Journal of Theoretical and Applied Information Technology
31st August 2024. Vol.102. No. 16

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AN AUTOMATED MULTIMODAL HYBRID SYSTEM FOR WEB CONTENT FACT-CHECKING BASED ON BERT LANGUAGE MODEL AND CONVOLUTIONAL NEURAL NETWORK

C. VISHNU MOHAN, N. V. CHINNASAMY

Research Scholar Department of Computer Science Karpagam Academy of Higher Education
Coimbatore, Tamil Nadu, India

Assistant Professor Department of Computer Science Karpagam Academy of Higher Education
Coimbatore, Tamil Nadu, India
vishnumohanc01@gmail.com, chinnasamy.nvaiyapuri@kahedu.edu.in

ABSTRACT

Over the last decade, people have been widely using online platforms for sharing information and for understanding the news that has been happening around them. Classification of social media texts, tweets etc., are one of the emerging areas of research in today's world, especially when it comes to information about political and entertainment sectors. However, there are certain challenges due to the fact that most commonly used Machine Learning techniques have not proven to be optimal, when considering both textual and image data for fake content detection. This study explores the efficacy of a hybrid deep learning architecture that leverages BERT for text representation along with Convolutional Neural Network (CNN) for classifying news as real or fake.

Keywords—*Machine Learning, classification, deep learning, BERT, CNN*

1. INTRODUCTION

The advances in technology over time and the widespread use of the Internet have changed the nature of the digital world and the way information is shared. The Internet has become a key tool for information research. Social media is the most popular reason for people to connect to the Internet. People's habits have altered much because of the fact that they use social media so often and has thus increased its popularity. Digital news has become most people's primary information source to know about the happenings around them. However, there are a large volume of online information that are questionable and often even meant to deceive. Also, a few false news stories are so close to the actual ones that it is challenging for people to distinguish them apart.

Due to their low cost, ease of use, and the viral nature, a number of online social media platforms, including WhatsApp, Facebook, Twitter, Instagram, YouTube, and many others, have grown in popularity. There are now a lot more people using the internet, and they utilize it for a variety of purposes. Internet-based news disseminates quickly and may be valid or invalid. People lack the intelligence to discern whether news is reliable or not. False news spreads quickly. Social media and word-of-mouth are two ways by which news can spread. News that is intentionally produced to deceive people is referred to as fake news. The term "fake news" refers to a phenomenon that has several definitions and takes many forms, ranging from exaggeration leading to fabrication [1]. This is even worse when they are even accepted by the society. False news has developed and evolved from time to time such that its frequency in online media is inappropriate and overwhelming [2].

Fake news has a negative effect on a person's, society's, or institution's reputation. The Presidential elections of the United States in 2016 marked a turning point in online misinformation, with a pro-candidate campaign spreading demonstrably false information over 37 million times on Facebook [3]. But even though it has recently grabbed a lot of attention, identifying fake news is a very difficult challenge. Fake news is typically produced by editing images, text, or videos, which emphasizes the importance of a multimodal detection.

Section II discusses briefly on the literature where researches are conducted on applying various techniques in the classification of content as fake or real. The Problem statement is given in Section III, and the Objectives of the research (Section IV) are

also identified based on the review of literature. A new methodology using a hybrid approach of BERT+CNN is proposed in the methodology section (Section V). Section VI discusses the implementation results of the new multimodal system.

2. RELATED WORKS

Academicians are now seriously considering the widespread dissemination of false information on the social media, as explained by Wu et al. [5]. Facebook, Twitter, Reddit, PolitiFact, Instagram, and other social media sites became increasingly popular, especially following the 2016 US Presidential election campaigns. Contrary to misinformation, which may be unintentional, disinformation is typically the false information that has been deliberately spread. The landscape of fake news detection has seen a surge in innovative approaches in recent times. This section includes pertinent research on spotting fake news on social media websites. The literature review that is now accessible indicates that machine learning models were frequently used to identify fake news, followed by deep learning models, and that transfer learning and pre-trained models are now also performing well in this domain.

Using n-gram analysis, Ahmed et al. [6] suggested a method to identify fake news. At first, the authors decided to use two feature generation techniques and tested them against 6 various machine learning classifiers. The researchers harnessed both Term Frequency (TF) and Term Frequency-Inverse Document Frequency (TFIDF) for extracting meaningful features from the text data, enabling effective model training. To identify the optimal model for the task, they compared the performance of K-Nearest Neighbors, Support Vector Machines, Decision Trees, Logistic Regression, and Linear Support Vector Machine classifiers. With an accuracy of 92%, they achieved the best result utilizing the feature extraction method Term Frequency-Inverse Document Frequency (TF-IDF) with the classifier Linear Support Vector Machine (LSVM). Although this study showed a great accuracy, this may be due to a Population Bias or Representation Bias, as the authors focused on n-gram analysis, as explained in the study conducted by Ninareh et al. [7]. Reliance solely on n-grams could be problematic, as we can see in Cruz et al. (2019) [8], because this feature extraction method may change based on media attention over time.

For the purpose of detecting fake news, Perez et al. [9] first introduced and discussed collecting, annotation, and validation procedures of two novel datasets. Second, the authors conducted a series of experiments and exploratory data analyses utilizing the datasets indicated above to pinpoint linguistic characteristics that are predominately present in the fake content. The authors used the Linear Support Vector Machine (LSVM) classifier and employed a fivefold cross-validation technique for classification. The highlight of this research is that the best possible combination of feature variables was selected, as opposed to the research published by Ahmed et al. [6] earlier, which puts more emphasis on finding the best feature variable generation and classification methods and less emphasis on the features themselves (features generated by n-grams). The authors also performed a number of experiments with various feature combinations in order to achieve this, including n-grams, punctuation characters, psycho-linguistic features, readability, and syntax. They created a fake news detector that performed at its peak with 78% accuracy when all features were utilized. The findings point to significant discrepancies in the substance of fake and real news. Some of these variations include the employment of more social and positive phrases, the expression of greater certainty, the emphasis on the current scenario and of future, and the presence of punctuation characters, verbs, and adverbs in false news articles.

Yaqing Wang et al. [10] have made use of the Multi-Modal Feature Extractor: EANN (Event Adversarial Neural Network) to detect fake news across many media channels. As they only learn event-specific properties that cannot be applied to unobserved events, the existing models struggle to distinguish between true and false reports on recently emergent and time-critical occurrences. However, this EANN can pick up on traits that are independent of the course of an event, which gives it the ability to spot fake news reports during live events. Their proposed model consisted of a multi-modal feature extractor, the event discriminator, along with a fake-news classifier that forms the model. Weibo and Twitter are only among a couple of the multimedia datasets that this study is built on. Using transferable characteristics depiction, the suggested system performs better than the current baseline methodologies.

Khan et.al. [11] conducted a conventional experiment on three extensive and diverse datasets in evaluating efficacy of the various ML algorithms. From 19 ML techniques, 8 employ conventional ML models, 6 utilize standard deep learning models, and 5 leverage state-of-the-art pre-trained models such as BERT. Across all datasets, it was observed that BERT-based systems outperformed alternative techniques in terms of both potential and performance. Furthermore, BERT-based methods demonstrated reliability, proving effective even with a small sample size. Naïve Bayes with N-Gram

models has yielded comparable results to neural network-based models on sufficiently large datasets.

To determine whether an event is real, Ma et al. [12] used a GRU with multiple layers and trained it using sequence of tweets based on time. A TFIDF score of 5000-dimension was fed to the model as the input from each tweet. When contrasted with non-deep learning methods such as Decision Trees, Random Forest and SVM classifications, this approach demonstrates a performance improvement in gain with 10%.

As per the information in the news articles, Fang et al. [13] recommended the utilization of a self-attention-based CNN, and they explained that the self-attention-based CNN produced greater accuracy than RNN-based models when given the task to identify articles that contain non-factual information. Their learning approach often uses features that are derived using linguistic techniques and static network analysis. However, it does not employ dynamic network information.

Rohit et al. observed that, on the Kaggle fake news data set, FNDNet performed better than feature engineering and conventional machine learning solutions [14]. The GloVe method was employed to transform words into a 100-dimensional vector, serving as the input for the FNDNet architecture, and the model in this instance only considers the features in the vector space. Constructed upon an enhanced Convolutional Neural Network (CNN) structure, this deep learning architecture combines three concurrent convolutional layers, followed by the addition of dense layers. When compared to traditional ML and deep learning utilizing CNN and LSTM models, FNDNet exhibited superior performance.

In the detection of fake news, deep learning techniques were used by Hiramath and Deshpande [15] to compare the classification algorithms Naïve Bayes, Random Forest, Logistic Regression, SVM, and Deep Neural Networks. They performed experiments on the LIAR dataset, employing common text preprocessing techniques derived from the field of Natural Language Processing, which include procedures like stop word removal and stemming. They thereby validated the FNDNet findings and observed that Deep Neural Networks outperform conventional machine learning techniques.

In the detection of fake news using deep neural networks, several models using Hashing Vectorizer in addition to TF-IDF as a vector space representation were analyzed [16]. The authors used K-Nearest Neighbors (KNN), Naïve Bayes (NB), Convolutional Neural Networks (CNN), Decision Tree (DT), Long Short-Term Memory (LSTM), and Random Forest (RF). The algorithms' performance accordingly declined in the order listed. Combining CNN and LSTM produced the greatest results, supporting the notion that deep learning models perform well. They combined a number of Kaggle datasets for the experiment.

The Text-mining-based fake news detection employing various types of Ensemble-based methods utilized comparable vector space representations and stylometric features [17]. Leveraging the richness of the stylometric data, researchers extracted three distinct sets of features to enhance model performance. The first one had a high character count (with or without whitespace), high complexity score, Gunning-Fog index, Flesch-Kincaid readability score, and a number of unique words. The second collection is based on a dataset for lie detection, and its features may be broken down to seven categories: vocabulary, uncertainty, quantity, Flesch-Kincaid score, and grammar. The last feature subset consisted of a write-print feature set that contain authorship attributes given in brief texts, which was divided into the following categories: Character, Word, Syntax, Structure, and Content. Various approaches for vector space representation were utilized, including TFIDF, skip-gram, bag-of-words, TF-IDF, and continuous bag-of-words which were employed to predict the next contextual word, and both Word2Vec and FastText tools. Both types of features underwent a feature selection process, with recursive elimination of less influential features ultimately resulting in the selection of stylometric features. Lemmatization, stemming, and Chi-square tests for feature selection were used to reduce the vocabulary in the word-vector space. They employed the Gaussian Naïve Bayes and Multinomial Naïve Bayes classifiers, Linear Regression, Random Forest Classifier, the kNN, SVM, Extra Trees Classifier, simple bagging, and bagging with AdaBoost - Gradient boosting, as classifiers. Armed with CboWWord2Vec features, Gradient Boosting reigned supreme among non-ensemble models, achieving the highest overall accuracy in classification. By harnessing the power of CboWWord2Vec's context-aware word representations, even run-of-the-mill non-ensemble techniques saw a significant accuracy boost, with Gradient Boosting shining the brightest.

In order to identify fake online book reviews, the usage of Rhetorical Structure Theory (RST) for Fake Online Review Detection is proposed by Olu [18]. The author used Deceptive Review as a dataset (DeRev). They created common macro-relations by grouping certain RST properties. According to the corpus analysis, the fake reviews have more macro-relations for Elaboration, Joint, and Background, whereas the genuine reviews possess relations attributed by Explanation, Evaluation, and Contrast.

It was also observed that the genuine reviews have relations for better comparison. This study demonstrates that reviewers who have been paid to write fake reviews frequently use the misleading pragmatics as seen in RTS method. They tend to mention the title, author, or substance, which is against genre norm.

Sentiment analysis, sentiments, and cosine similarity scores on Naïve Bayes, Random Forest classifiers trained on LIAR dataset are used in Fake News Detection in the work by Bhutani et. al [19]. They concluded that incorporating sentimental score improves the model's accuracy.

Through a combination of pre-processing techniques like stop word removal and feature extraction, Victoria et al. [20] built a sophisticated text analysis system utilizing TF-IDF vector space representation with unigrams and bigrams to identify nuanced language elements like irony, satire, and humour in various news feeds. Being an ML algorithm used for prediction, the SVM model was used. Punctuation extraction, absurdity using Part-of-Speech (PoS) tagging and Named Entity Recognition (NER), humour using knowledge-based punchline identification, grammar by counting the tags (PoS), and negative effect using the LIWC lexicon were among the attributes that were extracted. The detection was enhanced by each of these criteria, with humour features showing the weakest increase. The SVM was trained for a classification job using 10-fold cross-validation by the machine learning library sklearn.

Using collective user intelligence to detect fake news, Feng et al. [21] has developed a Neural User Response Generator. The two-level CNN with User Response Generator (TCNN-URG) is employed to determine the news's credibility based on both its substance and readers' responses to similar items in the past, as well as to predict how they would react to the latest information. When real-time user reactions are unavailable, this method can be used to identify fake news early. Both the Twitter dataset and Weibo dataset were used. The conditional variational autoencoder serves as the basis for the User Response Generator.

According to the researches conducted by Natali et al. [22], the authors have developed a hybrid model for fake news detection, where textual information is combined with user feedback from articles as well as data of the people who posted the news. This system operates in three phases: the first analyzes text and responses using an RNN network, while the second evaluates source reliability based on user and group data, and the third module combines these methods, tested on data sets from Weibo and Twitter.

The study conducted by Diego et. al [23] focuses on evaluating the legitimacy of the entire websites. It examines the current state of this field's research as well as recent setbacks, such as the price of external APIs and Google PageRank's discontinuation. They ignore user-based social variables due to the significant bias that these variables inject into the final model supported by the ANOVA test, and instead focuses on the online credibility model by using just content-based features. The final model was assessed using the Likert 5-star scale and two data sets - Microsoft Dataset and the Content Credibility Corpus, each consisting of many URLs. Readability, PageRank data, General Inquirer (a dictionary similar to LIWC), Vader Lexicon (sentiment), Lexical analysis algorithms like LSA, Authority data including the contact address, tags in social networks, webpage's HTML2seq feature in the form of a bag-of-tags were all content-based features they used. Regression and classification were the two configurations used for the credibility prediction. As a result, they put this model to test a real-world fact-checking problem and discovered that the model was able to distinguish between reputable and unreliable websites based on the assertions made in support of and opposition to each.

Similarity-Aware: SAFE Multi-Modal Fake News Detection, proposed by Xinyi et al. [24] uses multi-modal detection to identify fake news by using both textual and visual content. Although this has been done before, their method is innovative since it considers the similarities between textual and visual data and the technique that they implement to convert image data to text. They used the Linguistic Inquiry and Word Count (LIWC) for the textual data, and the VGG-19 – a convolutional neural network of 19 layers deep for the visual data, and the att-RNN network for the multi-modal data as baselines for their trials where all of which were outperformed.

By querying the knowledge graphs created for news stories from the knowledge base Dbpedia, it is possible to assess the credibility of the news based on the reliability of the content itself. This strategy was one of four ways listed in the survey [25] and used in [26][27][28][29]. It is regarded as an automatic fact-checking method.

Agrawal et al. [30] has considered time series into account on the Twitter news and employed a fake news classification method based on two algorithms - logistic regression, and a harmonic algorithm, and finally examined the performance. They inferred that the harmonic algorithm performed best with an accuracy of 90%. Ni et al. [31] has proposed a model that uses attention-based neural networks to study about fake news classification that spots the clues surrounding fake news and the trend by which they spread. For this, a Multi-View

Attention Network (MVAN) was being developed for detecting fake news on Twitter. This model had the ability to spot the clue words related to a particular news event.

Singhal et al. [32] introduced the FACTDRIL (Fact Checking Dataset for Regional Indian Languages), focusing specifically on the languages spoken in India such as Malayalam, Marathi, Bangla, Telugu, Tamil, and Sinhala. The dataset comprises 22,434 candidates from 10 Indian languages of Low Resource that have obtained accreditation from the International Fact Checking Network (IFCN). FACTDRIL stands as the inaugural extensive multilingual dataset designed to assess the accuracy of unverified claims in these low resource languages. It also presents a novel feature termed "Investigation reasoning through manual interference," addressing various methods employed by fact-checkers to determine the credibility of news.

Azer et al. [33] developed a machine learning-based classifier for verifying the credibility of news on the Twitter platform, focusing on two primary elements: content-based and user-based. Utilizing the Pheme dataset, and dividing them in a specified ratio, the authors applied 7 supervised ML approaches—Maximum Entropy, Support Vector Machine, Naïve Bayes classifier, K-Nearest Neighbor classifier, Random Forest, Conditional Random Forest, and Logistic Regression (LR). The dataset was partitioned in such a way that 80% was used for training, 10% for the need of testing, and the remaining 10% for the purpose of validation. The study's outcomes reveal that Random Forest (RF) exhibits the highest performance, achieving accuracy rates of 82% on user-specific features and 83% on combined features. Logistic Regression (LR) excels on content-based features, achieving an accuracy of 73.2%. Furthermore, the analysis indicates that user-based features exert a more significant influence than content-based features.

Sahoo and Gupta [34] have proposed an automated method to identify fake news based on a variety of data properties of Facebook using deep learning and machine learning techniques using a chrome environment. This proposed methodology uses certain additional information tied to the user's Facebook account and its news content for identifying fake tales. The Long Short-Term Memory (LSTM) algorithm, which is a deep learning technique, has proved with an exceptional performance of 99.4% when compared to the other learning methodologies.

A FakeBERT was proposed by Kaliyar et al. [35] combining multiple parallel blocks of a single-layer deep CNN along with the BERT model. The BERT is a deep learning technique that depends on bidirectional encoder representations from transformers. Understanding the ambiguities present in natural language ia a challenging aspect, which is handled well by this combination.

To extract attitude representations from a post and any accompanying replies, Xie et al. [36] suggested using the model – Stance Extraction and Reasoning Network (SERN). To accomplish binary fake news classification, they merged the posture representations and multimodal representation of both textual and visual content of a post.

The PHEME dataset and a condensed representation of the authors' own dataset from Fakeddit are used by the researchers Zubiaga et al. [37]. There are 5802 tweets in the PHEME dataset, 3830 of which are true and 1972 fake. The obtained accuracy rates are 76.53% and 96.63% respectively for Fakeddit and PHEME datasets.

The News Detection Graph (NDG), used by Kang et al. [38], is a heterogeneous graph that includes source nodes, domain nodes, review nodes and news nodes. Additionally, they suggested that implementing the Heterogeneous Deep Convolutional Network (HDCN) is beneficial in order to extract the news nodes' embeddings in the graph. Utilizing condensed versions of the Weibo and Fakeddit datasets, the authors assessed this approach. They achieved an F1 score of 96% for the Weibo dataset, 86% (three classes), 89% (binary classification), and 83% for the Fakeddit dataset (six classes).

Aum and Choe [39] propose an automated deep learning model called srBERT for classifying articles using the BERT (Bidirectional Encoder Representations from Transformers) algorithm. The srBERT is a pretrained model that is trained using the abstracts of various articles of Systematic Reviews (SRs). Two types of datasets were used – one comprising of more than 3200 articles under the theme – Moxibustion therapy and the second dataset comprising 400 case studies to verify how effective the treatments are for all types of diseases. The authors were able to showcase the performance of BERT when applied to text classification. An accuracy of 94.35% and F1 score of 66.12 was obtained using this model.

Mohammadi and Chapon [40] have developed various models to check how the outputs of each layer of BERT affects the performance of various classification tasks. A critical comparison of each of these models is conducted in terms of their performance. They conducted a thorough study on how the hyperparameter values affect the model performance. A series of tasks are performed such as Intent Classification, Answering user's questions,

Sentiment Analysis, and Topic classification. The results are given in Table. 1 below:

Task	Dataset	Accuracy			
		BERT-base	***BERT-CLS***	***BERT-Last***	***BERT-BiLSTM***
Intent Classification	30K-Intent	**64%**	60%	62%	64%
Sentiment Analysis	IMDB	**91%**	83%	87%	89%
Answering User Questions	Yahoo Answers	**71%**	65%	70%	63%
Topic Classification	AG's News	**94%**	89%	93%	90%

Table. 1 Classification Accuracy

The experiments finally interpreted that the BERT-base model having a fully connected layer for classification has outperformed all the other BERT models.

3. PROBLEM STATEMENT

Our purpose is to research the viability of automated methods to spot fake news spread on digital channels. While fact-checking is a crucial method for spotting fake news, it is ineffective even though simple. Therefore, an automatic fake news detection system may be used to help readers to identify whether a content is more likely to be false, while ultimate final decision is left for a professional.

Formally, the fake news prediction can be defined as – "to assess whether a series of news stories from social media that contain text and image information is fake or not". However, it is not that simple to recognize fake news automatically. First, it is intrinsically difficult for people to distinguish between true and false news [4], especially when it comes to touchy themes like politics, entertainment and health. The problem of identifying fake news is made even more difficult by the fact that news items are generated by several sources, each of which has a unique style of representing the news contents and inherent biasing. In addition, they are transmitted in many ways in various platforms.

Digital media and social networking platforms present a variety of research issues in identifying fake news. Firstly, the fact is that there are people who purposefully create fake news to confuse readers, such that the readers find it difficult to identify whether the news is real or not by just substantial reference. Thus, identification of fake news relying heavily on text data is always not productive. Second, additional data must be provided to improve detection, like the social interactions of users including the posts and their replies, and external knowledge bases [4]. However, the researches should be aware of the fact that using these supplementary data might affect the quality of data. Although information from various modalities can offer hints for fake news identification, it can raise concerns in drawing out the key aspects derivable from each modality and integrate them to an interpretable form.

The majority of studies are focused on unimodal data, however as information can come from various modalities, it is important to take into account both text and visual data for better fake news detection performance.

4. OBJECTIVES

The gap identified after reviewing similar studies in the area of fake news detection is pointed out below:

- A single modality feature makes it difficult to spot fake news.
- Numerous strategies to identify fake news have been developed using linguistic approaches. However, there hasn't been much work done on visual-based verification.
- Source verification is seen as a component that is absent from the current models.
- The size of the datasets used in the literature is rather small.
- Time-sensitive and recently occurring events have received less attention from the current methodologies.
- Dataset bias is a concern because the bulk of studies are concentrated on a specific category of news (such as political news).

The following are the objectives that are finally arrived at after a detailed review of literature related to fake news detection:

- To analyze the prediction performance of fake news detection solutions in the-state-of-art through review of literatures.
- To propose a model for automatically detecting fake news for both long and short series of text data, such as news articles and tweets.
- To build a system to identify fake images automatically.
- To assess the performance of the proposed approach using various news datasets.

5. METHODOLOGY

A. Dataset

This research uses the IFND (Indian Fake News Dataset) dataset for training and testing the hybrid model. The dataset encompasses both text and image data. It predominantly consists of news content related to events occurring between the years 2013 and 2021. To compile the dataset, the Parsehub tool was employed to scrape content from various sources. Samples from the dataset are shown in Figure. 1.1 and Figure 1.2.

	domain	img_url
0	i.redd.it	https://preview.redd.it/l0ga0tug17k31.jpg?widt...
1	i.imgur.com	https://external-preview.redd.it/VX7bXDu9G18UZ...
2	i.redd.it	https://preview.redd.it/bxp58zf01zy21.jpg?widt...
3	i.redd.it	https://preview.redd.it/ipfrolruml411.jpg?widt...
4	i.imgur.com	https://external-preview.redd.it/FPz_jid8GIOdf...
...	...	...
37108	msn.com	https://external-preview.redd.it/xT649IGK__Wyc...
37109	independent.ie	https://external-preview.redd.it/__1AjIER1EnF8...
37110	theguardian.com	https://external-preview.redd.it/Daruj1-0M6QDe...
37111	dailyjournal.net	https://external-preview.redd.it/rRJn2A584GGhv...
37112	i.redd.it	https://preview.redd.it/be71j19dltoz.jpg?width...

Fig. 1.1. IFND Image Dataset

	news_title	is_fake
0	My Xbox controller says hi	1
1	PsBattle: New image from The Mandalorian	1
2	This tree I found with a solo cup on it	1
3	Dude, I'd feel the same if I got a pole throug...	0
4	PsBattle: Japanese Black Pine tree	1
...	...	...
37108	Hero surgeon treks for three hours through sno...	1
37109	Half-baked burglar plunders family's pavlova	1
37110	“Russian bid” to influence “Brexit” vote detai...	1
37111	Nicaraguan bank sanctioned by US shuts down	1
37112	This column and emergency light	0

[37113 rows x 7 columns]

Fig. 1.2. IFND Text Dataset

B. Detailed Methodology

There are different machine learning methods currently available for automatically detecting fake news [45]. Deep learning, one of its more recent branches, began to gain increasing significance in the discipline over time as more researches were done on it. This is because deep learning approaches, which outperform traditional machine learning techniques in a number of sectors [46], have more than one hidden layer between the input and output.

Bidirectional Encoder Representations from Transformers (BERT) – based models such as BERT Base, BERT Large, DistillBERT, ALBERT and RoBERTa are among the deep learning techniques that makes use of a self-learning mechanism of NLP called transformers when trying to classify text. Researches using BERT for text classification is gaining its popularity because of BERT's self-attention mechanism, ability to better understand the context of text, and its existence as a pretrained model.

Classification using textual data alone cannot be seen trustworthy, a multimodal approach, including images corresponding to text is a better option. Convolutional Neural Networks (CNNs) are proven to be highly effective for image classification due to their ability to automatically learn relevant features from raw pixel data, their hierarchical representation learning, and their capacity to handle spatial hierarchies in images.

From the review of literatures, we can see that BERT can be used effectively for classification of textual content and CNN for classification of images. Here, we prefer a hybrid model (Figure. 2) because it would combine the image vector data and text vector data for classification thereby giving a better performance than most unimodal systems.

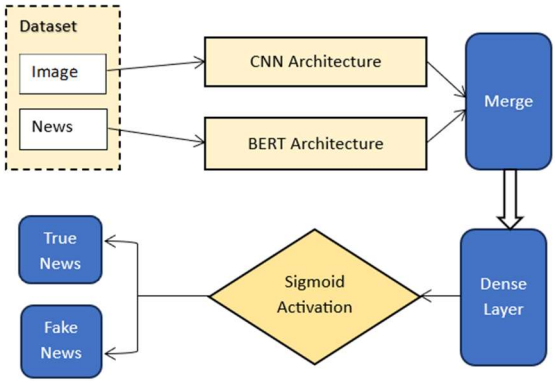

Fig. 2. High Level Representation of Proposed Methodology

The complete multimodal architecture of the methodology implemented is divided into three important sections:

- The self-attention-based text classifier - BERT
- The CNN-based image classifier
- A concatenation layer that is implemented for the final fake news classification.

Journal of Theoretical and Applied Information Technology
31st August 2024. Vol.102. No. 16

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The detailed diagram demonstrating the process of classification is shown in Figure 3.

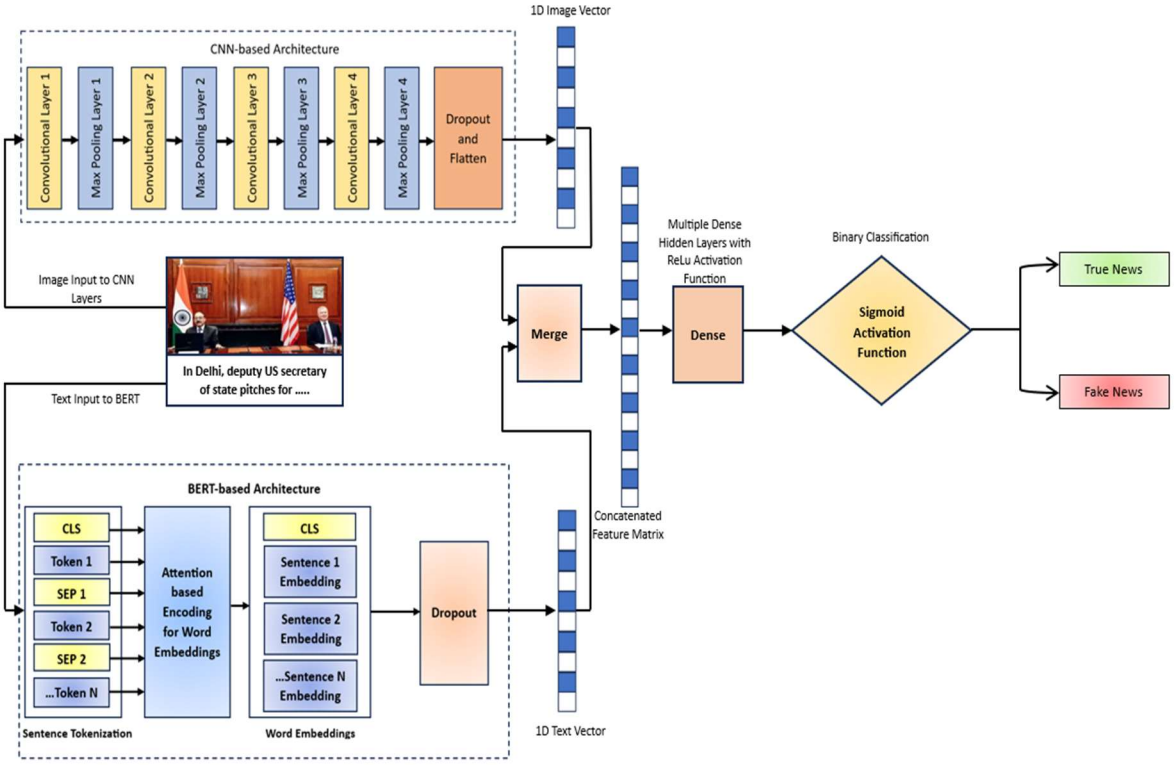

Fig. 3. Detailed Methodology

BERT is a pretrained model that first underwent a pretraining on an extensive text data using unsupervised learning. After this initial training, the model is fine-tuned for a particular text classification task, utilizing a labelled dataset specific to the target objective. However, as we consider the image classification also in addition to the text classification, we modify the general approach of BERT's classification algorithm. Equipped with BERT's pre-configured text preparation tools, the study converted the input news articles into uniform vector formats, laying the foundation for further analysis (Figure. 4).

As the first step, the text input is tokenized, breaking it into subwords or words, and special tokens like [CLS] and [SEP] are incorporated. The [CLS] token typically serves as the representation for the entire sequence. While performing the tokenization, a few characters like spaces, and punctuations are ignored and will not be a part of the final list of tokens. An example of how tokenization is performed is given in Figure. 5.

Next, BERT generates contextualized embeddings for each token in the input sequence. The [CLS] token's embedding is commonly employed as a comprehensive representation capturing the contextual information of the entire sequence. Transformers utilize self-attention mechanisms, where each word in the input sequence can attend to every other word. The attention scores signify the importance of each word in relation to the current word. Attention scores are computed for a given word with respect to all other words, indicating their relevance. These scores are dynamic and adapt to the relationships between words. The attention scores undergo softmax normalization, creating a probability distribution. This normalization ensures that the weights assigned to each word are proportional and sum to 1, enhancing the model's ability to focus on meaningful words. Now, the weighted sum of the word embeddings of all words based on their attention scores.

This process generates contextualized representations, accounting for the relationships between words in both left and right contexts. This attention mechanism is a fundamental aspect of the transformer architecture, allowing BERT to consider the entire context of a word by dynamically assigning weights to its relationships with other words in the sequence.

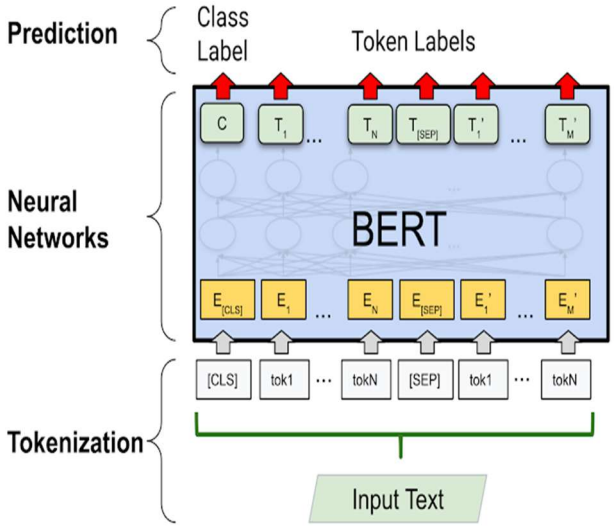

Fig. 4. BERT

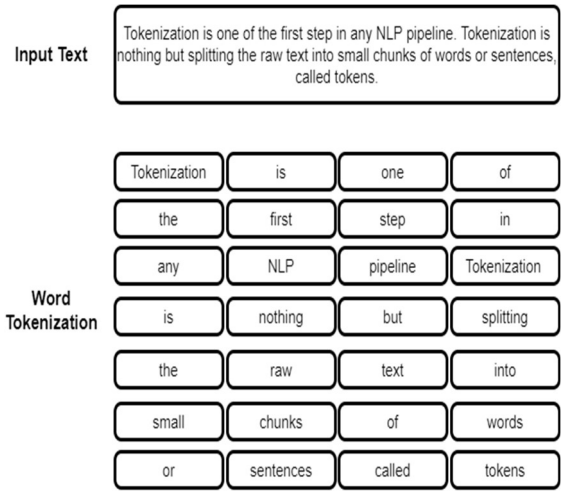

Fig. 5. Sample Tokenization

BERT's design leverages bidirectional context, allowing it to capture dependencies in both directions. This contrasts with traditional models that process language input in a unidirectional manner. It incorporates multiple layers of attention-based encoders. Each layer refines contextualized representations based on the information from the preceding layer. During pre-training, BERT employs attention masks to train the model to predict masked words in a sentence. This encourages the model to comprehend bidirectional context and acquire robust representations for each word. The output is a 1-dimensional feature matrix which is then passed to the dropout layer to avoid overfitting.

In the third phase, BERT utilizes a dropout layer as a regularization technique. Dropout, a common practice in neural network architectures, involves randomly deactivating a portion of the units (neurons) within a layer during each iteration of training. This serves as a preventive measure against overfitting and encourages the network to learn features that are more resilient and applicable across diverse scenarios. Within BERT's design, dropout is applied to multiple layers, encompassing both the attention mechanism layers and the subsequent feedforward neural network layers. The dropout probability, a tuneable hyperparameter, dictates the fraction of units to be randomly deactivated during training. Common values for dropout probability typically range between 0.1 and 0.5, but the optimal value may vary based on the specific architecture and dataset.

Finally, a dense layer employed with 768 neurons is utilized that makes use of the ReLu activation function. This produces the final 1-dimensional feature matrix vectors of the text input.

The CNN architecture is employed for the image analysis part. The corresponding image of the text is sent to the CNN layers for processing. The CNN used here consists of five layers and then forming a final vector representation of the image.

The input image is first sent to the convolution layer of the CNN architecture. This layer conducts the convolution operation, wherein a filter or kernel slides over the input data, extracting features at different spatial positions. The convolution operation entails systematically moving a small filter, also known as a kernel, across the input image. At each position, the filter calculates the dot product of its weights and the input values. This process is conducted independently at various spatial locations, capturing localized patterns. Multiple filters are employed in the convolutional layer, with each filter dedicated to detecting specific patterns or features in the input data. The outcome of applying a filter to the input is referred to as a feature map. Following the convolution operation, an activation function, commonly ReLU (Rectified Linear Unit), is applied element-wise to introduce non-linearity. This contributes to the network's ability to learn complex mappings.

The next layer is the max pooling layer that is used to reduce the spatial dimensions of feature maps while preserving essential information. Max pooling divides the input feature map into non-overlapping regions, often of size 2x2 or 3x3. Within each region, only the maximum value is retained, effectively downsizing the spatial dimensions of the feature map. Max pooling incorporates a striding parameter, determining the step size for moving the pooling window across the input. Retaining only the maximum values in each region helps preserve critical information while reducing computational demands in subsequent layers. The vector once again

passes through a convolution layer and a max pooling layer before dropping out and flattening. features. This volume then gets squeezed down to a single, 1D representation by a final dense layer,

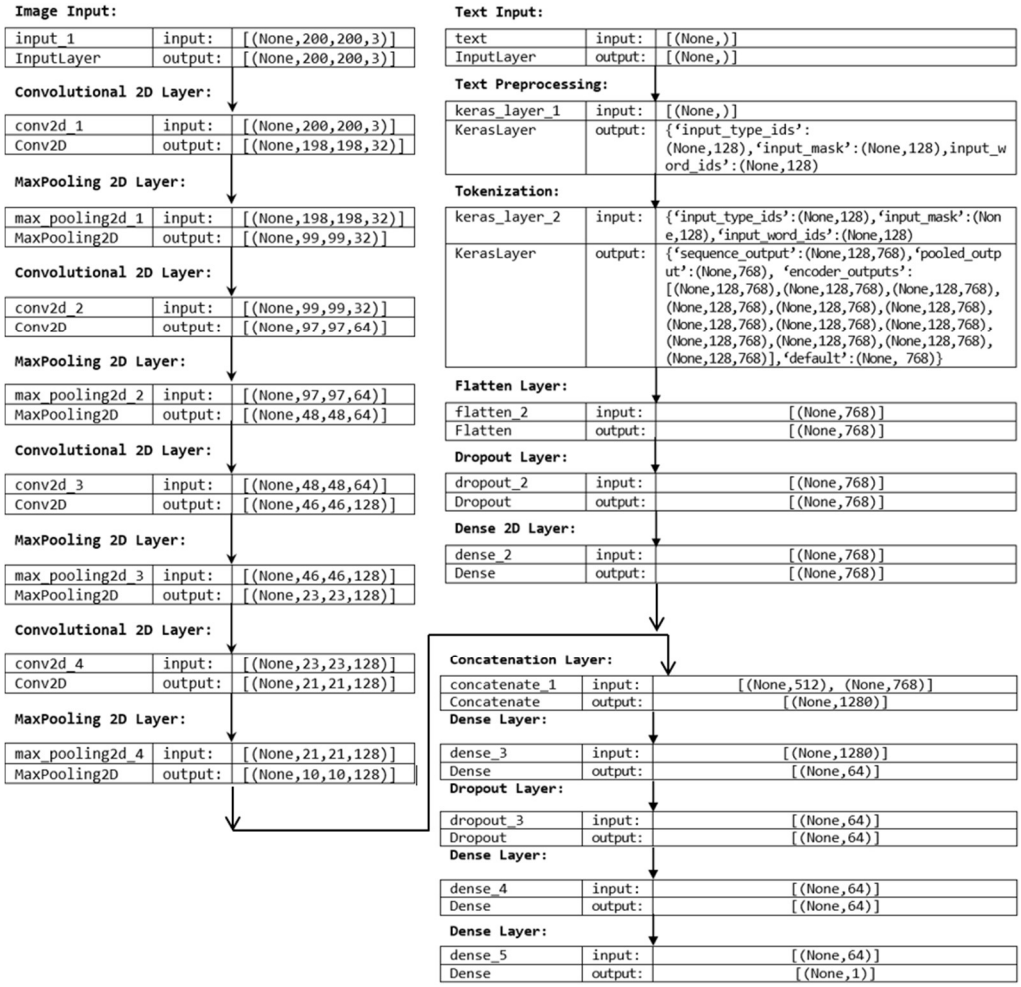

Figure. 6. Implementation Model Plot

The final layer in CNN includes a dropout and flatten layer. Typically positioned after convolutional and pooling layers and just before fully connected layers, the final dropout layer is implemented to mitigate overfitting. By randomly deactivating a portion of neurons during training, dropout encourages the network to generalize well to new data. Following the dropout layer, the flatten layer reshapes the output received from the preceding convolutional and max-pooling layers into a 1D vector representation. This is a flattened vector that becomes the input for the fully connected layers. The flatten function transforms the multi-dimensional tensor output from preceding layers into a one-dimensional vector, facilitating the transition to fully connected layers. After processing an image through a series of convolutional layers, the CNN model outputs a 3D volume capturing the extracted essentially summarizing the image's essential information.

The CNN architecture was constructed with four sequential building blocks, each comprising a convolutional layer and a max pooling layer. Within these blocks, 3x3 kernels were employed for feature extraction in the convolutional layer, while a 2x2 window was used in the max pooling layer for dimensionality reduction.

With both text and image vectors generated, the next critical phase involves their fusion. To achieve this seamless integration, a custom layer was meticulously designed, tailored to address the specific demands of the research problem. This layer expertly leverages the Keras functional API to orchestrate the concatenation of the two distinct 1D matrices representing text and image data. Within

this process, the Keras 'concatenate' function acts as the bridge, seamlessly merging the outputs of the CNN-based image classifier and the transformer-based text classifier. The resulting concatenated matrix stands as a unified representation of both modalities, encapsulating their collective insights. This integrated feature matrix then embarks on the next stage of processing, encountering a dense layer composed of 64 neurons.

The culmination of this fusion process produces a final, refined matrix that holds the key to the news classification task. This matrix is fed into a carefully crafted dense layer, designed to extract the most pertinent information for the decision-making process. Within this layer, a multitude of dense neurons work in concert, their collective efforts further enhanced by a thoughtfully positioned dropout layer that strategically prevents overfitting. The final verdict—real or fake—is delivered by a sigmoid layer, acting as the gatekeeper of news authenticity.

The plot of the hybrid model's implementation using the IFND dataset is given in Figure 6.

6. EXPERIMENTAL RESULTS

The experiment was conducted with a drop rate of 0.2 for the dropout layer. The model boasted roughly 7.4 million trainable parameters and harnessed the Adam optimizer for efficient learning. To tackle the binary classification task, a binary cross-entropy loss function was employed. To prevent overfitting during training, an early stopping mechanism with a 3-epoch patience window was implemented. Additionally, the model trained on batches of 64 samples for 10 epochs (Table 2).

With this experimental setup and configuration specified, the model was trained to produce the results as given in Figure 7. It tracks two key metrics: training accuracy and validation accuracy.

Training accuracy reflects the model's ability to correctly identify fake or real news within the training data, while validation accuracy measures its performance on a separate dataset specifically reserved for monitoring the training process. The x-axis shows the number of epochs, and y-axis the classification accuracy.

Table. 2. Experimental Setup And Model Configuration

Trainable Parameters	74,71,745
Optimizer	Adam
Learning Rate	0.001
Loss Function	Binary Cross Entropy
Early Stopping Patience	3
Batch Size	128
Epochs	10

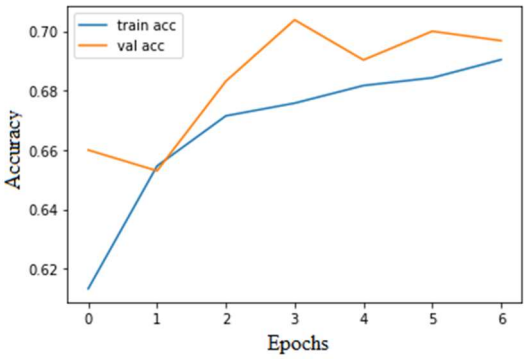

Fig. 7. Training vs Validation accuracy

The ideal scenario is for both lines to stay close on the chart. Here, we see both training and validation accuracy hovering around 70%, a commendable result. Additionally, the lines remain stable across epochs, implying the model avoids overfitting to the training data and retains its generalizability. The classifier report is interpreted in Figure. 8.

	precision	recall	f1-score	support
0	0.64	0.60	0.62	1225
1	0.75	0.78	0.77	1888

Fig. 8. Classifier Report

From the classifier report, it is seen that an F1 score of 0.619 is obtained for the class labelled "fake" i.e., fake news, and 0.765 for the class labelled "real", i.e., real news. The overall F1 score of 0.969 suggests a good overall balance between the precision and recall across both classes.

The model performance with the aid of a confusion matrix is given below (Figure. 9). The following are the points interpreted:

Accuracy: The model classified 73.3% of news articles correctly (4100 out of 5590).

True Positives (TP): The model rightly identified 2910 fake news articles.

True Negatives (TN): The model rightly identified 1190 real news articles.

False Positives (FP): The model incorrectly classified 200 real news articles as fake.

False Negatives (FN): The model incorrectly classified 1390 fake news articles as real.

When performing a class specific analysis, it was inferred that the model has a higher precision for fake news (82.1%) than recall (66.1%). This means the model is more likely to correctly classify a fake news article as fake (fewer FP) but misses some actual fake news articles (more FN). In addition, the model has a higher recall for real news (85.2%) than precision

(79.3%). This means the model is less likely to miss real news articles (fewer FN) but sometimes incorrectly classifies fake news articles as real (more FP).

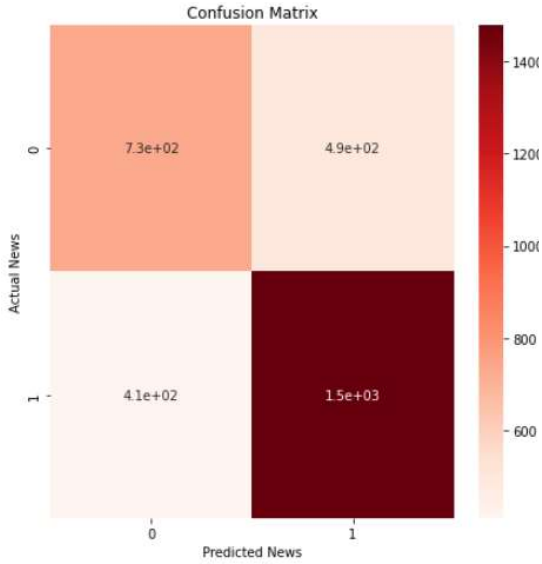

Fig. 9. Confusion Matrix

7. CONCLUSION

This study demonstrates the effectiveness of a multimodal approach combining BERT and CNN to detect fake news with higher accuracy than unimodal approaches covered in the literature survey, showcasing the importance of utilizing diverse information sources for reliable detection. Our hybrid model achieved an F1 score of 0.969 and accuracy of 73.3% in identifying fake news, suggesting the potential of multimodal analysis to significantly improve fake news detection capabilities. While the multimodal approach showed promising results, further research is needed to improve accuracy for specific types of fake news, like manipulated images or satirical content. Our findings suggest that incorporating additional modalities like audio or video analysis could further enhance the model's ability to discern genuine and fabricated content. Performing a comparative study with other news datasets is challenging and remains as a future work due to the fact that this hybrid model require both image and its corresponding textual content.

REFERENCES

[1] Rachel R. Mourão and Craig T. Robertson, "Fake news as discursive integration: An analysis of sites that publish false, misleading, hyperpartisan and sensational information", in Journalism Studies, 20(14):2077–2095, 2019.

[2] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu, "Fake news detection on social media: A data mining perspective", in ACM SIGKDD explorations newsletter, 19(1):22–36, 2017

[3] Ahmed, A. A. A., Aljabouh, A., Donepudi, P. K., & Choi, M. S., "Detecting Fake News using Machine Learning: A Systematic Literature Review". arXiv preprint arXiv:2102.04458, 2021, https://arxiv.org/abs/2102.04458.

[4] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu, "Fake news detection on social media: A data mining perspective", in ACM SIGKDD Explorations Newsletter, 19(1):22–36, 2017.

[5] Wu, L., Morstatter, F., Carley, K. M., & Liu, H., "Misinformation in Social Media: Definition, Manipulation, and Detection", in SIGKDD Explor. Newsl., 21(2), 80–90, 2019, https://doi.org/10.1145/3373464.3373475.

[6] Hadeer Ahmed, Issa Traore, and Sherif Saad, "Detection of online fake news using n-gram analysis and machine learning techniques", In International conference on intelligent, secure, and dependable systems in distributed and cloud environments. Springer, 2017.

[7] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan, "A survey on bias and fairness in machine learning". arXiv, 8 2019.

[8] André Cruz, Gil Rocha, Rui Sousa-Silva, and Henrique Lopes Cardoso, "Team fernandopessa at SemEval-2019 task 4: Back to basics in hyperpartisan news detection", In Proceedings of the 13th International Workshop on Semantic Evaluation, pages 999–1003, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics.

[9] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea, "Automatic detection of fake news", arXiv preprint arXiv:1708.07104, 2017.

[10] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, GuangxuXun, KishlayJha, Lu Su, and Jing Gao.2018, "EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection", in Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (London, United Kingdom) (KDD '18). Association for Computing Machinery, New York, NY, USA, 849–857. https://doi.org/10.1145/3219819.3219903.

[11] Khan, J. Y., Khondaker, M., Islam, T., Iqbal, A., &Afroz, S., "A benchmark study on machine learning methods for fake news detection", Computation and Language, 2021, https://arxiv.org/abs/1905.04749.

[12] Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B. J., Wong, K.-F., and Cha, M., "Detecting rumors from microblogs with recurrent neural networks", In Proceedings of the ACM International Joint Conference on Artificial Intelligence, pages 3818–3824, 2016.

[13] Fang, Y., Gao, J., Huang, C., Peng, H., and Wu, R., "Self multi-head attentionbased convolutional neural networks for fake news detection", PloS One, 2019, 14(9):e0222713.

[14] Kaliyar, Rohit; Goswami, Anurag; Narang, Pratik; Sinha, Soumendu, "FNDNet- A Deep Convolutional Neural Network for Fake News Detection", Cognitive Systems Research. 2020, vol. 61. Available from DOI: 10.1016/j.cogsys.2019.12.005.

[15] Hiramath, C. K.; Deshpande, G. C., "Fake News Detection Using Deep Learning Techniques", in: 2019 1st International Conference on Advances in Information Technology (ICAIT). 2019, pp. 411–415.

[16] Kaliyar, R. K., "Fake News Detection Using A Deep Neural Network", In: 2018 4th International Conference on Computing Communication and Automation (ICCCA). 2018, pp. 1–7.

[17] Reddy, Harita; Raj M, Namratha; Gala, Manali; Basava, Annappa, "Text-mining-based Fake News Detection Using Ensemble Methods", International Journal of Automation and Computing. 2020. Available from DOI: 10.1007/s11633-019-1216-5.

[18] POPOOLA, Olu, "Using Rhetorical Structure Theory for Detection of Fake Online Reviews", In: Proceedings of the 6th Workshop on Recent Advances in RST and Related Formalisms. Santiago de Compostela, Spain: Association for Computational Linguistics, 2017, pp. 58–63. Available from DOI: 10.18653/v1/W17-3608.

[19] BHUTANI, B.; RASTOGI, N.; SEHGAL, P.; PURWAR, A, "Fake News Detection Using Sentiment Analysis", In: 2019 Twelfth International Conference on Contemporary Computing (IC3). 2019, pp. 1–5.

[20] Rubin, Victoria; Conroy, Niall; Chen, Yimin; Cornwell, Sarah, "Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News", in: Proceedings of the Second Workshop on Computational Approaches to Deception Detection. San Diego, Cali fornia: Association for Computational Linguistics, 2016, pp. 7–17. Available from DOI: 10.18653/v1/W16-0802.

[21] Qian, Feng; Gong, Chengyue; Sharma, Karishma; Liu, Yan, "Neural User Response Generator: Fake News Detection with Collective User Intelligence", in: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18. International Joint Conferences on Artificial Intelligence Organization, 2018, pp. 3834–3840. Available from DOI: 10.24963/ijcai.2018/533.

[22] Ruchansky, Natali; Seo, Sungyong; Liu, Yan, "CSI: A Hybrid Deep Model for Fake News Detection", Proceedings of the 2017 ACM on Conference on Information and Knowledge Management - CIKM '17. 2017. ISBN 9781450349185. Available from DOI: 10.1145/3132847.3132877.

[23] Esteves, Diego; Reddy, Aniketh Janardhan; Chawla, Piyush; Lehmann, Jens, "Belittling the Source: Trustworthiness Indicators to Obfuscate Fake News on the Web", CoRR. 2018, vol. abs/1809.00494. Available from arXiv: 1809.00494.

[24] Zhou, Xinyi; Wu, Jindi; Zafarani, Reza, "SAFE: Similarity-Aware Multi-Modal Fake News Detection", 2020. Available from arXiv: 2003.04981.

[25] Zhou, Xinyi; Zafarani, Reza, "A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities", in ACM Computing Surveys. 2020. ISSN 1557-7341. Available from DOI: 10.1145/3395046.

[26] Ciampaglia, Giovanni Luca; Shiralkar, Prashant; Rocha, Luis M.; Bollen, Johan; Menczer, Filippo; Flammini, Alessandro, "Computational Fact Checking from Knowledge Networks", PLOS ONE. 2015, vol. 10, no. 6, pp. e0128193. ISSN 1932-6203. Available from DOI: 10.1371/journal.pone.0128193.

[27] Pan, Jeff; Pavlova, Siyana; Li, Chenxi; Li, Ningxi; Li, Yangmei; Liu, Jinshuo, "Content Based Fake News Detection Using Knowledge Graphs", in 17th International Semantic Web Conference, Monterey, CA, USA, October 8–12, 2018, Proceedings, Part I. In: 2018, pp. 669–683. ISBN 978-3-030-00670-9. Available from DOI: 10.1007/978-3-030-00671-6_39.

[28] Lin, Peng; Qi, Song; Shen, Jialiang; Wu, Yinghui, "Discovering Graph Patterns for Fact Checking in Knowledge Graphs", in book: Database Systems for Advanced Applications, 2018, pp. 783–801. ISBN 978-3-319-91451-0. Available from DOI: 10.1007/978-3-319-91452-7_50.

[29] Wang, Youze; Qian, Shengsheng; Hu, Jun; Fang, Quan; Xu, Changsheng, "Fake News Detection via Knowledge-Driven Multimodal Graph Convolutional Networks", In: Proceedings of the 2020 International Conference on Multimedia Retrieval. Dublin, Ireland: Association for Computing Machinery, 2020, pp. 540–547. ICMR '20. ISBN 9781450370875. Available from DOI: 10.1145/3372278.3390713.

[30] Agrawal, R., de Alfaro, L., Ballarin, G., Moret, S., Di Pierro, M., Tacchini, E., & Della Vedova, M. L., "Identifying Fake News from Twitter Sharing Data: A Large-Scale Study", 2019, ArXiv Preprint ArXiv:1902.07207. http://arxiv.org/abs/1902.07207.

[31] Ni, S., Li, J., & Kao, H.-Y., "MVAN: Multi-View Attention Networks for Fake News Detection on Social Media", IEEE Access, 2021, 9, 106907–106917. https://doi.org/10.1109/ACCESS.2021.3100245.

[32] Singhal, S., Shah, R. R., & Kumaraguru, P., "Factorization of Fact-Checks for Low Resource Indian Languages", 2021, arXiv preprint arXiv:2102.11276.

[33] Azer, M., Taha, M., Zayed, H. H., &Gadallah, M., "Credibility Detection on Twitter News Using Machine Learning Approach", in International Journal of Intelligent Systems & Applications, 2021, 13(3).

[34] Sahoo, S. R., & Gupta, B. B., "Multiple features based approach for automatic fake news Detection on social networks using deep learning", Applied Soft Computing, 2021, 100, 106983.

[35] Kaliyar, R. K., Goswami, A., & Narang, P., "FakeBERT: Fake news detection in Social media with a BERT-based deep learning approach", Multimedia tools and applications, 80(8), 2021, 11765-11788. 'doi: 10.1007/s11042-020-10183-2.

[36] Xie, J.; Liu, S.; Liu, R.; Zhang, Y.; Zhu, Y, "SERN: Stance Extraction and Reasoning Network for Fake News Detection", In Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual Conference, 6–12 June 2021; pp. 2520–2524.

[37] Zubiaga, A.; Liakata, M.; Procter, R., "Exploiting context for rumour detection in social media", In Proceedings of the International Conference on Social Informatics, Oxford, UK, 13–15 September 2017; pp. 109–123.

[38] Kang, Z.; Cao, Y.; Shang, Y.; Liang, T.; Tang, H.; Tong, L, "Fake News Detection with Heterogenous Deep Graph Convolutional Network", In Proceedings of the Advances in Knowledge Discovery and Data Mining, Virtual Event, 11–14 May 2021.

[39] Sungmin Aum and Seon Choe, srBERT: automatic article classification model for systematic review using BERT, in Systematic Reviews, (2021) 10:285, https://doi.org/10.1186/s13643-021-01763-w.

[40] Samin Mohammadi and Mathieu Chapon, Investigating the Performance of Fine-tuned Text Classification Models based on BERT, IEEE 22nd International Conference on High Performance Computing and Communications, 2020, 10.1109/HPCC-SmartCity-DSS50907.2020.00162.