https://doi.org/10.31436/iiumej.v26i3.3615

A DETECTOR FOR TEXTUAL-VISUAL FAKE NEWS USING TEXT SUMMARIZATION AND CONTRASTIVE LANGUAGE-IMAGE PRETRAINING EMBEDDING MODEL

IMAN QAYS ABDULJALEEL[1,2*], ISRAA H. ALI[1]

[1] Software Department, College of Information Technology, University of Babylon, Hilla, Iraq
[2] Department of Computer Science, College of Education for Pure Sciences, University of Basrah, Basrah, Iraq

**Corresponding author: imanqaysa.sw@student.uobabylon.edu.iq*

(Received: 17 February 2025; Accepted: 29 July 2025; Published online: 9 September 2025)

ABSTRACT: Recently, social media have become an influential means of spreading news in multiple forms, including text, images, audio, and video. With the development taking place today and the heavy reliance of society members on social media to share content online, social media platforms have become a means of spreading unfavorable stories. This study proposes a model to identify false information in various media and determine its authenticity. After pre-processing both textual and visual data independently, the proposed CLIPCrossAttGFN model is used to initialize the fusion feature vector passed to the fully connected layer as part of the classification step. The CLIPCrossAttGFN model starts by passing the image to the Contrastive Language-Image Pretraining (CLIP) visual encoder as a feature extraction method. Also, text, after being summarized to 512 words based on the Bidirectional Encoder Representations from Transformers (BERT), which can identify meaningful words based on the semantic relationship of the text as a whole, is passed to the CLIP encoder before sending the initial extracted features to several layers of CNN, LSTM, and a cross-attention mechanism as a final feature extraction technique. Finally, a multimodal feature is merged using a gate fusion network. The results reveal that the proposed model has better accuracy, ensuring reliable detection compared to other available approaches.

ABSTRAK: Baru-baru ini, media sosial telah menjadi saluran berpengaruh dalam penyebaran berita dalam pelbagai bentuk, termasuk teks, imej, audio dan video. Dengan perkembangan pesat berlaku hari ini, dan kebergantungan tinggi masyarakat pada media sosial bagi berkongsi konten secara atas talian, platform media sosial telah menjadi medium penyebaran maklumat yang tidak sahih. Kajian ini mencadangkan model bagi mengenal pasti maklumat palsu dalam beberapa media dan menentukan sama ada ianya benar atau palsu. Selepas kedua-dua data teks dan visual melalui proses pra-memprosesan secara berasingan, model CLIPCrossAttGFN yang dicadangkan diguna pakai bagi memulakan vektor ciri gabungan yang dihantar ke lapisan bersambung sepenuhnya sebagai langkah pengelasan. Model CLIPCrossAttGFN bermula dengan menghantar imej kepada pengekod visual Contrastive Language-Image Pretraining (CLIP) sebagai kaedah pengekstrakan ciri. Teks pula diringkaskan kepada 512 perkataan berdasarkan Perwakilan Pengekod Dwi Arah daripada Transformers (BERT), yang dapat mengenal pasti perkataan bermakna berdasarkan hubungan semantik teks secara keseluruhan, dihantar kepada pengekod CLIP sebelum ciri awal diekstrak ke beberapa lapisan CNN, LSTM dan mekanisme silang perhatian sebagai teknik pengekstrakan ciri akhir. Akhirnya, ciri multimodal digabungkan menggunakan rangkaian gabungan berpintu. Dapatan kajian mendapati bahawa model yang dicadangkan mempunyai ketepatan tertinggi, menjamin pengesanan yang boleh dipercayai berbanding pendekatan sedia ada.

KEYWORDS: *False news, Deep learning, Gate fusion, Multimodal data, BERT.*

1. INTRODUCTION

To simplify formatting, you can use the style menu located just below the standard menu. Every button has a name similar to the style name in brackets after each paragraph. (Normal). In the modern era, social media has spread terrifyingly, leading to the almost complete abandonment of other means of transmitting information, such as television, newspapers, and magazines. Therefore, any individual can be a source for spreading news, which may often be fake, to achieve a specific societal, political, or religious goal.

Fake news can be defined as misinformation disseminated through the deliberate imitation of prevailing news, distorting and fabricating it to make it appear real. Fake news often relies on real news after injecting it with false information, according to the intended purpose. This modification affects not only texts but also images and videos. Therefore, spreading fake news through social media has become a frightening threat to our real society. Fake news recorded its highest levels from 2020 to 2021, which coincided with the COVID-19 pandemic, immediately linked to quarantine and total reliance on social media as a source of information. Fake news began to be used after two political events: the 2016 US elections and Brexit [1].

The most famous social media on the Internet are Facebook, Twitter, Instagram, and TikTok, which have become indispensable sites for many individuals and companies to publish their news on the one hand and promote their businesses on the other. However, this freedom to publish content may lead to the publication of misleading content that relies on different multimedia, such as texts and images, to attract users and promote the publication of fake news. Because of this, detecting fake news becomes a big challenge for everyone to confront. For this reason, many algorithms built to detect fake news based on data sets from social media platforms: the first of which relies on the content of the published news to determine its characteristics according to the textual, image, or video content, while the other based on the social context, which based on the user's profiles on social media and his activity in responding, the activity of the post, how it is published and expanded, and the structures of dissemination to conclude the validity of the published news [2].

One of the primary reasons for analyzing news and identifying fake news is that some individuals or organizations, more specifically, seek to exploit the spread of fake news to achieve personal or political gains. Others rely on spreading fake news to mislead people and make them adopt false beliefs to achieve goals that destabilize and dissolve society at all levels. Others seek to erode people's confidence in the transmitted news, often reducing confidence in the country's governments and political leaders, thereby undermining security [3].

Deep learning (DL), a subset of artificial intelligence that deals with complicated problems and big datasets by automatically collecting the characteristics of multiple layers of neural networks in a short time, has grown in favor of various single- and multimodal false news detection models [4]. The most often used neural network models are recurrent neural networks (RNNs), convolutional neural networks (CNNs), and attention networks. Using these models to identify the distinctive features of fake news, whether its content is textual (Natural Language Processing (NLP)) or visual (Computer Vision), has yielded impressive detection results [9] by identifying special and different features according to the content. However, as most news is primarily text-based, text remains the primary source for identifying fake news features despite the limitations of text analysis models and their problems in capturing long-term dependencies between words in a single text to create improved textual representations of sentences [5]. However, sources that rely on multiple content, whether text and visual or text and video, are still few compared to research that relies on a unimodal. The main reason for the lack of work in the field of multimodal fake news detection is the difficulty of choosing methods for integrating multimodal content and the effectiveness of their implementation because most models fail to select accurate individual features specific to each pattern and contribute valuable information to detect fake news and exclude noise resulting from the fusion of multimodal features., which makes it one of the most significant challenges in this field [1]. The increasing diversity of fake news on social media, along with its coverage of various topics, poses a substantial challenge to those working in this field, which is often characterized by media diversity. Therefore, classifiers must choose a method to extract essential features from this multimodal data, whether visual or linguistic. Through numerous studies, relatively simple models have emerged that rely on encoding images and texts based on various models, each tailored to the medium's environment, followed by different fusion methods, such as cross-modal attention and concatenation. Recently, several models have emerged that analyze images and text as a linked pair to reveal the relationship between them, including the Contrastive Language-Image Pretraining (CLIP) and Bootstrapping Language-Image Pretraining (BLIP) models [6].

While various models focusing on language and vision harmony, such as the Vision-and-Language Bidirectional Encoder Representations from Transformers (ViLBERT) model, which uses Faster Region-based Convolutional Neural Network (R-CNN) to determine regions of interest in the image and BERT for text, it works by dividing both image and text handling into two separate channels. Then, Co-attention is used to merge the output of both channels. This structure is frequently employed in visual question answering because it works effectively in problems requiring a deep comprehension of visual content. The Universal Image-Text Representation (Uniter) model pre-combines the inputs from both channels, text, and areas of interest to generate a data vector. In contrast, ViLBERT preserves the independence of both text and image channels before building the correlation process. The BERT model receives this data vector and uses it to identify deep correlations between the inputs. So, it is perfect for text-to-image matching [7]. CLIP is one of the most important models for understanding multimodal data features. It is a learning model developed by OpenAI to understand the relationship between the contents of a digital image and its corresponding text words using contrastive learning on multiple pairs (text, image). Its work is defined by encoding the text and visual data in interconnected vector representations, thus enabling it to find the semantic relationship of its multimodal inputs with a high level of understanding [8]. The benefit of the CLIP model is that it offers a thorough understanding of the level of text-image consistency overall, sometimes eliminating the need to look for forgeries in individual visual objects or textual details as in ViLBERT or Uniter models. Thus, its use frees it from the difficulties of internal consistency of each text and image component, allowing for speed of development and simplicity of structured representation [8].

There are two main types of fake news detection methods: unimodal and multimodal. The first method detects bogus news using a single information, such as text, image, audio, or video. As a result, these methods rely on only a limited portion of the available information, restricting their general understanding of the news story and determining their capacity to distinguish between truth and falsehood. The second type overcomes the shortcomings of the first by detecting fake news using accessible data from various linguistic and visual domains. Despite the superiority of multimodal detection methods in detecting fake news, they are still confronting several challenges, the most important of which is the mechanism for integrating information from various media to benefit from the most significant number of features that distinguish the truth of the news from its falsehood. The third issue is a shortage of multimodal data sets that include various information types like text, image, video, and audio. This section will review the most relevant connected papers dealing with these two methodologies. In the paper by Kaliyar [9], the textual fake news framework structure is built on the BERT coding paradigm and utilizes convolutional layers to extract key text properties. Based on the Fake News dataset, the model accuracy is 98.9%. Rustam applied Artificial Intelligence (AI) mechanisms, including machine learning (ML) algorithms and deep learning, to a system for identifying false news. This approach converted textual news input into a visual representation, collecting accurate and crucial features that enhance fake news detection accuracy. The accuracy rate of 99.7% using the ISOT dataset, but the LIAR dataset is below 92% [10]. Work by Daik in [11] constructed a fake news detection system using a convolutional neural network. Still, this system struggled to collect enough image samples to train the model and maintain high accuracy.

In [12], complementary attention fusion is combined with a deep neural network to detect the associated features between text and image in the input dataset. The deep neural network consists of three fully connected layers (FCL) to extract high-level features that significantly detect fake news based on the annotations and texts extracted using CAF. The datasets used are GOSSIPCO, POLITIFACT, FAKEDDITE, and PHEME, and the model obtained an accuracy of 85%, 90%, 95%, and 90% respectively. In [13], the authors proposed a model that extracts text features using a multi-level encoding network, while image features are extracted using the VGG-19 network. Then, the resulting text and image features are passed into a relationship-aware attention network that can detect the similarity between information pieces in different media. Finally, the resulting fusion features are fed into a fake news detector. This approach is applied to three reference datasets: Weibo, Twitter, and PHEME, and the accuracy rates were 90%, 85.5%, and 87%, respectively. A multimodal detecting fake news model is presented in [8] based on contrastive learning. The model begins by extracting essential features from the input text and image at various levels of granularity through contrastive learning. Then, a multi-granular representation of multimedia news is applied, utilizing both image and text features in parallel. The model determines whether the news is real based on multi-head self-attention techniques. The RECOVERY, GOSSIPCOP, and MR2 datasets are used. The accuracy of each of the above datasets was 85.48%, 92.3%, and 89.45%, respectively. The research in [14] presents a method based on contrastive learning using CLIP and semantic alignment, applied to textual and image information, utilizing three datasets: PHEME, GOSSIPCOP, and MR2. The accuracy on each dataset was 91.29%, 91.98%, and 89.86% respectively.

The research [15] relies on quantum fusion in detecting multimodal fake news. Text features are extracted using Transformer-XL Networks (XLNet), and image features are extracted using VGG-19. Then, text and image features are combined and passed to a proposed quantum convolutional neural network to detect the truth of news from its falsity. Using two datasets, Gossip and POLITIFACT, with accuracies of 87.9% and 84.6%, respectively, we can observe the model's ability to detect and distinguish fake news. In [16], a multimodal news model relied on the representation of BERT to extract the text characteristics and used VGG-19 to extract the image characteristics. In addition, a multimodal compact bilinear pooling technique is used to merge textual and image attributes with an attention mechanism, resulting in one time and integrating textual and image attributes independently at another time. This proposed model was applied to two data groups, Twitter and Weibo, with accuracies of 68% and 81%, respectively. In the paper by Rashid in [17], a model based on a robustly optimized BERT approach (RoBERTa) is used to analyze the entrance text, and ResNet-50 is used to analyze the images associated with it as a first stage in the false news detection system which received an accuracy of 91 % using a Twitter dataset and 93 % using the Weibo dataset.

The main contributions of this research have been outlined as follows:

It offers a novel strategy for detecting multimodal false news based on the CLIP model, primarily for text and image pairs. This is followed by numerous parallel or sequential layers of Convolutional Neural Network (CNN), Long-Short-Term Memory (LSTM), and attention mechanisms.

1. Uses several techniques to clean the text by removing stop words and roots and adopting the spelling correction of words by 1-2 letters to retain the possible words distorted inadvertently.
2. Use the word summarization technique to extract the most important words in the text that are longer than 512 words. Adopt the BERT approach to summarize texts and detect the associations between words to determine their importance within the text.
3. The suggested architecture's efficacy is evaluated by accuracy, precision, recall, and the F1 score on two datasets: GOSSIPCO, RECOVERY, and Ti-CNN.

This paper has the following structure: In the second section, we give the background material used in this paper. Section three covers the proposal model by detailing the components of the suggested model. The fourth component covers the dataset employed, discusses the results of the proposal model, and compares it to other models. The fifth part described the work's results and the most crucial future ideas for overcoming the problems we encountered while working.

2. MATERIAL AND METHODS

2.1. Contrastive Language-Image Pretraining Embeddings Model

The proposed model begins by utilizing a pre-trained characteristics unit for managing multimodal data, termed CLIP, which combines text and image information to create a feature space. The CLIP relies on the transformer structure and operates on the principle of "Comparative Learning" based on both contrastive loss functions, to achieve a thorough grasp of the image and text through large-scale pre-training [11]. Thus, CLIP seeks semantic similarity between its two interrelated textual and visual inputs rather than precise information, distinguishing it from other models, particularly those dealing with multimodal inputs [9].

The CLIP model structure is divided into the image encoder and feature extraction, which are generated with pre-trained convolutional neural network models like ResNet-50 or Vision Transformer (ViT). The other part is for the text encoder, which is based on transformer principles such as BERT and Generative Pre-trained Transformer (GPT). In the training stage, input pairs (text-image) are matched to increase the cosine similarity of the matched pairings while reducing the cosine similarity of all other non-matched pairs. To minimize retraining on the subsequent dataset, the CLIP team collected 400 million and 1 billion matched pairs of images and texts for training. Since CLIP is pre-trained on some datasets like ImageNet, the text encoder can recognize rich natural language semantics, making it suitable for zero-shot prediction [11].

Zero-Shot Learning (ZSL) is a challenging machine learning paradigm [6]. Its goal is to recognize previously unseen objects or concepts. Traditional models require extensive datasets for each expected category. In contrast, ZSL uses supplementary data, such as semantic attributes, to infer the properties of unseen categories. The CLIP model's ability to embed images and text into a shared space naturally extends this concept. It enables effective knowledge transfer from seen to unseen categories [6]. The structure of CLIP comparative learning is visible in Figure 1.

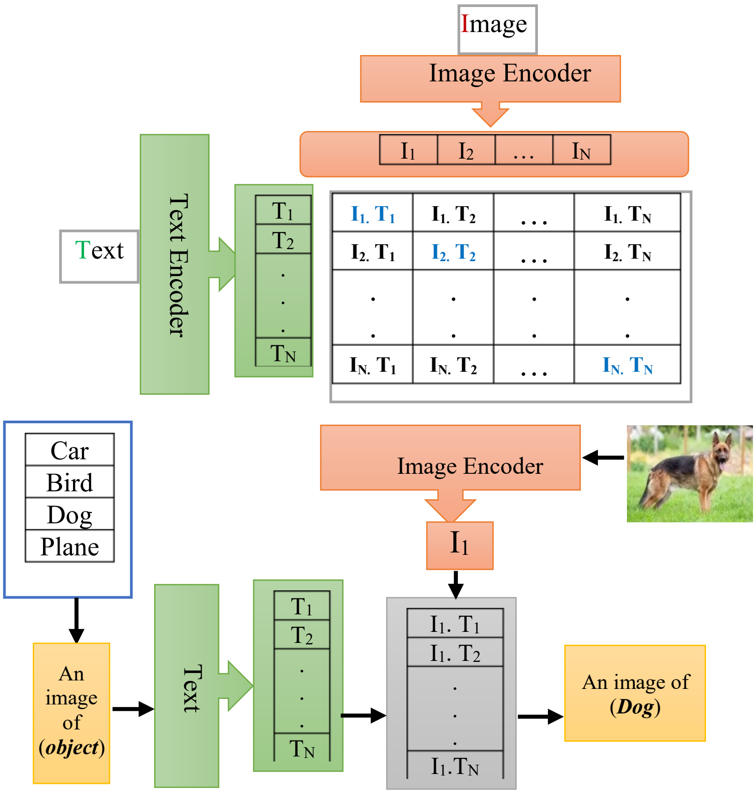

Figure 1. Structure of CLIP Contrastive Learning [6]: (a) Contrastive pre-training, (b) Dataset generation based on "text" and zero-shot prediction.

2.2. Cross-attention Mechanism

In recent years, attention mechanisms have gained popularity in various fields, although they originated in natural language processing, specifically in machine translation. The model by Bahdanau et al. [18] presented at that time could find all the information related to the specific words in the source sentence. Two popular types appeared in the multimodal datasets (image-text) models: co-attention and cross-attention. These two models could capture the unimodal and multimodal relationship with each other, and thus the ability to discover important features and detect fake news [13]. The cross-attention module is based on the self-attention module in its construction, which uses two inputs with different sequences (such as text and image), so the attention score is calculated based on both formats. Figure 2 describes the Architecture of the cross-attention mechanism.

The image-text features fusion process starts by initializing two vision feature vectors (VF) and one text feature vector (TF) as inputs to the cross-attention network to obtain text attention features (TFa). Also, the vision attention features (VFa) are obtained by passing two text feature vectors (TF) and one vision feature vector (VF) to the cross-attention network. They can be expressed mathematically by the following two equations, as in Eq. (1) and Eq. (2) [19] where TL is the transform layer, (Q, K, V) are the attention input, and $\sqrt{p}$ is the length of the aligned vectors.

$$TF_a = SoftMax\left(\frac{TL_Q(VF).TL_k(TF)^T}{\sqrt{P}}\right) TL_v(TF) \tag{1}$$

$$VF_a = SoftMax\left(\frac{TL_Q(TF).TL_k(VF)^T}{\sqrt{P}}\right)TL_v(VF) \tag{2}$$

The outputs of the two previous equations are then passed to a fully connected layer to process and improve the resulting representation. After that, layer normalization is applied to enhance the network's stability during training.

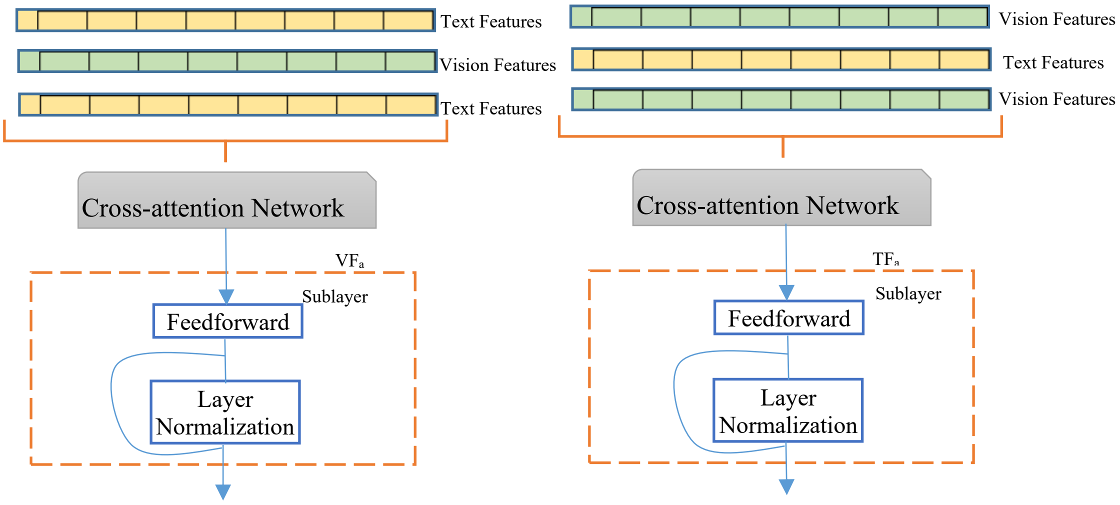

Figure 2. Architecture of the Cross-attention mechanism [19].

2.3. Gate Fusion Network (GFN)

A gate fusion module is used to understand the extracted features and how to refine them by removing noisy modality inputs that negatively affect fake news detection in a multimodal environment [6]. Its inputs are text and image extracts from previous stages, which control the selection of specific features by this gate fusion. There are several ways to combine gates described as follows [20]:

1. Independent Gates (IG): Two sets of gates are defined, one for extracting text features and the other for extracting image features. Each gate loads information specific to only one typical feature. In this paper, we used this type.
2. Cross-Gate (CRG): Here, two sets of gates are also defined, but here the two gates are multiplied by other feature maps, which increases the learning opportunity between multimodal patterns.
3. Complementary Gates (CG): Here, the connection between channel weights before the text and image feature fusion stage, and the complementary features are extracted based on the initialization of two separate gates. Here, it is evident that the two gates learn complementary information from the same concatenated weights.

3. METHODOLOGY

The work begins by preparing dataset totals in a way that the model can deal with in its bilateral form (text, image). Initially, text and image are processed separately by cleaning the text data and removing stop words before all words are returned to their root form using stemming and lemmatization. The long sentence text input is summarized at an optimal length before text embedding using CLIP. The image is resized to 224×224 pixels before being used with CLIP for feature extraction. After that, maintain the text linked with each image in the dataset to assist in dealing with it later using the proposed "CLIPCrossAttGFN" model to conclude the work with a result that distinguishes between true and disinformation. Figure 3 explains the general structure of the proposed model. Figure 4 describes a general diagram of the proposed multimodal fake news detection framework.

3.1. Image and Text Pre-processing

The dataset is pre-processed based on its type, where image data is analyzed by verifying the validity of image download links (URLs), then downloading images and resizing them to a uniform size of 224×224 pixels. As for text data, pre-processing involves removing all symbols, stop words, punctuation marks, links, and emojis to keep only words as a basis for assessing the importance of the existing text. After that, make all words lowercase, then root words by removing prefixes and suffixes to return them to their original form. Additionally, an attempt is made to correct the spelling of words to the minimum number of letters (two letters maximum) to retain the highest possible number of linguistically understandable words, which appears to be due to human error.

3.2. Text Summarization

To understand natural languages, models have recently emerged that can understand human languages, specifically using deep learning models. One is the BERT (Bidirectional Encoder Representations from Transformers) model. This model segments the input text into words and further segments the words into sub-words to understand the exact meaning of these smaller parts, which often include some that are not commonly used. BERT adds two tokens at the beginning and end of the input sequence to help better understand the language: CLS (at the start of the text input sequence) and SEP (at the end of the text input sequence). It then uses segment embeddings to identify the most distinctive words based on contextual information throughout the text, thus providing a comprehensive understanding before selecting the most important words [9]. After cleaning the pre-processing text, work on summarizing it to get the most influential words in the entered text with lengths compatible with what can be embedded using the CLIP model. For this, it is working on passing the words in the text to the BERT embedding technique, which calculates the weights of each existing word and the number of times it is duplicate, then choosing the most important ones based on the highest weight while allowing the word repeating in a proportion that is compatible with the number of times it is duplicate without neglecting other import words. Another problem is the appearance of unknown words in BERT. To address this, they are divided into sub-words to find similar words in BERT. However, some sub-words are converted into special symbols (like ##) because they are not found in BERT. Therefore, in the event of a similar situation, the parts of the word are re-merged and returned to their original state and dealt with as a new word with a specific weight.

3.3. The Proposed Model (CLIP-Cross-attention-GFN Model)

To deal with textual and visual news inputs, we aim to utilize transformers to capture the interactions within the patterns, whether they are textual or visual information. For textual data, it uses the pre-trained CLIP embedding (encoder) for texts to search for semantic information in the input news text words. As for image inputs, we also use the CLIP encoder, but we extract the image features of the news in the part related to dealing with visual information. For the text embeddings generated by a pre-trained CLIP model, we get 768-dimensional word embeddings, while for image embeddings, we also get 768-dimensional embeddings using the CLIP model.

The text features embedded using CLIP are passed into two models in parallel to obtain the spatial features of specific words, which indicate homogeneous meanings, such as a negative sensation. This can be obtained by concurrently sending the text features utilized via CLIP into multiple models. In addition, the second parallel model consists of two sequentially ordered layers of LSTM, which are applied to gather the text's temporal properties and contextual coherence from start to finish. Then the outputs of the two previous models (CNN and LSTM) and the text features embedded using CLIP are passed to a dense layer (a set of regular cells in a neural network to prevent overfitting). While the image features are embedded using CLIP, which are also passed to another dense layer, before the outputs of the text and image dense layers are passed to a cross-attention mechanism.

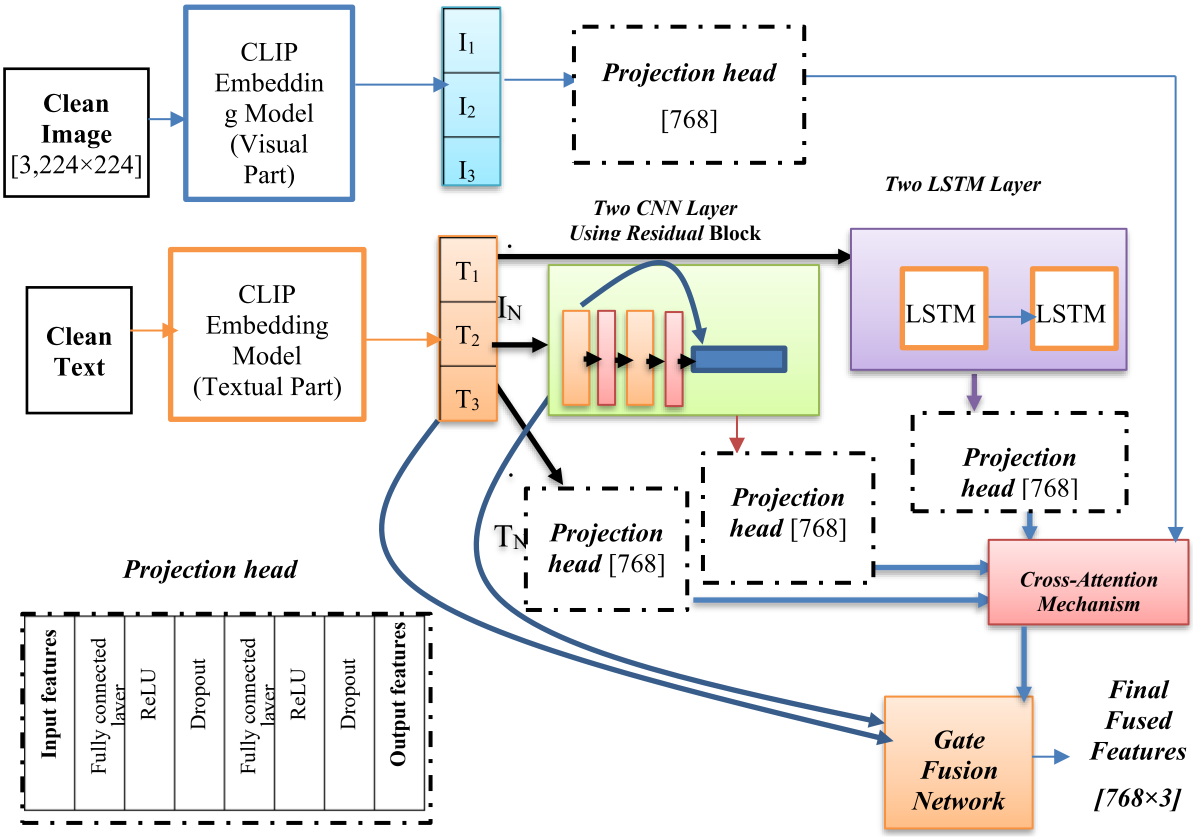

Figure 3. The proposal of the CLIPCrossAttGFN model.

Up to this stage, the most important features will be extracted. To obtain other features that we may not have been able to extract using the previous steps, we are now working on passing the outputs of the cross-attention mechanism in addition to the text features extracted using CNN and the text features embedded using CLIP as inputs to the GFN network to select the features based on the most influential ones, whether textual or visual. The final vector of features for this stage represents the vector of the extracted, combined textual and visual features. Leveraging DL methods at the appropriate locations to extract the most essential features during the next step gives the CLIPCrossAttGFN model its efficacy. First, it uses the pre-trained CLIP model, which begins by thoroughly depicting the semantic link between two input modalities (textual and visual). Second, it uses CNN and LSTM models to improve comprehension of information according to the news's structure and context. Finally, it uses a GFN network that retrieves the feature vector without distortion using cross-modal attention.

3.4. Classification Model

After merging the textual and visual features based on the CLIPCrossAttGFN model, we now rely on a fully connected layer, followed by a special function to predict the validity of the news with its two probabilities (True or False), which is the Sigmoid function. We split the fake news datasets in a ratio of 8:1:1 for training, validation, and testing. The number of LSTM neurons is 768, while CNN filters and kernel windows are 768 and 3, respectively. The following hyperparameters are organized to improve model training and testing: AdamW, initial learning rate, weight decay, number of epochs, batch size, and dropout. Their values are entered manually according to Table 1.

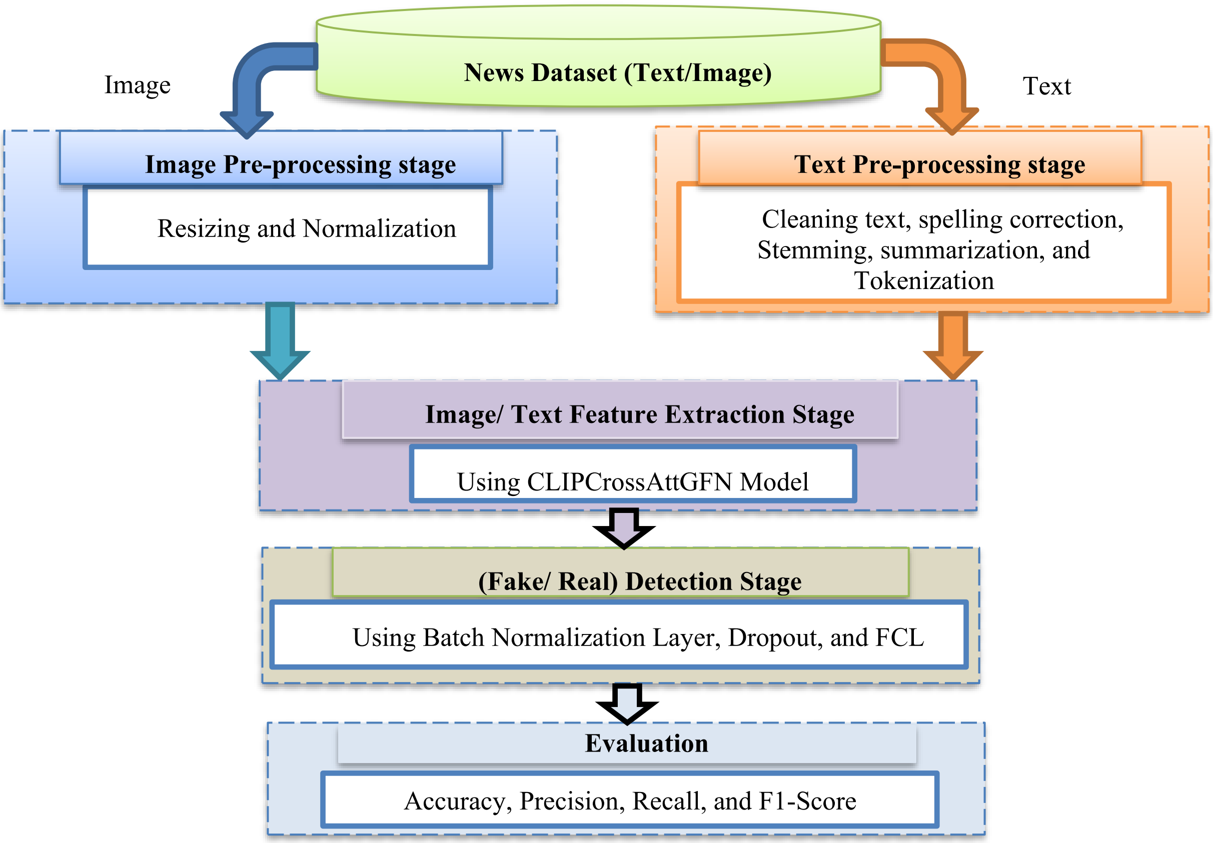
News Dataset (Text/Image)
Image
Text
Image Pre-processing stage
Resizing and Normalization
Text Pre-processing stage
Cleaning text, spelling correction, Stemming, summarization, and Tokenization
Image/ Text Feature Extraction Stage
Using CLIPCrossAttGFN Model
(Fake/ Real) Detection Stage
Using Batch Normalization Layer, Dropout, and FCL
Evaluation
Accuracy, Precision, Recall, and F1-Score

Figure 3. The general diagram of the proposed multimodal fake news detection Framework

Table 1. Test configuration of parameters following improvement

Hyperparameters	Value	Hyperparameters	Value
Optimizer	AdamW	Number of epochs	20
Initial learning rate	1e-4	Batch size	64
Weight decay	0.01	Dropout	0.5

3.5. Evaluation Metrics

Several evaluation metrics are used to demonstrate the efficiency of the proposed fake news model, the most important of which are the accuracy metric, followed by the precision, recall, and the F1 score. Accuracy alone may provide unrealistic results if the input data is unbalanced. The four metrics can be represented as follows in Eq. (3) to Eq. (6) [2].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100 \tag{3}$$

$$Precision = \frac{TP}{TP+FP} \times 100 \tag{4}$$

$$Recall = \frac{TP}{TP+FN} \times 100 \tag{5}$$

$$F1 - Score = 2 \times \frac{Recall \times Precision}{Recall + Precision} \tag{6}$$

4. EXPERIMENTAL RESULTS

4.1. Fake News-related Datasets

The GOSSIPCOP dataset from FakeNewsNet is the first dataset we use, which is based on English-language text websites and classified into two categories: real (label = 1) and fake (label = 0) [12, 8]. We also used Megan Risdal's Fake News dataset in Kaggle (Ti-CNN). It contains news articles of varying lengths, ranging from 5 to 3271 words, and URLs to images; unfortunately, some are corrupted due to the closure of these sites and their subsequent inoperability. Table 2 shows the number of fake and real news [8]. The third dataset we used is RECOVERY, which is considered one of the most essential datasets containing healthy multimodal information. Its articles were collected during the COVID-19 pandemic (i.e., during the first five months of 2022) based on several reliable and unreliable websites, a total of which reached 60 websites [12].

Table 2. Test configuration of parameters following improvement

Dataset	GOSSIPCOP dataset			RECOVERY			Ti-CNN		
	Train	Validate	Test	Train	Validate	Test	Train	Validate	Test
Separated size	10104	1263	1264	1142	143	143	2825	353	354
Total size		12631			1428			3532	
Maximum length of text		9137			9983			3271	
Class		2			2			2	

4.2. Weight Adjustment

Many multimodal fake news datasets suffer from an imbalance in the main category (like real and fake labels) due to a big difference in the number of samples for the categories that make up each dataset which done sometimes due to the damage to one item in the input pair (text, image), in addition to relying on a specific type of news (such as sport news, medical news, and art news) that is likely to contain a particular category less than the other category. During model training, these reasons led to the model being biased towards the most numerous categories. There are many ways to address the imbalance in the datasets, the most important of which are: oversampling (repeating samples for the least multiple categories in the dataset) and undersampling (reducing samples for the most abundant categories in the dataset). In this research paper, we used another type called "class weight adjustment", which depends on assigning different weights to samples from various categories, which in this way allows the model during training to focus on the categories that have the least number of samples and thus gives good results in its ability to learn from them [21].

4.3. Performance and Comparison Results

In this section, we evaluate the model's performance and compare its results with state-of-the-art models. Table 3 compares the performance of various networks for multimodal fake news identification. Existing approaches such as BERT-VGG16, BERT-VGG19, and BERT-ResNet50 are evaluated and compared to the suggested design. The CLIP for text-image data produces more accurate findings. The achieved outcomes based on the performance metrics: accuracy, precision, recall, and F1-score, with corresponding values of around 86%, 85.37%, 86%, and 85.57% in the GOSSIPCOP dataset. The results for the performance measures for the RECOVERY dataset are 90.21%, 89.58%, 90.21%, and 89.74%, whereas in the Ti-CNN dataset, they are 95.47%, 95.52%, 95.47%, and 95.4%, respectively. A comparative analysis was carried out using the dataset and considering the results obtained from the various performance indicators. Table 4 shows an analysis of the proposed technique and earlier models.

The proposal detection results based on the multimodal fake news model are presented for comparison with BERT-VGG16, BERT-VGG19, BERT-ResNet50, and BERT-EfficientNetB0 models, as well as with other research papers [1, 2, 12, 22, 23, 34], as explained in Tables 3 and 4. Datasets containing texts exceeding three thousand words were relied upon to test the ability of the proposed summarization technique based on BERT embedding to reduce text length before merging them with the associated images and sending them to the proposed fake news detection model. We also note that the data for the two patterns, real and fake, is unbalanced, which requires "Class Weight Adjustment" to help deal with this dataset. Another limitation is that this proposed model works only on text and image datasets, and other data types, such as video and audio, cannot be used because the CLIP embedding model currently available only deals with these two media (text and image). In addition, the currently available datasets are relatively small because they are derived from datasets built at previous times, which resulted in the loss of some of them due to the interruption of their sites. Although the proposed framework performs effectively, Table 5 shows that it occasionally fails to obtain accurate categorization. Furthermore, the nature of some previously pre-trained emotional phrases and images that the CLIP model perceives as real is very similar to that of some false news generated with the same emotional formula and visuals, making it more likely to be regarded as real at the news input level. Another factor is the imbalance between the quantity of false and true news articles, as seen in Table 2, which leads to poor generalization.

Table 3. Performance analysis of proposed models with various multimodal models

Models		**BERT-VGG16**	**BERT-VGG19**	**BERT-ResNet50**	**BERT-MobileNetV2**	**Ours Model**
RECOVERY dataset	**Accuracy**	74.13	73.43	79.02	81.12	90.21
	Precision	80.02	81.50	83.85	84.53	89.58
	Recall	74.13	73.43	79.02	81.12	90.21
	F1-score	76.58	76.50	80.90	82.49	89.74
GOSSOPCO dataset	**Accuracy**	61.23	62.10	68.32	75.14	86
	Precision	66.44	67.57	71.70	76.93	85.37
	Recall	61.23	62.10	68.32	75.14	86
	F1-score	63.48	64.18	70.36	76.88	85.57
Ti-CNN dataset	**Accuracy**	87.25	87.91	88.12	92.07	95.47
	Precision	87.55	87.73	89.86	92.64	95.52
	Recall	87.25	87.91	88.12	92.07	95.47
	F1-score	87.36	87.98	89.15	92.20	95.40

Table 4, Comparison evaluation of proposed and current methods for multimodal fake news

References	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-score
[22]	RECOVERY	79.03	72.51	45.76	56.11
	GOSSIPCOP	85.69	84.62	75.12	79.59
[12]	GOSSIPCOP	86.3	87	95	90.5
[23]	GOSSIPCOP	84.8	82.2	79.7	80.8
[24]	RECOVERY	81.72	72.68	66.41	69.40
	GOSSIPCOP	88.04	86.77	80.01	83.25
[2]	GOSSIPCOP	74.8	74.87	75.31	74.71
[1]	Ti-CNN	95	97	96	96
Ours Model	RECOVERY	90.21	89.58	90.21	89.74
	GOSSIPCOP	86	85.37	86	85.57
	Ti-CNN	95.47	95.52	95.47	95.40

Table 5. Samples of real and outcome labels

News URL	Downloaded Image	Title	Real Label	Outcome Label
www.dailymail.co.uk/tvshowbiz/article-5874213/Did-Miley-Cyrus-Liam-Hemsworth-secretly-married.html		Did Miley Cyrus and Liam Hemsworth secretly get married?	0	0
https://s.yimg.com/uu/api/res/1.2/rNLbr_QGKbcHp3fAhfgMXw--~B/aD0xNjY5O3c9MjUwMDtzbT0xO2FwcGlkPXl0YWNoeW9u/https://media.zenfs.com/en-US/nbc_news_122/b4cdec47217b1409c2ab2bd31284fdae		text voter afraid of contract the coronavirus can cast their ballot by mail in upcom elect fewer judge rule tuesday in page decks you district judge fred bieri wrote that the coronavirus tandem had left the world without immune and fear disable the decks came after the text democrat parti and individu voter file suit last month argue that the state denial of mail in ballot sure tandem was violet of their constitute right only people over the age of and those with disable that prevent them from vote in person are allow to cast mail in ballot in the state echo fall claim made by preside	0	1

5. CONCLUSION

This study proposes an efficient model to detect text-image fake news by improving the extraction from text and image inputs. The texts are handled in a way that facilitates the extraction of their influential words, especially in cases where the text length exceeds 512 words, which exceeds the capacity of most text representation models, such as BERT and CLIP. Thus, we overcome the issue of cutting these texts and not benefiting from words that may be important but come after the maximum length allowed to handle. The system for determining the reality news from its falsity uses many DL techniques. It combines them in an efficient way that reflects positively on the features extracted from the text and image pairs entered into the proposed system. After including the texts and images inserted using CLIP embedding, we relied on passing the CLIP embedding of text parallel to the CNN model based on the Residual Block and the LSTM model. The extracted features result (CNN features, LSTM features, and CLIP embedding features) are passed to a dense layer (text dense layer). Also, the CLIP embedding of the image is passed to the dense layer (image dense layer) before sending the result features of it with the other dense result (text dense layer) to the cross-attention mechanism layer to understand the relationship between the image and text features in parallel. The next stage involves using GFN by passing three inputs: cross-attention result, text CNN features, and text CLIP embedding features. Finally, we get the most influential features, which are transferred to a fully connected layer to determine whether the news is real. The model was tested on open-source English training sets RECOVERY, GOSSIPCOP, and Ti-CNN. The experimental results, in which we obtained accuracy values of 90 for the GOSSIPCOP dataset and 99 for the Ti-CNN dataset, indicate that the proposed model performed well compared to other classical models, demonstrating its ability to detect fake news in the future and trying to develop it after overcoming the most big obstacles it faced, the most important of which is the availability of a large dataset and modern data that has a diversity of languages and dialects as well as the news topics it contains. We also concluded that it can distinguish fake news whenever the input texts are shorter in length (Ti-CNN dataset, which is the shortest length text used here) due to its ability to summarize selected words in a length that is proportional to the maximum length allowed for the input text for the embedding mechanisms (like CLIP), which is 512 tokens. All these requirements enhance the proposed models' ability to detect fake news in live applications.

REFERENCES

[1] Yadav A, Gaba S, Khan H, Budhiraja I, Singh A, Singh KK (2024) ETMA: Efficient transformer-based multilevel attention framework for multimodal fake news detection. IEEE Trans Comput Soc Syst. 11(4):5015-5027. doi:10.1109/tcss.2023.3255242.

[2] Abduljaleel IQ, Ali IH (2024) Deep learning and fusion mechanism-based multimodal fake news detection methodologies: A review. Eng Technol Appl Sci Res. 14(4):15665-15675. doi:10.48084/etasr.7907.

[3] Nawaz MZ, Nawaz MS, Fournier-Viger P, He Y (2024) Analysis and classification of fake news using sequential pattern mining. Big Data Min Anal. 7(3):942-963. doi:10.26599/bdma.2024.9020015.

[4] Abbood EA, Al-Assadi TA (2022) A new convolution neural layer based on weights constraints. In: 2022 International Conference on Data Science and Intelligent Computing (ICDSIC). IEEE.

[5] Abduljaleel IQ, Ali, IH (2025) Detecting fake news using BERT word embedding, attention mechanism, partition and overlapping text techniques. TEM Journal, 1152–1165. https://doi.org/10.18421/tem142-16.

[6] Gu Y, Castro I, Tyson G (2024) Detecting Multimodal Fake News with Gated Variational AutoEncoder. In: ACM Web Science Conference. ACM; 129-138. doi: https://doi.org/10.1145/3614419.3643992.

[7] Lu J, Goswami V, Rohrbach M, Parikh D, Lee S (2020) 12-in-1: Multi-Task Vision and Language Representation Learning. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). doi: 10.1109/CVPR42600.2020.01045.

[8] Yan F, Zhang M, Wei B, Ren K, Jiang W (2024) FMC: Multimodal fake news detection based on multi-granularity feature fusion and contrastive learning. Alex Eng J.109:376-393. doi:10.1016/j.aej.2024.08.103.

[9] Kaliyar RK, Goswami A, Narang P (2021) FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. Multimed Tools Appl. ;80(8):11765-11788. doi:10.1007/s11042-020-10183-2.

[10] Rustam F, Aljedaani W, Jurcut AD, Alfarhood S, Safran M, Ashraf I (2024) Fake news detection using enhanced features through text to image transformation with customized models. Discov Computing.;27(1). doi:10.1007/s10791-024-09490-1.

[11] Dai K, Shao J, Gong B, Jing L, Chen Y (2024) CLIP-FSSC: A transferable visual model for fish and shrimp species classification based on natural language supervision. Aquacult Eng. ;107(102460):102460. doi:10.1016/j.aquaeng.2024.102460.

[12] Luvembe AM, Li W, Li S, Liu F, Wu X (2024) CAF-ODNN: Complementary attention fusion with optimized deep neural network for multimodal fake news detection. Inf Process Manag. 61(3):103653. doi:10.1016/j.ipm.2024.103653.

[13] Yang H, Zhang J, Zhang L, Cheng X, Hu Z (2024) MRAN: Multimodal relationship-aware attention network for fake news detection. Computer Stand Interfaces. 89(103822):103822. doi:10.1016/j.csi.2023.103822.

[14] Yan F, Zhang M, Wei B, Ren K, Jiang W (2024) SARD: Fake news detection based on CLIP contrastive learning and multimodal semantic alignment. J King Saud Univ - Computer Inf Sci. ;36(8):102160. doi:10.1016/j.jksuci.2024.102160.

[15] Qu Z, Meng Y, Muhammad G, Tiwari P (2024) QMFND: A quantum multimodal fusion-based fake news detection model for social media. Inf Fusion. 104(102172):102172. doi:10.1016/j.inffus.2023.102172.

[16] Jonnapalli TR, Selvi M (2024) Detecting fake news in social media networks with deep learning techniques. In: AIP Conference Proceedings. Vol 3086. AIP Publishing;030006. doi: https://doi.org/10.1063/5.0211567.

[17] Rashid J, Kim J, Masood A (2024) Unraveling the tangle of disinformation: A multimodal approach for fake news identification on social media. In: Companion Proceedings of the ACM Web Conference 2024. Vol 3. ACM; 1849-1853. doi: https://doi.org/10.1145/3589335.3651972.

[18] Bahdanau D, Cho K, Bengio Y (2015) Neural machine translation by jointly learning to align and translate. In: Proceedings of the 3rd International Conference on Learning Representations, pp. 1–15. doi: https://doi.org/10.48550/arXiv.1409.0473.

[19] Kou F, Wang B, Li H, Zhu C, Shi L, Zhang J, et al (2025) Potential Features Fusion Network for multimodal fake news detection. ACM Trans Multimedia Computer Communication Appl. doi:10.1145/3711866.

[20] Wei K, Dai J, Hong D, Ye Y (2024) MGFNet: An MLP-dominated gated fusion network for semantic segmentation of high-resolution multi-modal remote sensing images. Int J Appl Earth Obs Geoinf. 135(104241):104241. doi:10.1016/j.jag.2024.104241.

[21] Zhang C, Wu J (2024) Software defect prediction based on effective fusion of multiple features. IEEE Access. Published online 2024:1-1. doi:10.1109/access.2024.3409709.

[22] Singhal S, Shah RR, Chakraborty T, Kumaraguru P, and Satoh S (2019) Spotfake: A multi-modal framework for fake news detection. in: 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM), IEEE, pp. 39–47. Doi: https://doi.org/10.1109/BigMM.2019.00-44.

[23] Silva A, Luo L, Karunasekera S, and Leckie C (2021) Embracing domain differences in fake news: Cross-domain fake news detection using multi-modal data. Proc. Conf. AAAI Artificial Intelligence., vol. 35, no. 1, pp. 557–565. doi: https://doi.org/10.1609/aaai.v35i1.16134.

[24] Chen Y, Li D, Zhang P, Sui J, Lv Q, Tun L, and Shang L (2022) Cross-modal ambiguity learning for multimodal fake news detection. in: Proceedings of the ACM Web Conference 2022, pp. 2897–2905. Doi: https://doi.org/10.1145/3485447.3511968.