Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2023.037035
Article

Fake News Detection Based on Multimodal Inputs

Zhiping Liang*

Unversity of Melbourne, Melbourne, VIC 3010, Australia
*Corresponding Author: Zhiping Liang. Email: zhipliang@student.unimelb.edu.au
Received: 20 October 2022; Accepted: 15 January 2023

Abstract: In view of the various adverse effects, fake news detection has become an extremely important task. So far, many detection methods have been proposed, but these methods still have some limitations. For example, only two independently encoded unimodal information are concatenated together, but not integrated with multimodal information to complete the complementary information, and to obtain the correlated information in the news content. This simple fusion approach may lead to the omission of some information and bring some interference to the model. To solve the above problems, this paper proposes the Fake News Detection model based on BLIP (FNDB). First, the XLNet and VGG-19 based feature extractors are used to extract textual and visual feature representation respectively, and BLIP based multimodal feature extractor to obtain multimodal feature representation in news content. Then, the feature fusion layer will fuse these features with the help of the cross-modal attention module to promote various modal feature representations for information complementation. The fake news detector uses these fused features to identify the input content, and finally complete fake news detection. Based on this design, FNDB can extract as much information as possible from the news content and fuse the information between multiple modalities effectively. The fake news detector in the FNDB can also learn more information to achieve better performance. The verification experiments on Weibo and Gossipcop, two widely used real-world datasets, show that FNDB is 4.4% and 0.6% higher in accuracy than the state-of-the-art fake news detection methods, respectively.

Keywords: Natural language processing; fake news detection; machine learning; text classification

1 Introduction

With the rapid development of the Internet, social media, news websites, online newspapers, and other media have become closer to people. People can freely collect information and communicate with each other on these platforms, but it also brings the side effect of the rapid spread of fake news at the same time, which brings great harm to society. Therefore, to effectively reduce the impact of fake news, society needs an efficient and accurate method to identify and control it. Various methods

have been proposed to identify fake news. Current research on fake news detection can be categorized into two types: traditional machine learning methods and deep learning methods. Many traditional machine learning methods rely on using hand-crafted features to train the supervised classifier, such as user features, text content and propagation modes. Or use social contextual information to identify fake news [1–3]. Due to the complexity of the news content, it is difficult to capture all the information in the news with these handmade features. With the rapid development of this field of deep learning in recent years, the challenge of fake news detection has entered a new phase. By using deep learning neural networks, the researcher can enhance the ability to extract relevant features and more trivial information from news content to train the classifier.

However, most current methods emphasize unimodal tasks, such as text or images related to the news. Table 1 [4] shows the performance of unimodal and multimodal models in the fake news detection task. The unimodal models include a model that extracts textual features using Bidirectional Long Short-Term Memory (Bi-LSTM) and a model that extracts visual features using VGG-19. The features extracted by these two models are fed into a 32-dimensional fully connected layer for making the prediction. While the multimodal models include Recurrent Neural Network (RNN) with an attention mechanism (att-RNN) [5], Event Adversarial Neural Network (EANN) [6], and Multimodal Variational Autoencoder (MVAE) [4]. Based on the results of Table 1, it is known that the unimodal model performs worse than the multimodal model in the field of fake news detection. Among them, the Similarity-Aware FakE news detection method (SAFE) [7] proposed by Zhou et al. have shown superior performance on fake news detection tasks. Based on the above information, this paper decides to focus on the area of multimodal fake news detection. Nevertheless, most of these models encode information from different modalities independently and then classify the input according to the concatenated feature representation, so the models will likely miss correlated information between different modalities. Besides, there will be cases where the contextual information and the hierarchical semantics in multimodal features are not fully utilized.

Table 1: Classification results on Weibo datasets [4]

	Method	Accuracy
Weibo	Textual	0.643
	Visual	0.608
	att-RNN	0.772
	EANN	0.782
	MVAE	0.824

To solve the problems mentioned above, this paper proposes FNDB, a multimodal fake news detection network based on the pre-trained BLIP model. The FNDB is a multimodal fake news detection model and composed of three modules: (1) The feature extraction layer which is used to extract different types of feature representation. (2) The feature fusion layer for a more effective fusion of the extracted features for information complementation. (3) The fake news detector that use fused features to discriminate whether the news story is fake or not.

The main contributions of this work are as follows:

- This paper applies a multimodal pre-trained model to the field of fake news detection.
- This paper proposes a novel neural network model FNDB—a multimodal fake news detection network based on the pre-trained BLIP model for multimodal fake news detection. By

incorporating multimodal features extracted from the pre-trained BLIP model and using the cross-modal attention module to fuse multimodal, textual, and visual features with each other, information between different modalities can complement each other and help the model better understand the relevant information of different modalities.
- This paper performs experiments on two real-world datasets—Weibo and Gossipcop. The results demonstrate that FNDB can outperform state-of-the-art fake news detection methods by 4.4% and 0.6% in terms of accuracy on the two datasets, respectively.

The rest of this paper is organized as follows: Section 2 is a literature review on recent work related to fake news detection, as well as describes some applications of attention mechanisms in fake news detection. In Section 3, the proposed model and its different components are presented. In Section 4, the dataset, experimental setup, and detailed analysis of the experimental results, including ablation experiments, are discussed. Finally, in Section 5, based on the research presented in this paper, a brief conclusion is provided.

2 Related Works

2.1 Fake News Detection

A large number of research have been reported to address the important challenge of fake news detection, while most of the existing research regards fake news detection as a binary classification task. Many of these studies have tried to solve this issue by training relevant fake news detection classifiers using features extracted from news textual content or news image. As mentioned above, the current research directions can be broadly divided into traditional machine learning methods and deep learning methods. In the early research stages, researchers trained classifiers with hand-crafted or selected features and applied them to the fake news detection task. For example, some studies have used sentiment signals extracted from texts to discriminate fake news [8,9] or used word-level features and writing styles to make predictions [10]. However, hand-crafted or manual selection of various features is a time-consuming task, which cannot effectively extract deep semantic information from the news.

Nowadays, a lot of neural network methods have been applied to the field of fake news detection. For instance, Convolutional Neural Networks (CNN) are used to extract textual features from news texts [11], or recurrent neural networks are used to extract latent features from texts [12] to detect fake news. The propagation path of news is also modeled as a multivariate time series, and a combined RNN and CNN model is used to detect fake news [13]. In addition to applying neural networks to textual content for fake news detection, some research also focused on images in the news, like distinguishing fake news based on the clarity of images in the news [14]. Although these methods perform well in the fake news detection tasks, they all focus on only one type of modal information.

Through the great succession of deep learning methods in learning visual and textual feature representations, some researchers have realized that the multimodal information contained in the news, such as images and text, can complement each other and help the model to analyze the input content [15,16]. As a result, more research groups are proposing to use multimodal deep learning methods in fake news detection tasks. For example, the MVAE and SAFE models mentioned in the introduction section. In addition to these models, the EANN proposed by Wang et al. to learn event invariant features of news representing different topics and domains is impressive [6]. Although multimodal detection methods have achieved satisfactory performance in the field of fake news detection, there are still some issues that need attention. For example, the correlation between news textual content

and images is not sufficiently explored. Moreover, multimodal contextual information that is not processed or simply processed may not be sufficient to exploit the multimodal contextual information and hierarchical semantics of the text content.

In this context, the introduction of the Contrastive Language-Image Pretraining (CLIP) model [17], was designed to bridge the gap between computer vision and natural language processing. And the BLIP model [18], which uses the Multimodal mixture of Encoder-Decoder (MED), has provided a new way forward for vision-language tasks. These large models, pre-trained with large scales data, have sufficient ability to learn multimodal relationships between image-text pairs and support various downstream tasks. The BLIP model currently achieves state-of-the-art results for a wide range of vision-language tasks. In addition, the BLIP model exhibits strong generalization capabilities.

2.2 Attention Mechanisms

Attention mechanisms have been proven to be effective in natural language processing tasks. Many models have introduced attention mechanisms or made various adaptations to improve their performance. In fake news detection, many methods using attention mechanisms or other adaptations of attention mechanisms have naturally been proposed. For example, Jin et al. [5] created att-RNN and combined textual, visual, and social contextual features by using attention mechanisms. The multimodal fake news detector proposed by Qian et al. [19] uses the multimodal contextual attention mechanism to fuse textual and visual features of posts efficiently. Shu et al. use the sentence-comment co-attention mechanism to learn the correlation between news content and related comments [20], and Chen et al. [21] assisted the model in learning a series of hidden representations of posts by using a soft attention mechanism. These studies show that the attention mechanism can be applied to fake news detection tasks.

3 Methodology

This paper proposes a novel multimodal method, the FNDB model to address the problem of fake news detection. The basic idea of FNDB is to supply information on independent encoding features with information on image-text pairs contained in multimodal features. The overall architecture of the FNDB model is shown in Fig. 1. It has three main components: (1) Feature extraction layer, which extracts the information from news articles and images by using three different sub-modules. (2) Feature fusion layer, it is using a cross-modal attention module to process the output of the feature extraction layer, and the information between multiple feature representations can be complementarily fused to facilitate classification. (3) Fake news detector, which uses the fused feature representation (latent vector) to predict whether the news is fake or not.

3.1 Feature Extraction Layer

The feature extraction layer consists of 3 different submodules. The three submodules are used to encode the input news article, the image related to the news, and image-text pairs to obtain different types of feature representation.

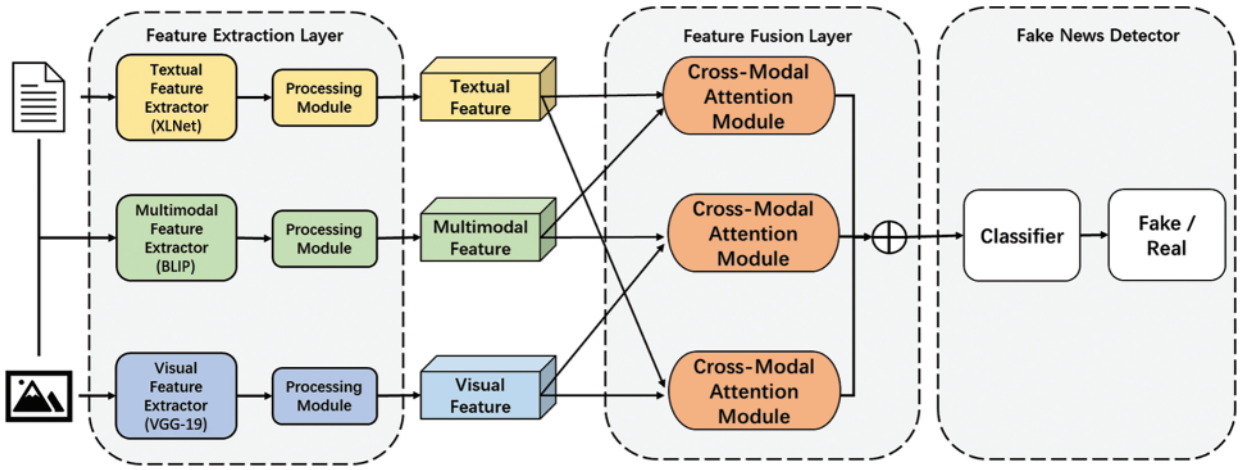

Figure 1: The architecture of the proposed FNDB method. Different modalities of news are extracted by 3 feature extractors and processed by the processing module. A cross-modal attention module is used to fuse these features to adapt the classifier to classify fake news

3.1.1 Textual Feature Extractor

The news is composed of a varying number of sentences. Most of the text content-based fake news detection approaches use traditional word vector models, which perform well in the modal analysis of unambiguous sentences. To accurately extract the textual feature representation within the news article precisely, the pre-trained XLNet model is used as the core of the textual feature extractor. Because the XLNet model with the permutation language model as the core idea can obtain broader contextual information.

The input of the textual feature extractor is the sequential list of sentences. These sentences are obtained by segmenting the entire news article with the sent tokenizer, i.e., $N_t = \{S_1, S_2 \ldots S_n\}$, where N_t is the collection of the processed sentences, n is the number of sentences, and each sentence $S_i \in N$ in the new article. The sentences processed by the sent tokenizer are fed into the pre-trained XLNet model for feature extraction and generating textual embeddings corresponding to the input. The overall textual embeddings of the news content will be:

$$\mathrm{E}_{\mathrm{t}} = \{E_1, E_2 \ldots E_i\} = XLNet(N_t) \tag{1}$$

The E_t is the collection of the textual embeddings of the news content. E_i is the hidden representation corresponding to sentence i generated from the pre-trained XLNet model. Since the length of each news text is different, the length of news textual embedding will be standardized, and those exceeding the set value will be truncated, and those less than the set value will be filled with zero matrices. To maintain the quality of textual embedding, the textual feature extractor will load XLNet model that has been pre-trained on different language corpora. Because the language of the datasets used in this paper is not the same. When FNDB conducts experiments on the Gossipcop dataset, since the language of the dataset is English, it loads the pre-trained "xlnet-base-cased" model trained on the English corpus [22]. When FNDB conducts experiments on the Weibo dataset, since the language of the dataset is Chinese, it loads the "hfl/chinese-xlnet-mid" model trained on the Chinese corpus [23]. Appropriate pre-trained models are selected based on the language to ensure the quality of the extracted features and the model's performance.

3.1.2 Visual Feature Extractor

Considering that the images in the news contain information that can be used to help identify whether the news is fake or not, a visual feature extractor will be set up to do the job of feature extraction, since some news content may contain more than one image. In this case, the top image in the news article will be chosen as the input of the visual feature extractor, based on the assumption that the top image in the news content is highly relevant to the current content. Aiming to extract features from the top image, this paper employs the VGG-19 model as the core of the visual feature extractor. This pre-trained convolutional network architecture has been trained over the ImageNet database, and the output of the third last layer of VGG-19 convolutional network is used as the initial visual representation. The dimension of the initial visual representation needs to be adjusted for subsequent fusion operations. The adjustment operation will be done by the processing module in the feature extraction layer. The final visual feature representation that matches the news content will be represented by Eq. (2), where E_v is the final visual representation, E_{VGG} is the visual feature representation extracted from the VGG-19 convolutional network, W is the weight matrix in the fully connected layer and σ is the activation function used.

$$\mathrm{E_v} = \sigma(W \cdot E_{VGG}) \tag{2}$$

3.1.3 Multimodal Feature Extractor

In addition to containing meaningful semantic information, multimodal features also reflect the correlation between the two modalities. Although XLNet and VGG-19 models are pre-trained on unimodal tasks and can extract semantic features of their respective modalities, they also focus on more trivial clues in the respective modal data. However, there are still significant cross-modal semantic differences between these independently encoded textual and visual feature representations. If two independently encoded feature representations are directly concatenated, the model will not be able to learn sufficient multimodal semantics from the concatenated feature representations to discriminate the input data. Most of the current approaches also use this form to obtain multimodal information in news and use it to identify fake news.

If the multimodal information can be directly extracted from the news content, the information in the multimodal features can be used to a certain extent to complement the information in the other unimodal features, so that the model can better learn the various features in the news content. To extract the multimodal feature from the news content, the BLIP model is used as the core of the multimodal feature extractor. The BLIP model, as a multimodal model which pre-trained on a large-scale of image-text pairs, has sufficient capacity to perform the task of extracting multimodal features from the input news content, compared to various popular unimodal pre-trained models. That multimodal feature will be able to cooperate with other unimodal features, so that the model can better understand the feature representation of the news content for more accurate identification.

3.1.4 Processing Module

After the feature extraction steps, the model yields three feature representations of different modalities and dimensions. In order to make these feature representations suitable for the subsequent feature fusion operations, each feature extractor is followed by a separate processing module to process the feature representation. These processing modules are composed of Multi-Layer Perceptrons (MLP), which are mainly used to adjust the dimension of rough features provided by feature extractors, ensure that the final dimensions of the three feature representations are consistent, and help filter out redundant information. These modules have subtle structural differences and do not share weights

with each other. For example, textual and multimodal features need to be flattened by a Flatten layer before they can be adjusted. In addition to the Flatten layer, each processing module contains three groups of layers, each group including a fully connected layer that applies the ReLU activation function, a batch normalization layer, and a dropout layer.

3.2 Feature Fusion Layer

Through processing, the feature extraction layer produces three different feature representations with the same dimensionality, i.e., a textual feature representation, a visual feature representation, and a multimodal feature representation. However, as mentioned above, if these feature vectors are simply concatenated together, there may be a lack of correlation between the features. It cannot provide reliable information to the model to assist in the identification task, as independently encoded feature representations (such as textual and visual feature representation) lack cross-modal relationships. If the model uses these features directly, there may be some noise and performance is affected. To solve this problem and to effectively accomplish information complementarity between features. In this paper, a cross-modal attention module, which was inspired by recent research [24], is used in the feature fusion layer. This novel module can cooperatively realize feature learning and correspondence, facilitating the interaction between the various feature representation. For example, the textual feature representation, R_t generated by the XLNet model can be complemented with the multimodal feature representation, R_m generated by the BLIP model by performing the following operations in the feature fusion layer, as shown in Eq. (3).

$$\begin{aligned}\text{Cross} - \text{modal Feature} = &concatenation\,(Multi - head\ attention\,(R_t, R_m),\ Multi \\ &-head\ attention\,(R_m, R_t))\end{aligned} \tag{3}$$

The cross-modal attention module helps the model capture information's correspondence in different modalities. Unlike the self-attention mechanism, the cross-modal attention module can assist in establishing the correspondence between different modal features. When two different modal features are input to the cross-modal attention module, they will be fused by cross-matching through a multi-head attention mechanism [25] and two sets of adjusted feature vectors will be obtained. These vectors will be concatenated together to form a cross-modal feature. The three modal features in FNDB are combined in a group of two modalities to obtain multiple feature pairs and input into the cross-modal attention module to generate three sets of cross-modal features. Finally, the cross-modal feature vectors are concatenated and used as input to the fake news detector. The model is able to learn from the cross-modal features and make more accurate judgments.

3.3 Fake News Detector

When the feature fusion layer finishes fusing the outputs of the feature extraction layer, it gets three sets of cross-modal features that are concatenated as the final news content representation, X, as shown in Eq. (4).

$$\begin{aligned}\text{X} = &Cross - modal\ feature(Text\&Multimodal) \oplus Cross - modal\ feature(Text\&Image) \oplus Cross \\ &- modal\ feature(Image\&Multimodal)\end{aligned} \tag{4}$$

The cross-modal feature is the output processed by the feature fusion layer, and the content inside the cross-modal feature is the input of the feature fusion layer, e.g., *Cross − modal Feature* (*Text*&*Multimodal*) is the fused feature representation obtained by inputting textual and multimodal

features to the feature fusion layer. These concatenated feature vectors are input to the fake news detector for the final identification task. The fake news detector consists of two fully connected layers. The last fully connected layer with the SoftMax activation function applied, will act as a classification layer to predict the input and obtain the specific labels of the news content.

4 Experiment

4.1 Dataset

4.1.1 FakeNewsNet

The dataset is from the FakeNewsNet repository [26] (https://github.com/KaiDMML/FakeNewsNet), which contains datasets from two different domains—the political domain and the entertainment domain. The news in these datasets and the labels corresponding to the news content were collected from two websites, PolitiFact and GossipCop. PolitiFact focuses on political news, while GossipCop focuses on entertainment news. Experts review the labels corresponding to the news in the dataset in the respective fields, which means the data is reliable. In addition to the actual news text and news-related images, FakeNewsNet also provides a lot of additional information, such as the social background of the user and space information. In this paper, the Gossipcop dataset is chosen to conduct experiments.

4.1.2 Weibo

The Weibo dataset was published by Jin et al. [5] and used widely in several multimodal fake news detection tasks. The news data contained in the dataset were collected from the authoritative news sources in China, such as the Sina Weibo social platform and the Xinhua News Agency. The news data in these datasets are collected and then verified by Weibo's official disinformation system and Xinhua News Agency.

This paper uses the traditional dataset partitioning ratio, which was divided into three sub-datasets, the training set, the validation set and the test set. The ratio of the three sub-datasets is 7:1:2. The information about those datasets is shown in Table 2, and the values in the table indicate the number of samples that are suitable for use after the data has been processed.

Table 2: The statistics of the real-world datasets

Method	Weibo	Gossipcop
#Of fake news	4121	2549
#Of real news	1054	10206

4.2 Implementation Details

News textual content and related images are inputs to the feature extraction layer to complete the extraction steps. When extracting textual feature representations, the textual feature extractor loads different pre-trained XLNet models to maintain the quality of the extracted features since the language of the datasets used in the experiments is different. For the Gossipcop dataset (English), FNDB load the pre-trained "xlnet-base-cased" model. For the Weibo dataset (Chinese), FNDB load the pre-trained "hfl/chinese-xlnet-mid" model. To extract the visual feature representation, this paper employs the VGG-19 convolutional network which is pre-trained on ImageNet, and the initial visual feature is the output of the third last layer of this model. Finally, for the multimodal features, the news

content will be organized as image-text pairs. And put it into the pre-trained BLIP model to complete the multimodal feature extraction. Since the BLIP model uses Bidirectional Encoder Representations from Transformers (BERT) to extract textual features. To ensure the quality of the features, BERT models that have been pre-trained in different corpora should be used, according to the language of the dataset. Depending on the language of the dataset, BLIP will load different pre-trained BERT models [27], such as load the "bert-base-chinese" model for the Weibo dataset and load the "bert-base-uncased" model for the Gossipcop dataset.

After the feature extraction step, three different modal features were obtained. These feature representations passed through the processing modules with no shared weights before being fed to the feature fusion layer. These processing modules contains three groups of layers, each group including a fully connected layer that applies the ReLU activation function, a batch normalization layer, and a dropout layer. The size for each fully connected layer is 1000, 500 and 100. These fully connected layers apply L2 regularization with a value of 0.01 that imposes a penalty on the weights of the network layer. And perform the kernel initialization by using the he_normal mechanism. With these processing modules, the dimension of all feature representations will be set to 100. To enhance the generalization capability of the model, the module has been set up the dropout layers with parameters 0.5, 0.3 and 0.2 for each layer.

The final processed feature representations are fused using the cross-modal attention module to enable a certain degree of information complementarity between these features. As a result, the model can make more accurate decisions. For the multi-head attention mechanism inside the cross-modal attention module, the number of heads is set to 4. The size of each attention head for query and key is set to 50. The output of the feature fusion layer will be concatenated and used as input to the fake news detector. The fake news detector consists of a fully connected layer that applies the ReLU activation function and a classification layer using the SoftMax activation function. The parameters of the first fully connected layer apply L2 regularization with a value of 0.01, and the size is set to 50. During the training period, the model applies an early stopping mechanism to prevent overfitting of the model.

4.3 Evaluation Metrics

Since fake news detection can be considered a binary classification task, this paper uses accuracy as the main evaluation metric to measure the overall performance of the FNDB model in the fake news detection task. However, considering the possible class imbalance and insufficient data in the dataset, this paper also applies the precision, recall and F1-score to evaluate the performance of FNDB in two classes, i.e., real news and fake news. The precision and recall are used to display which classes FNDB predicts correctly. And which are wrong when identifying real and fake news. And F1-score is the harmonic mean of the model's precision and recall, which can be used to measure the overall performance of FNDB in different classes. With this setup, FNDB can be evaluated more comprehensively.

4.4 Result

In this section, the model proposed in this paper is compared with some state-of-the-art multi-modal fake news detection methods in terms of accuracy, precision, recall and F1-score. Considering that two different datasets are used in this paper, the choice of the model approach will be different. For the Weibo dataset, the FNDB model proposed in this paper will be compared with the following six fake news detection methods. These methods are briefly described below.

- EANN [6], An end-to-end event adversarial neural network framework. It has the ability to learn transferable representations for unseen events to identify fake news.
- MVAE [4], a multimodal fake news detection model using a bimodal variational autoencoder with the capability of capturing multimodal representations.
- Multimodal Knowledge-aware Event Memory Network (MKEMN) [28], a multimodal neural network, through two techniques, Multimodal Knowledge-aware Network (MKN) and Event Memory Network (EMN), this model have the ability to identify fake news by exploiting the connection between multimodal content and external knowledge.
- SAFE [7], a multimodal fake news detection model that can fed the relevance between news textual and visual information to fake news classifier.
- Multimodal Consistency Neural Network (MCNN) [29], a multimodal approach that uses the social media information to identify fake news while considering the consistency of multimodal data.
- CAFE [30], an ambiguity-aware multimodal fake news detection method. This method can learn unimodal features and cross-modal correlations. And this model has the ability to correct the misclassifications caused by modal differences.

For the Gossipcop dataset in the FakeNewsNet repository, the FNDB model proposed in this paper will be compared with the following five fake news detection methods. Since SAFE and CAFE have already been introduced, only three other multimodal fake news detection methods are described below.

- Long Short-Term Memory (LSTM) with self-attention mechanism (LSTM-ATT) [31], an attention-based LSTM model with 134 hand-selected features extracted from each article for the fake news detection task.
- Spotfake+ [32], uses VGG-19 and XLNet to extract image and textual features, respectively, and simply concatenates the two features. The concatenated features are used in the fake news detection task.
- DistilBert [33], the model takes into account the correlation between user-generated content and user-shared content, and guides model learning through potential representations in news articles and user-generated content.

The detailed comparison results will be shown in the following table, which contains the performance of several state-of-the-art multimodal fake news detection models in the field of fake news detection.

The table above shows the performance of FNDB on different datasets. FNDB can achieve the highest accuracy of 88.8% on the Weibo dataset and 87.3% on the Gossipcop dataset. Its performance exceeds the current state-of-the-art method by 4.4% and 0.6% respectively. Moreover, Table 3 shows that the FNDB achieves the first and the second highest performance on both datasets in terms of precision in distinguishing real news, indicating that the FNDB can identify real news accurately to some extent. Based on these experimental results, it can be shown that FNDB is a reliable method for detecting fake news and can discriminate between real and fake news in a given multilingual and multidomain news. Compared with other state-of-the-art multimodal fake news detection methods, the accuracy of FNDB improves on the Weibo and Gossipcop datasets. It also demonstrates that FNDB can enhance the understanding and learning ability of the model for multimodal information by complementing the information between different modalities, enabling it to perform better in the fake news detection task. That result proves the effectiveness of FNDB in the task of multimodal fake news detection, and Table 3 also shows that FNDB is effective on the Chinese news dataset. The selected fake

news detection methods, such as EANN, MVAE, and SpotFake+, rely on fused features that are simply concatenated together when identifying fake news. Although these fused features perform well in the identification task, they do not allow the model to focus on the correlation between two unimodal features because the textual and visual features, which are separately encoded, are not in the same semantic space. In contrast, CAFE, a state-of-the-art multimodal fake news detection model, uses cross-modal alignment to train an encoder that maps text and images into the same semantic space and fuses them for classification. This approach facilitates the process of information complementation among different modalities to a certain extent. However, the encoder does not encode well due to the limitation of the number of datasets and the quality of the training labels. The semantic gap between the textual and visual features obtained by the model is still large. Based on these analyses, the FNDB model can obtain multimodal features generated in the same semantic space with rich semantic information by a multimodal feature extractor in the feature extraction layer. The multimodal features obtained by the multimodal feature extractor can provide complementary information to other unimodal features in the FNDB model. The cross-modal attention module in the feature fusion layer enables the FNDB model to perform the feature fusion operation better. These fused features are used further to improve the classification performance of the fake news detector.

Table 3: Performance comparison between FNDB and other baseline methods

			Fake news			Real news		
	Method	Accuracy	Precision	Recall	F1-score	Precision	Recall	F1-score
Weibo	EANN	0.795	0.806	0.795	0.800	0.752	0.793	0.804
	MVAE	0.824	0.854	0.769	0.809	0.802	**0.875**	0.837
	MKEMN	0.814	0.823	0.799	0.812	0.723	0.819	0.798
	SAFE	0.816	0.818	0.815	0.817	0.816	0.818	0.817
	MCNN	0.823	0.858	0.801	0.828	0.787	0.848	0.816
	CAFE	0.840	0.855	0.830	0.842	0.825	0.851	**0.837**
	FNDB	**0.888**	**0.891**	**0.972**	**0.930**	**0.869**	0.610	0.717
Gossipcop	SAFE	0.838	0.758	0.558	0.643	0.857	0.937	0.895
	SpotFake+	0.856	–	–	–	–	–	–
	LSTM-ATT	0.842	**0.845**	**0.842**	**0.844**	0.839	0.842	0.821
	DistilBert	0.857	0.805	0.527	0.637	0.866	0.960	0.911
	CAFE	0.867	0.732	0.490	0.587	**0.887**	0.957	0.921
	FNDB	**0.873**	0.790	0.440	0.565	0.883	**0.973**	**0.926**

4.5 Ablation Experiments

This section explores the impact of each module in FNDB on the structure of the experiments. In each class of experiments, different components are removed, and the model is trained from scratch to evaluate the model's performance under different settings. FNDB's various model variants will be broadly classified into four classes, with specific configurations as follows.

Class One: This class variant model only uses one type of modal feature for fake news detection experiments. It will be compared with class two and class four models. This task aims to

explore whether providing more information to the model can improve the model's classification performance.

- FNDB-OT: Remove the multimodal feature extractor, visual feature extractor and cross-modal attention module. And only use the textual feature to classify.
- FNDB-OI: Remove the multimodal feature extractor, textual feature extractor and cross-modal attention module. And only use the visual feature to classify.
- FNDB-OM: Remove the textual feature extractor, visual feature extractor and cross-modal attention module. And only use the multimodal feature to classify.

Class Two: Class Two model chooses two features from the following three features: textual, visual, and multimodal features to conduct fake news detection experiments. The model in Class Two will be compared with Class One and four to investigate whether providing more information can improve the model's classification performance. In addition, it will be compared with class three to investigate whether adding a cross-modal attention module to the model can improve the model performance while using the same information of both modal features.

- FNDB-NTNA: Remove the textual feature extractor and cross-modal attention module. And only use the visual and multimodal features to classify.
- FNDB-NINA: Remove the visual feature extractor and cross-modal attention module. And only use the textual and multimodal features to classify.
- FNDB-NMNA: Remove the multimodal feature extractor and cross-modal attention module. And only use the textual and visual features to classify.

Class Three: Class Three model chooses two features from the following three features: textual, visual, and multimodal features to conduct fake news detection experiments. Meanwhile, the model will use the cross-modal attention module to fuse those feature representations. The result will be compared with Class Two and Class Four. This task aims to explore the contribution of the cross-modal attention module and whether providing more information to the model can improve the model's classification performance.

- FNDB-NT: Remove the textual feature extractor, only use the visual and multimodal features to classify. Those features will be fused by using cross-modal attention module.
- FNDB-NI: Remove the visual feature extractor, only use the textual and multimodal features to classify. Those features will be fused by using cross-modal attention module.
- FNDB-NM: Remove the multimodal feature extractor, only use the textual and visual features to classify. Those features will be fused by using cross-modal attention module.

Class Four: Class Four model is highly similar to the FNDB model, but without using the cross-modal attention module for the fusion operation. The purpose is to investigate whether applying the cross-modal attention module can enhance the classification ability of the model when using three modal feature information. It also explores whether adding new features without using the cross-modal attention module is helpful for classification or not.

- FNDM-Natt: Remove the cross-modal attention module. All three-feature representations will be concatenated as the final news content representation to classify.

The results shown in Table 4 provide conditions to analyze the impact of the different components of FNDB on the fake news detection task. It can be found that FNDB outperforms the FNDB variant model using other configurations, as described below.

(1) Based on the comparison results of Class One, Class Two, and Class Four models. It shows that the fake news detector can select only one of the modalities to feature in the news content to discriminate whether the input news is fake or not. However, these models only focus on a single modal feature and cannot thoroughly learn the information and relationships of other modal features in the news content. As a result, the Class One model has a lower classification performance than several other variant models and FNDB.

(2) Based on the comparison results of Class One, Class Two, and Class Four models. It shows that when more modal features are provided to the model for classification, even if the modal information is simply concatenated together. The model can still understand the news content better and learn more information to complete the classification task. But according to the results of the comparison between the Class Three model and the FNDB model. It shows that these models using simple concatenated features cannot focus well on the correlation between text and images, so the performance is worse than the FNDB model.

(3) Based on the comparison between the Class Four model and FNDB, it can be concluded that the model's performance is improved when the model simply concatenates these three different modal features and uses them for classification. However, is still slightly lower than that of the FNDB model. Since FNDB uses a cross-modal attention module to fuse different modal features, it can reduce the impact of the model on the noise of some modal features. Moreover, these fused features are used for better classification.

Table 4: Results of ablation experiments performed by FNDB on two datasets

			Fake news			Real news		
	Method	Accuracy	Precision	Recall	F1-score	Precision	Recall	F1-score
Weibo	**FNDB**	**0.888**	0.891	**0.972**	**0.930**	**0.869**	0.610	0.717
	FNDB-OI	0.743	0.794	0.896	0.842	0.410	0.236	0.300
	FNDB-OT	0.845	0.885	0.916	0.901	0.689	0.610	0.647
	FNDB-OM	0.759	0.820	0.878	0.848	0.477	0.366	0.414
	FNDB-NTNA	0.759	0.815	0.886	0.849	0.475	0.339	0.396
	FNDB-NINA	0.861	0.889	0.936	0.912	0.745	0.614	0.673
	FNDB-NMNA	0.860	0.919	0.896	0.908	0.685	0.740	0.711
	FNDB-NT	0.772	0.794	0.950	0.865	0.532	0.187	0.276
	FNDB-NI	0.871	**0.921**	0.909	0.915	0.714	**0.744**	**0.728**
	FNDB-NM	0.864	0.890	0.938	0.913	0.753	0.618	0.679
	FNDB-Natt	0.873	0.893	0.949	0.920	0.788	0.625	0.697
Gossipcop	**FNDB**	**0.873**	**0.790**	0.440	0.565	0.883	**0.973**	**0.926**
	FNDB-OI	0.763	0.369	0.374	0.372	0.855	0.853	0.854
	FNDB-OT	0.799	0.470	**0.584**	0.521	**0.898**	0.848	0.873
	FNDB-OM	0.851	0.618	0.530	**0.570**	0.895	0.924	0.909
	FNDB-NTNA	0.825	0.541	0.423	0.474	0.873	0.917	0.895
	FNDB-NINA	0.837	0.581	0.465	0.517	0.882	0.922	0.902
	FNDB-NMNA	0.829	0.543	0.563	0.553	**0.898**	0.891	0.894
	FNDB-NT	0.826	0.552	0.372	0.444	0.865	0.930	0.896
	FNDB-NI	0.848	0.628	0.458	0.529	0.882	0.937	0.909

(Continued)

Table 4: Continued

Method	Accuracy	Fake news			Real news		
		Precision	Recall	F1-score	Precision	Recall	F1-score
FNDB-NM	0.842	0.604	0.456	0.520	0.881	0.931	0.905
FNDB-Natt	0.863	0.735	0.423	0.537	0.879	0.965	0.920

Next, by analyzing the contribution or impact of different modes on the task, it can be found that the model performs worst when just using visual feature representation to detect, according to the results in Table 4. On the Weibo dataset, the scores of all three evaluation criteria related to fake news are deficient. It indicates, to some extent, that visual feature representation provides fewer cues than other modalities for fake news detection. Based on the experimental results, it can be found that the models can perform well with only one modality. Such as FNDB-OT on the Weibo dataset and FNDB-OM on the Gossipcop dataset. However, the accuracy of these models is lower than other models that use two or three modalities. Furthermore, the results also demonstrate that when the model uses additional modal information, it can effectively fuse this information and facilitate the complementation of information between different modalities through the cross-modal attention module in the feature fusion layer. Afterward, the model could improve its discrimination ability by understanding and learning these fused features.

5 Conclusion

This paper proposes a novel multimodal fake new detection model, FNDB. Many multimodal fake news detection methods use different modal features by concatenating them. It would limit these independently encoded modal features to informatively complement each other for more effective fusion, although such methods perform well in fake news detection tasks. To overcome this limitation, this paper proposes the FNDB model. The FNDB contains 3 contributions: (1) Introduce BLIP, a Vision-Language Pre-training (VLP) model, to the fake news detection task. (2) Proposes a novel neural network model incorporating multimodal features extracted from the pre-trained BLIP model and using a cross-modal attention module to fuse multimodal, textual, and visual features. (3) A comparison of the experimental results shows that the FNDB model can outperform state-of-the-art fake news detection methods by 4.4% and 0.6% in terms of accuracy on the two datasets. The results also indicate that the FNDB model can facilitate the fusion of multimodal information to a certain extent and help the model to learn the features of news content. Furthermore, since both datasets are collected from the real world, this indicates, to some extent, that the FNDB model can identify real and fake news in real-world social platforms. However, there is still room for improvement in the fusion technique of the FNDB model, and this will be the direction of subsequent research to improve further the performance of FNDB in the field of fake news detection.

Funding Statement: The author received no specific funding for this study.

Conflicts of Interest: The author declares that he has no conflicts of interest to report regarding the present study.

References

[1] C. Castillo, M. Mendoza and B. Poblete, "Information credibility on twitter," in *Proc. WWW'11*, Hyderabad, India, pp. 675–684, 2011.

[2] N. Ruchansky, S. Y. Seo and Y. Liu, "CSI: A hybrid deep model for fake news detection," in *Proc. CIKM'17*, Singapore, pp. 797–806, 2017.

[3] F. Qian, C. Y. Gong, K. Sharma and Y. Liu, "Neural user response generator: Fake news detection with collective user intelligence," in *Proc. IJCAI'18*, Stockholm, Sweden, pp. 3834–3840, 2018.

[4] D. Khattar, J. S. Goud, M. Gupta and V. Varma, "MVAE: Multimodal variational autoencoder for fake news detection," in *Proc. WWW'19*, San Francisco, CA, USA, pp. 2915–2921, 2019.

[5] Z. W. Jin, J. Cao, H. Guo, Y. D. Zhang and J. B. Luo, "Multimodal fusion with recurrent neural networks for rumor detection on microblogs," in *Proc. MM'17,* Mountain View, California, USA, pp. 795–816, 2017.

[6] Y. Q. Wang, F. L. Ma, Z. W. Jin, Y. Yuan, G. X. Xun *et al.,* "EANN: Event adversarial neural networks for multi-modal fake news detection," in *Proc. KDD'18*, London, UK, pp. 849–857, 2018.

[7] X. Y. Zhou, J. D. Wu and R. Zafarani, "SAFE: Similarity-aware multi-modal fake news detection," in *Proc. PAKDD 2020*, Singapore, pp. 354–367, 2020.

[8] O. Ajao, D. Bhowmik and S. Zargari, "Sentiment aware fake news detection on online social networks," in *Proc. ICASSP 2019*, Brighton, UK, pp. 2507–2511, 2019.

[9] A. Giachanou, P. Rosso and F. Crestani, "Leveraging emotional signals for credibility detection," in *Proc. SIGIR'19*, Paris, France, pp. 877–880, 2019.

[10] S. Ghosh and C. Shah, "Toward automatic fake news classification," *Proceedings of the Association for Information Science and Technology*, vol. 55, no. 1, pp. 805–807, 2018.

[11] R. K. Kaliyar, A. Goswami, P. Narang and S. Sinha, "FNDNet—A deep convolutional neural network for fake news detection," *Cognitive Systems Research*, vol. 61, no. C, pp. 32–44, 2020.

[12] J. Ma, W. Gao, P. Mitra, S. Kwon, B. J. Jansen *et al.,* "Detecting rumors from microblogs with recurrent neural networks," in *Proc. IJCAI'16*, New York, NY, USA, pp. 3818–3824, 2016.

[13] Y. Liu and Y. -F. Wu, "Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks," in *Proc. AAAI'18/IAAI'18/EAAI'18*, New Orleans, LA, USA, vol. 32, no. 1, pp. 354–361, 2018.

[14] Z. W. Jin, J. Cao, Y. D. Zhang, J. S. 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, 2017.

[15] S. Liu, S. S. Qian, Y. Guan, J. W. Zhan and L. Ying, "Joint-modal distribution-based similarity hashing for large-scale unsupervised deep cross-modal retrieval," in *Proc. SIGIR'20*, Virtual Event, China, pp. 1379–1388, 2020.

[16] X. Wu, C. -W. Ngo and A. G. Hauptmann, "Multimodal news story clustering with pairwise visual near-duplicate constraint," *IEEE Transactions on Multimedia*, vol. 10, no. 2, pp. 188–199, 2008.

[17] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh *et al.,* "Learning transferable visual models from natural language supervision," in *Proc. ICML 2021*, Online, vol. 139, pp. 8748–8763, 2021.

[18] J. N. Li, D. X. Li, C. M. Xiong and S. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proc. ICML 2022*, Baltimore, Maryland, USA, pp. 12888–12900, 2022.

[19] S. S. Qian, J. G. Wang, J. Hu, Q. Fang and C. S. Xu, "Hierarchical multi-modal contextual attention network for fake news detection," in *Proc. SIGIR'21*, Virtual Event, Canada, pp. 153–162, 2021.

[20] K. Shu, L. M. Cui, S. H. Wang, D. W. Lee and H. Liu, "DEFEND: Explainable fake news detection," in *Proc. KDD'19*, Anchorage, AK, USA, pp. 395–405, 2019.

[21] T. Chen, X. Li, H. Z. Yin and J. Zhang, "Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection," in *Proc. PAKDD 2018*, Melbourne, Australia, pp. 40–52, 2018.

[22] Z. L. Yang, Z. H. Dai, Y. M. Yang, J. Carbonell, R. Salakhutdinov *et al.,* "XLNet: Generalized autoregressive pretraining for language understanding," in *Proc. NIPS'19*, Vancouver, BC, Canada, pp. 5753–5763, 2019.

[23] Y. M. Cui, W. X. Che, T. Liu, B. Qin, S. J. Wang *et al.,* "Revisiting pre-trained models for chinese natural language processing," in *Proc. EMNLP 2020*, Online, pp. 657–668, 2020.
[24] X. R. Song, H. T. Guo, X. N. Xu, H. Q. Chao, S. Xu *et al.,* "Cross-modal attention for MRI and ultrasound volume registration," in *Proc. MICCAI 2021*, part IV, Strasbourg, France, pp. 66–75, 2021.
[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones *et al.,* "Attention is all you need," in *Proc. NIPS'17*, Long Beach, CA, USA, vol. 30, pp. 5998–6008, 2017.
[26] K. Shu, D. Mahudeswaran, S. H. Wang, D. W. Lee and H. Liu, "FakeNewsNet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media," *Big Data*, vol. 8, no. 3, pp. 171–188, 2020.
[27] J. Devlin, M. -W. Chang, K. T. Lee and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," ArXiv:1810.04805, 2018.
[28] H. W. Zhang, Q. Fang, S. S. Qian and C. S. Xu, "Multi-modal knowledge-aware event memory network for social media rumor detection," in *Proc. MM '19*, Nice, France, pp. 1942–1951, 2019.
[29] J. X. Xue, Y. B. Wang, Y. C. Tian, Y. F. Li, L. Shi *et al.,* "Detecting fake news by exploring the consistency of multimodal data," *Information Processing and Management*, vol. 58, no. 5, pp. 102610, 2021.
[30] Y. X. Chen, D. S. Li, P. Zhang, J. Sui, Q. Lv *et al.,* "Cross-modal ambiguity learning for multimodal fake news detection," in *Proc. WWW'22*, Virtual Event, Lyon, France, pp. 2897–2905, 2022.
[31] J. Lin, G. Tremblay-Taylor, G. Y. Mou, D. You and K. Lee, "Detecting fake news articles," in *Proc. 2019 IEEE Int. Conf. on Big Data (Big Data)*, Los Angeles, CA, USA, pp. 3021–3025, 2019.
[32] S. Singhal, A. Kabra, M. Sharma, R. R. Shah, T. Chakraborty *et al.,* "SpotFake+: A multimodal framework for fake news detection via transfer learning (student abstract)," in *Proc. AAAI-20 Student Tracks*, New York, NY, USA, vol. 34, no. 10, pp. 13915–13916, 2020.
[33] L. Allein, M. -F. Moens and D. Perrotta, "Like article, like audience: Enforcing multimodal correlations for disinformation detection," ArXiv:2108.13892, 2021.