machine learning models for online fake news detection. *Mach. Learn. Appl.* **2021**, *4*, 100032. [[CrossRef](#)]
- Ahmed, H.; Traore, I.; Saad, S. Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques. In Proceedings of the Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments, Vancouver, BC, Canada, 26–28 October 2017; pp. 127–138.
- Gravanis, G.; Vakali, A.; Diamantaras, K.; Karadaï, P. Behind the cues: A benchmarking study for fake news detection. *Expert Syst. Appl.* **2019**, *128*, 201–213. [[CrossRef](#)]
- Kaliyar, R.K.; Goswami, A.; Narang, P. FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. *Multimed. Tools Appl.* **2021**, *80*, 11765–11788. [[CrossRef](#)] [[PubMed](#)]
- Abedalla, A.; Al-Sadi, A.; Abdullah, M. A Closer Look at Fake News Detection. In Proceedings of the 2019 3rd International Conference on Advances in Artificial Intelligence, Istanbul, Turkey, 26–28 October 2019; pp. 24–28.
- Choudhary, A.; Arora, A. ImageFake: An Ensemble Convolution Models Driven Approach for Image Based Fake News Detection. In Proceedings of the 2021 7th International Conference on Signal Processing and Communication (ICSC), Noida, India, 25–27 November 2021; pp. 182–187.
- 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; pp. 518–527.
- Zhou, Y.; Yang, Y.; Ying, Q.; Qian, Z.; Zhang, X. Multimodal fake news detection via clip-guided learning. In Proceedings of the 2023 IEEE International Conference on Multimedia and Expo (ICME), Brisbane, Australia, 10–14 July 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 2825–2830.
- Wei, Z.; Pan, H.; Qiao, L.; Niu, X.; Dong, P.; Li, D. Cross-Modal Knowledge Distillation in Multi-Modal Fake News Detection. In Proceedings of the ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; pp. 4733–4737.
- 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.
- Xue, J.; Wang, Y.; Tian, Y.; Li, Y.; Shi, L.; Wei, L. Detecting fake news by exploring the consistency of multimodal data. *Inf. Process. Manag.* **2021**, *58*, 102610. [[CrossRef](#)] [[PubMed](#)]
- Zhang, X.; Dadkhah, S.; Weismann, A.G.; Kanaani, M.A.; Ghorbani, A.A. Multimodal Fake News Analysis Based on Image–Text Similarity. *IEEE Trans. Comput. Soc. Syst.* **2023**, *11*, 959–972. [[CrossRef](#)]
- Guo, B.; Ding, Y.; Yao, L.; Liang, Y.; Yu, Z. The Future of False Information Detection on Social Media. *ACM Comput. Surv.* **2020**, *53*, 1–36. [[CrossRef](#)]
- Ahmad, I.; Yousaf, M.; Yousaf, S.; Ahmad, M.O.; Uddin, M.I. Fake News Detection Using Machine Learning Ensemble Methods. *Complexity* **2020**, *2020*, 1–11. [[CrossRef](#)]
- Reis, J.C.S.; Correia, A.; Murai, F.; Veloso, A.; Benevenuto, F. Supervised Learning for Fake News Detection. *IEEE Intell. Syst.* **2019**, *34*, 76–81. [[CrossRef](#)]
- Wang, J.; Mao, H.; Li, H. FMFN: Fine-Grained Multimodal Fusion Networks for Fake News Detection. *Appl. Sci.* **2022**, *12*, 1093. [[CrossRef](#)]
- Cao, J.; Guo, J.; Li, X.; Jin, Z.; Guo, H.; Li, J. Automatic rumor detection on microblogs: A survey. *arXiv* **2018**, arXiv:1807.03505.
- Lu, Y.J.; Li, C.T. GCAN: Graph-aware Co-Attention Networks for Explainable Fake News Detection on Social Media. *arXiv* **2020**, arXiv:2004.11648.
- Yang, R.; Wang, X.; Jin, Y.; Li, C.; Lian, J.; Xie, X. Reinforcement Subgraph Reasoning for Fake News Detection. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, 14–18 August 2022; pp. 2253–2262.
- Shu, K.; Cui, L.; Wang, S.; Lee, D.; Liu, H. dEFEND. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 4–8 August 2019; pp. 395–405.

27. Jin, Y.; Wang, X.; Yang, R.; Sun, Y.; Wang, W.; Liao, H.; Xie, X. Towards fine-grained reasoning for fake news detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Online, 22 February–1 March 2022; pp. 5746–5754.
28. Kaliyar, R.K.; Goswami, A.; Narang, P.; Chamola, V. Understanding the Use and Abuse of Social Media: Generalized Fake News Detection with a Multichannel Deep Neural Network. *IEEE Trans. Comput. Soc. Syst.* **2022**, 1–10. [\[CrossRef\]](#)
29. Qian, F.; Gong, C.; Sharma, K.; Liu, Y. Neural User Response Generator: Fake News Detection with Collective User Intelligence. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 13–19 July 2018; pp. 3834–3840.
30. Xu, W.; Liu, Q.; Wu, S.; Wang, L. Counterfactual Debiasing for Fact Verification. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Toronto, ON, Canada, 9–14 July 2023; Volume 1, pp. 6777–6789.
31. Jaiswal, R.; Singh, U.P.; Singh, K.P. Fake News Detection Using BERT-VGG19 Multimodal Variational Autoencoder. In Proceedings of the 2021 IEEE 8th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON), Dehradun, India, 11–13 November 2021; pp. 1–5.
32. 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.
33. Jing, J.; Wu, H.; Sun, J.; Fang, X.; Zhang, H. Multimodal fake news detection via progressive fusion networks. *Inf. Process. Manag.* **2023**, *60*, 103120. [\[CrossRef\]](#)
34. Liu, J.; Wang, C.; Li, C.; Li, N.; Deng, J.; Pan, J.Z. DTN: Deep triple network for topic specific fake news detection. *J. Web Semant.* **2021**, *70*, 100646. [\[CrossRef\]](#)
35. Hua, J.; Cui, X.; Li, X.; Tang, K.; Zhu, P. Multimodal fake news detection through data augmentation-based contrastive learning. *Appl. Soft Comput.* **2023**, *136*, 110125. [\[CrossRef\]](#)
36. Ying, Q.; Hu, X.; Zhou, Y.; Qian, Z.; Zeng, D.; Ge, S. Bootstrapping Multi-view Representations for Fake News Detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023.
37. Yu, C.; Ma, Y.; An, L.; Li, G. BCMF: A bidirectional cross-modal fusion model for fake news detection. *Inf. Process. Manag.* **2022**, *59*, 103063. [\[CrossRef\]](#)
38. Sun, M.; Zhang, X.; Ma, J.; Xie, S.; Liu, Y.; Philip, S.Y. Inconsistent Matters: A Knowledge-guided Dual-consistency Network for Multi-modal Rumor Detection. *IEEE Trans. Knowl. Data Eng.* **2023**, *35*, 12736–12749. [\[CrossRef\]](#)
39. Li, P.; Sun, X.; Yu, H.; Tian, Y.; Yao, F.; Xu, G. Entity-Oriented Multi-Modal Alignment and Fusion Network for Fake News Detection. *IEEE Trans. Multimed.* **2022**, *24*, 3455–3468. [\[CrossRef\]](#)
40. Song, C.; Ning, N.; Zhang, Y.; Wu, B. A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks. *Inf. Process. Manag.* **2021**, *58*, 102437. [\[CrossRef\]](#)
41. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog* **2019**, *1*, 9.
42. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
43. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning, PMLR, Online, 18–24 July 2021; pp. 8748–8763.
44. Zhen, L.; Hu, P.; Wang, X.; Peng, D. Deep supervised cross-modal retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 10394–10403.
45. Hong, Y.; Mundt, M.; Park, S.; Uh, Y.; Byun, H. Return of the normal distribution: Flexible deep continual learning with variational auto-encoders. *Neural Netw.* **2022**, *154*, 397–412. [\[CrossRef\]](#)
46. Zhang, H.; Qian, S.; Fang, Q.; Xu, C. Multimodal Disentangled Domain Adaption for Social Media Event Rumor Detection. *IEEE Trans. Multimed.* **2021**, *23*, 4441–4454. [\[CrossRef\]](#)
47. Nan, Q.; Cao, J.; Zhu, Y.; Wang, Y.; Li, J. MDFEND: Multi-domain Fake News Detection. In Proceedings of the 30th ACM International Conference on Information & Knowledge Management, Gold Coast, Australia, 1–5 November 2021; pp. 3343–3347.
48. Maigrot, C.; Claveau, V.; Kijak, E.; Sicre, R. Mediaeval 2016: A multimodal system for the verifying multimedia use task. *MediaEval 2016: “Verifying Multimedia Use” Task*. 2016. Available online: <https://hal.science/hal-01394785/> (accessed on 22 July 2024).
49. Zhang, G.; Giachanou, A.; Rosso, P. SceneFND: Multimodal fake news detection by modelling scene context information. *J. Inf. Sci.* **2022**, *50*, 355–367. [\[CrossRef\]](#)
50. Shu, K.; Wang, S.; Liu, H. Exploiting tri-relationship for fake news detection. *arXiv* **2017**, arXiv:1712.07709.
51. Bibal, A.; Delchevalerie, V.; Frénay, B. DT-SNE: T-SNE discrete visualizations as decision tree structures. *Neurocomputing* **2023**, *529*, 101–112. [\[CrossRef\]](#)
52. Wang, Y.; Ma, F.; Jin, Z.; Yuan, Y.; Xun, G.; Jha, K.; Su, L.; Gao, J. Eann. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK, 19–23 August 2018; pp. 849–857.

53. Singhal, S.; Shah, R.R.; Chakraborty, T.; Kumaraguru, P.; Satoh, S.I. Spottfake: A multi-modal framework for fake news detection. In Proceedings of the 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM), Singapore, 11–13 September 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 39–47.
54. Zhang, H.; Fang, Q.; Qian, S.; Xu, C. Multi-modal Knowledge-aware Event Memory Network for Social Media Rumor Detection. In Proceedings of the 27th ACM International Conference on Multimedia, New York, NY, USA, 21–25 October 2019; pp. 1942–1951.

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.