frontiers | Frontiers in Artificial Intelligence

TYPE Original Research
PUBLISHED 06 November 2025
DOI 10.3389/frai.2025.1665798

OPEN ACCESS

EDITED BY
Jiaqi Gong,
University of Alabama, United States

REVIEWED BY
Antonio Sarasa-Cabezuelo,
Complutense University of Madrid, Spain
Xiaoming Guo,
University of Alabama System, United States

*CORRESPONDENCE
Jinrong Hu
✉ hjr@cuit.edu.cn

RECEIVED 14 July 2025
ACCEPTED 20 October 2025
PUBLISHED 06 November 2025

CITATION
Chen H, Yu Y, Guo H, Hu B, Hu S, Hu J, Lyu S, Wu X, Lin C-S and Wang X (2025) A self-learning multimodal approach for fake news detection. *Front. Artif. Intell.* 8:1665798. doi: 10.3389/frai.2025.1665798

A self-learning multimodal approach for fake news detection

Hao Chen[1], Yue Yu[2], Hui Guo[3], Baochen Hu[4], Shu Hu[5], Jinrong Hu[1]*, Siwei Lyu[6], Xi Wu[1], Ching-Sheng Lin[7] and Xin Wang[8]

[1]School of Computer Science, Chengdu University of Information Technology, Chengdu, China, [2]CAACSRI, Chengdu, China, [3]Department of Mathematics & Statistics, University at Albany, New York, NY, United States, [4]Dropbox Inc., California, CA, United States, [5]School of Applied and Creative Computing, Purdue University, West Lafayette, IN, United States, [6]Department of Computer Science and Engineering, University of Buffalo, New York, NY, United States, [7]Master Program of Digital Innovation, Tunghai University, Taichung, Taiwan, [8]Department of Epidemiology and Biostatistics, College of Integrated Health Sciences, and AI Plus Institute, University at Albany, New York, NY, United States

The rapid growth of social media has resulted in an explosion of online news content, leading to a significant increase in the spread of misleading or false information. While machine learning techniques have been widely applied to detect fake news, the scarcity of labeled datasets remains a critical challenge. Misinformation frequently appears as paired text and images, where a news article or headline is accompanied by a related visuals. In this paper, we introduce a self-learning multimodal model for fake news classification. The model leverages contrastive learning, a robust method for feature extraction that operates without requiring labeled data, and integrates the strengths of Large Language Models (LLMs) to jointly analyze both text and image features. LLMs are excel at this task due to their ability to process diverse linguistic data drawn from extensive training corpora. Our experimental results on a public dataset demonstrate that the proposed model outperforms several state-of-the-art classification approaches, achieving over 85% accuracy, precision, recall, and F1-score. These findings highlight the model's effectiveness in tackling the challenges of multimodal fake news detection.

KEYWORDS

fake news, contrastive learning, large language model, multimodal, machine learning

1 Introduction

The emergence of social media platforms has profoundly transformed news dissemination, offering immediate and widespread access to diverse information (Wang et al., 2023). However, this increased accessibility has inadvertently facilitated the rapid spread of misinformation, a problem exacerbated by technologies such as deepfakes (Chadha et al., 2021). A piece of misinformation is illustrated in Figure 1, where it is evident that the images are paired with misleading textual content, which can be easily fabricated or manipulated using AI-driven tools (Guo et al., 2022a). As a result, such misinformation can rapidly spread across the vast expanse of the digital landscape. In recent years, the widespread distribution of false narratives has become a critical social issue, causing negative impacts within digital environments and broader social contexts. This situation has raised substantial concerns across various demographic groups, leading to increased anxiety and a decrease in public trust in media sources (Pu et al., 2022). Therefore, there is an urgent need for the development and implementation of effective detection systems to combat the spread of fake news on social media platforms.

FIGURE 1
An example of fake news (mismatching image-text) from dataset (reproduced from Nakamura et al., 2020, European Language Resources Association (ELRA), licensed under CC-BY-NC).

The laborious and time-consuming nature of manual fact-checking has spurred the development of automated approaches to address the widespread issue of fake news. Among these, machine learning techniques-particularly supervised classification models (Bagozzi et al., 2024; Kaliyar et al., 2020; Jiang et al., 2020) have attracted significant attention. However, the efficacy of these models largely depends on the availability of high-quality labeled datasets. Unfortunately, such datasets are often challenging to obtain and are typically insufficient to capture the full diversity inherent in fake news content due to their limited scope. In contrast, using weakly supervised or unsupervised methods mitigates the need for large volumes of labeled data and offers distinct advantages over traditional supervised approaches.

While existing approaches primarily focus on textual semantic and syntactic similarities and have achieved some success in fake news detection, they often fail to account for the intricate interactions between different data modalities, particularly the subtle and complex relationships between images and text. Currently, Large multimodal models (e.g., Gemini, GPT) demonstrate strong image–text understanding through extensive pre-training, but their computational cost and general-purpose design limit applicability in domain-specific tasks such as fake news detection. In this paper, we propose a novel methodology for multi-modal fake news detection to address these shortcomings. Our approach uses contrastive learning to mitigate the challenge of limited labeled data. Additionally, we incorporate a large language model to integrate and analyze image and text features, enhancing the model's ability to assess the integrity of information. This design enables effective multimodal reasoning on smaller datasets while allowing targeted fine-tuning for domain-relevant cues. Hence, our approach provides a practical balance between performance and accessibility, complementing rather than competing with large generalist models. The primary contribution of this work lies in the architectural integration and synergistic design of several advanced yet complementary techniques to enhance multimodal fake news detection. Specifically:

1. Incorporate contrastive learning into the visual feature extraction pipeline, enabling the model to better capture image semantics and improve generalization. This architectural choice increase detection performance, particularly under conditions of limited labeled data.
2. Build upon existing large language model (LLM) architectures, we design a framework that integrates textual and visual modalities through learnable queries and prompt-based alignment. This strategic combination allows for more coherent multimodal reasoning, leading to notably higher detection accuracy.
3. Design a dynamic optimization strategy for the loss function that adapts to the evolving state of the LLM during fine-tuning. This mechanism ensures stable convergence and maintains high detection performance throughout the training process.

It is important to note that most of the techniques we introduce are general and can be applied to various classification tasks. Specifically, our use of contrastive learning proves advantageous in scenarios with a scarcity of labeled training data. The paper is organized as follows: Section 2 reviews related work, and Section 3 describes our proposed method. Our experimental evaluation is presented in Section 4. The conclusion and future work are presented in Section 6.

2 Related work

In recent years, the widespread dissemination of fake news on social media platforms has resulted in significant detrimental effects, thereby motivating the scholarly community to engage in extensive investigations into fake news detection. Initially, Bondielli and Marcelloni (2019) conducted a rigorous classification of information, distinguishing between fake news and rumors based on whether credible sources had thoroughly verified the content. Subsequently, both Meel and Vishwakarma (2020) and Guo et al. (2019) provided comprehensive analyses of the various terminologies related to misinformation prevalent on social media, including disinformation, fake news, and misinformation, among others (Guo et al., 2022b). Instead of concentrating on the subtle and complex distinctions among these definitions, our research is primarily focused on the machine learning methodologies employed for detection. Furthermore, our study predominantly examines news articles that feature paired text and images. Within this framework, the present paper categorizes existing fake news detection techniques into two main types—unimodal and multimodal—based on the nature of the data utilized.

2.1 Unimodal classification

A substantial body of research has leveraged supervised learning algorithms, such as Support Vector Machines (SVM) (Bagozzi et al., 2024), Naïve Bayes (Granik and Mesyura, 2017), and Logistic Regression (Sudhakar and Kaliyamurthie, 2023), for the detection of fake news. These models are trained on

annotated datasets that classify news articles as either authentic or deceptive. The performance of these algorithms, however, is highly contingent on the quality and diversity of the training data provided. Moving beyond conventional machine learning methods, neural networks have gained prominence in this domain (Guo et al., 2022c). For instance, convolutional neural networks (CNN) were employed by Kaliyar et al. (2020), recurrent neural networks (RNN) were used by Jiang et al. (2020), and Nasir et al. (2021) implemented a hybrid of CNN and RNN. Recently, researchers have explored pre-trained language models. BERT (Devlin et al., 2019) and RoBERTa (Zhuang et al., 2021) are employed to analyze news content, achieving notable advancements. Nevertheless, due to their relatively modest model sizes, the capacity of these pre-trained models to extract complex knowledge and perform advanced reasoning is constrained, limiting their effectiveness in handling fake news that requires deeper content analysis and inference. In contrast, large language models (LLMs), such as GPT (Brown et al., 2020), exhibit superior performance in natural language processing (NLP) tasks by employing deeper neural architectures and significantly larger parameter counts. These models rely on extensive textual datasets from diverse fields and topics, constructing a comprehensive knowledge base and contextual understanding that bolsters their reasoning capabilities. Consequently, LLMs require minimal additional data for fine-tuning to effectively differentiate authentic news from misinformation (Li et al., 2024; Su et al., 2023; Teo et al., 2024). However, most existing studies have predominantly focused on either text or image data in isolation, rather than integrating both modalities for a more comprehensive approach.

2.2 Multimodal classification

In the wake of the incessant development of social media, the news that is now being widely spread predominantly incorporates information such as text and images. As a result, scholars have stepped up their endeavors in the detection of multimodal fake news. Singhal et al. (2019) introduced a multimodal model which harnesses text and visual features. Likewise, Giachanou et al. (2020) combined the image features extracted by the VGG (Simonyan and Zisserman, 2014) model and the text features extracted by the BERT (Devlin et al., 2019) model to detect image-text misinformation. Aneja et al. (2022) centered their attention on "Cheapfakes" produced by employing free artificial intelligence methods (e.g., filtering). They made use of multimodal embedding to predict whether image-caption pairs are mismatched. Fact verification, on the other hand, necessitates an additional information base. In light of this circumstance, some scholars have focused on the use of LLMs with pre-trained on the prior knowledge. Zhu et al. (2023) introduced the multimodal large language model MiniGPT-4, which achieves alignment between image and linguistic features by employing the Q-Former module. Liu et al. (2024) introduced a fake news detection model, FakeNewsGPT-4, leveraging the MiniGPT-4 framework. This model advances the prompting capabilities of large language models by integrating both prior and dynamically generated knowledge, thereby achieving superior performance across multiple domains.

3 Methodology

In this paper, we propose an innovative model for the task of multimodal fake news detection, as illustrated in Figure 2. The overall structure comprises three core components: the contrastive module, the multimodal fusion module, and the classification module. To overcome the lack of training data, a contrastive learning mechanism is incorporated. In this mechanism, the image feature is acquired through augmentation where the model is trained by maximizing the similarity between positive pairs (e.g., different augmentations of the same data instance) and minimizing the similarity between negative pairs (from different data instances). Once the image features has been learned, we then use the pre-trained large-scale model to align the text content with the image features, which we called the infuse module. Rather than directly using the image features extracted from contrastive learning, we introduced a multi-task learning approach namely Q-Former to dynamically adjust the feature weights of image, thus to achieve the co-optimization of the image feature encoder and the text encoder. Subsequently, the pre-trained large-scale model, which is structured upon MiniGPT-4, is employed to effect a profound combination of image and text features. The multimodal features along with the appropriate prompts are fed as input of large language model. The output is followed by the linear layer and finally is trained to classify an accurate inference regarding the authenticity of news. We will explain each module in detail as follows:

3.1 The contrastive learning module

Within the contrastive learning module, each of the image undergoes augmentation procedures. Subsequently, the augmented images are inputted into an image encoder, which is then followed by a fully connected layer. This sequential process is designed to learn the image feature and generate a dense vector. For the purpose of training the contrastive learning model, all the augmented images originating from the same sample are regarded as positive instances. Meanwhile, images that are randomly selected from other samples within the dataset are considered as negative instances. To be specific, the image is augmented (e.g., rotation, flip, scaling etc.) and then fed to the encoder for feature extraction. We adopt the pre-trained ViT model (Granik and Mesyura, 2017) as the backbone network.

Rather than directly training ViT for image feature extraction, we harness the power of momentum mechanism to smooth and stabilize the image encoding. As shown in Figure 3, the input image Z, along with the augmentation $\tilde{Z}$, are fed to the the image encoder and momentum encoder. Both of them are identical initially while performing training procedure individually. The difference is that the parameters of the momentum encoder are not updated synchronously with those of the encoder. Updating the momentum encoder parameters using a smaller value ensures the stability of the training process, as shown in Equation 1:

$$y_t = m y_{t-1} + (1 - m) x_t \tag{1}$$

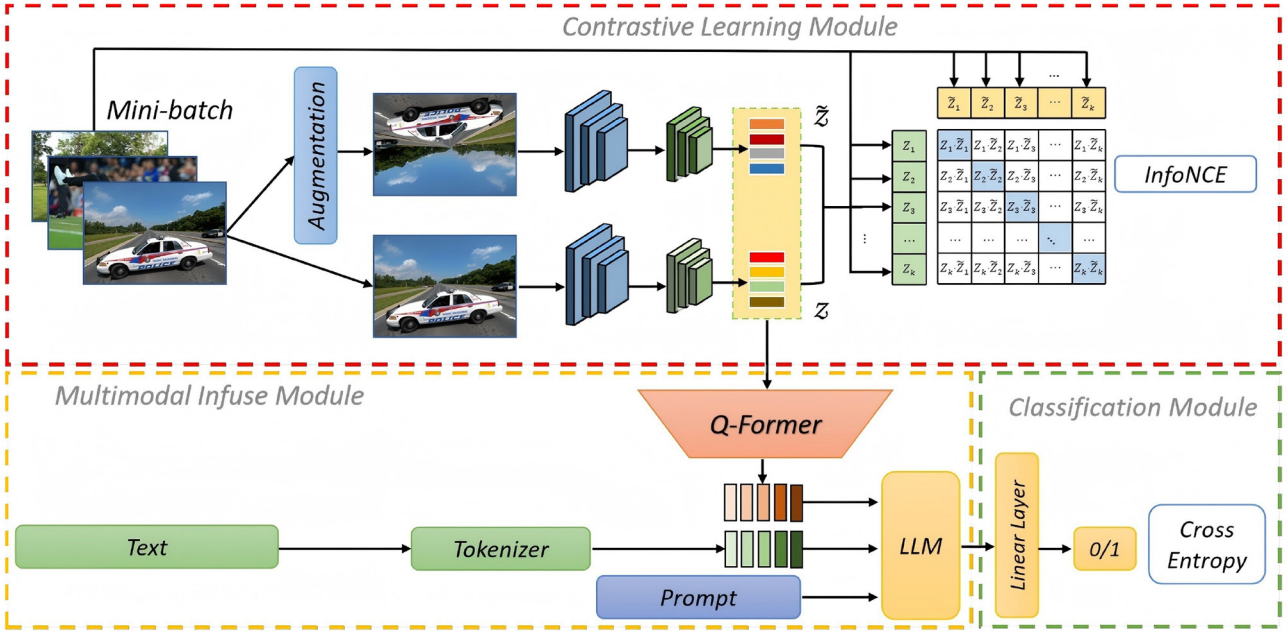

FIGURE 2
The overall structure of multimodal fake news detection (images reproduced from Nakamura et al., 2020, the Fakeddit dataset, https://github.com/entitize/Fakeddit). The model is composed of three components, contrastive learning module is for learning the image feature using a small sample of training data, infusing module aims to align text and image feature and then apply the large language model for the multimodal combination, the classification module is for the prediction of fake news.

where y_{t-1} is the parameter at the previous moment, x_t is the parameter at the current moment, and m is the updated threshold of the parameter. The image features from two encoders are conducted scalar product respectively to get two matrices. Then, the two matrices are added to get final matrix for contrastive learning. We use mini-batch for training. The elements of the diagonal of the matrix indicate the similarity between the positive samples within a mini-batch, and the other elements indicate the similarity between the positive samples and the negative samples. The overall loss function of contrastive learning is using InfoNCE as shown in Equation 2:

$$L_1 = -\log \frac{\exp\left(q \cdot k^+/\tau\right)}{\exp\left(q \cdot k^+/\tau\right) + \sum_{k^-} \exp\left(q \cdot k^-/\tau\right)} \tag{2}$$

where τ is the temperature coefficient that controls the softness or sharpness of the probability distribution over the positive and negative samples when calculating the InfoNCE loss. q is an anchor sample, k^+ is the positive sample, and k^- is the negative sample. Mathematically, the InfoNCE loss is given by a formula that involves taking the logarithm of the ratio of the exponential of the score of the positive sample to the sum of the exponentials of the scores of all samples (both positive and negative).

3.2 Multimodal fusion module

Concurrently, with regard to the text content, the Byte-Pair Encoding algorithm is initially employed as a tokenizer for the purpose of transforming sentences into tokens. Subsequently, the multimodal model predicated on MiniGPT-4 amalgamates the text along with the image features that have been encoded during the pre-training phase. Given that the large language model (LLM) utilized in our paper has already undergone pre-training with an extensive volume of data in advance, the downstream task merely necessitates fine-tuning with a relatively small quantity of data to fulfill the specified task requirements.

Instead of using the image features from contrastive learning module, we would like to leverage Q-Former (Query Transformer) (Li et al., 2023), a core component of MiniGPT-4 (Nakamura et al., 2020), to bridge the relationship between image and text. It employs a small set of learnable query tokens that interact with visual encoder outputs through cross-attention mechanisms, enabling the extraction of semantically relevant features. The resulting query embeddings form a compact and interpretable representation that is compatible with large language models, thereby facilitating efficient and effective multimodal fusion. In detail, the module extracts features from a frozen image encoder and aligns them with a large language model. The key component is a learnable query which is a set of vectors that are designed to interact with the input data in a way that helps extract relevant information and establish meaningful connections. The structure is illustrated in Figure 4. Q-Former's alignment is divided into two phases, the first phase mainly involves the learning of image features, through which Q-Former learns the most relevant image feature representation to the current text by the learnable queries. The second stage is generative learning, which combines the output of Q-Former with a pre-trained frozen large language model to achieve visual-to-language generative learning. The LLM [Vicuna (Chiang et al., 2023)] is used to understand and describe the visual expression features of

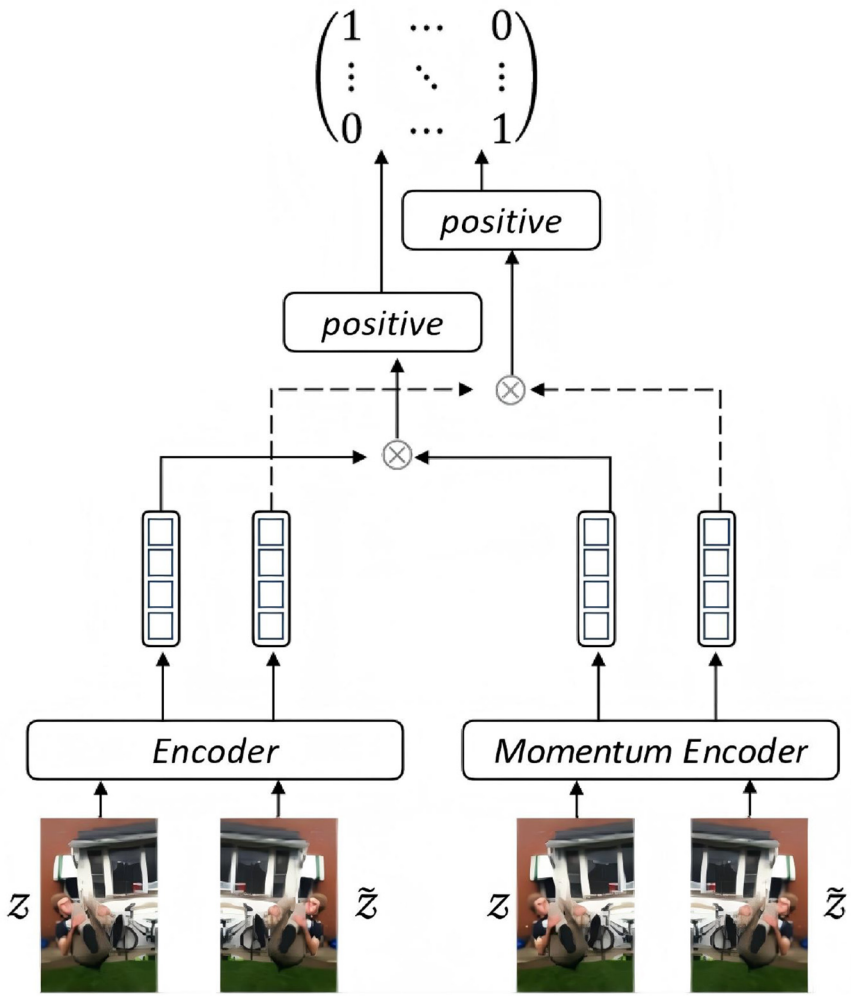

FIGURE 3
Momentum configuration for contrastive learning (image reproduced from Nakamura et al., 2020, the Fakeddit dataset, https://github.com/entitize/Fakeddit).

the Q-Former output, thus building a relationship between visual information and linguistic description.

In the end of multimodal fusion module, text feature e_{text} is obtained through the text embedding layer, image feature e_{img} is obtained in a pre-trained image encoder, and function f that can be interpreted by the large language model are obtained by using Q-Former with approapriate prompt, and these features are connected together to obtain the final mixed feature representation E. The equations are shown in Equations 3, 4.

$$e_{\text{prompt}} = e_{\text{text}} + f_{\text{Q-Former}}\left(e_{\text{img}}\right) \tag{3}$$

$$E = LLM\left(e_{\text{prompt}}\right) \tag{4}$$

3.3 Classification module

After performing multimodal fusion module to obtain the hidden layer features E, these features are then input into the fake news classifier. Subsequently, the final authenticity of the news is output. The fake news classifier is composed of two linear layers along with a GELU activation function. The specific details of the classification process are presented in Equation 5 below:

$$L_2 = MLP(E \mid \theta) \tag{5}$$

where θ is all the model parameters and l is the specific news category which is true or false.

3.4 Prompt

Within the LLM domain, prompts hold a vital position as they significantly influence the model's actions and resultant outputs. Essentially, prompts are text-based instructions or input indications furnished to the LLM, with the aim of communicating the task or context that the model is required to manage. In our task, we crafted a variety of prompts allow the model be able to differentiate the authentic of news from different perspectives.

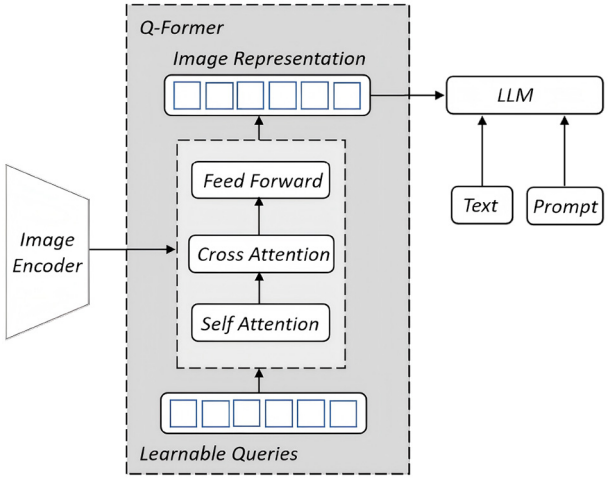

FIGURE 4
Q-Former structure adopted from Li et al. (2023).

The methodology put forward within this chapter employs four distinct Prompts for the training of the model. These Prompts are designed in line with the session format of the Vicuna language model and are randomly chosen to be inputted into the model. Sepecifically, the prompt is beginning with *<Img><ImageHere></Img><Text><TextHere></Text>*, where *<ImageHere>* represents the visual features generated by the image encoder, while *<TextHere>* represents the corresponding news text content. It followed by the different questions as follows:

- Determine if the text and images of this news story are lying?
- Determine if this text and images are describing a rumor?
- Determine if this news story is false?
- Determine if this text and image information is untrue?

3.5 Loss function

In the model previously described, both the contrastive learning and classification components necessitate training procedures. As a result, the overall loss function is constituted by two sub-functions, namely the contrastive learning loss L_1 and the classification loss L_2. Given the varying optimization priorities between these two individual tasks, employing fixed weights does not yield optimal results, and manually adjusting these weights is time-consuming. To address this issue, we implemented the Automatic Weighted Loss (AWL) method Kendall et al. (2018) to augment multi-task learning and concurrently optimize multiple loss functions. The details of this implementation are illustrated in Equation 6.

$$L \approx \frac{1}{2\sigma_1^2} L_1 + \frac{1}{\sigma_2^2} L_2 + \log(1+\sigma_1) + \log(1+\sigma_2) \quad (6)$$

Where σ_1 and σ_2 are learnable parameters representing the uncertainty of the corresponding task, the higher the uncertainty of the task, the smaller the weight of its loss function. Through this dynamic adjustment, the model can learn the weights suitable for different tasks and avoid the negative impact of fine-tune of pre-trained model. Additionally, during the preliminary experiment, we found that the loss function was getting smaller, but the model was not improved. To avoid this issue, we add 1 to each of learnable parameters. To be clarified, we train contrastive learning module based on L_1 for all images in the dataset. Then we fix the contrastive learning module and train L_2. The weights of the whole model are updated iteratively as described in Chen et al. (2023).

4 Experimental settings

4.1 Datasets

The data (Nakamura et al., 2020) obtained from social media platforms generally comprises a variety of elements. Among them, invalid information such as URLs and stop words are frequently encountered. This study aims to remove such information. However, certain proper names, including those of individuals, locations, and countries, which are considered as key information, are deliberately retained on account of the implementation of LLMs. For the image data, a series of augmentation operations are meticulously devised. These encompass operations such as horizontal flipping, hue transformation, and grayscale conversion. During the training phase of the model, one of these augmentation operations is randomly selected and applied. In addition, the data associated with entries lacking attached images and those accompanied by invalid images are removed from the dataset. Following this pre-processing, a complete and refined dataset is constructed. This dataset comprises nearly 563,600 training samples, 59,000 validation samples, and 59,500 testing samples. The statistical details of the dataset are presented in Table 1 below.

4.2 Implementation details

The image encoder of our network employed off-the-shelf ViT model where its block size is set to 14. All input images were scaled to a fixed resolution of 224 × 224. Followed by the encoding process, the feature vector dimension was set up to the batch size × 12 × 1,048. In terms of large language model we used, it is worth noting that the Vicuna model applied in this paper is version 7B. During the training process, a total of 100 training epochs were conducted, which we proved it's the best practices based on a variety of preliminary experiments, we extended the number of samples in each batch to 96 by the gradient accumulation method. The assessment methods in this paper are chosen as the accuracy, precision, recall, and F1-score, which are widely used in the field of

TABLE 1 Statistical information of the dataset.

Data category	True	False
Training set	222.1k	341.5k
Validation set	23k	36k
Testing set	23.5k	36k

fake news detection, as the metrics for evaluating the performance of the model.

4.3 Baselines

To assess the efficacy of our proposed model, this study selects a series of state-of-the-art classification methods as the baselines for comparing. All the chosen models are widely employed in image-text paring tasks and achieved strong ability to detect fake news. All models were trained using identical publicly available datasets and tested on the same data. Various evaluation metrics were carried out. Furthermore, we investigated model performance under conditions of limited training data, specifically utilizing only 10% of the Fakeddit training set. The baseline methods are: **EANN** (Wang et al., 2018) employs the pre-trained VGG for extracting image features, followed by Text-CNN. EANN incorporates auxiliary tasks such as an event discriminator, which outputs event categories of news to aid in decision-making. **CAFE** (Jin et al., 2021) leverages pre-trained BERT and ResNet-34 models to extract features from text and image respectively. It used a unified embedding space in order to integrate diverse modalities. **SpotFake** (Singhal et al., 2019) utilizes a pre-trained VGG network to extract features from images and BERT to capture textual features. SpotFake's notable advantage lies in its simplicity, avoiding complex auxiliary training tasks often seen in other models. **SpotFake+** (Singhal et al., 2019) modified SpotFake framework by the use of additional layers with attention mechanisms, advanced regularization techniques, and more sophisticated training process. **MVAE** (Qi et al., 2019) The Multimodal Variational Autoencoder extends variational autoencoders for multimodal data (text, images, audio). It learns shared latent representations to address tasks like classification. **HMCAN** (Qian et al., 2021) is short for Hierarchical Memory Compressed Attention Network which introduces a hierarchical structure with memory-compressed attention to capture both local and global contexts, optimizing long-sequence processing in NLP tasks. **VERITE** (Papadopoulos et al., 2023): "VERification of Image-TExt pairs" uses CLIP (Contrastive Language–Image Pre-training) as the feature extractor for images and texts, followed by the encoding layer of the Transformer. It has demonstrated a strong fake news detection performance.

5 Results

5.1 General performance

The Table 2 presents a comparative analysis of various models based on four evaluation metrics: accuracy, precision, recall, and F1-score. The models compared include EANN, CAFE, SpotFake, SpotFake+, MVAE, HMCAN, VERITE, and the proposed model referred to as "Ours." EANN achieved the lowest accuracy of 72.27% and an F1-score of 70.12%. In contrast, CAFE performed significantly better, with an accuracy of 84.14% and an F1-score of 85.32%, indicating robust performance in both precision and recall. Comparing SpotFake and its enhanced version, SpotFake+ showed notable improvement with SpotFake+ reaching an accuracy of 83.08% and an F1-score of 85.62%, outperforming SpotFake's 77.29% accuracy and 71.20% F1-score. HMCAN demonstrated superior precision, recall, and F1-score values, all around 84%, compared to MVAE's lower performance in these metrics. VERITE model also showed high performance with an accuracy of 84.72% and an F1-score of 84.85%, closely aligning with CAFE and SpotFake+ in terms of overall effectiveness.

TABLE 2 The results of three models over the Accuracy, Precision, Recall and F1-score.

Model	*Accuracy*	*Precision*	*Recall*	*F1-Score*
EANN (Wang et al., 2018)	72.27	78.43	63.4	70.12
CAFE (Jin et al., 2021)	84.14	85.39	85.27	85.32
SpotFake (Singhal et al., 2019)	77.29	71.63	70.77	71.20
SpotFake+ (Singhal et al., 2019)	83.08	86.38	84.87	85.62
MVAE (Qi et al., 2019)	70.24	76.53	74.75	75.63
HMCAN (Qian et al., 2021)	82.89	84.03	84.04	84.03
VERITE (Papadopoulos et al., 2023)	84.72	85.34	84.37	84.85
Ours	**88.88**	**86.40**	**85.40**	**85.90**

The proposed model ("Ours") outperformed all baseline models with the highest accuracy of 88.88%, precision of 86.40%, recall of 85.40%, and F1-score of 85.90%. This indicates the superior capability of the our method in handling the experimental tasks, achieving a balanced and high performance across all evaluation metrics. The observed improvements in accuracy and recall of our model indicate a significant reduction in the likelihood of misclassifying real news articles as fake, as well as a substantial decrease in the incidence of false positives.

To further analyzing results, EANN and MVAE exhibit only slight performance differences, reflecting the limitations of their basic multimodal fusion approaches. EANN utilizes an event discriminator for classification but does not address the semantic relationships between text and image features. MVAE, while using decoding structures to improve performance, still struggles with misalignment between modalities. SpotFake outperforms both by using pre-trained BERT for text encoding, although the direct concatenation of features limits its potential. SpotFake+ achieves even higher accuracy with XLNet, though it faces the same fusion constraints. HMCAN, CAFE, and VERITE outperform others, but fall short compared to the proposed method. HMCAN, about 6% less accurate than ours, applies multimodal contextual attention to fuse features. CAFE improves accuracy by aligning features in a unified space, reducing semantic gaps. VERITE stands out by combining CLIP for feature extraction with attention-based fusion, making it the top performer among the comparative models. To assess the effectiveness of the proposed method, statistical

significance tests were performed across all evaluation metrics. The Friedman test indicated a significant overall difference among the eight compared models. Although subsequent Bonferroni-adjusted *post-hoc* comparisons did not reveal individually significant differences, the consistent gains observed in Accuracy, Precision, Recall, and F1-Score suggest a robust performance advantage. Notably, the proposed model achieved the highest Accuracy with a statistically significant margin ($p < 0.05$, one-tailed t-test), demonstrating superior effectiveness and stability relative to existing approaches. Based on the above experimental results, this paper believes that the main factors that affect classification performance are as follows:

- The approach of feature fusion exerts a substantial influence on the performance of the model. While the direct concatenation of features appears to be a straightforward method, it fails to facilitate the interaction among data from diverse modalities. In contrast, models engineered with feature fusion techniques tend to attain favorable accuracy levels.
- Model scale also serves as a crucial metric. Apart from the model introduced in this study, the parameter scales of other comparative models are relatively limited. This smaller parameter size constrains the model's inferential abilities, impeding its capacity to make precise judgments regarding text and image features.

5.2 Ablation study

In this section, a series of ablation experiments were carried out. These experiments strictly adhered to the principle of controlling variables, ensuring that only one module was altered each time, thereby enabling the individual assessment of the impact of each module on the overall performance. Specifically, two types of experiments were designed, namely the comparison of different modal data and comparison of individual module. The specific settings of the experiments are as follows:

5.2.1 Comparison of different modal data

To verify the impact of different modal data on fake news detection, three experiments were designed respectively, namely single-modal image, single-modal text, and the multimodal combination of image and text. According to the experimental results presented in Table 3, in the experiment where only single-modal image data was utilized, the lowest accuracy, precision, and F1-score were 79.68%, 68.76%, and 72.75% respectively. In contrast, the model using single-modal text data achieved relatively higher accuracy, precision, and F1-score, which were 83.18%, 79.71%, and 78.64% respectively. This indicates that text information is of greater significance in determining the authenticity of news compared to image information. The data of the multimodal combination of image and text yielded the highest accuracy of 88.88%, which is approximately 7% and 5% higher than that of single-modal image and text respectively. Moreover, the accuracy, recall rate, and F1-score were all at relatively high levels, thereby demonstrating the effectiveness of multimodal fusion in fake news detection.

TABLE 3 Ablation comparison of different modals.

Modal type	Accuracy	Precision	Recall	F1-score
Image	79.68	68.76	89.38	72.75
Text	83.18	79.71	77.25	78.46
Image and text	88.88	86.40	85.40	85.90

TABLE 4 Ablation comparison of different modules.

Model	Accuracy	Precision	Recall	F1-score
Experiment A	87.16	78.60	92.91	85.16
Experiment B	88.21	81.19	91.45	86.02
Experiment C	88.88	86.40	85.40	85.90

5.2.2 Comparison of individual modules

In order to precisely gauge the contributions of each module within the proposed methodology to the performance of the model, three sets of comparative experiments were meticulously designed for this ablation study. The detailed experimental configurations are presented as follows:

- Experiment A: Employed the large language model along with the fully connected layer.
- Experiment B: Incorporated the contrastive learning module on top of the setup in Experiment A.
- Experiment C: Added the multimodal infuse module based on the configuration of Experiment B.

As shown in Table 4, Experiment A displayed the weakest performance in the context of fake news detection. Its accuracy rate was merely 87.16%, which was lower than that attained by Experiment B and Experiment C. The models that utilized contrastive learning and multimodal learning approaches demonstrated an advantage over the LLM model across various metrics. The improvement could be ascribed to the data augmentation procedure, which broadened the data samples and thus improved the generalization and robustness of the model. When analyzing the precision and recall rates, it was found that the multi-task learning method maintained a relatively stable performance, while Experiments A and B showed relatively higher recall rates.

6 Conclusion

This study aimed to establish an innovative multimodal classification framework for verifying the authenticity of news shared on social media platforms. In light of limited labeled data, especially for image-based content, we employed contrastive learning to improve feature representation. Additionally, we demonstrated the effectiveness of Large Language Models (LLMs)

in facilitating the seamless integration of text and image features. Instead of a simplistic multimodal fusion, we introduced a learnable alignment module that significantly improved the model's accuracy by aligning text-image features. Key contributions of this work include: (1) We developed an enhanced fake news detection model grounded in contrastive learning, a self-supervised approach utilizing data augmentation during training. Experimental results strongly indicated the model's superiority in scenarios with limited labeled data. (2) Our focus on news items as paired image-text combinations revealed that dynamically infusing features from different data formats substantially improved fake news detection, achieving an accuracy rate close to 89%. This multimodal approach considerably outperformed single-modal text- or image-only analyses. (3) By integrating contrastive learning with LLMs through a carefully designed feature infusion mechanism for multimodal classification, we conducted extensive comparative experiments. The findings highlighted the robust detection capabilities of our proposed model relative to numerous existing models, and demonstrated that larger LLMs further enhance detection accuracy.

A primary limitation of this study is the scope of data analysis, as our current focus excludes potentially valuable supplementary information, such as social networks, geographic data, and event context. While our model achieves state-of-the-art results purely from the content, its full potential for misinformation detection on social media platforms remains untapped without these relational features. Furthermore, a significant limitation lies in the omission of computational efficiency metrics. Although our model demonstrates superior performance in Accuracy and F1-score, we did not report its run-time, inference speed, or the total computational resources (e.g., GPU hours) required for training. This prevents a complete evaluation of the crucial trade-off between performance and deployment cost. Future research will focus on two main directions: first, developing robust integration strategies to effectively combine our model with supplementary social and event-based information to further boost predictive accuracy; and second, conducting comprehensive optimization and benchmarking to reduce run-time, improve inference speed, and ensure the computational viability of our approach for real-world application.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://github.com/entitize/Fakeddit.

Ethics statement

This study utilizes the publicly available Fakeddit dataset, which comprises Reddit posts collected in accordance with Reddit's content and API usage policies. All personal identifiers and user metadata were removed to ensure anonymity and prevent re-identification. Although no direct human participation was involved, data use adhered to privacy protection and data minimization principles consistent with responsible research standards. As Fakeddit reflects the biases of its source community, care was taken to mitigate potential effects on model fairness and interpretation. The research aims to advance academic understanding of misinformation detection, not to develop operational moderation systems. Ethical use was guided by transparency, accountability, and awareness of potential societal implications such as bias reinforcement or trust erosion.

Author contributions

HC: Conceptualization, Writing – original draft, Formal analysis, Funding acquisition, Methodology, Writing – review & editing. YY: Supervision, Writing – original draft. HG: Data curation, Investigation, Writing – original draft. BH: Methodology, Writing – review & editing. SH: Supervision, Validation, Writing – review & editing. JH: Methodology, Visualization, Writing – review & editing. SL: Methodology, Supervision, Writing – review & editing. XWu: Supervision, Validation, Writing – review & editing. C-SL: Methodology, Supervision, Writing – review & editing. XWa: Conceptualization, Project administration, Supervision, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This research was funded by the Chengdu University of Information Technology Program (No. KYTZ2023053).

Conflict of interest

BH was employed by Dropbox Inc.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

References

Aneja, S., Midoglu, C., Dang-Nguyen, D.-T., Khan, S. A., Riegler, M., Halvorsen, P., et al. (2022). Acm multimedia grand challenge on detecting cheapfakes. *ArXiv, abs/2207.14534*.

Bagozzi, B. E., Goel, R., Lugo-De-Fabritz, B., Knickmeier-Cummings, K., and Balasubramanian, K. (2024). A framework for enhancing social media misinformation detection with topical-tactics. *Dig. Threat.* 5, 1–29. doi: 10.1145/3670694

Bondielli, A., and Marcelloni, F. (2019). A survey on fake news and rumour detection techniques. *Inf. Sci.* 497, 38–55. doi: 10.1016/j.ins.2019.05.035

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. (2020). Language models are few-shot learners. *ArXiv, abs/2005.14165*.

Chadha, A., Kumar, V., Kashyap, S., and Gupta, M. (2021). "Deepfake: an overview," in *Proceedings of second international conference on computing, communications, and cyber-security: IC4S 2020* (Springer), 557–566. doi: 10.1007/978-981-16-0733-2_39

Chen, H., Zheng, P., Wang, X., Hu, S., Zhu, B., Hu, J., et al. (2023). "Harnessing the power of text-image contrastive models for automatic detection of online misinformation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 923–932. doi: 10.1109/CVPRW59228.2023.00099

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., et al. (2023). Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, march 2023. Available online at: https://lmsys.org/blog/2023-03-30-vicuna 3:6.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). "Bert: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–4186.

Giachanou, A., Zhang, G., and Rosso, P. (2020). "Multimodal multi-image fake news detection," in *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, 647–654. doi: 10.1109/DSAA49011.2020.00091

Granik, M., and Mesyura, V. (2017). "Fake news detection using naive bayes classifier," in *2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)*, 900–903. doi: 10.1109/UKRCON.2017.8100379

Guo, B., Ding, Y., Yao, L., Liang, Y., and Yu, Z. (2019). The future of misinformation detection: new perspectives and trends. *ArXiv, abs/1909.03654*.

Guo, H., Hu, S., Wang, X., Chang, M.-C., and Lyu, S. (2022a). "Eyes tell all: irregular pupil shapes reveal gan-generated faces," in *ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE), 2904–2908. doi: 10.1109/ICASSP43922.2022.9746597

Guo, H., Hu, S., Wang, X., Chang, M.-C., and Lyu, S. (2022b). Open-eye: an open platform to study human performance on identifying ai-synthesized faces. *arXiv preprint arXiv:2205.06680*.

Guo, H., Hu, S., Wang, X., Chang, M.-C., and Lyu, S. (2022c). Robust attentive deep neural network for exposing gan-generated faces. *IEEE Access* 10, 32574–32583. doi: 10.1109/ACCESS.2022.3157297

Jiang, T., Li, J. P., Haq, A. U., and Saboor, A. (2020). "Fake news detection using deep recurrent neural networks," in *2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, 205–208. doi: 10.1109/ICCWAMTIP51612.2020.9317325

Jin, X., Chen, P.-Y., Hsu, C.-Y., Yu, C.-M., and Chen, T. (2021). "Cafe: catastrophic data leakage in vertical federated learning," in *Advances in Neural Information Processing Systems*, eds. M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan (Curran Associates, Inc.), 994–1006.

Kaliyar, R. K., Goswami, A., Narang, P., and Sinha, S. (2020). Fndnet – a deep convolutional neural network for fake news detection. *Cogn. Syst. Res.* 61, 32–44. doi: 10.1016/j.cogsys.2019.12.005

Kendall, A., Gal, Y., and Cipolla, R. (2018). "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7482–7491. doi: 10.1109/CVPR.2018.00781

Li, J., Li, D., Savarese, S., and Hoi, S. (2023). "Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models," in *International Conference on Machine Learning* (PMLR), 19730–19742.

Li, X., Zhang, Y., and Malthouse, E. C. (2024). Large language model agent for fake news detection. *ArXiv, abs/2405.01593*.

Liu, X., Li, P., Huang, H., Li, Z., Cui, X., Liang, J., et al. (2024). "Fka-owl: advancing multimodal fake news detection through knowledge-augmented lvlms," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 10154–10163. doi: 10.1145/3664647.3681089

Meel, P., and Vishwakarma, D. K. (2020). Fake news, rumor, information pollution in social media and web: a contemporary survey of state-of-the-arts, challenges and opportunities. *Expert Syst. Appl.* 153:112986. doi: 10.1016/j.eswa.2019.112986

Nakamura, K., Levy, S., and Wang, W. Y. (2020). "Fakeddit: a new multimodal benchmark dataset for fine-grained fake news detection," in *International Conference on Language Resources and Evaluation*.

Nasir, J. A., Khan, O. S., and Varlamis, I. (2021). Fake news detection: a hybrid CNN-RNN based deep learning approach. *Int. J. Inf. Manag. Data Insights* 1:100007. doi: 10.1016/j.jjimei.2020.100007

Papadopoulos, S.-I., Koutlis, C., Papadopoulos, S., and Petrantonakis, P. C. (2023). Verite: a robust benchmark for multimodal misinformation detection accounting for unimodal bias. *Int. J. Multim. Inf. Retr.* 13:4. doi: 10.1007/s13735-023-00312-6

Pu, W., Hu, J., Wang, X., Li, Y., Hu, S., Zhu, B., et al. (2022). Learning a deep dual-level network for robust deepfake detection. *Patt. Recognit.* 130:108832. doi: 10.1016/j.patcog.2022.108832

Qi, P., Cao, J., Yang, T., Guo, J., and Li, J. (2019). "Exploiting multi-domain visual information for fake news detection," in *2019 IEEE International Conference on Data Mining (ICDM)*, 518–527. doi: 10.1109/ICDM.2019.00062

Qian, S., Wang, J., Hu, J., Fang, Q., and Xu, C. (2021). "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*. doi: 10.1145/3404835.3462871

Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *CoRR, abs/1409.1556*.

Singhal, S., Shah, R. R., Chakraborty, T., Kumaraguru, P., and Satoh, S. (2019). "Spotfake: a multi-modal framework for fake news detection," in *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, *pages* 39–47. doi: 10.1109/BigMM.2019.00-44

Su, J., Cardie, C., and Nakov, P. (2023). Adapting fake news detection to the era of large language models. *ArXiv, abs/2311.04917*. doi: 10.18653/v1/2024.findings-naacl.95

Sudhakar, M., and Kaliyamurthie, K. (2023). *Fake News Detection Approach Based on Logistic Regression in Machine Learning*. Cham: Springer, 55–60. doi: 10.1007/978-981-19-9304-6_6

Teo, T. W., Chua, H. N., Jasser, M. B., and Wong, R. T. (2024). "Integrating large language models and machine learning for fake news detection," in *2024 20th IEEE International Colloquium on Signal Processing Its Applications (CSPA)*, 102–107. doi: 10.1109/CSPA60979.2024.10525308

Wang, X., Guo, H., Hu, S., Chang, M.-C., and Lyu, S. (2023). GAN-generated faces detection: a survey and new perspectives. *arXiv preprint arXiv:2202.07145*.

Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., et al. (2018). "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*. doi: 10.1145/3219819.3219903

Zhu, D., Chen, J., Shen, X., Li, X., and Elhoseiny, M. (2023). Minigpt-4: enhancing vision-language understanding with advanced large language models. *ArXiv, abs/2304.10592*.

Zhuang, L., Wayne, L., Ya, S., and Jun, Z. (2021). "A robustly optimized BERT pre-training approach with post-training," in *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, eds. S. Li, M. Sun, Y. Liu, H. Wu, K., W. Che, S. He, and G. Rao (Huhhot, China: Chinese Information Processing Society of China), 1218–1227.