FOREX Publication
Open Access | Rapid and quality publishing

International Journal of Computer Science Research (IJCSR)
Survey Article | Volume 2, Issue 4 | Pages 96-105 | ISSN XXXX-XXXX

Multi-modal Fake News Detection using Multimodal Approach of BERT and ResNet110

Chetan Agrawal[1*], Anjana Pandey[2] and Sachin Goyal[3]
[1]*Computer Science & Engineering, University Institute of Technology Rajiv Gandhi Proudyogiki Vishwavidyalay (UIT - RGPV), Bhopal, India , Chetan.agrawal12@gmail.com*
[2]*Information Technology, University Institute of Technology Rajiv Gandhi Proudyogiki Vishwavidyalay (UIT - RGPV), Bhopal, India , anjanapandey@rgtu.net*
[3] *Information Technology, University Institute of Technology Rajiv Gandhi Proudyogiki Vishwavidyalay (UIT - RGPV), Bhopal, India , sachingoyal@rgtu.net*

***Correspondence:** *Chetan.agrawal12@gmail.com*

ABSTRACT- Globally, the usage of social media has significantly increased and has become the most common way for people to deplete news. The easy sharing of multimedia content on social media has caused the fake news dimension, which threatens the stability as well as security of the society. Fake news detection (FND) in social media becomes challenging, because of which various tools are developed to detect them. Multi-modal FND aims to determine fake data by text as well as images. Most commonly, researchers identify fake news only as text, but not as images. This research proposes a multimodal approach for detecting fake news in the formats of both text and image, and for classifying news as real or fake. The proposed multimodal-based convolutional neural network (CNN) combines the designs of both text and image of fake news. This method utilizes two classification methods named bidirectional encoder representations from transformers (BERT) for text, and ResNet110 for images. This method uses the Fakeddit dataset to estimate and evaluate the performance. The experimental results of the proposed ResNet110+BERT model achieves respective accuracy, precision, recall and F1-score values of about 0.931, 0.944, 0.942, and 0.946, which is superior when compared to the existing methods, recurrent CNN (RCNN) and fine-grained multimodal fusion network (FMFN). From the analysis, it is proven that the proposed method ResNet110-BERT achieves an accuracy of 0.931, and hence shows better results for overall metrics when compared to the existing methods of RCNN and FMFN.

Keywords: bidirectional encoder representations from transformers; convolutional neural network; multimodal fake news detection; ResNet110, social media.

ARTICLE INFORMATION
Author(s): Chetan Agrawal, Anjana Pandey and Sachin Goyal;
Received: 03 /10/2023; **Accepted**: 12/12/2023; **Published**: 30/12/2023;
e-ISSN: XXXX-XXXX;
Paper Id: IJCSR-020401;
Citation: 10.37391/IJCSR.020401

Publisher's Note: FOREX Publication stays neutral with regard to Jurisdictional claims in Published maps and institutional affiliations.

1. INTRODUCTION

Recently, people obtain updated news from various platforms including online sources and social media sites [1]. The social media resources Facebook, Twitter and Sina Weibo have become the top sources of news as thousands of users read news through these social media platforms. Social media has aided users to acquire news, indicate perspectives and communicate personal judgements with others [2]. The huge evolution of web technology has made it possible for users to post both real and fake news on social media [3] [4]. The blogs, headlines and social media messages are deliberately put forward as ambiguous for various reasons [5].

Many communities progressively receive information and devour news through social media. It allows the community to share information, where their information is in the form of both text and images [6]. The news arise from various platforms making it a challenging task to identify the credibility of the posts and news [7]. The information provided in the images is stronger than on text [8]. Social networks provide the benefit of transmission of multimedia information easily at low costs with a faster distribution which easily and rapidly benefits and attracts people [9]. A growing number of users have begun fabricating fake news on social media with the intent of luring people and communities [10]. Fake news is rapidly increasing, causing a negative impact and constituting a significant menace to the individuals, society, economy and health [11].

Automatic fake news detection (FND) has become a major concern because it is challenging to detect fake news manually. Hence, automatic FND has turned out to be a recent research subject [12]. The FND methods effectively determine false news and supportive for an administrator to remove fake news from the social media [13]. However, the FND can provided a fine-grained multi-classification problem during implementation because of developing a number of categories for collections of data [14]. The FND supports readers in determining the fake news and bias in a news article, and hence reduces the spreading of fake news [15] [16]. The initial process of FND is to identify if the fake news is a text or image. For this, the multimodal approach is essential to provide correlative benefits for FND [17]. The features of multimodal are anticipated to the most favourable in FND than the existing

International Journal of Computer Science Research (IJCSR)
Survey Article | Volume 2, Issue 4 | Pages 96-105 | ISSN XXXX-XXXX

models [18]. The existing research suggest that the ensemble learning and machine learning (ML) algorithms can train the exact high-level representations obtained from the postings on social media sources for recognition [19]. Few existing researchers have tried to physically develop a series of features that are provided to the ML for identifying fake news, however this method still consumes more time and has poor conception [20]. For network administrators, manually rejection of fake news one by one is expensive and laborious. These methods can also be endured from a flexibility characteristic and solves the overfitting problem. To overcome these problems, a multimodal approach is proposed for FND in the formats of both text and images. The proposed multimodal-based convolutional neural network (CNN) combines the design of text and image for fake news. This method utilizes two classification methods namely, bidirectional encoder representations from transformers (BERT) for text, and ResNet110 for images. The primary contributions of this research are discussed as follows:

• This research proposes Multimodal BERT and ResNet110 architecture for FND to classify the news as real or fake.

• This method utilizes two classification methods namely, BERT for text and ResNet110 for image. The multimodal fake news detection makes use of the Fakeddit dataset for the effective performance of the model.

• The algorithm of error level analysis (ELA) focuses on the malignant spliced and fake image attributes which is better than sending the image directly into the frequency domain by evaluation.

The rest of the paper is arranged as follows: Section 2 discusses the recent research on task assignment problems and Section 3 provides the proposed work of this paper. The assessment results are discussed in *Section 4*. The discussion and limitations of the proposed method are explained in *Section 5*, while *Section 6* presents the overall summary and future work of this research.

2. LITERATURE SURVEY

Liu et al. [21] implemented a semantic gap bridging among text and image by utilizing the caption-based approach to capture semantic information from images. This model optimized the use of image by combining the entity features with global features to enhance the accuracy. It also leveraged image caption technology to produce image data and integrate it into the original text so as to bridge the gaps. This method provided better performance and significant improvement than the other methods. However, the model gave rise to insufficient data due to its expensive annotation.

Segura-Bedmar and Alonso-Bartolome [22] implemented a fake news fine-grained classification on the dataset of Fakeddit, utilizing both the approaches of single and multiple models. The multiple modals used an architecture of CNN which combined the data of image and text, whereas the single modal only utilized the text. The advantage of this modal is that it provided better accuracy while the image information expanded the scope for a good fake news detection performance, but still, aggressive training gave rise to time complexity.

Ying et al. [23] developed a novel method for end-to-end multimodal Cross-attention networks with multiple stages of textual content. This method jointly integrated the duplicate relationships of text and visual data, as well as the variant modalities of social media news in a unified method. ResNet and BERT classification were pretrained for work so as to produce effective likeness for the regions. This method provided better performance than the other methods by combining the multi-level models. However, the method was unable to make the central hidden state capture the ample textual linguistics.

Guo and Song [24] introduced multiple model fake news detections with the attention and pooling methods of neural networks. This method initially utilized two multimodal learning stages, averaging pooling and multimodal fusion using the dot-product attention. This method utilized the hidden knowledge merge by both the temporal and spatial effects. The model provided better results than the other methods by utilizing the attention and pooling blocks, but this model was ineffective in ensuring a similar relation in attention.

Wang et al. [25] implemented a fine-grained multimodal fusion network (FMFN) as a nuanced method to fuse textual and visual features for effective detection of fake news. This model used scaled dot-product attention method to enhance both features and fuse the improved features, thereby capturing the province among features. This presented method showcased significant improvements in the detection of fake news, but it had a number of representations which were complex to distinguish.

Jing et al. [26] implemented a multimodal progressive fusion network (MPFN) for multiple FND. The MPFN minimized fake news with determinate baseline techniques by using image superficial information and deep information consideration. This method was most widely used for multimodal fused features and modalities interactions in fake news detection. This method improved the modal's performance by utilizing the features of the modalities, but the manual labelling performance was not supportive for the fake news identification in dataset results.

Ekbal and Kumari [27] implemented attention-based multimodal factorized bilinear pooling (AMFB). This framework utilized the post-textual and image data as input, to determine whether that post was real or fake. An attention based multi-level CNN-recurrent neural network (ABM-CNN-RNN) was utilized for the extraction of features from an image. This method provided better feature extraction (FE) and fusion performance than the previous methods, but could not extract post-specific features from a difficult post.

Rai et al. [28] developed an approach for fake news classification based on the titles of news, with the use of content-based classification method. This approach deployed the hybrid model of BERT and long short-term memory (LSTM) for classification to classify whether the news was

International Journal of Computer Science Research (IJCSR)
Survey Article | Volume 2, Issue 4 | Pages 96-105 | ISSN XXXX-XXXX

legal or fake. This method classified the news based on semantic attributes of news articles or reports including grammatical, semantical and synthetical perspectives. This method utilized the FakeNews Net dataset consisting of two subsets, PolitiFact and GossipCop for training and validation. This method achieved better classification performance because of LSTM having a greater capability to catch the semantic, as well as long-distance relationships. Nonetheless, this model had enormous due to corpus data as well as structure of training.

Li et al. [29] presented a new method of semantic-enhanced multimodal fusion network for FND. This method contained multiple subnetworks such as multimodal fusion, fake news detector, and adaptation network of the event domain. It employed CNN to fuse the multimodal data, and later embraced a domain adaptation network to learn the transferred attributes among events. This method also examined the distribution of language statistics of various social media to traverse an optimal selection of the pre-trained BERT. This method effectively captured the mutual attributes between events, and hence brought profit for fake news detection.

Mehta et al. [30] introduced a natural language processing framework of BERT for fake news classification. This method fine-tuned a BERT for specific domain datasets, aside from employing human justification and metadata for the extended performance of this method. This method analysed that the deep-contextualizing nature of BERT was efficient for this work, also acquiring an efficient improvement on the binary classification. It effectively classified the 6 label classification models. Nonetheless, this method consumed more time for training.

Palani et al. [31] implemented the CapsNet and BERT (CB) for an automated detection of fake news. The CB-Fake model utilized both visual and textual data from the social media news articles for estimating as real or fake. This method included BERT to extract the features of text which protected the semantic relationships among the words. The CapsNet caught the significant visual attributes from the image. Then, those attributes were integrated to acquire the abundant information that supported to examine if the news was real or fake, yet this model had high computational cost.

Kaliyar et al. [32] presented a BERT-based deep learning method for FND. This method integrated various parallel blocks of single-layer deep CNN which had various kernel sizes and filters with BERT for efficient learning. This method was developed on bidirectional transformer encoder top based pre-trained word embedding model (BERT). It achieved better performance and did not need handwritten features, but demanded more computation due to its size.

Aslam et al. [33] presented an ensemble-based deep learning approach for classification of fake news. This method utilized two DL methods namely, bi-directional LSTM (Bi-LSTM), and gated recurrent unit (GRU) for textual features, while for the balance features, dense DL methods were utilized. This method applied the natural language processing (NLP) techniques on the statement features and utilized the LIAR dataset to classify whether or not the news was fake. This method achieved the better performance compared to the other methods. Nevertheless, the Bi-LSTM approach consumed more time to train the model than the traditional LSTM.

Choudhary and Arora [34] developed a linguistic feature-based driven DL approach for effective detection and classification of fake news. The linguistic approach was developed to identify the content effects which introduced the language-driven features. This approach extracted the synthetic, sentimental, readability and grammatical aspects of the particular news. It achieved commendable outcomes by deploying neural-based sequential learning. Yet, the language driven model needed a method to control the time-consumption, as well as handcrafted feature problems cannot control the imprecation of dimensionality problem.

Xue et al. [35] implemented a multimodal consistency neural network (MCNN) for identifying the text and image of the fake news. This method employed ranch network for extracting the visual semantic features in visual perception of fake news. In visual tampering FE, the ELA and CNN were developed to examine the intellect and originality of the pictures given manually in fake news. The visual tampering FE module effectively detected the malicious tampered images of fake news. But this method needed huge labelled data for model training.

Guo [36] introduced multimodal fake news detection using a Mutual Attention Neural Network (MANN) that learnt the relationship between different modalities on the Weibo dataset. The model comprised four components: a multimodal feature extractor, mutual attention fusion, fake news detector, and irrelevant event discriminator. But, the features on the VGG16 and VGG19 layers were tensors with dimensions (channels, width, height). Additionally, some challenges arose due to the varying sizes across different layers.

Yang et al. [37] developed the Multi-modal transformer for fake news detection using Thermogravimetric Analysis (TGA) on Weibo dataset and Twitter dataset. TGA, a transformer-based multimodal approach, consisted of four key components: a text feature extractor, an image feature extractor, late fusion, and a classifier. Nevertheless, in the TGA, the mass loss of volatiles was not equivalent to the formation of degradants, significantly impeding its ability to provide consistent universal indicators of the actual extent of degradation.

Sastrawan et al. [38] introduced a CNN–RNN-based fake news detection model which underwent testing and training on datasets of ISOT, fake news dataset, fake or real news dataset, and fake news detection dataset. To address data imbalances between classes, the datasets underwent augmentation process using the back-translation method. However, a challenge encountered was the difficulty with long sequences. In dealing with very long sequences, RNNs faced memory limitations, potentially impacting their ability to effectively process extensive information.

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Computer Science Research (IJCSR)
Survey Article | Volume 2, Issue 4 | Pages 96-105 | ISSN XXXX-XXXX

Al Obaid et al. [39] introduced an approach to improve the existing methods' performances through the utilization of ensemble DL-based on attention mechanisms. In this approach, the achievement of ensemble approach depended on various learners. The loss function enforced every learner to perform various news content parts and obtained better classification accuracy. Furthermore, the learners were developed on basic deep feature extractor but differed from attention approaches. The parameter numbers were efficiently minimized and overfitting was solved.

Uppada et al. [40] presented a framework that flagged the fake news by embedding visual data with text. The suggested framework performed on the data obtained from a benchmark dataset. The suggested approach had various architectures to learn visual, as well as linguistic approaches from the individual news. The various datasets were analysed, while the features extracted from those perceptions were text-based supported data.

2.1 Review

The above section showcases some limitations such as manual labelling performance that does not stand well for fake news identification as it becomes time consuming to train the model, aside from increased computational complexity. To overcome these, a multi-modal fake news detection technique is suggested in this manuscript. The proposed method classifies news into real or fake by using BERT and RsNet110 for the both formats of text and image with superior accuracy.

3. PROPOSED METHOD

In this work, the proposed multi-modal FND is fitted by giving text and image news. The model's goal is to determine if the news is real or fake. This proposed method is categorized into three main parts: textual features, visual extractor, as well as feature fusion of textual and visual data. *Figure 1* depicts the block diagram of the proposed multi-model FND.

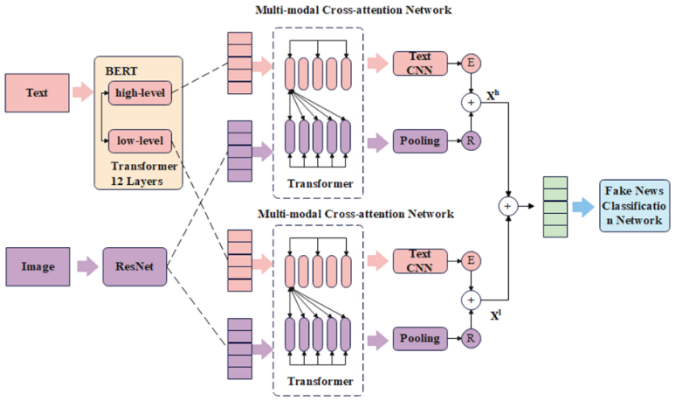

Figure 1. Proposed method for Multimodel fake news detection

3.1. Dataset

Primarily, this research obtains the standard multimodal fake news detection dataset of Fakeddit. This dataset involves comments, text piece, related images, context features as well as ground truth labels. Then, the collected dataset is provided for the pre-processing step.

3.2. Pre-processing

In this step, this research pre-processes the collected data of both text and image. Primarily, this process eliminates the instances which involve either text or image for achieving the better result in multimodal data. In textual data, the methods of tokenization, stemming and lemmatization are utilized to remove the stop words, punctuation marks and so on. Tokenization is the process of splitting out text strings into a list of tokens that are later used for lexical analysis, where the data fields are converted to tokens. For image data, the normalization technique is used for converting raw data into a suitable format as the dataset from various resources contain incomplete data. Next, normalization is employed to perform linear data transformation. It is also called min-max normalization where all the attribute values are in the range of 0 and 1. Further, the pre-processed data are provided as input to the feature extraction process.

3.3. Textual Feature Extraction

The FND based text approaches employ original vector system. The approach is effectively employed for the identification of minimal and dissimilar sentences. Yet, the processing is complex for identifying the actual contents. The polysemy model is used for solving the above discussed problem and to identify the relations of words before and after. To address this problem, this research applies pre-trained BERT model [41][42] for feature extraction, as represented by *equation (1).*

$$h_i^t = BERT(t_i) \quad (1)$$

Where, h_i^t is the textual feature vector and t_i – ith is the input sentence.

The word2vec is a Machine Learning (ML) approach which is used for performing text embedding. Every word in text is depicted as a vector, which permitting to estimate the similarity degree between the words and distance of two vectors. The aim of this approach is to combine maximum words but not merge the words. This technique involves two models: continuous bag-of-words (CBOW), for forecasting the target word wt, which is illustrated in *equation (2)* and another one is skip-gram which uses wt for anticipation of framework, which is illustrated in *equation (3)* as:

$$L = \sum_{w_t \epsilon c} \log p\left(w_t \middle| Context(w_t)\right) \quad (2)$$

$$L = \sum_{w_t \epsilon c} \log p(Context(w_t)|w_t) \quad (3)$$

Where, c represents the entire words in a training set. By utilizing the BERT model, the textual features are significantly extracted. After that, the image feature is extracted by ResNet110 architecture and the detailed information is provided below.

3.4. Image Feature Extraction

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Computer Science Research (IJCSR)
Survey Article | Volume 2, Issue 4 | Pages 96-105 | ISSN XXXX-XXXX

The FE technique is significantly implemented to extract crucial data features for the reduction of dimensionality. The aim of this approach is to reduce the large number of vital sources to explain the numerous data. The greater amount of data supports the quality supplied for FE. The high-level image outcomes like shape, colour and structure are used for FE. The CNN architecture contains pooling and convolutional layer which have become most popular in computer vision, and are used for FE. For CNN, the feature map numbers are acquired by utilizing the convolutional task of the kernel as well as identifying the image perception features [43, 44]. The number of image features wholly fuse the textual features which are illustrated through the feature vectors. This approach utilizes the outcome of CNN as low-level image feature set in this phase, which is then fused through altering the detection approach to distinguish the image's physical level determination.

This approach utilizes the pre-trained ResNet110 architecture [45] as input for image encoding which is illustrated in equation (4) as:

$$h^v = ResNet110(v) \quad (4)$$

Where, v is the actual input image and h^v is the extracted visual semantic feature using ResNet110.

Further, the semantic features are forwarded to the attention module to target the image regions. Therefore, the image feature distributes the weight for representing the model's importance when acquiring a visual model illustration, which is as expressed in *equations (5) to (7).*

$$u^v = U^T \tanh(W^v h^v + b^v) \quad (5)$$

$$\alpha^v = \frac{\exp(u^v)}{\sum_i \exp(u^v)} \quad (6)$$

$$s^v = \sum_i \alpha_i h^v \quad (7)$$

Where, W^v is the weight matrix, b^v is the bias term, U^T is the transposed weight vector, and u is the scoring function which identifies the individual significance of the vector.

The word2vec [46, 47] is utilized for obtaining a better semantic data in image-to-sequence model and is often used to develop visual characteristics for an image's features organization by text. This approach is typical for the utilization of embedding layer in text determination which forwards the language feature image to language level sequence, as illustrated in *equation (8).*

$$f^v = word2vec(s^v) \quad (8)$$

Where, s^v is the extracted language linguistic representation using ResNet110, f^v is the visual language sequence by word2vec. The image features are significantly extracted using the ResNet110. Then, visual tampering feature modules are extracted using the ELA algorithm and this is briefly explained in the following section.

3.5. Visual tampering feature extraction module

In comparison to the image of a real news, a fake news image is repeatedly maliciously spliced or has undergone a recompression many times because of its proliferation. The algorithm of ELA [48] focuses on a malignant spliced as well as fake image attributes which is better than sending the image directly into the frequency domain by evaluation. The ELA is used to compress the image processing that is gradually changing. In the visual tampering FE module, the ELA transformation of an image is utilized. Later, ResNet110 extracts the image features as numerically shown in *equations (9) and (10)*.

$$v^{ela} = ELA(v) \quad (9)$$

$$h^{ela} = ResNet110(v^{ela}) \quad (10)$$

Where, v denotes the original input image, v^{ela} denotes an original image refined with ELA, h^{ela} denotes the tampered feature by ResNet110. The visual tampering feature modules are efficiently extracted by the ELA algorithm. The extracted features of the text, image and visually tampered feature outcomes are then provided to the similarity measurement module.

3.6. Similarity Measurement Module

In FND, false news is distinguished by measuring the cooperation between the data of text and images. The similarity module measurement is used to evaluate fake news of text and image likeness. The sub-networks of text and visual semantic features are certified by learning the basic representation patterns of image and text space. This method applies the fully connected layer to the sub-networks' final layer and treats sub-networks for transforming the last layer loads. This method efficiently generates the same depiction to a similar group of image and text samples. Later, cosine similarity is applied to measure the similarity among the text and image [49], which is represented by *equations (11) and (12).*

$$s = \frac{s^t . s^v}{\|s^t\| \times \|s^v\|} \quad (11)$$

Where, s^t and s^v respectively denote a linguistic sequence of images and text.

$$p^s = sigmoid(s) \quad (12)$$

Where, sigmoid is the activation method deployed to map the range between 0 and 1. The similarity measurement module measures the similarity of the image and text data. Then, the outcome of this approach is provided to the multimodal approach for identifying the fake data in both text and image, further briefly explained in the following section.

3.7. Multimodal Approach

After obtaining the textual and visual feature depiction, this research implements the feature fusion mechanism in order to attain a shared depiction. The fusion mechanism enhances the relationship among textual and visual feature depictions. It is assumed that $(R_T) \in R_m$ for text feature vector, and $(R_V) \in R_n$

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Computer Science Research (IJCSR)
Survey Article | Volume 2, Issue 4 | Pages 96-105 | ISSN XXXX-XXXX

for visual feature vector. The general multimodal bilinear model is mathematically expressed in *equation (13).*

$$R_{TV} = R_T^T W_i R_V \quad (13)$$

Where, $W_i \in R^{mn}$ is the projection matrix, and R_{TV} is the bilinear model outcome. To minimize the parameter numbers, W_i is factorized as a low rank matrix which is mathematically expressed in *equation (14).*

$$R_{TV} = 1^T (U_i^T R_T \circ V_i^T R_V) \quad (14)$$

Where, U_i^T and V_i^T are the element-wise multiplication of two vectors, and they are reformulated as two-dimensional matrices, $U' \in R^{m*ko}$ and $V' \in R^{n*ko}$, which is mathematically expressed in *equation (15).*

$$R_{TV} = \frac{R_{TV}^T}{\|R_{TV}\|} \quad (15)$$

In that, the feature fusion mechanism which enhances the relationship between the text and image features, furthermore supplies a suitable alignment.

The CNN-based multimodal approach utilizes the text and image's similar news for FND. Once the pre-processed data are forwarded to CNN, a number of tasks are imposed to the multimodal data. The CNN is performed in the unimodal structure which ignores the last two dense layers of the rectified linear activation function (ReLU). CNN primarily labels the information provided to the convolutional layer and then forwards it to the hidden layer. It is eventually forwarded to the further input standard. The filter size (5 × 5) is used with stalk from 1 to 0 padding. A convolutional layer's outcome is forwarded to ReLU activation function and then, the max-pooling layer is acquiring the filter size (2 × 2) and resulted as (278 × 278). Then the outcome of maxpooling is forwarded to the convolutional layer, which is an input for outcome channels of 6 and 3. These outcomes are then forwarded to the succeeding layer when the stride length, padding and filter sizes are equal. The ReLU and max-pooling are required for the subsequent feature maps from the convolutional layer. Therefore, the given input data acquires multiple feature maps (137 × 137) and eventually, the feature maps are flattened into a vector length of 56307. The textual data is set through CNN rather than dense outcome to the softmax layer, hence the output vector illustrates the text which is associated by the vector obtained from CNN. Then, the vector is forwarded to dense layers among ReLU. Eventually, a logsoftamx layer is required and the possibilities logarithm is used for the provided input class examination. Therefore, fake news is significantly classified by using CNN. Finally, the output size of CNN is identical to the number of hidden units that is 256 with a batch size of 32. The model is trained by using 10 epochs.

3.7.1. Loss function (L)

The cross-entropy is an efficient probability distribution, which compares the performance of prediction values. If the cross-entropy gets minimum and the prediction gets maximum accuracy, it becomes 0 and the prediction is efficient. In binary classification, where the number of classes (M) equals 2, cross-entropy is estimated as given in *equation (16).*

$$L = -(y log(p) + (1 - y) \log(1 - p)) \quad (16)$$

Where, y is the binary indicator (0 or 1), and p is the predicted probability observation o of class c.

3.8. Dropout

Dropout is a regularization approach for minimizing overfitting in neural networks that avoids the intricate co-adaptation of training data. This procedure results in the neural network's decreased units. The premise of this approach is to randomly drop units from the network during training. This approach is effective in establishing model averaging with the neural networks. Each hidden neuron's fully-connected layer output is established to zero with a 0.5 probability. In this manner, the neurons are "dropped out", thereby not contributing to the backpropagation (BP). The fully-connected layers of neurons are employed during training however, their outcomes are multiplied by 0.5 outcomes.

Further, the evaluation and implementation of the suggested model is provided in the below section.

Algorithm: Convolution Neural Network
Input: Normalized data
Output: Fake news detection
Begin
Step 1: Set input data weight: W_i, W_f, W_c, W_o
Step 2: Set Recurrent data Weight: R_i, R_f, R_c, R_o
Step 3: Set peephole weight: $V \epsilon R^N$
Step 4: Set Offset: $b_i, b_f, b_c, b_o \epsilon R^N$
Step 5: At time t, x_t is the input and y_t is the output of the node
Step 6: perform convolution operation at time t.
Step 7: $i_t = \sigma(W_i x_t + R_i h_{t-1} + b_i)$ is the output of the input data at time t.
Step 8: $\acute{C}_t$, C_t is the input and cell structure of the node at time t, respectively, which are expressed as:
$\acute{C} = tanh(W_c x_t + R_c h_{t-1} + b_c)$
$C = i_t \odot \acute{C}_t + f_t \odot c_{t-1}$
Step 9: $O_t = (W_o x_t + R_o h_{t-1} + b_o)$ is the output of the CNN network
Step 10: The final output h_t of the node is expressed: $h = O_t \odot \tanh(C_t)$
End

4. EXPERIMENTAL RESULT

In this section, the performance and effectiveness of the proposed multi-modal FND is evaluated with a fake news dataset named Fakeddit for both text and image of fake news.

4.1. Dataset Description

In this suggested method, the fake news dataset named Fakeddit [50] is used for training and testing the model, consisting of 1 million examples. The samples are based on 2-way, 3-way and 6-way classification categories through distant supervision.

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Computer Science Research (IJCSR)
Survey Article | Volume 2, Issue 4 | Pages 96-105 | ISSN XXXX-XXXX

This dataset includes a number of posts that are collected from the users of Reddit, containing a number of images, text, metadata and comments. The dataset is used to perform a fine-grained fake news classification, as well as to determine and identify if the news is real or fake. In this Fakeddit, every example has a label which distinguishes fake news in five categories. The dataset is classified into the partitions of training, testing and validation. This dataset contains two variety versions called unimodal, which contains only text, while multimodal contains both text and images. This dataset contains 682,661 images of news and 2,90,000 text news. It encompasses a ratio of 2:3 for real and fake news. This denotes that the proposed approach is capable to simplify efficiently the various levels of complexity and various types of fake news. Furthermore, the suggested method for detection of fake news through combining various features is significant and outstrips the existing others. The dataset has imbalance problem which also affects the classification task as it is difficult for those classes with fewer instances. Hence, the dataset is split into a ratio of 70:30 for both text and image. The 5,63,523 samples are utilized for training, 59,283 are used for validation, and 59,257 are utilized for testing.

4.2. Evaluation Metrics

Several performance metrics are utilized to validate the proposed method of multi-modal FND. The employed hyperparameters have an initial learning rate of 0.001 and 0.01, the number of layers in the CNN are 256, feature vector size of 32, Adam optimizer with learning rate 0.01 is used to optimize the model, momentum of 0.0 and 0.2, and decay rate of 0.001. The performance metrics accuracy, precision, recall, F1-score, Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) are estimated from the detection of all fake news. The mathematical expressions of these metrics are given in *equation (17) to (22)* below:

Accuracy: The ratio of all correct classifications to total number of classifications.

$$Accuracy = \frac{TP}{TP+FP} \times 100 \quad (17)$$

Precision: The ratio of true positive over the classifications of all positive.

$$Precision = \frac{TP}{TP+FP} \times 100 \quad (18)$$

Recall: The proportion of original positives that are correctly classified.

$$Recall = \frac{TP}{TP+FN} \times 100 \quad (19)$$

F1-score: This combines precision and recall to give the average value of weight.

$$\text{F1-score} = \frac{2x\ Precision\ x\ Recall}{Precision + recall} \times 100 \quad (20)$$

MAE: It is determined by the absolute difference between the predicted value and actual values.

$$MAE = \frac{1}{K}\sum_{i=1}^{K} |\widehat{y_i} - y_i| \quad (21)$$

RMSE: It is evaluated by the average size of the error and is concerned with the variations from the actual value.

$$RMSE = \sqrt{\frac{1}{k}\sum_{i=1}^{k} \widehat{(y_i} - y_i)^2} \quad (22)$$

Where, TP is True Positive, FP is False Positive, TN is True Negative, and FN is False Negative.

4.2. Quantitative and Qualitative Analysis

This section provides the performance analysis of an image CNN model in terms of achievable sum rate. The experimental model is performed using the CNN model for VGG16, VGG19, ResNet 50, and ResNet110.

Table 1. Performance analysis of proposed method with various methods network methods for image data

Methods	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	MAE	RMSE
VGG16	92.91	91.65	91.68	91.75	0.84	0.91
VGG19	94.87	93.25	93.09	94.54	0.76	0.87
ResNet50	96.90	96.48	96.71	96.59	0.54	0.68
ResNet110	98.85	97.89	97.01	98.32	0.44	0.57

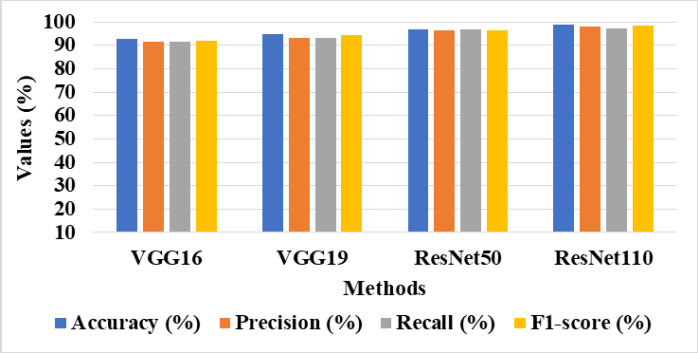

Figure 2. Graphical representation of image network methods

Table 1 and *figure 2* represent the performance analysis of various networks for FND of the image data. The existing methods such as VGG16, VGG19 and ResNet 50 are estimated and compared with the proposed ResNet110 architecture. The ResNet110 for the image data enables much faster training at each layer and also achieves better accuracy results. The acquired outcomes prove that the ResNet110 attains commendable results based on the performance metrics, accuracy, precision, recall and F1-score with corresponding values of about 98.85%, 97.89%, 97.01%, and 98.32%, alongside MAE of 0.54 and RMSE of 0.57.

Table 2. Performance analysis of proposed method with various methods network methods for Text data

Methods	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	MAE	RMSE
MVAE	86.34	85.24	84.89	85.56	0.88	0.89

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Computer Science Research (IJCSR)
Survey Article | Volume 2, Issue 4 | Pages 96-105 | ISSN XXXX-XXXX

TextGCN	88.79	87.25	87.19	87.89	0.76	0.88
RNN	91.40	89.47	89.20	91.02	0.67	0.62
BERT	92.95	91.39	90.69	92.90	0.53	0.59

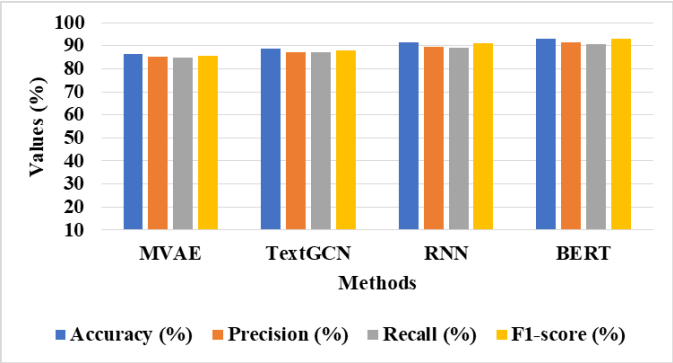

Figure 3. Graphical representation of Text network methods

Table 2 and *figure 3* exhibit the performance analysis of various networks for the FND of the text data. The existing methods which are MVAE, TextGCN, and RNN, are estimated and compared with the proposed BERT. The BERT for the text data enables much faster training at each layer and also achieves better accuracy results. The acquired outcomes show that the BERT attains superior results on the basis of the performance metrics, accuracy, precision, recall and F1-score, with corresponding values of about 92.95%, 91.39%, 90.69% and 92.90%, alongside MAE of 0.53 and RMSE of 0.59.

Table 3. Performance analysis of proposed method with various methods network methods for Multimodal data

Methods	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	MAE	RMSE
NB+RF	71.98	70.28	70.13	71.09	0.87	0.91
NB+SVM	74.06	73.25	82.19	73.89	0.72	0.83
BiLSTM+CNN	76.26	74.90	74.98	75.65	0.64	0.77
BERT+ResNet110	93.15	94.49	94.26	94.66	0.51	0.61

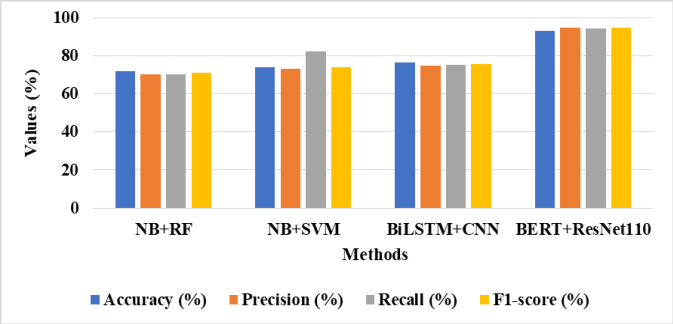

Figure 4. Graphical representation of Multimodal Network methods

Table 3 and *figure 4* present the performance evaluation of various networks for multimodal FND. The existing methods: NB+RF, NB+SVM and BiLSTM+CNN are analyzed and contrasted with the introduced BERT+ResNet110. The BERT+ResNet110 for the image and text data enables much faster training at each layer with commendable accuracy results. The acquired outcomes prove that the BERT+ResNet110 attains preferable results on the performance metrics accuracy, precision, recall and F1-score with values of about 93.15%, 94.49%, 94.26% and 94.66%, alongside MAE of 0.51 and RMSE of 0.61.

Table 4 shows the outcomes of computation time as consumed by the proposed method, in contrast with the existing methods. It is hence derived that the BERT+ResNet110 method achieves a minimum computational time of 300ms which is preferable in contrast to the existing methods.

Table 4. Performance analysis of computational time

Methods	Computational time (ms)
NB+RF	600
NB+SVM	500
BiLSTM+CNN	400
BERT+ResNet110	300

4.4. Comparative Analysis

This section demonstrates the comparative analysis of the developed method against the existing methods. The comparative analysis is carried out with the utilization of the dataset, with the consideration of the accomplished results on the metrics of accuracy, precision, recall, and F1-score. *Table 5* demonstrates the comparative analysis of the proposed method alongside the previous models.

Table 5. Comparative Analysis of proposed method with existing methods for fake news

Author	Method	Data set	Accuracy	Precision	Recall	F1-score
Liu et al. [21]	RCNN	Weibo	0.888	0.862	0.920	0.893
		Fakeddit	0.925	0.938	0.937	0.937
Ying et al. [23]	MMCN	Weibo	0.879	0.886	0.861	0.879
		Pheme	0.872	0.837	0.780	0.807
Wang et al. [25]	FMFN	Weibo	0.885	0.878	0.851	0.864
Proposed BERT+ResNet110	BERT+ResNet110	Fakeddit	0.931	0.944	0.942	0.946

5. DISCUSSION

The obtained classification outcomes of the developed model are presented in Tables 1 to 3, while the graphical representation of these results is provided in Figures 2 to 5. The existing method, RCNN [21] utilizes two datasets named, Weibo and Fakeddit. On the Weibo dataset, it achieves an accuracy of 0.888, precision of 0.862, recall of 0.920 and F1-score of 0.893. Likewise, on the Fakeddit dataset, it achieves an accuracy of 0.925, precision of 0.9444, recall of 0.942 and F1-score of 0.946. The MMCN [23] method makes use of the Weibo and Pheme fake news dataset, where the obtained results on the Weibo dataset is 0.879 of accuracy, 0.886 of precision, 0.861 of

International Journal of Computer Science Research (IJCSR)
Survey Article | Volume 2, Issue 4 | Pages 96-105 | ISSN XXXX-XXXX

recall and 0.879 of F1-score. Likewise, on the Pheme dataset the obtained results are: 0.872 of accuracy, 0.837 of precision, 0.780 of recall and 0.807 of F1-score. Furthermore, the FMFN [25] method utilizes the Weibo dataset. The obtained results on the Weibo dataset are: 0.885 of accuracy, 0.878 of precision, 0.851 of recall and 0.864 of F1-score. The number of posts on various events is imbalanced on the Fakeddit dataset. Due to this issue, the learned text features majorly focus on particular events only. It was complex for earlier textual modality approaches to extract the features of transferability among various events. Therefore, the textual performance is seen to be minimum for every approach. As a result, the BERT is proposed as a powerful tool for feature extraction in text. The pre-trained BERT architecture follows the transformer model in which multi-head attention is deployed for preserving the semantic relations between words. The proposed method is proven to be efficient in overcoming the event imbalance impact through utilizing the event domain adaptation network that estimates the variation among various events and ignores the individual characteristics of every event, so as to build the model's performance. The developed multimodal BERT+ResNet110 deploys the Fakeddit fake news dataset and achieves better results on various performance metrics given as: 0.931 of accuracy, 0.944 of precision, 0.942 of recall and 0.946 of F1-score. The proposed BERT+ResNet110 achieves better classification results when compared to the existing methods. However, developing an effective hand-crafted feature needs great knowledge of the related areas and of particular events. In the meantime, this method depends on hand-crafted features, where the acquired feature vector's robustness is not sufficient since it has no knowledge of fake news detection.

5.1. Limitations

The proposed BERT+ResNet110 method is introduced for FND in social media. This method is improved by integrating the name entity identification techniques which find the significant features present in the fake news via differentiating them from the original data. The proposed BERT+ResNet110 method extracts only the significant knowledge from both image and text fake news. The limitation identified for this method is that it does not perform on various real-time applications, but only considering the FND. This limitation of the proposed method can be overcome in future work.

6. CONCLUSION

The FND in social media becomes challenging, because of which various tools are developed for the detection of fake news. In this work, the multi-modal FND is proposed for determining the correctness of fake news. This method utilizes the multimodal approach on both text and image news for identifying if the news is real or fake. This method employs two classification methods, BERT classifier for text classification and ResNet110 for image classification. This developed model makes use of the Fakeddit dataset for an effective multimodal detection of fake news. As compared to the previous methods, the proposed multimodal-based CNN combines text and image plans to detect the fake news. The proposed model achieves a classification accuracy of 0.931, precision of 0.944, recall of 0.942, F1 score of 0.946 respectively. In the future, the developed method will be encompassed to optimize a method for feature fusion approach in numerous applications.

Author Contributions: Conceptualization, X.X. and Y.Y.; methodology, X.X.; software, X.X.; validation, X.X., Y.Y. and Z.Z.; formal analysis, X.X.; investigation, X.X.; resources, X.X.; data curation, X.X.; writing—original draft preparation, X.X.; writing—review and editing, X.X.; visualization, X.X.; supervision, X.X.; project administration, X.X.; funding acquisition, Y.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Department of XXXX, Place, Country under Grant ABC123456.

Conflicts of Interest: The authors declare no conflict of interest.

REFERENCES

[1] Katariya, P.; Gupta, V.; Arora, R.; Kumar, A.; Dhingra, S.; Xin, Q.; Hemanth, J. A deep neural network-based approach for fake news detection in regional language. International Journal of Web Information Systems 2022, 18(5/6), 286-309.

[2] Fu, L.; Peng, H.; Liu, S. KG-MFEND: an efficient knowledge graph-based model for multi-domain fake news detection. The Journal of Supercomputing 2023, 79, 18417-18444.

[3] Dixit, D.K.; Bhagat, A.; Dangi, D. An accurate fake news detection approach based on a Levy flight honey badger optimized convolutional neural network model. Concurrency Comput. Pract. Exper. 2023; 35(1), e7382.

[4] Ravish; Katarya, R.; Dahiya, D.; Checker, S. Fake News Detection System Using Featured-Based Optimized MSVM Classification. IEEE Access 2022, 10, 113184-113199.

[5] Lai, C.M.; Chen, M.H.; Kristiani, E.; Verma, V.K.; Yang, C.T. Fake news classification based on content level features. Applied Sciences 2022, 12(3), 1116.

[6] Liao, Q.; Chai, H.; Han, H.; Zhang, X.; Wang, X.; Xia, W.; Ding, Y. An integrated multi-task model for fake news detection. IEEE Trans. Knowl. Data Eng. 2022, 34(11), 5154-5165.

[7] Meel, P., Vishwakarma, D.K. HAN, image captioning, and forensics ensemble multimodal fake news detection. Inf. Sci. 2021, 567, 23-41.

[8] Singhal, S., Pandey, T., Mrig, S., Shah, R.R., Kumaraguru, P. Leveraging Intra and Inter Modality Relationship for Multimodal Fake News Detection. In Companion Proceedings of the Web Conference 2022, ACM, 2022; pp. 726-734.

[9] Papadopoulos, S.I.; Koutlis, C.; Papadopoulos, S.; Petrantonakis, P.C. VERITE: a Robust benchmark for multimodal misinformation detection accounting for unimodal bias. International Journal of Multimedia Information Retrieval 2024, 13(1), 4.

[10] Raza, S.; Ding, C. Fake news detection based on news content and social contexts: a transformer-based approach. Int. J. Data. Sci. Anal. 2022, 13(4), 335-362.

[11] Cui, X.; Yang, L. Fake News Detection in Social Media based on Multi-Modal Multi-Task Learning. International Journal of Advanced Computer Science and Applications 2022, 13(7), 912-918.

[12] Dong, D.; Lin, F.; Li, G.; Liu, B. Similarity-Aware Attention Network for Multimodal Fake News Detection. 2020, 76-85.

[13] Ying, Q.; Hu, X.; Zhou, Y.; Qian, Z.; Zeng, D.; Ge, S. Bootstrapping Multi-view Representations for Fake News Detection. In Proceedings of the AAAI conference on Artificial Intelligence, arXiv:2003.04981, 2023; pp. 5384-5392.

[14] Hakak, S.; Alazab, M.; Khan, S.; Gadekallu, T.R.; Maddikunta, P.K.R.; Khan, W.Z. An ensemble machine learning approach through effective feature extraction to classify fake news. Future Gener. Comput. Syst. 2021, 117, 47-58.

[15] Meel, P., Vishwakarma, D.K. A temporal ensembling based semi-supervised ConvNet for the detection of fake news articles. Expert Syst. Appl. 2021, 177, 115002.

[16] Mughaid, A.; Al-Zu'bi, S.; Arjan, A.A.L.; AL-Amrat, R.; Alajmi, R.; Zitar, R.A.; Abualigah, L. An intelligent cybersecurity system for detecting fake news in social media websites. Soft Comput. 2022, 26(12), 5577-5591.

[17] Zhou, Y.; Yang, Y.; Ying, Q.; Qian, Z.; Zhang, X. Multi-modal Fake News Detection on Social Media via Multi-grained Information Fusion. In ICMR '23: Proceedings of the 2023 ACM International Conference on Multimedia Retrieval, 2023; pp. 343-352.

[18] Raj, C.; Meel, P. ConvNet frameworks for multi-modal fake news detection. Appl. Intell. 2021, 51(11), 8132-8148.

[19] Kaliyar, R.K.; Goswami, A.; Narang, P. EchoFakeD: improving fake news detection in social media with an efficient deep neural network. Neural Comput. Appl. 2021, 33(14), 8597-8613.

[20] Song, C.; Ning, N.; Zhang, Y.; Wu, B. A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks. Inf. Process. Manage. 2021, 58(1), 102437.

[21] Liu, P.; Qian, W.; Xu, D.; Ren, B.; Cao, J. Multi-Modal Fake News Detection via Bridging the Gap between Modals. Entropy 2023, 25(4), 614.

[22] Segura-Bedmar, I.; Alonso-Bartolome, S. Multimodal fake news detection. Information 2022, 13(6), 284.

[23] Ying, L.; Yu, H.; Wang, J.; Ji, Y.; Qian, S. Multi-level multi-modal cross-attention network for fake news detection. IEEE Access 2021, 9, 132363-132373.

[24] Guo, Y.; Song, W. A Temporal-and-Spatial Flow Based Multimodal Fake News Detection by Pooling and Attention Blocks. IEEE Access 2022, 10, 131498-131508.

[25] Wang, J.; Mao, H.; Li, H. FMFN: Fine-grained multimodal fusion networks for fake news detection. Applied Sciences 2022, 12(3), 1093.

[26] Jing, J.; Wu, H.; Sun, J.; Fang, X.; Zhang, H. Multimodal fake news detection via progressive fusion networks. Inf. Process. Manage. 2023, 60(1), 103120.

[27] Kumari, R.; Ekbal, A. AMFB: Attention based multimodal factorized bilinear pooling for multimodal fake news detection. Expert Syst. Appl. 2021, 184, 115412.

[28] Rai, N.; Kumar, D.; Kaushik, N.; Raj, C.; Ali, A. Fake News Classification using transformer based enhanced LSTM and BERT. Int. J. Cognit. Comput. Eng. 2022, 3, 98-105.

[29] Li, S.; Yao, T.; Li, S.; Yan, L. Semantic-enhanced multimodal fusion network for fake news detection. Int. J. Intell. Syst. 2022, 37(12), 12235-12251.

[30] Mehta, D.; Dwivedi, A.; Patra, A.; Kumar, A.M. A transformer-based architecture for fake news classification. Social Network Anal. Min. 2021, 11(1), 39.

[31] Palani, B.; Elango, S.; Viswanathan, K.V. CB-Fake: A multimodal deep learning framework for automatic fake news detection using capsule neural network and BERT. Multimedia Tools Appl. 2022, 81(4), 5587-5620.

[32] Kaliyar, R.K., Goswami, A.; Narang, P. FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. Multimedia Tools Appl. 2021, 80(8), 11765-11788.

[33] Aslam, N.; Khan, I.U.; Alotaibi, F.S.; Aldaej, L.A.; Aldubaikil, A.K. Fake detect: A deep learning ensemble model for fake news detection. Complexity 2021, 2021, 5557784.

[34] Choudhary, A.; Arora, A. Linguistic feature based learning model for fake news detection and classification. Expert Syst. Appl. 2021, 169, 114171.

[35] Xue, J.; Wang, Y.; Tian, Y.; Li, Y.; Shi, L.; Wei, L. Detecting fake news by exploring the consistency of multimodal data. Inf. Process. Manage. 2021, 58(5), 102610.

[36] Guo Y. A mutual attention based multimodal fusion for fake news detection on social network. Appl. Intell. 2023, 53(12), 15311-15320.

[37] Yang, P.; Ma, J.; Liu, Y.; Liu, M. Multi-modal transformer for fake news detection. Math. Biosci. Eng. 2023, 20(8), 14699-14717.

[38] Sastrawan, I.K.; Bayupati, I.P.A.; Arsa, D.M.S. Detection of fake news using deep learning CNN–RNN based methods. ICT Express 2022, 8(3), 396-408.

[39] Al Obaid, A.; Khotanlou, H.; Mansoorizadeh, M.; Zabihzadeh, D. Multimodal Fake-News Recognition Using Ensemble of Deep Learners. Entropy 2022, 24(9), 1242.

[40] Uppada, S.K.; Patel, P.; Sivaselvan, B. An image and text-based multimodal model for detecting fake news in OSN's. J. Intell. Inf. Syst. 2023, 61, 367-393.

[41] Keya, A.J.; Wadud, M.A.H.; Mridha, M.F.; Alatiyyah, M.; Hamid, M.A. AugFake-BERT: Handling imbalance through augmentation of fake news using BERT to enhance the performance of fake news classification. Applied Sciences 2022, 12(17), 8398.

[42] Shishah, W. Fake news detection using BERT model with joint learning. Arabian J. Sci. Eng. 2021, 46(9), 9115-9127.

[43] Liu, Y.; Pu, H.; Sun, D.W. Efficient extraction of deep image features using convolutional neural network (CNN) for applications in detecting and analysing complex food matrices. Trends in Food Science & Technology 2021, 113, 193-204.

[44] Barbhuiya, A.A.; Karsh, R.K.; Jain, R. CNN based feature extraction and classification for sign language. Multimedia Tools Appl. 2021, 80(2), 3051-3069.

[45] Guo, N.; Gu, K.; Qiao, J.; Bi, J. Improved deep CNNs based on Nonlinear Hybrid Attention Module for image classification. Neural Networks 2021, 140, 158-166.

[46] Zhao, W.; Zhu, L.; Wang, M.; Zhang, X.; Zhang, J. WTL-CNN: A news text classification method of convolutional neural network based on weighted word embedding. Connect. Sci. 2022, 34(1), 2291-2312.

[47] Zhang, F. A hybrid structured deep neural network with Word2Vec for construction accident causes classification. Int. J. Construct. Manage. 2022, 22(6), 1120-1140.

[48] Zhang, Y.J., Shi, T.T., Lu, Z.M. Image splicing detection scheme based on error level analysis and local binary pattern. Journal of Network Intelligence 2021, 6(2), 303-312.

[49] Singh, R.; Singh, S. Text similarity measures in news articles by vector space model using NLP. J. Inst. Eng. India Ser. B 2021, 102(2), 329-338.

[50] Fakeddit dataset link: https://paperswithcodecom/dataset/fakeddit