# Deep Learning and Fusion Mechanism-based Multimodal Fake News Detection Methodologies: A Review

**Iman Qays Abduljaleel**
Software Department, College of Information Technology, University of Babylon, Iraq | Department of Computer Science, College of Education for Pure Sciences, University of Basrah, Iraq
imanqaysa.sw@student.uobabylon.edu.iq (corresponding author)

**Israa H. Ali**
Software Department, College of Information Technology, University of Babylon, Iraq
israa_hadi@itnet.uobabylon.edu.iq

*Received: 21 May 2024 | Revised: 1 June 2024 | Accepted: 3 June 2024*

*DOI: https://doi.org/10.48084/etasr.7907*

## ABSTRACT

**Today, detecting fake news has become challenging as anyone can interact by freely sending or receiving electronic information. Deep learning processes to detect multimodal fake news have achieved great success. However, these methods easily fuse information from different modality sources, such as concatenation and element-wise product, without considering how each modality affects the other, resulting in low accuracy. This study presents a focused survey on the use of deep learning approaches to detect multimodal visual and textual fake news on various social networks from 2019 to 2024. Several relevant factors are discussed, including a) the detection stage, which involves deep learning algorithms, b) methods for analyzing various data types, and c) choosing the best fusion mechanism to combine multiple data sources. This study delves into the existing constraints of previous studies to provide future tips for addressing open challenges and problems.**

***Keywords-misinformation; attention mechanism; fusion methods; social media; vision transformer***

## I. INTRODUCTION

Over the past few decades, fake news has become ubiquitous to the point of deceiving the public. When this kind of information becomes available, it causes social divisions and suspicions in the ruling environment and among individuals [1-3]. When data about a specific event (correct or incorrect) are disseminated, they changes people's beliefs, typically emphasizing certain prejudices. Furthermore, deceptive or manipulative news seeks to feed widespread ignorance and greed to benefit individuals or groups at the expense of society [4]. Recently, many social networks have become the first choice for transmitting knowledge and exchanging information and events, providing platforms for sharing opinions and beliefs with others around the world [5-6]. Several studies have focused on fake news detection. As a result, specific components have been developed, using some classic datasets, to provide insight into their issue of interest [7]. Some distinctive examples of fake news are the "Zinoviev Letter" [8], the fake news on the 2016 elections in the United States [9-10], and the untrue environmental report on the spread of fires in the Amazon rainforest in 2018 [11].

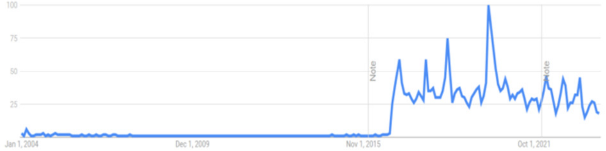

Fig. 1. Google fake news trends [10].

Social networks have become an ideal setting for the spread of rumors, threatening network order, people's health, and social stability [12-13]. Social networks and live streaming platforms have become an essential part of daily life. Several dictionaries have defined the term fake news [14], which can be defined more broadly based on its authenticity or intent [15]. One possible explanation for the widespread transmission of fake news is a lack of basic knowledge and skills within the population. The public is not informed of the legitimacy of the information sources and the veracity of the news it reads. Another factor is that there is a lack of automatic fact-checking procedures. Although few websites have made significant efforts to detect fake news, most of them rely on time-consuming manual methods. It is too difficult to prevent fake news since the extensive use of social networks allows the fast propagation of disinformation [7, 16].

Fake news detection is an ongoing study subject that can be interpreted from several angles. It aims to mitigate the negative effects of such news by creating a system that recognizes it using techniques, such as Machine Learning (ML), language proficiency, optimization algorithms, Deep Learning (DL), and others [5, 7]. However, since ML-based systems have several constraints, involvig generating a large training dataset and selecting appropriate features to best capture the deception, DL algorithms have been applied to detect fake news. In particular, attention mechanisms have emerged as one of the most potent strategies in Natural Language Processing (NLP). They are primarily used alongside Recurrent Neural Networks (RNNs) to anticipate the most significant information in an input sequence, either textual or visual [17]. Fake news providers frequently employ written content and visuals or distort facts to appeal to readers' psychology and entice and mislead them, allowing for quick diffusion. In general, themes on social hotspots or disputes include detailed textual descriptions of their emotional expression and visual influence on pictures [9]. Multimodal knowledge is more difficult to handle than single-modal knowledge since it requires information fusion procedures. Data fusion, decision-making, features, and other approaches are examples of information fusion. These approaches contain two steps: combining data, information, and features from multiple data sources and then processing them. As a result, they can provide an additional accurate and reliable data representation [18].

Table I portrays the most important abbreviations used in this paper. This study explored recent suggestive literature on fake news detection. In particular, the former focused on developing detection systems based on specific characteristics of multimodal fake news. The papers were obtained by searching for the keywords "fake news" through the search engines observed in Table II. Several review studies exist in this domain, as evidenced in Table III. The main contributions of this review study can be summarized as:

- Provides knowledge about the specific fake news attributes and their corresponding terms.
- Focuses on detecting multimodal fake news and explaining these system's methods to compare them in all stages from the perspective of description to detection.
- Focuses briefly on the DL methods deployed in fake news detection models, such as attention mechanisms, CNN, ResNet, etc.

## II. NATURAL LANGUAGE PROCESSING

NLP systems include morphological traits, lexical classes, syntactic categories, semantic connections, etc. In principle, statistical NLP models can be implemented to determine the relevance of these aspects, and so gain a greater understanding of the model. In contrast, it is more difficult to explain what occurs in a neural network model. Much of the analytical work therefore seeks to understand how language ideas, often used as features in NLP systems, are captured in neural networks. NLP techniques employ attention mechanisms to increase text classification accuracy. The attention model aims to improve efficiency by predicting the result based on only a few words of the input series rather than the complete phrase [19]. Furthermore, the development of pre-trained language models (e.g., BERT, RoBERTa, and GPT) and their utilization in NLP has opened up new ways to categorize fake news [18].

TABLE I. ABBREVIATIONS USED IN THIS PAPER

| Abbreviation | Description |
|---|---|
| CNN | Convolutional Neural Network |
| ResNet | Residual Neural Network |
| RNN | Recurrent Neural Network |
| ViT | Vision Transformers |
| BERT | Bidirectional Encoder Representation of Transformer |
| LSTM | Long Short-Term Memory |
| NLP | Natural Language Processing |
| GPT | Generative Pre-trained Transformers |
| POS | Parts of Speech Tagging |
| TF-IDF | Term Frequency Inverse Document Frequency |
| BoW | Bag of Words |
| GRU | Gated Recurrent Unit |
| ALBERT | A Lite Bidirectional Encoder Representation of Transformer |
| DeBERTa | Decoding-enhanced Bidirectional Encoder Representation of Transformer with Disentangled Attention |
| RoBERTa | Robustly optimized Bidirectional Encoder Representation of Transformer Pretraining approach |
| VGG | Visual Geometry Group |
| MLP | Multi-Layer Perceptron |
| DenseNet | Densely Connected Convolutional Networks |
| Glove | Global Vectors |

TABLE II. SEARCH ENGINES

| Search engine | Number of results | Selected references | Type |
|---|---|---|---|
| ACM Digital Library | 10,375 | [20, 21] | Journals |
| | | [22] | Conference |
| Science Direct | 1043 | [23, 24] | Journals |
| Google Scholar | 7854 | [9, 25-26] | Journals |
| ResearchGate | 216 | [16, 27] | Journals |
| Scopus | 361 | [18, 28-30] | Journals |
| IEEE | 75 | [31-33] | Conferences |
| | | [34-36] | Journals |
| MDPI | 79 | [11, 37] | Journals |
| Springer | 102 | [38] | Journal |

TABLE III. EXISTING REVIEWS ON DETECT MULTIMODAL FAKE NEWS

| Reference | Datasets | Word Embedding | Fusion Mechanism | Deep Learning Techniques | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | CNN | RNN | ViT | Attention | BERT | LSTM |
| [1] | √ | √ | √ | √ | √ | × | √ | √ | √ |
| [2] | √ | √ | × | √ | √ | × | √ | √ | √ |
| [3] | √ | √ | √ | √ | √ | × | √ | √ | √ |
| [39] | √ | √ | × | √ | √ | × | √ | √ | √ |
| [14] | √ | √ | × | √ | √ | × | √ | √ | √ |
| **This study** | √ | √ | √ | √ | √ | √ | √ | √ | √ |

Within NLP, the word and token features are similar. When applying ML algorithms to extract text, it is critical to identify the best features. The goal of identifying these traits is to develop effective indications that can be generalized for text classification. Some of them are mentioned below [40].

- n-grams is used to record the dependencies between all words that appear sequentially in a sentence structure. However, n-grams does not maintain the syntactical or semantic relationships of the words.
- Parts Of Speech (POS) Tagging: POS tagging distinguishes the grammatical meaning of words in a sentence putting into service particular tags, such as noun, pronoun, verb, adjective, adverb, conjunction, etc.
- TF-IDF: Its value increases linearly with the number of times a word appears in the document, but is offset by the term's frequency in the body. Although this vectorization is effective, the semantic meaning of the words is lost in the attempt to convert them to digits.
- BoW: This approach treats a single news story as a document and calculates the frequency count of each word to provide a numerical representation of the data. Along with data loss, this strategy has other drawbacks. The relative position of the words is ignored and contextual information is removed. This loss can sometimes be significant when weighed against the benefits of processing a pleasant level of usage.
- Word2Vec provides a set of model designs and optimizations to extract word embeddings from large datasets. Word embeddings learned through Word2Vec are more effective in collecting word semantics and leveraging word relatedness.

## III. DEEP LEARNING (DL)

DL networks are given sensory information such as texts, photographs, movies, or sounds to simulate the human learning process. These networks outperform other cutting-edge approaches in several tasks, and as a result, the field has expanded enormously [41]. CNN, RNN, LSTM, and GRU are some of the conventional DL models used to identify fake news. CNN-based techniques can extract relevant information from tiny areas but are incapable of dealing with larger structural links. Time-series techniques examine the sequential spread of misinformation using temporal structural elements while ignoring the broader structural characteristics of fake news. More importantly, these approaches cannot recognize many modes concurrently. For example, existing designs limit the ability to expand the detection to other modalities. Current fusion algorithms are not particularly sophisticated and cannot effectively integrate multi-modal advantages while avoiding noises offered by other sources [34]. Transfer learning has proven to be indispensable in DL training, as it transfers previously learned context knowledge to new designs that solve different issues [19].

### *A. Attention Mechanisms*

Attention mechanisms try to deal with input in the same way as the human brain/vision would. Human eyesight does not analyze the full image at once, instead it concentrates on individual areas. This allows the concentrated areas of the human visual space to be experienced in high resolution, while the surroundings appear in low resolution. Instead of analyzing the entire vision space, the brain can examine and narrow down the most important elements in a precise and efficient manner. This aspect of human eyesight led researchers to design the attention mechanism [42]. Attention mechanisms work by assigning varying weights to various types of information. Thus, assigning more weight to important information draws the focus of the DL model. Attention mechanism methods can be classified based on four criteria [16]:

- Softness of attention (deterministic attention): To generate the final context vector, the network calculates the average of each input weight item. The context vector is a high-dimensional vector that represents the components or sequence of the input factors, and the attention mechanism generally seeks to add more contextual information to the final context vector. Hard attention (stochastic attention) computes the final context vector by choosing pieces arbitrarily from the sample set. This decreases the computation time. In addition, global and local attention is often deployed in computer vision tasks. Global attention is like soft attention in that it evaluates all input items. However, the former improves soft attention by using the output of the current time step rather than the previous one, while local attention combines soft and hard attention. This technique evaluates a subset of input components at a time, overcoming the drawback of hard attention (i.e., being non-differentiable) while remaining computationally efficient.
- Attention mechanisms' ability can be classified according to their input requirements: item-wise and location-wise. Item-wise attention necessitates inputs that are directly known to the model or generated through pre-processing. However, location-wise attention is not implied because the model must deal with difficult-to-distinguish input objects.
- Attention models can work with single and multiple inputs. The overall processing strategy for the inputs varies between the created models. Most contemporary attention networks utilize a single input and process it in two separate sequences (i.e., a distinctive model). Certain connections exist within sources when recognizing multimodal systems (including images and text). Rather than simply splicing source features, the co-attention method is followed to simulate intense interactions between source features via sharing information and generates an attention-pooled feature for one modality (e.g., text) based on another one (e.g., image). The similarity of data pairs between sources is utilized to link them. A self-attention network computes attention solely based on model input, reducing the reliance on external data. This improves the model's performance in images with complicated backgrounds by focusing more on certain locations. The hierarchy attention mechanism computes weights based on the initial input and several of its levels. This attention mechanism is often referred to as fine-grained attention in image classification.

- Attention structures usually utilize a single output form. It processes one characteristic at a time and calculates weight ratings. There are two more multidimensional and multi-head attention systems. Multi-head attention evaluates inputs linearly in several groups before combining them to compute the final attention weights. This is especially advantageous when deploying the attention mechanism in conjunction with CNN approaches. Multidimensional attention, which is mostly employed for NLP, calculates weights utilizing a matrix representation of the characteristics rather than vectors.

Different types of attention mechanisms for computer vision can be classified into different categories [36]:

- Channel attention: This category assumes that in deep CNNs, distinct channels in various feature maps frequently represent various objects. As a result, channel attention is responsible for automatically calibrating the weight of each channel.
- Spatial Attention: This category is similar to channel attention. In this case, the attention mechanism is responsible for flexibly calibrating the weight of each part of the image. This system functions as an adaptive spatial area selection process, selecting where to focus.
- Temporal attention: This category considers data to have a time component. Thus, in computer vision tasks, this form of attention mechanism is commonly used for video analysis. This system operates as a dynamic temporal selection process, selecting when to pay attention.
- Branch attention: This category covers multi-branched DL architectures. Branch attention is to adapt to the weight of each branch. This mechanism functions as a dynamic branch selection process, deciding which branches to pay attention to.
- Channel and spatial attention: This approach functions as a dynamic spatial area and object choice procedure, deciding what and where to focus attention.
- Spatial and temporal attention: This system functions as a dynamic geographic area and time-frame process to select where and when to focus.

### B. Transformers

Transformers primarily deploy the self-attention mechanism to extract fundamental characteristics and have enormous promise for widespread use in AI [43]. Transformers, compared to RNNs, can attend to full sequences, and thus learn long-term connections. Transformers parse text in parallel implementing a powerful attention mechanism, producing complex and meaningful word descriptions. This approach looks at the relationships between textual phrases or entities. Many competing models of neural pattern transmission contain an encoder-decoder component. The encoder turns an endless flow of symbols from the input to a continuous output. The decoder then generates an output series involving one symbol at a time, using the encoder's continuous form [44]. BERT is an encoder layer with a transformer design. Instead of a static periodic function in the transformer, BERT learns the embedding location. This increases learning effort in the relevant step, but additional efforts could be almost completely avoided given the number of trainable parameters in the encoder [16]. In certain recent-related tasks, BERT-based models outperform RNN and CNN networks. The Swin transformer broadens the usefulness of the transformer, transferring its outstanding performance to visual surroundings, addresses the shortage of CNNs for global information feature extraction, and, with its unique window mechanism, substantially reduces the computational cost of self-attention and solves the challenge of secured token scale, which has become the general core of computer vision research [37]. ALBERT is a more portable form of BERT to address the drawbacks of the huge number of parameters and the lengthy training time [44]. DeBERTa is an improved BERT with disentangled attention and has two new features. First, the model suggests a disentangled attention. In DeBERTa, each token in the input is represented by two separate vectors that encode its word embedding and place. Attention weights among words are acquired utilizing disentangled matrices in this paired form. Second, an Enhanced Mask Decoder (EMD) is employed to forecast the masked tokens during the pre-training phase. Although BERT depends on relative places, EMD enables DeBERTa to make more accurate predictions since the syntactic functions of words are greatly influenced by their current location within the sentence. In an equivalent spirit, the BERTweet approach shares a similar architecture to BERT and was trained adopting the RoBERTa pre-training process [45]. Vision transformers break the image into 2D patches and feed them into the framework. However, vision transformers face several hurdles, including computational cost, dimensions, scalability to huge datasets, understanding, resilience to adversarial attacks, and generalization accuracy [46].

## IV. FAKE NEWS DETECTION

Fake news detection models can be categorized according to the following strategies: Strategies based on knowledge, features, and modality [47]. From a knowledge viewpoint, an impartial fact-checker reviews news stories and assigns an actual value to statements. The three kinds of fact-checking are expert-oriented, assessing the accuracy of information by relying on domain-matter experts who analyze data and documents and draw conclusions, crowd-sourcing-oriented, allowing users to discuss and comment on the accuracy of specific news resources, and computational-oriented, an intelligent system that classifies a news item as having true or false matter.

AI-based algorithms to detect fake news rely on a variety of important criteria, including content-based, network-based, and user-based attributes. However, combining all these variables may not increase the classifier's performance. Many studies relied only on content features or content-based characteristics (textual and visual) in conjunction with using additional characteristics to detect fake news. Existing fake news identification research is divided into two groups: single-modal and multimodal.

### *A. Single-Modal Fake News Approaches*

In general, text and image characteristics can be employed to detect fake news alone, while other features are typically deployed as supplementary to help identification. The single-model-based technique utilizes only one characteristic to detect. In [48], the relevance of an image component was used for automatic false news detection on social media to address this issue. It has been established that authentic and false news events have different image distribution patterns. In [49], a co-attention technique was followed to identify the top K most significant phrases in a news story and the top K most important user evaluations for the final classification. In [50], a CNN-based capsule network model with pre-trained word embeddings was implemented to classify false news in the ISOT and LIAR datasets. In [51], a generative model was proposed to extract new patterns and aid in the identification of fake news by examining previous relevant user reactions. In [52], n-grams were applied with TF-IDF word embedding to obtain content characteristics, and LSTM and BERT models were trained to deal with contextual information. Then a feedforward neural network was utilized for classification. However, this technique did not account for the complete use of different textual characteristics.

### *B. Multi-Modal Fake News*

In general, social media postings featuring photos and graphics receive far more retweets and comments and spread much faster than those having only text. Images are spread widely, captivating people's emotions, and expressing a sense of reality. Images related to a post may have been edited or simply taken out of context. It is not uncommon to distort images for political or personal motives, as well as to use photo editing software to change an image. As a result, when analyzing both text and images, photo captions are critical to identifying clickbait and false captions [29].

#### *1) Datasets*

Table IV lists multimodal datasets applied in various studies, and Figure 3 describes the most popular dataset dimensions.

#### *2) Textual and Visual Preprocessing*

Pre-processing in text starts with cleaning the input datasets by extracting excess, extreme, and duplicative text parts. Word embedding methods keep only meaningful tokens that are transformed into vectors. The text is stemmed/lemmatized, normalized, and tokenized. Stemming and lemmatization remove words and symbols without meaning. Normalization transforms text into canonical form. Stemming cuts off the ends of input words to lower their inflection and convert them into their core structures. In general, the canonical form of the original input word is deployed. It is very important to normalize text in the web scope and social media data, as it contains a lot of noise, such as abbreviations, misspellings, and words that are out of vocabulary. Image data are pre-processed by reviewing that all URLs are correct. It is also important to normalize the size of images and divide them for training and testing. Textual and visual data are pre-processed individually and then merged to complete each instance in terms of its three parameters: title, text, and vision [20].

TABLE IV. IMPORTANT MULTIMODAL DATASETS

| Dataset | Description | Study |
|---|---|---|
| MedialEval (2015) | Contains 15000 items, including 176 images in 5,008 real news tweets and 185 misused images in 7032 fraudulent tweets. | [33] |
| Twitter (2016) | It contains 7898 fake news, 6026 real news, and 514 images. | [9, 11, 21, 23, 30, 36] |
| MediaEval (2016) | Contains 17,000 unique tweets on various events. One-third are real and the remaining are fake news. | [32] |
| Weibo (2017) | Consists of 4749 fake and 4779 real news. | [9, 23, 27, 33, 36, 37] |
| Fakeddit (2019) | Multimodal standard dataset of 1,063,106 samples. | [16, 38] |
| Gossip (2020) | News stories include text, news image link, publishing time, author name, and social media responses. | [9, 21, 26] |
| Politifact (2020) | Contains text, news image location, publishing time, and remarks made on social networks. | [9, 18, 21] |
| All Data (2020) | Contains 11,941 fake and 8,074 real news. | [19, 34] |
| ReCOVery (2020) | Contains 2,029 news articles shared on social media, most of which (2,017) have both textual and visual information. | [53] |
| Twitter Indian Dataset v3 (2021) | Contains a list of fake and accurate news stories covered primarily from politics, Bollywood, and religion. | [29] |
| Ti-CNN (2021) | 20,000 articles from websites, including over 11,000 fake and more than 8,000 real news items. | [9, 21] |
| Fake news sample by Guilherme Pontes (2021) | 45,569 news, 25,343 are real the remaining are fake news articles. | [20] |
| Twitter_database (2023) | Includes 5 partitions to perform 5-fold cross-validation. | [26] |

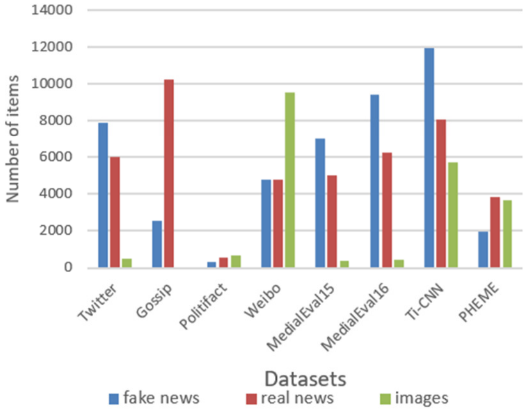

Fig. 2. Distribution of multimodal fake news datasets.

#### *3) Textual and Visual Feature Extraction and Selection*

Textual properties can be obtained at several levels in the hierarchy, such as word, sentence, and message. The most basic lexical characteristics are the overall total of characters, the number of different words, the average length of words, and so on. In the meantime, the semantics of linguistic characteristics, such as the proportion of first/third person pronouns, the number of news detection by pooling and attention blocks, and positive or negative emoji symbols, are all accessible options. Unlike linguistic characteristics, syntactic features improve the aim of feature extraction to a significant level: emotion score or part-of-speech labeling. Recently, various complicated models, namely BoW, Word2Vec, and other embedding techniques, have been used to recognize fake news. Image extraction provides additional visual information. Several studies employed the BERT pre-trained model to extract text characteristics. However, the BERT model has many parameters and a slower training speed. Furthermore, visual and text characteristics are in separate semantic feature spaces, resulting in heterogeneity [31].

### 4) *Fusion Mechanism*

The combination of textual content and images is one of the widely utilized features for multimodal fake news detection. The intuition behind this cue is that some fake news spreaders deploy tempting images, e.g., exaggerated, dramatic, or sarcastic graphics, that are far from the textual content to attract users' attention. Information fusion techniques have an original ability to manage input data with their multimodal nature. Many experiments have proven the benefit of these techniques and that their full exploitation leads to improved performance [31, 38]. Several techniques combine textual and visual information into a single representation, ignoring their associations, which might lead to poor results. Fusion can be classified according to different times as follows [53]:

- Early fusion (feature fusion): Feature vectors from multiple modalities are combined and fed into a model for prediction. Due to the fusion of pre-processed features from different modalities at the input layer, working with features with higher granularity becomes tedious (Figure 3).
- Late fusion (decision-level fusion or kernel-level fusion) combines results from various modalities using summation, maximization, averages, or weighted average methods. Most late fusion solutions employ handcrafted rules, prone to human bias and far from real-world peculiarities (Figure 3).
- Intermediate fusion (mid-fusion) involves combining units from several modality-specific paths into a single shared layer. It is possible to create a representation layer either by mapping multiple channels at the same time or by combining different modal sets at various levels.

Fusion mechanisms can be divided according to the technology followed to merge textual and visual attributes in [36]:

- Simple operation-based: DL combines vectorized features from several data sources using fundamental algorithms such as concatenation or weighted addition. As models based on DL techniques are trained concurrently, the features of high-level standards could be extracted at a level that accommodates both activities. Such processes often have minimal or no correlation factors.
- Attention-based: Fusion often involves attention processing. Different outputs are frequently used to provide different sets of changing weights for summing, preserving more information by merging the results from each peek.
- Bilinear pooling-based: This is achieved by adding the external product of both vectors (text and image input vectors) to increase and multiply the exchanges between all elements of both vectors. This process is more expressive.

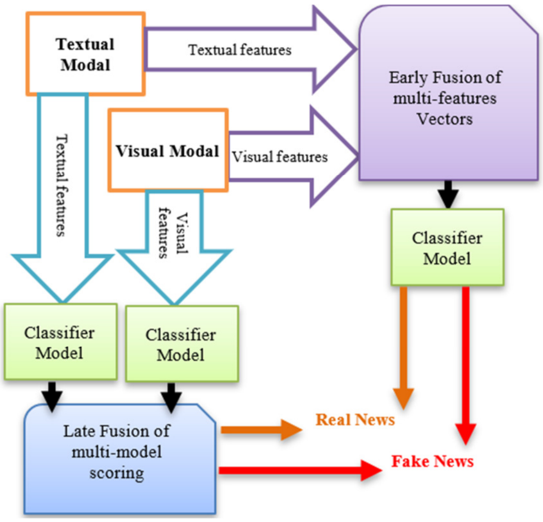

Fig. 3. Early and late fusion mechanism structures.

### 5) *Model Evaluation Metrics*

A confusion matrix serves as the basis for evaluating a classification model. True Positives (TP) indicate news that was projected to be true and was true, False positives (FP) indicate news projected to be true but was fake, True Negatives (TN) indicate news that was projected to be false and was untrue, and False Negatives (FN) indicate news projected as untrue but was accurate. [52]. The efficiency of a model is evaluated by [54-56]:

$$Accuracy\ (Acc) = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Precision\ (Pre) = \frac{TP}{TP+FP} \quad (2)$$

$$Recall\ (Rec) = \frac{TP}{TP+FN} \quad (3)$$

$$F1 = \frac{2\times Precision\times Recall}{Recall+Precision} \quad (4)$$

## C. *Studies on Multimodal Fake News Detection*

In [32], a scaled dot product attention mechanism was implemented to capture the relationship between the text features extracted by BERT and the image features extracted by VGG-19. In [33], another model based also on BERT and VGG19 was proposed, accepting both text and picture input. Subsequently, the pair of embedding was joined and subjected to a multi-modal variation autoencoder to obtain the common latent representation. A multimodal cross-attention network was designed to fuse the resulting features. In [23], four distinct submodules made up a fake news system: feature fusion based on multi-modal factorized bilinear pooling, two attention mechanisms, one for textual description combined with Stacked BiLSTM and the other for visual feature extraction, combined with multi-level CNN–RNN, and MLP for classification. In [20], visual picture attributes were extracted using image captioning and forensic evaluation, and textual hidden patterns were extracted employing a Hierarchical

Attention Network (HAN). In [21], a multi-modal coupled ConvNet architecture was presented, combining textual and visual data modules from three datasets and utilizing a late fusion mechanism. In [58], a detection framework was proposed, deploying Word2Vec to fuse text input to the embedding layer and passing image input to a cross-modal attention residual and multi-channel CNN. The multi-channel CNN was implemented as a reducer to the amount of trash data produced by cross-modal fusion parts. In [9], an MCNN was proposed, considering the consistency of multi-modal data and capturing the overall characteristics of social media information based on an early fusion mechanism. This model used BERT in the text feature extraction module and the attention mechanism with ResNet-50 in the visual semantic feature extraction module. In [34], three modalities were evaluated: text, image, and image attributes. Additionally, a model based on dual attention fusion networks was applied to combine features. Initially, the model extracts image (based on ResNet-50 V2) and text modalities (based on BERT). In the end, the features were combined to create a feature vector that can be used for classification. In [28], news post images were converted from their spatial dimension to their frequency domain utilizing machine learning. Subsequently, a multi-layer CNN model was engaged to extract the characteristics of the frequency picture, and MML was deployed to retrieve image-related web pages on Google. Simultaneously, MML uses the evidence veracity classification task to support the false news detection task by selecting evidence. This part involved feeding the evidence and the claim into a BERT-based encoder, followed by learning evidence representations employing claim-evidence correlation representations. Ultimately, the co-attention process fuses the representations of the image with relevant evidence. In [35], a model was presented based on two principles, blocking and fusion. This model determined the spatial and temporal location of the data in the fusion mechanism for the visual and textual attributes. In [37], text features were extracted from bidirectional encoder representations of transformers, image features were extracted from Swin-transformers, and then deep autoencoding was used as an early fusion technique by merging text and visual attributes. In [38], the proposed framework was based on the BERT and Xception models to learn visual and linguistic models. In [31], the ALBERT model was combined with a multi-modal circulant fusion technique to detect fake news. This system included a textual feature extractor (ALBERT), a visual feature extractor (VGG-19), a feature fusion, a fake news detector, and domain classification modules. In [26], multimodal pre-processing of both words and images was performed. Glove embedding and Word2vector approaches were deployed to extract the text characteristics and the Adaptive Water Strider Algorithm (A-WSA) was applied to extract the best characteristics from both text and image data. Feature fusion receives the optimized features, which are obtained by the same A-WSA optimization process based on the weight factor. Lastly, O-BiLSTM was utilized for fake news classification. In [27], a model based on BLIP (FNDB) was proposed. XLNet and VGG-19-based feature extractors were engaged to extract textual and visual feature representations, respectively, and the BLIP-based multimodal feature extractor was put into service to obtain multimodal feature representation in news content. Then, the feature fusion layer combined these features with the help of the cross-modal attention module to promote various modal feature representations to complement the information. In [29], a multi-modal DL technique was proposed to use and process visual and textual features, employing EfficientNet-B0 and a sentence transformer. Feature embedding was performed on individual channels, while fusion was performed on the last classification layer. Late fusion was applied to mitigate the noisy data generated by multi-modalities. In [11], TLFND was proposed, which was based on a three-phase feature-matching distance technique to detect fake news. An attention-guiding module was devised to assist in aggregating the cross-modality correlations and the aligned unimodal representations in an effective and interpretable manner. In [30], a model based on transformers and multi-modal fusion was introduced. This model extracts text and image features using different transformers, and fuse features implementing attention mechanisms. In [18], a quantum-based standard was proposed for multimedia data fusion to identify fake news. This system extracted features in both textual and visual forms and sent them to the convolutional-quantum network to achieve classification.

## V. DISCUSSION

DL has begun to be strongly involved in multimodal fake news detection systems at all stages, whether it is engaged in extracting features of textual and visual inputs, in the mechanisms of fusing features extracted from multimodal data, or in the classification of fake news. It is possible to detect fake news adopting these strategies but some restrictions limit their accuracy, involving the requirement for a huge dataset containing diverse data in all fields of life (political, economic, technological, technical, and health, etc.), in addition to the inability to fuse the extracted features efficiently, take advantage of the most multimodal important features, and measure the extent of interconnection between them. Some studies focused on a single social network, such as Twitter, Weibo, or Facebook, but future fake news detection systems must be applicable on different websites and social networks to acquire knowledge deeply and detect fake news quickly. Many studies used BERT word embedding [9, 33-35, 37-38, 57] and depreciated traditional techniques, such as GloVe [20-21] and Word2Vec [24, 58] in their textual feature extraction model. BERT can discover the implicit associations within the sentence words and texts in which the system is trained, but that has not prevented a recent trend toward including derived models, like RoBERTa [11, 16], ALBERT [31], distilBERT [29], and XLNet [18, 27]. Although all proposed multimodal fake news detection systems still use CNN [21, 25, 26], VGG [31-33], and ResNet [34-35] neural networks in a visual feature extraction stage, there is a new strategy deploying ViT for textual and visual feature extraction. Regarding fusion techniques, it is clear that in recent years there was no clear interest in examining how to benefit from extracted features and how to choose, as concatenation [32, 33, 38] of extracted features is the common fusing operation in early or late fusion mechanisms. However, there is interest in the technology of attention mechanisms [23, 32, 36] and their strong entry during the past two years to support the approved fusion mechanisms.

TABLE V. A COMPARISON OF SEVERAL MULTIMODAL FAKE NEWS DETECTION APPROACHES

| Ref. | Datasets | Evaluation metrics (%) | Feature extraction | | Fusion mechanism | DL model | Drawback | Future scope |
|---|---|---|---|---|---|---|---|---|
| | | | Textual | Visual | | | | |
| [57] | Twitter | Acc=83, Pre=81, Rec=63, F1=71 | BERT | VGG-19 | Concatenation | FCL | The only image system scored lower than the only text one on the Twitter dataset. | Apply probabilistic method and deep model to evaluate if the image pertains to the written content. |
| | Weibo | Acc=84.2, Pre=83, Rec=87, F1=85 | | | | | | |
| [9] | Ti-CNN | Acc=96.3, Pre=97.2, Rec=96.4, F1=96.8 | BERT | ResNet-50 + Attention | Early fusion | FCL | High time cost. | Enhance the approach at the feature fusion phase for a better fit of multimodal features in different locations. |
| | Weibo | Acc=94.7, Pre=95.2, Rec=94.2, F1=94.6 | | | | | | |
| | Twitter | Acc=78.4, Pre=85, Rec=81.4, F1=83.1 | | | | | | |
| | PolitiFact | Acc=88.4, Pre=97.3, Rec=86.7, F1=91.7 | | | | | | |
| [20] | All Data | Acc=95.5, Pre=94.5, Rec=94.4, F1=94.4 | GloVe+ Hierarchical Attention Network | Error Level Analysis | Concatenation | Max voting ensemble technique | Few features | Add new features to improve efficacy. |
| | FND by Jruvika | Acc=94.7, Pre=95.6, Rec=93.1, F1=94.4 | | | | | | |
| | Fake News Sample | Acc=95.9, Pre=97.8, Rec=94.6, F1=96.25 | | | | | | |
| [21] | TI-CNN | Acc=96.2, Pre=95.7, Rec=96, F1=95.89 | Glove+ Text-CNN | Image-CNN | Late Fusion | FCL | Cannot extract deep characteristics. | Effective classification algorithm based on CNN with fine-tuned hyperparameters to improve fake news detection. |
| | Emergent | Acc=93.5, Pre=94.1, Rec=89.3, F1=93.12 | | | | | | |
| | MICC-F220 | Acc=95.1, Pre=95.1, Rec=78.2, F1=85.88 | | | | | | |
| [23] | Twitter | Acc=88.3, Pre=89, Rec=95, F1=92 | Attention+ Bi-LSTM | Attention + 2-Level CNN–RNN | Multimodal factorized bi-linear pooling | MLP | Can lose some critical information depending on the feature extraction approach. | Semantic connections of text and images to improve fusion methods. |
| | Weibo | Acc=83.2, Pre=82, Rec=86,F1=84 | | | | | | |
| [24] | Pheme | Acc=87.2, Pre=83.7, Rec=78, F1=80.7 | Word2vec+Bi LSTM | VGG-19 | Hierarchical attention mechanism | FCL | High time cost. | Detect forged images. |
| | Weibo | Acc=83.4, Pre=86.3, Rec=78, F1=82.4 | | | | | | |
| [32] | MediaEval (2016) | Acc=81.2, Pre=81.3, Rec=87.4, F1=84.3 | BERT+ CNN+ Attention mechanism | VGG-19+ Self-attention | Concatenation | FCL | Small dataset. | Use many images to identify fake news. |
| [33] | MediaEval (2016) | F1=92.4 | BERT | VGG-19 | Concatenation | FCL | Ensure data quality taken from an image or text before fusion. | Context-dependent latent representations such as image captioning. |
| | Weibo | F1=65.6 | | | | | | |
| [34] | All Data | Pre=97.8, Rec=98.2, F1=98.07 | BERT | ResNet-50 V2 | Dual attention mechanism | FCL | Small dataset. | Add other modalities. |
| [35] | Weibo | Acc=87.9, Pre=88.6, Rec=87.1, F1=87.9 | BERT | ResNet-50 | Cross-attention | FCL | Dataset constraints. | Easier way to leverage prior experience in deep networks. |
| [58] | Twitter | Acc=84.2, Pre=85.4, Rec=61.9, F1=71.8 | Word2Vec+Self-attention +Cross modal attention | VGG-19 +Cross-modal attention | Concatenation | FCL | Need testing on real-time fake information. | Investigate event-level multimodal fake news detection using visual data. |
| | Weibo A | Acc=85.3, Pre=89.1, Rec=81.4, F1=85.1 | | | | | | |
| | Weibo B | Acc=86.9, Pre=93.5, Rec=79.6, F1=86 | | | | | | |
| | Weibo C | Acc=92.2, Pre=89, Rec=96.5, F1=92.6 | | | | | | |
| [16] | Fakeddit | Acc=88.1, Pre=87.1, Rec=87.9, F1=87.51 | RoBERTa | DenseNet-161 | Co-attention | FCL | Improved recognition of changed material, improper connections, and reality compared to fraudulent content. | Detect extra fake data, such as imposter material, and satire/parody. |
| [25] | Fakeddit | Pre=75, Rec=79, F1=77 | BERT | CNN | Early fusion+ Concatenation | Linear layer | Inaccuracy in extracting the most relevant features | Other DL approaches (GRU) and ways to merge visual and textual representations. |
| [28] | CCMR | Acc=92.2, Pre=91.6, Rec=92.6, F1=92.15 | BERT | multiCNN | Co-attention | MLP | Using Google search takes more time. | Ideal extraction of images and text dataset. |
| [31] | Twitter | Acc=85, Pre=84.2, Rec=65.4, F1=73.6 | ALBERT | VGG-19 | Multi-modal circulant fusion | FCL | Extracting attributes is not deep enough. | Stronger visual information technique. |
| | Weibo | Acc=86.1, Pre=85.5, Rec=88.5, F1=87 | | | | | | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| [36] | Weibo | Acc=65.4, Pre=66.4, Rec=66.8, F1=66.6 | BiLSTM | VGG-19 +pooling | Attention mechanism | FCL | Failure of fine-tuned feature extraction. | Other fusion methods based on attention mechanisms. |
| [37] | Twitter | Acc=75.6, Pre=72.8, Rec=97.7, F1=83.4 | BERT | Swin-transformer | Trained deep autoencoder | FCL | If a post has many images, only one may be used to identify it. | Reduce the model's difficulty to ensure its use on small devices. |
| | Weibo | Acc=59.7, Pre=56.4, Rec=99.4, F1=71.9 | | | | | | |
| [38] | Fakeddit | Acc=91.8, Pre=93.3, Rec=93.2, F1=93.2 | BERT | Xception | Concatenation | FCL | Small dataset. | Use post content and comments, together with user-related data. |
| [11] | Politifact | Acc=94.4, Pre=97.4, Rec=96.6, F1=97 | RoBERTa | VGG-19+ BiLSTM | Concatenation | FCL | A fusion approach that focuses on the contents of the extracted features. | Adapting new areas and improving technology while testing the proposed model is underway. |
| | Gossipcop | Acc=90.9, Pre=93.2, Rec=94.7, F1=93.9 | | | | | | |
| | Twitter | Acc=83.1, Pre=85.2, Rec=82.4, F1=83.7 | | | | | | |
| [15] | Twitter | Acc=86.8, Pre=83.1, Rec=75.4, F1=79.1 | BERT+ two-text-branch | ResNet-50 +two-image-branch | Multi-modal bilinear pooling +Self-attention mechanism | FCL | The use of multiple techniques negatively affects execution time | Combine textual information with several photos. |
| | Weibo | Acc=90.4, Pre=94.3, Rec=87.1, F1=90.5 | | | | | | |
| [26] | Twitter | Acc=96.5, Pre=88.8, Rec=96.2, F1=92.41 | Word2vec +Glove | ResNet-50 +VGG-16 | Adaptive feature fusion | O-BiLSTM based on optimized WSA | Not extracting textual features efficiently. | Use audio signals and captions to detect false news videos. |
| [27] | Weibo | Acc=88.8, Pre=89.1, Rec=97.2, F1=93 | XLNet | VGG-19 | Cross-modal attention | FCL | Use a more effective model to extract features. | Improve the fusion process. |
| | Gossipcop | Acc=87.3, Pre=79, Rec=44, F1=56.5 | | | | | | |
| [29] | MediaEval | Acc=86.4, Pre=84, Rec=93, F1=88 | DistilBERT | Efficient-Net-B0 | Late fusion | ANN | High-resolution photos with only a small altered area appeared to be poorly detected. | Detect satirical news and the text that is placed over the photos. |
| | Weibo | Acc=81.4, Pre=80.3, Rec=86.3, F1=83.6 | | | | | | |
| | Twitter Indian Dataset v3 | Acc=67.1 | | | | | | |
| | Fakeddit | Acc=88.8, Pre=85, Rec=87, F1=86 | | | | | | |
| [30] | Twitter | Acc=93.5, Pre=96.5, Rec=93.7, F1=95.1 | BERT+ BiLSTM | VGG-19 | Concatenation | MLP | Cannot be used directly when one of the modalities is lacking. | Improve feature extraction to counteract intentionally deceiving photos. |
| | Weibo | Acc=91.5, Pre=91.3, Rec=91.3, F1=91.3 | | | | | | |
| [59] | Twitter | Acc=91.8, Pre=91.2, Rec=85.4, F1=91.8 | GloVe+ Transformer | ViT | Late fusion based on attention mechanism | MLP | Time complexity | Enhance the model for cross-domain news detection. |
| | Weibo | Acc=92.2, Pre=96.9, Rec=88.6, F1=92.5 | | | | | | |
| [18] | Gossip | Acc=87.9, Pre=95.8, Rec=89.9, F1=92.8 | XLNet | VGG-19 | Quantum multimodal fusion | FCL | Quantum circuit with time-based complexity | Apply quantum fuzzy neural networks. |

## VI. RESEARCH GAPS AND CHALLENGES

Fake news is fundamentally multimodal and multilingual, taking visual, auditory, or literary forms and expressing itself in a language that readers may not be familiar with. A new viewpoint can be developed to make deep systems more acceptable. Additionally, appropriate feature collection and classification techniques can improve the detection of fake news. Studies must investigate whether the classification approach is most appropriate for certain features: textual or visual feature extractors. As a result, greater attention must be paid to feature choice and fusion to improve performance. The challenges in multimodal fake news detection approaches can be summarized as:

- Existing techniques often employ a basic concatenate strategy to fuse inter-modal information, yielding mediocre detection results.
- There is a significant difference between image similarities and sentences in most fake news, but existing algorithms do not fully capitalize on this.
- The lack of large and rich multimodal fake news datasets negatively affects system development. In addition, datasets are limited to the economic or political field only. In addition, the lack of multilingual datasets supports the possibility of developing fake news detection systems in several languages and different dialects of the same language as well.
- Not relying on psychological data, combined with the contextual features of texts and images of published news, saves a great deal of time in contacting people responsible for false information sharing and revealing their purposes.

## VII. CONCLUSION

After studying the literature on fake news analysis methods, this paper summarized the basic features of multimodal fake news detection systems, including datasets, visual and textual preprocessing, feature extraction, fusion mechanisms, and fake news detection stages, as well as related techniques such as BERT, transformer, ViT, and attention mechanisms. A brief review of important multimodal fake news detection systems was performed, with different deep learning methods in different stages. Future studies could focus on modern attention mechanisms in fake video detection systems. In addition, efficient early detection mechanisms must be developed.

## REFERENCES

[1] S. Hangloo and B. Arora, "Combating multimodal fake news on social media: methods, datasets, and future perspective," *Multimedia Systems*, vol. 28, no. 6, pp. 2391–2422, Dec. 2022, https://doi.org/10.1007/s00530-022-00966-y.

[2] L. Hu, S. Wei, Z. Zhao, and B. Wu, "Deep learning for fake news detection: A comprehensive survey," *AI Open*, vol. 3, pp. 133–155, Jan. 2022, https://doi.org/10.1016/j.aiopen.2022.09.001.

[3] C. Comito, L. Caroprese, and E. Zumpano, "Multimodal fake news detection on social media: a survey of deep learning techniques," *Social Network Analysis and Mining*, vol. 13, no. 1, Aug. 2023, Art. no. 101, https://doi.org/10.1007/s13278-023-01104-w.

[4] D. Gifu, "An Intelligent System for Detecting Fake News," *Procedia Computer Science*, vol. 221, pp. 1058–1065, Jan. 2023, https://doi.org/10.1016/j.procs.2023.08.088.

[5] J. Li and M. Lei, "A Brief Survey for Fake News Detection via Deep Learning Models," *Procedia Computer Science*, vol. 214, pp. 1339–1344, Jan. 2022, https://doi.org/10.1016/j.procs.2022.11.314.

[6] A. Gandhi, K. Adhvaryu, S. Poria, E. Cambria, and A. Hussain, "Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions," *Information Fusion*, vol. 91, pp. 424–444, Mar. 2023, https://doi.org/10.1016/j.inffus.2022.09.025.

[7] M. Nirav Shah and A. Ganatra, "A systematic literature review and existing challenges toward fake news detection models," *Social Network Analysis and Mining*, vol. 12, no. 1, Nov. 2022, Art. no. 168, https://doi.org/10.1007/s13278-022-00995-5.

[8] A. Figueira, N. Guimaraes, and L. Torgo, "Current State of the Art to Detect Fake News in Social Media: Global Trendings and Next Challenges:," in *Proceedings of the 14th International Conference on Web Information Systems and Technologies*, Seville, Spain, 2018, pp. 332–339, https://doi.org/10.5220/0007188503320339.

[9] J. Xue, Y. Wang, Y. Tian, Y. Li, L. Shi, and L. Wei, "Detecting fake news by exploring the consistency of multimodal data," *Information Processing & Management*, vol. 58, no. 5, Sep. 2021, Art. no. 102610, https://doi.org/10.1016/j.ipm.2021.102610.

[10] "Google Trends," *Google Trends*. https://trends.google.com/trends/explore?date=today%205-y&q=fake%20news&hl=en (accessed May 30, 2024).

[11] J. Wang, J. Zheng, S. Yao, R. Wang, and H. Du, "TLFND: A Multimodal Fusion Model Based on Three-Level Feature Matching Distance for Fake News Detection," *Entropy*, vol. 25, no. 11, Nov. 2023, Art. no. 1533, https://doi.org/10.3390/e25111533.

[12] K. Liu and M. Hai, "Rumor Detection of Covid-19 Related Microblogs on Sina Weibo," *Procedia Computer Science*, vol. 221, pp. 386–393, Jan. 2023, https://doi.org/10.1016/j.procs.2023.07.052.

[13] S. Ahmed, K. Hinkelmann, and F. Corradini, "Combining Machine Learning with Knowledge Engineering to detect Fake News in Social Networks - a survey." arXiv, Jan. 20, 2022, https://doi.org/10.48550/arXiv.2201.08032.

[14] Y. Shen, Q. Liu, N. Guo, J. Yuan, and Y. Yang, "Fake News Detection on Social Networks: A Survey," *Applied Sciences*, vol. 13, no. 21, Jan. 2023, Art. no. 11877, https://doi.org/10.3390/app132111877.

[15] Y. Guo, H. Ge, and J. Li, "A two-branch multimodal fake news detection model based on multimodal bilinear pooling and attention mechanism," *Frontiers in Computer Science*, vol. 5, Apr. 2023, https://doi.org/10.3389/fcomp.2023.1159063.

[16] L. Qian, R. Xu, and Z. Zhou, "MRDCA: a multimodal approach for fine-grained fake news detection through integration of RoBERTa and DenseNet based upon fusion mechanism of co-attention," *Annals of Operations Research*, Dec. 2022, https://doi.org/10.1007/s10479-022-05154-9.

[17] A. M. Luvembe, W. Li, S. Li, F. Liu, and X. Wu, "CAF-ODNN: Complementary attention fusion with optimized deep neural network for multimodal fake news detection," *Information Processing & Management*, vol. 61, no. 3, May 2024, Art. no. 103653, https://doi.org/10.1016/j.ipm.2024.103653.

[18] Z. Qu, Y. Meng, G. Muhammad, and P. Tiwari, "QMFND: A quantum multimodal fusion-based fake news detection model for social media," *Information Fusion*, vol. 104, Apr. 2024, Art. no. 102172, https://doi.org/10.1016/j.inffus.2023.102172.

[19] F. A. O. Santos, K. L. Ponce-Guevara, D. Macêdo, and C. Zanchettin, "Improving Universal Language Model Fine-Tuning using Attention Mechanism," in *2019 International Joint Conference on Neural Networks (IJCNN)*, Budapest, Hungary, Jul. 2019, pp. 1–7, https://doi.org/10.1109/IJCNN.2019.8852398.

[20] P. Meel and D. K. Vishwakarma, "HAN, image captioning, and forensics ensemble multimodal fake news detection," *Information Sciences*, vol. 567, pp. 23–41, Aug. 2021, https://doi.org/10.1016/j.ins.2021.03.037.

[21] C. Raj and P. Meel, "ConvNet frameworks for multi-modal fake news detection," *Applied Intelligence*, vol. 51, no. 11, pp. 8132–8148, Nov. 2021, https://doi.org/10.1007/s10489-021-02345-y.

[22] L. Wang, C. Zhang, H. Xu, Y. Xu, X. Xu, and S. Wang, "Cross-modal Contrastive Learning for Multimodal Fake News Detection," in *Proceedings of the 31st ACM International Conference on Multimedia*, Ottawa, Canada, Nov. 2023, pp. 5696–5704, https://doi.org/10.1145/3581783.3613850.

[23] R. Kumari and A. Ekbal, "AMFB: Attention based multimodal Factorized Bilinear Pooling for multimodal Fake News Detection," *Expert Systems with Applications*, vol. 184, Dec. 2021, Art. no. 115412, https://doi.org/10.1016/j.eswa.2021.115412.

[24] J. Zeng, Y. Zhang, and X. Ma, "Fake news detection for epidemic emergencies via deep correlations between text and images," *Sustainable Cities and Society*, vol. 66, Mar. 2021, Art. no. 102652, https://doi.org/10.1016/j.scs.2020.102652.

[25] I. Segura-Bedmar and S. Alonso-Bartolome, "Multimodal Fake News Detection," *Information*, vol. 13, no. 6, Jun. 2022, Art. no. 284, https://doi.org/10.3390/info13060284.

[26] V. Kishore and M. Kumar, "Enhanced Multimodal Fake News Detection with Optimal Feature Fusion and Modified Bi-LSTM Architecture," *Cybernetics and Systems*, Jan. 2023, https://doi.org/10.1080/01969722.2023.2175155.

[27] Z. Liang, "Fake News Detection Based on Multimodal Inputs," *Computers, Materials & Continua*, vol. 75, no. 2, pp. 4519–4534, 2023, https://doi.org/10.32604/cmc.2023.037035.

[28] X. Cui and Y. Li, "Fake News Detection in Social Media based on Multi-Modal Multi-Task Learning," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 13, no. 7, 31 2022, https://doi.org/10.14569/IJACSA.2022.01307106.

[29] D. K. Sharma, B. Singh, S. Agarwal, H. Kim, and R. Sharma, "FakedBits- Detecting Fake Information on Social Platforms using Multi-Modal Features," *KSII Transactions on Internet and Information Systems (TIIS)*, vol. 17, no. 1, pp. 51–73, Jan. 2023.

[30] L. Wu, P. Liu, and Y. Zhang, "See How You Read? Multi-Reading Habits Fusion Reasoning for Multi-Modal Fake News Detection," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37,

no. 11, pp. 13736–13744, Jun. 2023, https://doi.org/10.1609/aaai.v37i11.26609.

[31] X. Wang, X. Li, X. Liu, and H. Cheng, "Using ALBERT and Multi-modal Circulant Fusion for Fake News Detection," in *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Prague, Czech Republic, 2022, pp. 2936–2942, https://doi.org/10.1109/SMC53654.2022.9945303.

[32] N. M. Duc Tuan and P. Quang Nhat Minh, "Multimodal Fusion with BERT and Attention Mechanism for Fake News Detection," in *2021 RIVF International Conference on Computing and Communication Technologies (RIVF)*, Hanoi, Vietnam, Aug. 2021, pp. 1–6, https://doi.org/10.1109/RIVF51545.2021.9642125.

[33] R. Jaiswal, U. P. Singh, and K. P. Singh, "Fake News Detection Using BERT-VGG19 Multimodal Variational Autoencoder," in *2021 IEEE 8th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, Dehradun, India, Nov. 2021, pp. 1–5, https://doi.org/10.1109/UPCON52273.2021.9667614.

[34] H. Yang *et al.*, "Multi-Modal fake news Detection on Social Media with Dual Attention Fusion Networks," in *2021 IEEE Symposium on Computers and Communications (ISCC)*, Athens, Greece, Sep. 2021, pp. 1–6, https://doi.org/10.1109/ISCC53001.2021.9631256.

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

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

[37] Y. Liang, T. Tohti, and A. Hamdulla, "False Information Detection via Multimodal Feature Fusion and Multi-Classifier Hybrid Prediction," *Algorithms*, vol. 15, no. 4, Apr. 2022, Art. no. 119, https://doi.org/10.3390/a15040119.

[38] S. K. Uppada, P. Patel, and S. B., "An image and text-based multimodal model for detecting fake news in OSN's," *Journal of Intelligent Information Systems*, vol. 61, no. 2, pp. 367–393, Oct. 2023, https://doi.org/10.1007/s10844-022-00764-y.

[39] S. K. Hamed, M. J. Ab Aziz, and M. R. Yaakub, "A review of fake news detection approaches: A critical analysis of relevant studies and highlighting key challenges associated with the dataset, feature representation, and data fusion," *Heliyon*, vol. 9, no. 10, Oct. 2023, Art. no. e20382, https://doi.org/10.1016/j.heliyon.2023.e20382.

[40] D. S. Asudani, N. K. Nagwani, and P. Singh, "Impact of word embedding models on text analytics in deep learning environment: a review," *Artificial Intelligence Review*, vol. 56, no. 9, pp. 10345–10425, Sep. 2023, https://doi.org/10.1007/s10462-023-10419-1.

[41] J. Egger, A. Pepe, C. Gsaxner, Y. Jin, J. Li, and R. Kern, "Deep learning—a first meta-survey of selected reviews across scientific disciplines, their commonalities, challenges and research impact," *PeerJ Computer Science*, vol. 7, Nov. 2021, Art. no. e773, https://doi.org/10.7717/peerj-cs.773.

[42] K. Han *et al.*, "A Survey on Visual Transformer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 87–110, Jan. 2023, https://doi.org/10.1109/TPAMI.2022.3152247.

[43] A. Choudhary and A. Arora, "Assessment of bidirectional transformer encoder model and attention based bidirectional LSTM language models for fake news detection," *Journal of Retailing and Consumer Services*, vol. 76, Jan. 2024, 103545, https://doi.org/10.1016/j.jretconser.2023.103545.

[44] S. F. N. Azizah, H. D. Cahyono, S. W. Sihwi, and W. Widiarto, "Performance Analysis of Transformer Based Models (BERT, ALBERT and RoBERTa) in Fake News Detection." arXiv, Aug. 09, 2023, https://doi.org/10.48550/arXiv.2308.04950.

[45] D. Tomás, R. Ortega-Bueno, G. Zhang, P. Rosso, and R. Schifanella, "Transformer-based models for multimodal irony detection," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 6, pp. 7399–7410, Jun. 2023, https://doi.org/10.1007/s12652-022-04447-y.

[46] J. Maurício, I. Domingues, and J. Bernardino, "Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review," *Applied Sciences*, vol. 13, no. 9, Jan. 2023, Art. no. 5521, https://doi.org/10.3390/app13095521.

[47] Z. Jin, J. Cao, Y. Zhang, J. Zhou, and Q. Tian, "Novel Visual and Statistical Image Features for Microblogs News Verification," *IEEE Transactions on Multimedia*, vol. 19, no. 3, pp. 598–608, Mar. 2017, https://doi.org/10.1109/TMM.2016.2617078.

[48] K. Shu, L. Cui, S. Wang, D. Lee, and H. Liu, "dEFEND: Explainable Fake News Detection," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Anchorage, AK, USA, Apr. 2019, pp. 395–405, https://doi.org/10.1145/3292500.3330935.

[49] T. Chen, X. Li, H. Yin, and J. Zhang, "Call Attention to Rumors: Deep Attention Based Recurrent Neural Networks for Early Rumor Detection," in *Trends and Applications in Knowledge Discovery and Data Mining*, Melbourne, Australia, 2018, pp. 40–52, https://doi.org/10.1007/978-3-030-04503-6_4.

[50] M. H. Goldani, S. Momtazi, and R. Safabakhsh, "Detecting fake news with capsule neural networks," *Applied Soft Computing*, vol. 101, Mar. 2021, Art. no. 106991, https://doi.org/10.1016/j.asoc.2020.106991.

[51] F. Qian, C. Gong, K. Sharma, and Y. Liu, "Neural User Response Generator: Fake News Detection with Collective User Intelligence," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, Stockholm, Sweden, Jul. 2018, pp. 3834–3840, https://doi.org/10.24963/ijcai.2018/533.

[52] N. Kausar, A. AliKhan, and M. Sattar, "Towards better representation learning using hybrid deep learning model for fake news detection," *Social Network Analysis and Mining*, vol. 12, no. 1, Nov. 2022, Art. no. 165, https://doi.org/10.1007/s13278-022-00986-6.

[53] S. Abdali, S. Shaham, and B. Krishnamachari, "Multi-modal Misinformation Detection: Approaches, Challenges and Opportunities." arXiv, Mar. 27, 2024, https://doi.org/10.48550/arXiv.2203.13883.

[54] M.-H. Guo *et al.*, "Attention mechanisms in computer vision: A survey," *Computational Visual Media*, vol. 8, no. 3, pp. 331–368, Sep. 2022, https://doi.org/10.1007/s41095-022-0271-y.

[55] B. Ahmed, G. Ali, A. Hussain, A. Baseer, and J. Ahmed, "Analysis of Text Feature Extractors using Deep Learning on Fake News," *Engineering, Technology & Applied Science Research*, vol. 11, no. 2, pp. 7001–7005, Apr. 2021, https://doi.org/10.48084/etasr.4069.

[56] H. M. Al-Dabbas, R. A. Azeez, and A. E. Ali, "Two Proposed Models for Face Recognition: Achieving High Accuracy and Speed with Artificial Intelligence," *Engineering, Technology & Applied Science Research*, vol. 14, no. 2, pp. 13706–13713, Apr. 2024, https://doi.org/10.48084/etasr.7002.

[57] T. Zhang *et al.*, "BDANN: BERT-Based Domain Adaptation Neural Network for Multi-Modal Fake News Detection," in *2020 International Joint Conference on Neural Networks (IJCNN)*, Glasgow, UK, Jul. 2020, pp. 1–8, https://doi.org/10.1109/IJCNN48605.2020.9206973.

[58] C. Song, N. Ning, Y. Zhang, and B. Wu, "A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks," *Information Processing & Management*, vol. 58, no. 1, Jan. 2021, Art. no. 102437, https://doi.org/10.1016/j.ipm.2020.102437.

[59] P. Yang, J. Ma, Y. Liu, and M. Liu, "Multi-modal transformer for fake news detection," *Mathematical biosciences and engineering*, vol. 20, no. 8, pp. 14699–14717, Jul. 2023, https://doi.org/10.3934/mbe.2023657.