Received: September 29, 2023. Revised: November 28, 2023. 654

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Deep Learning Based Semantic Model for Multimodal Fake News Detection

Sanjeev Mewalal Dwivedi[1]* Sunil B. Wankhade[2]

[1]*Vidyalankar Institute of Technology, Wadala, Mumbai, India*
[2]*Rajiv Gandhi Institute of Technology, Andheri, Mumbai, India*
* Corresponding author's Email: sanjeev.dwivedi@vit.edu.in

Abstract: Social media has gradually become the primary news source for people in recent years, providing convenience but also leading to the spread of false information. With the rise of media-rich social media platforms, fake news has evolved from single-text to multimodal formats, prompting increased attention to multi-modal fake news detection. However, most existing methods rely on representation-level features that are closely tied to the dataset, resulting in insufficient modelling of semantic-level features and a limited ability to generalize to new data. To address this issue, we propose a semantically enhanced multimodal fake news detection method that utilizes pre-trained language models to capture implicit factual knowledge and explicitly extracts visual entities to better understand the deep semantics of multimodal news. We also extract visible features at different semantic levels, use a text-guided attention mechanism to model semantic interactions between text and images, and integrate multimodal features. Experimental results on real datasets based on Weibo news demonstrate that the proposed method outperforms other methods with an accuracy of 0.895 in multimodal fake news detection.

Keywords: Social media fake news detection, Multimodality, Knowledge fusion, Attention mechanism.

1. Introduction

Social media platforms like Twitter, Instagram, and Facebook have emerged as the primary channels for accessing news information, enabling real-time, open, convenient, and interactive communication. However, the ease of participation and information sharing on these platforms has also facilitated the rapid spread of false information, including fake news, within the online space. The detrimental effects of internet fake news extend beyond misleading audiences. They erode the authority and credibility of mainstream media and pose risks in various domains, such as the economy and politics. As social media increasingly incorporates rich media content, with users sharing multimedia formats combining images and text, fake news publishers have adapted their strategies by employing deceptive and manipulated images to capture readers' attention and propagate misinformation. Consequently, detecting multi modal fake news on platforms like Twitter, Instagram, and Facebook has become a prominent research area, aiming to mitigate the spread of false information and preserve the integrity of news dissemination in the digital age [1].

Existing research indicates that there are significant differences between fake news and real news at the surface level [2]. Fake news often exhibits stronger emotional appeals, subjectivity [3], and frequently includes high-frequency phrases like "urgent notice" or "share quickly" [4]. The images accompanying fake news tend to have low quality but possess strong visual impact [5]. In contrast, real news tends to be more objective and rigorous, with higher-quality accompanying visuals. Current multimodal approaches [6-8] commonly employ recurrent neural networks (RNNs) and convolutional neural networks (CNNs) to capture the characteristics of fake news in terms of both textual and visual modalities at the surface level. However, the surface-level characteristics of fake news are highly dataset-dependent, which makes methods that perform well on specifc datasets often struggle to generalize effectively to new datasets and are prone to

misjudging fake news with less obvious surface-level characteristics.

In fact, for the task of fake news detection, focusing solely on how news is expressed, i.e., the surface-level characteristics of news, is not sufficient. It is also important to consider the specific content that the news describes, i.e., the semantic-level characteristics of news. At the semantic level, fake news often involves controversial topics or discrepancies between text and images. Compared to the surface level, capturing the semantic-level characteristics of fake news is more challenging. On one hand, as news is a special genre of storytelling, it often contains named entities such as personal names, place names, organizational names, and other proper nouns. Understanding these entities plays a vital role in modelling the semantic-level characteristics of fake news. However, their meanings are not easily understood through context alone and require the introduction of external factual knowledge. On the other hand, in the semantic understanding of multimodal news, the image modality often provides crucial information about key entities (such as celebrities, landmarks, or fags) that can aid the model's predictions. For example, we can infer the credibility of a news article by checking the consistency of the identities of people depicted in the text and images. However, general visual feature representations mostly remain at the perceptual level and fail to uncover and adequately model the deep semantics behind these visual entities. Furthermore, general visual semantic features and textual semantic features exist in different feature spaces, leading to semantic gaps and feature heterogeneity. Therefore, adequately modelling the semantic interactions between text and images is also a crucial aspect that requires careful consideration.

To address the challenges mentioned above, a semantic-enhanced multimodal fake news detection method is proposed. Firstly, the vast amount of factual knowledge implicitly embedded in pre-trained language models is leveraged to achieve a better understanding of entity concepts within multimodal news. Secondly, general visual feature vectors are extracted, and external models are employed to explicitly extract visual entities and embed textual information from news images, resulting in visual features at different semantic levels. Lastly, a text-guided attention mechanism is employed to model the semantic interaction between text and visual features at different levels, thereby obtaining a unified representation of multimodal features.

The main contributions of this paper can be summarized in three aspects:

1. A novel semantic-enhanced multimodal fake news detection method is pro-posed. By integrating external knowledge and explicit visual entity extraction, a better understanding of entity semantics within multimodal news is achieved, leading to a more comprehensive exploration of semantic clues in multimodal fake news.
2. The use of text-guided attention mechanism models the semantic interaction between text and visual features at different levels, effectively integrating heterogeneous multimodal features.
3. The proposed method is validated on a real-world Weibo dataset. Compared to current state-of-the-art methods, our model significantly improves the accuracy of fake news detection.

The rest of the paper is organized as; section 2 illustrate related work; section 3 discuss the proposed method and 3.1- 3.2 contain a detailed description text and visual Semantic Encoder and classification model; section 4 contains experimental setup and result analysis, Finally, section 5 concludes the overall proposed work.

2. Related work

Based on the different research targets, fake news detection can be divided into event-level detection and Weibo-level detection. Event-level detection involves assessing the credibility of a news event by considering the collective information from all Weibo posts related to that event. However, events typically take time to develop, and some major fake news stories may have already spread widely on social media before the event fully forms, causing significant negative impact within a very short period. Weibo-level detection, on the other hand, focuses on determining the credibility of individual Weibo messages. In comparison to event-level detection, this approach enables real-time detection, making it highly relevant for practical applications. This study specifically focuses on Weibo-level fake news detection.

Most existing research on fake news detection utilizes textual content and the social context generated during the dissemination process [9]. Text-based detection methods primarily rely on modelling the specific language style associated with fake news, including early approaches that extract linguistic features, topic features, and other handcrafted features [10-13], as well as more recent methods that leverage deep models to automatically learn high-level features from the data [10]. Context-based methods, on the other hand, include approaches based

on user behaviour credibility [14-16] and methods based on the propagation network [17-20].

In recent years, some studies have started to focus on the role of visual modality in fake news detection [21-25]. Fake news images can be broadly categorized into two types: manipulated images and misused images [5]. Manipulated images refer to those that have been intentionally altered at the pixel level using tools or are non-realistic images generated by algorithms. Misused images, on the other hand, generally refer to real images taken from other events or images whose content has been misinterpreted. Existing research on visual modality primarily utilizes evidential features [17], semantic features [5], distribution features [21], and contextual [18, 19] of images for fake news detection.

The textual and visual modalities provide distinct and complementary information for fake news detection. Therefore, there is a growing interest in methods that combine multimodal information for fake news detection. Shishah et al. [6] used deep neural networks to incorporate multimodal information into fake news detection. They proposed an attention-based recurrent neural network that integrates textual, visual, and social context information. To improve the model's generalization performance on new fake news events, Ali et al. [7] introduced an adversarial learning approach by incorporating an auxiliary task of event classification to guide the model in learning more generalized multimodal features unrelated to specific events. Shahid et al. [8] present an encoder-decoder structure that employed to construct the feature representation of multimodal news. These approaches have shown certain effectiveness in multimodal fake news detection. However, due to the lack of sufficient factual knowledge, they fail to fully understand the deep semantics of multimodal news events.

To address this issue, Ilie et al. [26] extracted concept knowledge corresponding to text entities from external knowledge graphs and incorporated it into the multimodal representation to achieve better semantic understanding. Ying et al. [27], proposed a graph neural network model that create interactions between text, knowledge, and objects in images. These methods enhance the understanding of textual semantics by incorporating external knowledge graphs. However, there are still limitations in modelling the semantic information of images and integrating heterogeneous multimodal features.

Key research gaps in the realm of fake news detection include addressing the temporal discrepancy between event-level and Weibo-level detection, enhancing the integration of textual and visual modalities for a deeper understanding of multimodal content, developing robust methods for incorporating external knowledge to improve semantic comprehension, and focusing on real-time detection to mitigate the rapid spread of false information. These gaps highlight the need for innovative approaches and methodologies to more effectively counter the dissemination of fake news in today's digital information landscape.

Therefore, to address the limitations of existing work, we propose a semantic-enhanced multimodal fake news detection method that not only leverages external knowledge to gain a deeper understanding of the semantic information in both text and images but also fully integrates heterogeneous features from different modalities.

3. Semantic-enhanced fake news detection method

The authenticity of a given single multimodal news piece, distinguishing between real and fake, is the objective of this study. The innovative model for semantic- enhanced multimodal fake news detection, encompassing four key components: textual semantic encoder, visual semantic encoder, multimodal feature fusion, and classification, is showcased in Fig. 1.

3.1 Textual semantic encoder

Text, as the narrative body of news events, contains rich information that provides clues at different levels for determining the credibility of news. Existing methods mostly employ recurrent neural networks and similar approaches to model the contextual information of input text, capturing patterns at the surface level of the text [6, 8, 26]. However, due to the lack of involvement of corresponding factual knowledge in the feature extraction process, these methods have limited understanding of named entities within news text, thereby making it difficult to fully capture semantic-level clues of fake news.

Recent Han et al. [28] has indicated that pre-trained language models, with BERT (bidirectional encoder representations from transformers) as a representative example, have strong modelling capabilities. By learning from extensive pre-training corpora, these models have acquired certain syntactic and common sense knowledge. A knowledge-enhanced semantic representation model called ERNIE (enhanced representation from knowledge integration) has been proposed by Ying et al. [29], which shares a similar structure with BERT, utilizing multilayer transformers [30-31] as the basic encoders for modelling contextual information through self-

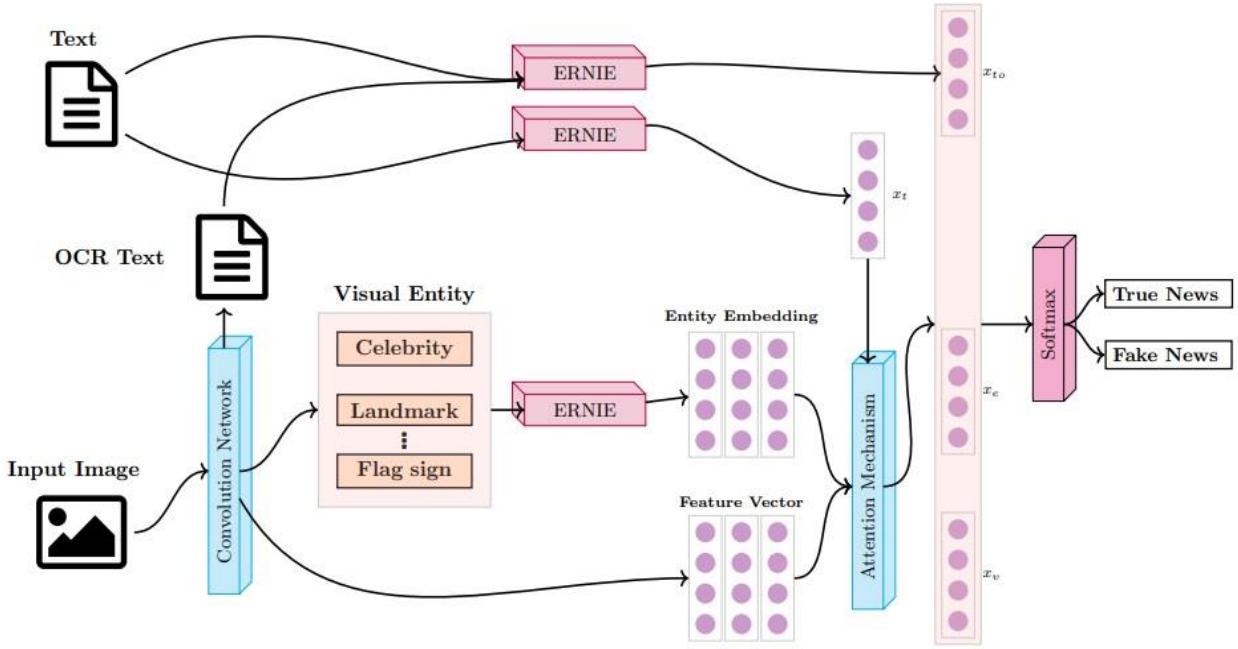

Figure. 1 Proposed framework for fake news identification

attention mechanisms. Unlike BERT, ERNIE masks semantic units such as words and entities and extends pre-training on word corpora rich in knowledge, allowing for better modelling of entity concepts and other prior semantic knowledge, thereby further enhancing the model's semantic representation capabilities. ERNIE can serve not only as a context encoder for generating sentence expressions but also as a knowledge repository, implicitly utilizing a vast amount of stored factual knowledge during sentence generation. Therefore, ERNIE is used as the feature extractor for the textual modality, simultaneously capturing the characteristics of text at both the surface and semantic levels.

Specifically, we first fine tune ERNIE on the dataset of the fake news classification task. For an input sentence $T=[\ w_1, w_2, \cdots, w_n]$ where w_i represents the i^{th} word in the sentence, ERNIE encodes it by adding [MASK], [SEP], [CLS] and other tags, and then undergoes training. We extract the 768-dimensional feature vector corresponding to [CLS] as the final semantic representation of the input sentence as shown in Eq. (1):

$$x_{\mathrm{t}} = \mathrm{ERNIE}(T)\,, x_{\mathrm{t}} \in \mathrm{R}^{768} \tag{1}$$

Furthermore, there are many news articles on social media that primarily consist of text-based images, where the main text of the news is represented in the form of an image. We utilize the Baidu pre-trained OCR text detection model to extract text information from the images. After data pre-processing, the recognized text in the image can be represented as a sequence of words, denoted as $O = [w_1, w_2, \cdots, w_n]$,, where w_i represents the i^{th} word in the sentence. To fully model the semantic interaction between the input text T and the image text O, we concatenate them into a single sequence, separated by [SEP], and feed it into the ERNIE network to obtain the corresponding semantic representation:

$$x_{\mathrm{to}} = \mathrm{ERNIE}(T[\mathrm{SEP}]O) \tag{2}$$

3.2 Visual semantic encoder

In contrast to authentic news visuals, false news images tend to possess lower image quality while exuding a striking visual impact and stirring emotional style [5]. Consequently, current approaches predominantly employ convolutional neural networks to extract hierarchical visual features like colour, edges, and textures, aiming to model the quality and stylistic aspects of these images. Nonetheless, these methods fall short in capturing the profound underlying semantics of news visuals, as they lack the incorporation of external knowledge. Thus, the representation of these visual features remains confined to the perceptual realm, failing to fully grasp the deep layers of meaning conveyed by news images.

In actuality, news images often encompass highly newsworthy visual entities, including celebrities, landmarks, flags, and sensitive targets. The accurate identification of these entities contributes to a more thorough understanding of the semantic aspects of

multimodal news, thereby enhancing the detection of clues pertaining to false news. For instance, through the recognition of celebrities and landmarks in an image, discrepancies between the portrayed individuals or locations and the textual description of the news can be revealed. By identifying sensitive symbols and objects within the image, the emphasis on relevant entities mentioned in the text is heightened, facilitating a better comprehension of the controversial points within multimodal news. Consequently, to comprehensively model the semantic characteristics of false news images, visual feature vectors are extracted to model their quality and stylistic attributes, while external models are introduced to explicitly extract the visual entities within the images and model their deep-level semantics.

Specifically, in order to capture the characteristics of image quality and style, the previous work is referred to, and the VGG19 network [32] is utilized for extracting visual feature vectors from the images. It has been observed through preliminary experiments that VGG19 demonstrates more stable performance on image datasets for false news classification tasks compared to models such as ResNet [33] and Inception [34]. Taking into consideration the inconsistent information density and importance across different spatial regions in the images, the input images are subjected to block-wise feature extraction. Firstly, the VGG19 network pre-trained on ImageNet [35] is fine-tuned using the dataset for false news classification tasks. For an input image *I*, a feature map of size 7×7×512 can be obtained from the last convolutional layer of the VGG19 network. Subsequently, this feature map is further represented as a sequence of feature vectors $V = [v_1, v_2, \cdots, v_n]$, where $v_i \in \mathrm{R}^{512}$ represents the visual feature vector corresponding to the i^{th} image block, and $n = 49$.

In order to accurately identify named entities such as celebrities and landmarks appearing in the images, the corresponding annotated dataset for the task can be used to pre-train a target detection model, which is then employed to detect entities in the news images. Due to the limited availability of large-scale word annotated datasets related to the aforementioned tasks, the visual entity recognition interface provided by baidu AI platform is utilized. Specifically, the recognition models, including the celebrity detection model, landmark detection model, flag detection model, and sensitive target detection model, are utilized for the recognition of various entities. The celebrity detection model can recognize famous political and public figures, the landmark detection model can identify renowned landmarks both domestically and internationally, the flag detection model can recognize national flags, party emblems, police badges, ethnic costumes, and various symbols of reactionary organizations, and the sensitive target detection model can identify firearms, military weapons, instances of bloodshed, disease manifestations, explicit content, acts of terrorism, explosions, fires, and car accidents, among other visually sensitive objects. By performing entity recognition on the input image *i*, a corresponding list of entities is obtained. To comprehensively understand the underlying semantic information behind these entities, the entity name list is in- putted into the ERNIE network, resulting in the generation of the corresponding entity representation sequence $E = [e_1, e_2, \ldots e_n]$, where $e_i \in \mathrm{R}^{768}$ represents the semantic representation of the i^{th} visual entity identified in the image.

3.3 Multimodal feature fusion

Up to this point, the text representation x_1, the joint representation of text and image x_{to}, the representation of visual entity sequences *E*, and the representation of visual feature vector sequences *V* have been obtained. In this section, an explanation will be provided on how to integrate the aforementioned heterogeneous features to achieve a unified multimodal representation.

Multiple visual entities may exist within the image, yet not all detected entities contribute to the task of false news classification. Incorporating all entity information could potentially introduce redundancy or noise. Based on observations, it has been noticed that visual entities which correspond to the text tend to hold greater significance. Hence, a fusion technique employing a text-guided attention mechanism is applied to the multiple visual entities $E = [e_1, e_2, \ldots, e_n]$. Initially, the importance of each visual entity e_i is calculated with respect to the text feature x_t:

$$f(x_\mathrm{t}, e_i) = f\left(x_\mathrm{t}^\mathrm{T} W e_i\right), i \in [1, n], \tag{3}$$

where *W* denotes a parameter matrix that is randomly initialized and jointly optimized during the training process, and *f(.)* represents an activation function. Subsequently, the weights are normalized as follows:

$$\alpha_{e_i} = \frac{\exp\left(F(x_\mathrm{t}, e_i)\right)}{\sum_{i=1}^{n} \exp\left(F(x_\mathrm{i}, e_i)\right)} \tag{4}$$

Finally, a weighted summation is performed on the different visual entity representations based on the

obtained weights, yielding the ultimate visual entity representation:

$$x_{\mathrm{e}} = \sum_{i=1}^{n} \alpha_{e_i} e_i \quad (5)$$

Similarly, different regions of an image hold varying degrees of importance for semantic understanding. Therefore, we perform fusion using a text-guided attention mechanism on the feature vectors of different regions in the image, denoted as $V = [v_1, v_2, \ldots, v_n]$, to obtain the final visual feature vector representation:

$$F(x_t, v_i) = f\left(x_{\mathrm{t}}^{\mathrm{T}} W v_i\right), i \in [1, n], \quad (6)$$

$$\alpha_{v_i} = \frac{\exp\left(F(x_{\mathrm{t}}, v_i)\right)}{\sum_{i=1}^{n} \exp\left(F(x_{\mathrm{t}}, v_i)\right)} \quad (7)$$

$$x_{\mathrm{v}} = \sum_{i=1}^{n} \alpha_{v_i} v_i \quad (8)$$

After the aforementioned operations, x_{t0} representation of the original text and image text, the representation of visual entities in the image x_{e}, and the representation of visual feature vectors in the image x_{v}. These features model the semantic information of the input multimodal news from different perspectives, providing complementary information. We concatenate these features together to obtain the final multimodal representation of the news:

$$x = x_{\mathrm{t0}} \oplus x_{\mathrm{e}} \oplus x \quad (9)$$

where $\oplus$ denotes the concatenation operation.

3.4 Classification

After obtaining the multimodal representation x of the input news, we pass it through a fully connected layer and generate classification label distribution using a softmax layer:

$$p = \mathrm{softmax}\left(W_C x + b_C\right), \quad (10)$$

Where W_C and b_C are the model's parameters. We use cross-entropy as the model's loss function:

$$L = -\sum\left[y^f \log\, p^f + \left(1 - y^f\right)\log\left(1 - p^f\right)\right], \quad (11)$$

where y^f represents the true label of the sample, where 1 indicates the sample is fake news, and 0 indicates the sample is true news; p^f represents the predicted probability of the sample being fake news.

Table 1. Statistical indicators of the data set

Dataset	Training Set	Validation Set	Test Set	Total
Fake News	**2849**	**950**	**950**	4749
True News	**2879**	**950**	**950**	4779
Total	**5728**	**1900**	**1900**	9528

4. Experimental detail and result analysis

4.1 Dataset

In the current realm of false news research, there is a limited availability of publicly accessible multimodal datasets. As a result, in the subsequent experiments of this paper, the performance evaluation primarily focuses on the microblog dataset. This is due to the model's primary emphasis on the extraction and interaction of deep-level semantics from text and images, which is not significantly influenced by the specific linguistic forms of the text. Further validation of the model's impact on language forms will be conducted in future work.

The false news dataset utilized in this study was constructed by Shishah et al. [6] based on the Sina Weibo platform. The dataset encompasses all news messages officially certified as false on the Weibo official rumour reporting platform from May 2012 to January 2022, as well as microblog messages of authentic news collected from News Agency's hot news discovery system during the same period. To ensure dataset quality, steps were taken by Jin et al. to remove duplicate, excessively small, and irrelevant images, considering the presence of noise and redundancy on social media platforms

To enhance the assessment of the model's generalization capability across new news events, the data was clustered and subsequently divided at the event level, ensuring that the training, validation, and testing data do not contain news from the same event. Given the relatively small size of the overall dataset, a division ratio of 3:1:1 was employed to allocate the final training, validation, and testing sets. Detailed data metrics are presented in Table 1.

4.2 Experimental setup

This paper employs accuracy, F1-score in the false news category, precision, and recall as evaluation metrics. In the context of model implementation, we utilize the pre-trained ERNIE model obtained from the open-source Transformers project on GitHub [36]. During the fine-tuning of

VGG19, we apply data augmentation techniques, including image flipping, to enhance the model's generalization performance. Hyper parameters for the model include a maximum sentence length of 128, a batch size of 64, and the utilization of the ReLU function as the non-linear activation function. The optimization process is carried out using the Adam optimization method [37] to optimize the loss function.

4.3 Experiment 1: Comparison of false news detection performance

4.3.1. Comparison methods

To validate the effectiveness of the proposed method, we implemented three representative methods for performance comparison. The attRNN method was provided by the authors of the Shishah et al. [38], while the other methods were reproduced by the authors of this paper based on the descriptions provided in the respective papers.

4.3.1.1. Single-text modality:

a) **Text CNN:** This method utilizes a convolutional neural network (CNN) proposed by Galende et al. [38] for text classification. It employs three different sizes of convolutional kernels with heights of 3, 4, and 5, respectively. The number of filters for each kernel size is set to 100.
b) **BiLSTM+GAtt:** Recurrent neural networks, such as LSTM, are a classic modelling approach for text classification tasks. Sansonetti et al. [39] present a two-layer LSTM with a stacked attention mechanism is selected as a comparative method. The hidden units of the network are set to 128.
c) **BERT:** Pre-trained language models have shown superior performance in various natural language processing tasks in recent years. In this study, a fine-tuned BERT model on the task-specific dataset is employed for comparison. The pre-trained BERT model, "bert-base-word," used Shahbazi and Byun [36] and sourced from the open-source project Transformers on GitHub.
d) **ERNIE:** The _ne-tuned ERNIE model on the task-specific dataset is used as a comparative method. The pre-trained ERNIE model, "nghuyong/ernie- 1.0," utilized Shahbazi and Byun [36] and sourced from the open-source project Trans- formers on GitHub.

4.1.3.2. Single visual modality

a) **VGG19 [40]** In the current research on multimodal false news, VGG19 is widely used as a visual feature extractor. In this paper, the VGG19 model pre-trained on the ImageNet dataset (Bahurmuz et al. 2022) is _ne-tuned on the task-specific dataset.
b) **ResNet152 [33]** The ResNet152 model pre-trained on the ImageNet dataset is _ne-tuned on the task-specific dataset in this paper.

4.1.3.3. Multimodal fusion

a) **AttRNN:** Shishah et al. [6] proposes an attention-based recur- rent neural network (RNN) that integrates features from three modalities: text, visual, and social context. The text modality is modelled using LSTM, while the visual modality utilizes pre-trained VGG19 for feature extraction. To ensure fair comparison, the social feature processing part is removed in the specific implementation.
b) **EANN:** Ali et al. [7] introduces a neural network based on event adversarial mechanism. By incorporating an event classifier as an auxiliary task, the model learns multimodal features that are independent of the event. This model utilizes Text CNN and pre-trained VGG19 for text and visual feature extraction, respectively. The features from these two modalities are concatenated to form the multimodal representation of false news, which is then inputted into the false news classifier and event classifier.
c) **MVAE:** Shahid et al. [8] proposes a multi-task model that combines a multimodal variational autoencoder (VAE) with a false news detector. The text and image features are extracted separately using bidirectional LSTM and pre-trained VGG19, respectively. The concatenated features are encoded into an intermediate representation used for feature reconstruction and false news classification.
d) **KMGCN:** Ying et al. [27] introduces a knowledge- guided multimodal graph convolutional network. This method lever- ages external knowledge from a knowledge graph to extract concepts corresponding to named entities in the text. For each input multimodal news, a graph is constructed with nodes representing words in the text, concepts corresponding to text entities, and object names identified in the image. The nodes are initialized with pre-trained Word2Vec word embeddings, and the edge weights are set based on the Pointwise Mutual Information (PMI) between two words. Graph convolutional networks and max pooling are

Table 2. Performance comparison of different methods

Classification	Method	Accuracy	F1-Score	Precision	Recall
Single Text Modal	Textcnn	0.764	0.722	0.88	0.612
	Bilstm-Att	0.785	0.763	0.851	0.692
	BERT	0.83	0.798	0.977	0.675
	ERNIE	0.852	0.83	0.97	0.725
Single Vision Mode	VGG19	0.73	0.698	0.789	0.626
	Resnet152	0.688	0.675	0.705	0.647
Multimodal	Ttrnn	0.808	0.787	0.882	0.711
	EANN	0.803	0.776	0.899	0.682
	MVAE	0.797	0.787	0.827	0.751
	KMGCN	0.714	0.677	0.599	0.777
	Proposed	0.895	0.89	0.936	0.847

employed to obtain the graph representation for false news classification.

4.3.1.4. Result analysis

The results of the comparative experiments are presented in Table 2 and Figs. 2-4, from which the following conclusions can be drawn:

a) The effectiveness of our proposed method in enhancing the detection of false news is evident as it significantly surpasses the other comparative methods in classification accuracy. It is observed that our model can detect false news that is overlooked by existing methods by thoroughly exploring multimodal semantic clues, particularly evident in the recall of false news, where our method outperforms others by more than 7 percentage points.

b) Among the multimodal methods, KMGCN exhibits significantly lower performance compared to other methods. This can be attributed to the limited modelling capability of GCN in handling short texts such as microblogs, which consequently hampers the utilization of external knowledge. Additionally, KMGCN's reliance solely on object label information from images leads to inadequate semantic modelling of images.

c) Methods based on single-text modality demonstrate superior performance over those based on single-visual modality, highlighting the predominant role of textual cues in false news detection. Furthermore, multimodal methods outperform single-modality methods with identical subnetwork structures, emphasizing the complementary nature of text and image modalities in providing clues for false news detection. Notably, our proposed method exhibits a 4.3 percentage point improvement in accuracy compared to ERNIE, underscoring the significance of semantic features derived from images.

d) Within the single-text modality methods, pre-trained language models out- perform traditional text modelling methods such as CNN and RNN. This improvement stems from the greater modelling capacity of Transformers and the linguistic knowledge acquired from extensive pre-trained corpora. ERNIE demonstrates superior performance compared to BERT, indicating that the incorporation of entity concept knowledge enhances the semantic understanding of news, thereby elevating the effectiveness of false news detection.

4.4 Experiment 2: Ablation Analysis

4.4.1. Comparison methods

To investigate the impact of different model components on the experimental results, we designed five variations of the model for ablative analysis.

a) **ERNIE removal:** The bidirectional LSTM combined with attention mechanism replaces ERNIE for modelling text and image text. Pre-trained Word2Vec word vectors are used instead of the word vectors generated by ERNIE for representing visual entities.

b) **OCR text removal:** The extraction and processing of text from images are removed. In this case, the multimodal representation of input information is composed of the original text feature representation and the concatenation of visual feature vectors and visual entity vectors guided by the original text.

c) **Visual entity removal:** The extraction and processing of visual entities from images are removed. In this case, the multimodal representation of input information is composed of the joint representation of the original text and image text, as well as the

Table 3. Elimination analysis

Method	Accuracy	F1-Score	Precision	Recall
Proposed Approach	0.895	0.89	0.936	0.847
Remove ERNIE	0.806	0.799	0.83	0.771
Remove OCR text	0.873	0.872	0.877	0.866
Remove visual entities	0.877	0.87	0.929	0.817
Remove the eigen vectors	0.881	0.868	0.971	0.784
Remove the attention mechanism	0.881	0.87	0.963	0.793

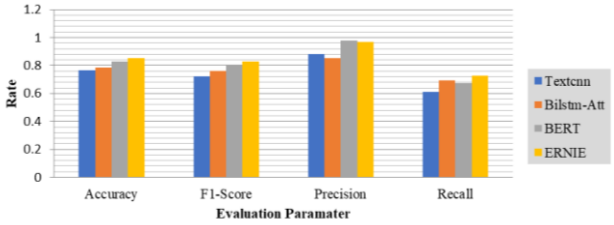

Figure. 2 Performance comparison single text modal

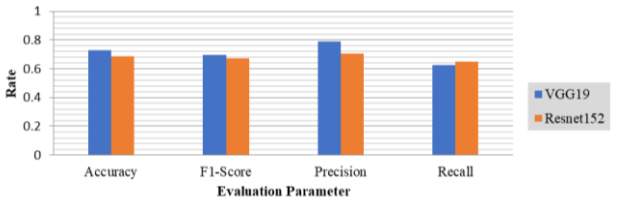

Figure. 3 Performance comparison single vision modal

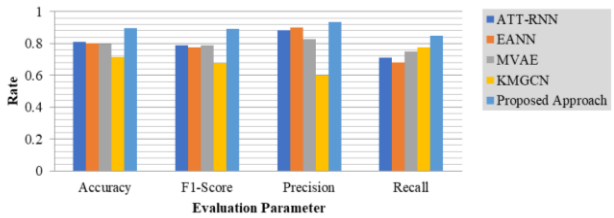

Figure. 4 Performance comparison multimodal

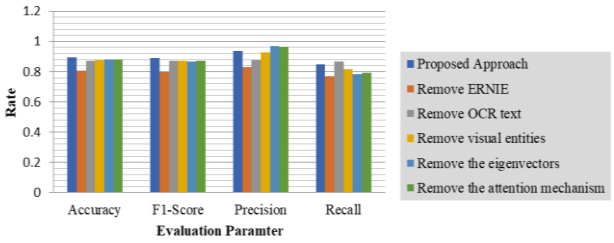

Figure. 5 Comparative elimination analysis

concatenation of visual feature vectors guided by the original text.

d) **Feature vector removal:** The processing of visual feature vectors from im-ages is removed. In this case, the multimodal representation of input information is composed of the joint representation of the original text and image text, as well as the concatenation of visual entity vectors guided by the original text.

e) **Attention mechanism removal:** The attention mechanism for visual entities and visual feature vectors guided by text is removed. In this case, multiple visual entity vectors and visual feature vectors are fused separately through averaging.

4.4.2. Results analysis

Table 3 and Fig. 5 presents the experimental results of the ablative analysis, from which two conclusions can be drawn:

a) Removing any component of the model results in a certain degree of decrease in classification accuracy, indicating the effectiveness of each model element.

b) Based on the extent of decrease in classification accuracy after removal, the importance of each model component can be ranked as follows:
ERNIE>Image Text>Visual Entities > Visual Feature Vectors = Attention Mechanism. This suggests that, for the task of fake news detection, text plays a more important role than images, and high-level semantic information in images is more crucial than low-level semantic information.

5. Conclusion

In response to the limited semantic understanding capabilities of existing methods for multimodal news, this paper presents a novel approach that enhances the semantic comprehension for detecting fake news across multiple modalities. By leveraging the vast amount of factual knowledge stored in external models, our method achieves a deeper understanding of the underlying semantics in multi-modal news. Through the extraction of distinct semantic levels of visual features and the utilization of a text-guided attention mechanism, we effectively integrate heterogeneous multimodal features. The experimental results demonstrate that our proposed method significantly surpasses the current state-of-the-art approaches in terms of accuracy, highlighting the effectiveness of our semantic-enhanced approach.

Conflicts of interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

Author contributions

Co-authors jointly formulated the research paper's conception and design in close collaboration. This encompassed the development of research questions, hypotheses, and the overall study design.

References

[1] A. Gupta, N. Kumar, P. Prabhat, S. Tanwar, G. Sharma, and P. N. Bokaro, "Combating Fake News: Stakeholder Interventions and Potential Solutions", *IEEE Access*, Vol. 10, pp. 78268-78289, 2022, doi: 10.1109/ACCESS.2022.3193670.

[2] H. Saleh, A. Alharbi and S. H. Alsamhi, "OPCNN-FAKE: Optimized Convolutional Neural Network for Fake News Detection", *IEEE Access*, Vol. 9, pp. 129471-129489, 2021, doi: 10.1109/ACCESS.2021.3112806.

[3] M. Kang, J. Seo, C. Park, and H. Lim, "Utilization Strategy of User Engagements in Korean Fake News Detection", *IEEE Access*, Vol. 10, pp. 79516-79525, 2022, doi: 10.1109/ACCESS.2022.3194269.

[4] D. Rohera et al., "A Taxonomy of Fake News Classification Techniques: Survey and Implementation Aspects", *IEEE Access*, Vol. 10, pp. 30367-30394, 2022, doi: 10.1109/ACCESS.2022.3159651.

[5] Y. Guo and W. Song, "A Temporal-and-Spatial Flow Based Multimodal Fake News Detection by Pooling and Attention Blocks", *IEEE Access*, Vol. 10, pp. 131498-131508, 2022, doi: 10.1109/ACCESS.2022.3229762.

[6] W. Shishah, "JointBert for Detecting Arabic Fake News", *IEEE Access*, Vol. 10, pp. 71951-71960, 2022, doi: 10.1109/ACCESS.2022.3185083.

[7] H. Ali, M. S. Khan, A. Alghadhban, M. Alazmi, A. Alzamil, K. Al-Utaibi, and J. Qadir, "All Your Fake Detector are Belong to Us: Evaluating Adversarial Robustness of Fake-News Detectors Under Black-Box Settings", *IEEE Access*, Vol. 9, pp. 81678-81692, 2021, doi: 10.1109/ACCESS.2021.3085875.

[8] W. Shahid, Y. Li, D. Staples, G. Amin, S. Hakak, and A. Ghorbani, "Are You a Cyborg, Bot or Human?—A Survey on Detecting Fake News Spreaders", *IEEE Access*, Vol. 10, pp. 27069-27083, 2022, doi: 10.1109/ACCESS.2022.3157724.

[9] Q. Liao, H. Chai, H. Han, X. Zhang, X. Wang, W. Xia, and Y. Ding, "An Integrated Multi-Task Model for Fake News Detection", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 34, No. 11, pp. 5154-5165, 1 Nov. 2022, doi: 10.1109/TKDE.2021.3054993.

[10] K. Xu, F. Wang, H. Wang, and B. Yang, "Detecting fake news over online social media via domain reputations and content understanding", *Tsinghua Science and Technology*, Vol. 25, No. 1, pp. 20-27, 2020, doi: 10.26599/TST.2018.9010139.

[11] P. Wei, F. Wu, Y. Sun, H. Zhou, and X. Y. Jing, "Modality and Event Adversarial Networks for Multi-Modal Fake News Detection", *IEEE Signal Processing Letters*, Vol. 29, pp. 1382-1386, 2022, doi: 10.1109/LSP.2022.3181893.

[12] M. Marra, M. Umer, S. Sadiq, A. Eshmawi, H. Karamti, A. Mohamed, and I. Ashraf, "Selective Feature Sets Based Fake News Detection for COVID-19 to Manage Infodemic", *IEEE Access*, Vol. 10, pp. 98724-98736, 2022, doi: 10.1109/ACCESS.2022.3206963.

[13] N. Ebadi, M. Jozani, K. K. R. Choo, and P. Rad, "A Memory Network Information Retrieval Model for Identification of News Misinformation", *IEEE Transactions on Big Data*, Vol. 8, No. 5, pp. 1358-1370, 1 Oct. 2022, doi: 10.1109/TBDATA.2020.3048961.

[14] H. Choi and Y. Ko, "Using Adversarial Learning and Biterm Topic Model for an Effective Fake News Video Detection System on Heterogeneous Topics and Short Texts", *IEEE Access*, Vol. 9, pp. 164846-164853, 2021, doi: 10.1109/ACCESS.2021.3122978.

[15] Ravish, R. Katarya, D. Dahiya, and S. Checker, "Fake News Detection System Using Featured-Based Optimized MSVM Classification", *IEEE Access*, Vol. 10, pp. 113184-113199, 2022, doi: 10.1109/ACCESS.2022.3216892.

[16] T. Jiang, J. P. Li, A. U. Haq, A. Saboor, and A. Ali, "A Novel Stacking Approach for Accurate Detection of Fake News", *IEEE Access*, Vol. 9, pp. 22626-22639, 2021, doi: 10.1109/ACCESS.2021.3056079.

[17] L. Wu, Y. Rao, C. Zhang, Y. Zhao and A. Nazir, "Category-Controlled Encoder-Decoder for Fake News Detection", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 35, No. 2, pp. 1242-1257, 1 Feb. 2023, doi: 10.1109/TKDE.2021.3103833.

[18] M. Umer, Z. Imtiaz, S. Ullah, A. Mehmood, G. S. Choi and B. W. On, "Fake News Stance

Detection Using Deep Learning Architecture (CNN-LSTM)", *IEEE Access*, Vol. 8, pp. 156695-156706, 2020, doi: 10.1109/ACCESS.2020.3019735.

[19] M. F. Mridha, A. J. Keya, M. A. Hamid, M. M. Monowar, and M. S. Rahman, "A Comprehensive Review on Fake News Detection With Deep Learning", *IEEE Access*, Vol. 9, pp. 156151-156170, 2021, doi: 10.1109/ACCESS.2021.3129329.

[20] S. Ni, J. Li, and H. Y. Kao, "MVAN: Multi-View Attention Networks for Fake News Detection on Social Media", *IEEE Access*, Vol. 9, pp. 106907-106917, 2021, doi: 10.1109/ACCESS.2021.3100245.

[21] D. Li, H. Guo, Z. Wang, and Z. Zheng, "Unsupervised Fake News Detection Based on Autoencoder", *IEEE Access*, Vol. 9, pp. 29356-29365, 2021, doi: 10.1109/ACCESS.2021.3058809.

[22] M. Abdulqader, A. Namoun and Y. Alsaawy, "Fake Online Reviews: A Unified Detection Model Using Deception Theories", *IEEE Access*, Vol. 10, pp. 128622-128655, 2022, doi: 10.1109/ACCESS.2022.3227631.

[23] P. Li, X. Sun, H. Yu, Y. Tian, F. Yao and G. Xu, "Entity-Oriented Multi-Modal Alignment and Fusion Network for Fake News Detection", *IEEE Transactions on Multimedia*, Vol. 24, pp. 3455-3468, 2022, doi: 10.1109/TMM.2021.3098988.

[24] P. K. Verma, P. Agrawal, I. Amorim, and R. Prodan, "WELFake: Word Embedding Over Linguistic Features for Fake News Detection", *IEEE Transactions on Computational Social Systems*, Vol. 8, No. 4, pp. 881-893, Aug. 2021, doi: 10.1109/TCSS.2021.3068519.

[25] G. Shan, B. Zhao, J. R. Clavin, H. Zhang and S. Duan, "Poligraph: Intrusion-Tolerant and Distributed Fake News Detection System", *IEEE Transactions on Information Forensics and Security*, Vol. 17, pp. 28-41, 2022, doi: 10.1109/TIFS.2021.3131026.

[26] V. I. Ilie, C. O. Truică, E. S. Apostol, and A. Paschke, "Context-Aware Misinformation Detection: A Benchmark of Deep Learning Architectures Using Word Embeddings", *IEEE Access*, Vol. 9, pp. 162122-162146, 2021, doi: 10.1109/ACCESS.2021.3132502.

[27] L. Ying, H. Yu, J. Wang, Y. Ji, and S. Qian, "Fake News Detection via Multi-Modal Topic Memory Network", *IEEE Access*, Vol. 9, pp. 132818-132829, 2021, doi: 10.1109/ACCESS.2021.3113981.

[28] B. Han, X. Han, H. Zhang, J. Li, and X. Cao, "Fighting Fake News: Two Stream Network for Deepfake Detection via Learnable SRM", *IEEE Transactions on Biometrics, Behavior, and Identity Science*, Vol. 3, No. 3, pp. 320-331, 2021, doi: 10.1109/TBIOM.2021.3065735.

[29] L. Ying, H. Yu, J. Wang, Y. Ji, and S. Qian, "Multi-Level Multi-Modal Cross-Attention Network for Fake News Detection", *IEEE Access*, Vol. 9, pp. 132363-132373, 2021, doi: 10.1109/ACCESS.2021.3114093.

[30] M. Babaei, J. Kulshrestha, A. Chakraborty, E. M. Redmiles, M. Cha, and K. P. Gummadi, "Analyzing Biases in Perception of Truth in News Stories and Their Implications for Fact Checking", *IEEE Transactions on Computational Social Systems*, Vol. 9, No. 3, pp. 839-850, 2022, doi: 10.1109/TCSS.2021.3096038.

[31] X. Dong, U. Victor, and L. Qian, "Two-Path Deep Semisupervised Learning for Timely Fake News Detection", *IEEE Transactions on Computational Social Systems*, Vol. 7, No. 6, pp. 1386-1398, 2020, doi: 10.1109/TCSS.2020.3027639.

[32] T. H. Do, M. Berneman, J. Patro, G. Bekoulis, and N. Deligiannis, "Context-Aware Deep Markov Random Fields for Fake News Detection", *IEEE Access*, Vol. 9, pp. 130042-130054, 2021, doi: 10.1109/ACCESS.2021.3113877.

[33] N. R. D. Oliveira, D. S. V. Medeiros, and D. M. F. Mattos, "A Sensitive Stylistic Approach to Identify Fake News on Social Networking", *IEEE Signal Processing Letters*, Vol. 27, pp. 1250-1254, 2020, doi: 10.1109/LSP.2020.3008087.

[34] G. Shrivastava, P. Kumar, R. P. Ojha, P. K. Srivastava, S. Mohan, and G. Srivastava, "Defensive Modeling of Fake News Through Online Social Networks", *IEEE Transactions on Computational Social Systems*, Vol. 7, No. 5, pp. 1159-1167, 2020, doi: 10.1109/TCSS.2020.3014135.

[35] N. O. Bahurmuz, G. A. Amoudi, F. A. Baothman, A. T. Jamal, H. S. Alghamdi, and A. M. Alhothali, "Arabic Rumor Detection Using Contextual Deep Bidirectional Language Modeling", *IEEE Access*, Vol. 10, pp. 114907-114918, 2022, doi: 10.1109/ACCESS.2022.3217522.

[36] Z. Shahbazi and Y. C. Byun, "Fake Media Detection Based on Natural Language Processing and Blockchain Approaches", *IEEE*

Access, Vol. 9, pp. 128442-128453, 2021, doi: 10.1109/ACCESS.2021.3112607.

[37] Z. Jin, J. Cao, Y. Zhang, J. Zhou, and Q. Tian, "Novel Visual and Statistical Image Features for Microblogs News Verification", *IEEE Transactions on Multimedia*, Vol. 19, No. 3, pp. 598-608, March 2017, doi: 10.1109/TMM.2016.2617078.

[38] B. A. Galende, G. H. Peñaloza, S. Uribe, and F. Á. García, "Conspiracy or Not? A Deep Learning Approach to Spot It on Twitter", *IEEE Access*, Vol. 10, pp. 38370-38378, 2022, doi: 10.1109/ACCESS.2022.3165226.

[39] G. Sansonetti, F. Gasparetti, G. D'aniello, and A. Micarelli, "Unreliable Users Detection in Social Media: Deep Learning Techniques for Automatic Detection", *IEEE Access*, Vol. 8, pp. 213154-213167, 2020, doi: 10.1109/ACCESS.2020.3040604.

[40] J. C. Neves, R. Tolosana, R. V. Rodriguez, V. Lopes, H. Proença, and J. Fierrez, "GANprintR: Improved Fakes and Evaluation of the State of the Art in Face Manipulation Detection", *IEEE Journal of Selected Topics in Signal Processing*, Vol. 14, No. 5, pp. 1038-1048, 2020, doi: 10.1109/JSTSP.2020.3007250.