EasyChair Preprint

№ 11811

Identification of Misinformation Using Word Embedding Technique Word2Vec, Machine Learning and Deep Learning Models

Arati Chabukswar, P Deepa Shenoy and K R Venugopal

January 19, 2024

Identification of Misinformation using Word Embedding Technique Word2Vec, Machine Learning and Deep Learning Models

Arati Chabukswar[1[0000-0003-3147-1385]], P Deepa Shenoy[2[0000-0002-4513-8949]], Venugopal K R [3[0000-0003-0113-1419]]

[1] Department of Computer Science & Eng, UVCE, Bangalore University, Bengaluru, India, arati.chabukswar0106@gmail.com
[2] Department of Computer Science & Eng, UVCE, Bangalore University, Bengaluru, India, shenoypd1@gmail.com
[3] Department of Computer Science & Eng, UVCE, Bangalore University, Bengaluru, India, venugopalkr@uvce.ac.in

Abstract. Real-time news is widely disseminated through the internet on a global scale. One of the factors contributing to its success is the simple and speedy spread of news. Social networking platforms have a huge user base that includes people of all ages, genders, and social backgrounds. Considering these positive aspects, a serious drawback is the propagation of misinformation, as most individuals read and spread information without giving any thought to its veracity. Researching techniques for news authenticity is so essential. To address this problem, a fake news identification system is created by training the COVID-19 tweets with roughly 12427 records taken from Kaggle and GitHub repository from three different sets, annotated manually as Fake (0) and Real (1) by cross-checking through websites that verify facts using machine learning classifiers like RF, SVM, LR, NB, and Deep Learning classifiers LSTM and Bi-LSTM. The feature extraction process makes use of the Word2Vec word embedding technique. According to the findings, Bi-LSTM performed better than all the other models in terms of accuracy, scoring 87.3%.

Keywords: COVID-19 Tweets, Deep Learning, Fake news, Machine Learning, NLP, Text Classification, Word2vec embedding technique.

1 Introduction

The term "fake news" describes information that is false or deceitful but presents as authentic. It is often shared on a range of media platforms, includes news portals, social media networks, and well-known news organizations. False information could exist in articles, images, videos, and audio recordings. It aims to deceive viewers or readers by mimicking reputable news sources or using enticing language to unique attention and garner clicks or views. Social networking websites are popular since it is convenient and simple to find and share content with others. However, the rapid dissemination of false information enables its widespread dissemination, particularly propaganda, which is harmful to the society at large and its inhabitants [1]. An illustration of the degree to which incorrect information: The International Fact-Checking Network (IFCN)

published a study that stated that the epidemic has spread. Over a hundred organizations that mostly conduct fact-checking are gathered under ICFN [2].

Table 1. List of Abbreviations

Abbreviations	**Definition**
NLP	Natural Language Processing
Word2Vec	Word to Vectors
ML	Machine Learning
DL	Deep Learning
SVM	Support Vector Machine
RF	Random Forest
NB	Naïve Bayes
LG	Logistic Regression
LSTM	Long-Short Term Memory
Bi-LSTM	Bidirectional LSTM

People now prefer to get and post news on social networking sites rather than through conventional media sources because of social media's increase in digitalization. For instance, 53% of American people will frequently or occasionally look used for information on the internet in 2020, up from 47% in 2018 and 44% in 2016. People can receive facts as quickly as possible due to the immediate nature of the internet. Still, a sizable portion of misinformation has evolved into a weapon towards swaying public sentiment at the expense of economic and political advantage. For instance, during the COVID-19 outbreak, bogus news generated a great deal of needless concern [3].

Detailed research has been done on a wide range of misinformation, including rumors, false news, clickbait content, and hoaxes, with real-world instances; an inventory of all misinformation propagators and vendors of services; a record of freely accessible data sets for false news in several dissimilar formats, including texts, images, and videos; key bibliometric indicators; and more [4]. As a result, it is possible to predict the type of information that will be spread, where it will occur, and what effects it. The authors of [5] conducted an extensive examination of false news and its sources, as well as its attributes, including news content, social context, makers, and their intentions; its news content, including language and semantic evaluation; knowledge-driven analysis (automatic or manual fact-checking); style-based analysis (text and image-based news); feature analysis for its identification; and the identification approaches, including third-party services, ML and DL methods, Geometric DL analysis. It provides a thorough route for future researchers to understand the source of information on fake news detection.

A novel framework called semantic graph attention-based representation learning was developed on the TALLIP fake news dataset in [6] to address the problem of recognizing bogus news in languages with limited resources and extracting conceptual and contextual information from documents in a specific multiple languages text corpus.

To identify false news stories, every article's presentation of the claim in the title was divided into one of the following four distinct groups: agrees, discusses, disagrees, and irrelevant. Investigators in [7] employed the FNC-1 (Fake News Challenge) dataset for misleading information identification and compared the results utilizing large-scale data technologies (spark) and machine learning. N-grams, Hashing TF-IDF, and count vectorizer are used for gathering information.

Due to low cost and ease of access, online social media networks (OSMs) are seeing an increase in the spread of incorrect information. It has a gravely detrimental effect on both the individual and society. Due to the rapid diffusion of information, identifying it might be difficult. Authors in [8][9] have worked on stance detection model and the fabricated content classifier using ML and DL models on different datasets.

Several embedding techniques, including one-hot encoding, Bag-Of-Words, TF-IDF, Word2vec, Glove, Fasttext, transformer based-BERT; ML, DL, Federated learning, Blockchain technology, etc., helps in detecting fake news using feature selection techniques on various types of information pertaining to political news, COVID-19 tweets, entertainment news, sports news, etc. Genetic and evolutionary feature selection (GEFeS), one of these feature selection techniques, is employed in the false news detection system in [10].

Synonymous and analogous associations within phrases are captured by Word2Vec embeddings. Negative sampling is a method used for training the Word2Vec models. In negative sampling, the target word is trained to be distinguished from randomly selected "negative" terms. This aids in the model's learning about how to distinguish among phrases which occur in related situations. Authors in [11] made use of Word2Vec, FastText, and BERT methods on Albanian language for identification of bogus news.

In this study, a false new detecting system that uses word embedding technology called Word2Vec in conjunction with ML models including SVM, LR, NB, RF, and DL models LSTM and Bi-LSTM using social media data pertaining to COVID-19 tweets is proposed. Refer Table 1. for list of abbreviations used throughout the paper.

Several major components make up this paper. The work summary for the suggested model is shown in Section II. The framework of the system and a description are provided in Section III. The analysis and findings of the investigation are in Section IV. The conclusion and discussion of the upcoming works are presented in Section V.

2 Related Work

Numerous academics have investigated this issue using a range of techniques to determine which strategy is most successful and yields the greatest results. From a data mining, NLP, ML, and DL standpoint, a few articles have looked at feature extraction and model construction as false news detection strategies. The authors of [12] developed a brand-new embedding method known as link2vec-based model, a modified version of word2vec, towards identifying instances of false information on two datasets in languages including English and Korean. In comparison to existing Text-based models

and Text+Whitelist-based models, the Link2vec-based identification approach outperformed them all, revealing a whole new approach to the identification of fake news.

Word embedding approaches such as BOW, n-grams, count-vectorizer, and TF-IDF are employed and skilled samples through five distinct machine learning (ML) classification methods in [13] but could not work with word2vec embedding technique.

Using bag-of-words and the word2vec embedding techniques, researchers in [14] have studied five distinct datasets that included text and postings on social networking sites such as Facebook and Twitter in three distinct languages, which are Germanic, Latin, and Slavic. The data was then trained on four machine learning classifiers.

The authors of [15] collaborated on both image and text data to determine the underlying semantic knowledge of published news article. They did this by employing the cosine similarity index (CSI) to forecast the news trustworthiness and by attaining an upper limit of greater than 0.62 for real news. They further trained the model using deep learning techniques and word2vec, which converts words to vectors.

In [16], an innovative framework known as UPFD (User Preference-aware false News Identification) was put forth to recognize false news by taking advantage of user preferences. Using feature set GCNFN (word2vec) as the graph (text) encoder, it jointly models content and graphs to capture many signals from user preferences in order to work on GNN-based fake news identification.

In reference [17], authors employed BERT-based machine learning with a Light gradient boosting (LGB) model to identify misleading information on a variety of machine learning and deep learning classifiers, including MNB (Multinominal Naïve Bayes), LSVM (Linear Support Vector Machine), LSTM using TF-IDF, Glove, and BERT-based word embedding techniques on three separate datasets: ISOT, TI-CNN, and FNC.

To eliminate data inequalities among classes, the research in [18] employs DL models like CNN, Bi-LSTM, and ResNet on a variety of word embedding approaches like Word2Vec, GloVe, and fastText. The models used undergo training on four distinct datasets through a procedure of data augmentation utilizing the the reverse translation approach. In comparison with other models, the Bi-LSTM model worked well.

Grid search and hyper optimization techniques are used to compare an optimized CNN model, called OPCNN-FAKE, with RNN, LSTM, and six ML classifiers to determine whether the news is genuine or not. DL models employ Glove word embedding for character representation after features are collected from the datasets using N-gram and TF-IDF [19].

A data modelling technique which vectors the triples using Word2vec and Glove; TF-IDF and Counter Vectorizer by presenting an approach which collectively gathers the triples with named entity tags is employed by the researchers in [20] to utilize Multi-layer Perceptron to classify whether the triple is real or fake.

A literature review was conducted by investigators in [21], with an emphasis on natural language processing (NLP) tasks that involve text preprocessing, which involves tokenization, stopword removal, and lemmatization; feature extraction using TF-IDF, BoW; word embedding techniques for word to vector conversion, such as Word2vec, BERT, FastText, and GloVe on publicly available online information sets;

and performance of models is evaluated through data training using a variety of machine learning (ML) and deep learning (DL) models.

The authors of [22] have suggested work on Deep Learning models CNN and Bi-LSTM with LIAR and Kaggle datasets on NLP for false news identification using One-hot encoding, TF-IDF, Word2Vec, and Doc2Vec word embedding approaches. On the Kaggle dataset, Bi-LSTM with Doc2vec performed well together.

By using a meta-heuristic algorithm to choose the features and training a Deep Neural Network, the researchers of [23] present OptNet-Fake, a unique method for detecting fake news on social media. The Modified Grasshopper Optimization technique is utilized to extract the d-D feature vectors using TF-IDF. At last, the vectors are processed and supplied into the CNN framework for training with various filter sizes to extract the n-gram characteristics of the text.

Background Work

In [13] authors have proposed the fake news detection system using BOW, n-grams, count vectorizer, TF-IDF and trained the model on five different classifiers such as NB, LR, SVM, RF, Stochastic Gradient classifier considering Precision, Recall, F1-score performance metrics. As a future work, Word2Vec word embedding technique will be included. Authors in [24] used Keras neural network model trained with N-gram; Bi-LSTM; Tensorflow framework, achieving an overall accuracy of 84% which is comparatively less than the proposed work.

3 Proposed Methodology

The architectural diagram of proposed model is as shown in Fig. 1. Dataset for fake news identification is collected from different Kaggle and GitHub repository related to COVID-19 tweets, annotated as 0 for fake and 1 for real manually by checking it through fact checking websites (few) and then checked for NaN, na values, and missing values (if any). These values are handled by importing pandas Dataframe and employing mean on the obtained dataset within fillna function to avoid losing important tweets. After data preparation process, text preprocessing is required, because unwanted words present in the text will create noise and reduces the performance of the model. So, it is necessary to pre-process the data before feeding it to the model. Text-preprocessing is done by importing NLTK toolkit which is an open-source NLP package which is used in converting text into smallcase letters, tokenization process, stop words removal and stemming/lemmatization process. Before feeding the text directly to the model, text must be converted in the form of vectors (numbers), which is done using Word2Vec word embedding technique to convert words into vectors in the form of matrix. It is widely used embedding technique, adopts importance of many terms used in the document by importing Gensim model.

The two versions of word2vec CBOW and Skip-Gram; the former version aims to predict a target word based on its context and later one takes a target word and tries to predict the context words surrounding it. Next process is feeding vectors to the model

for training and testing process. ML Models such as SVM, RF, NB, Logistic regression, and DL models such as LSTM and Bi-LSTM are used for training and testing the data.

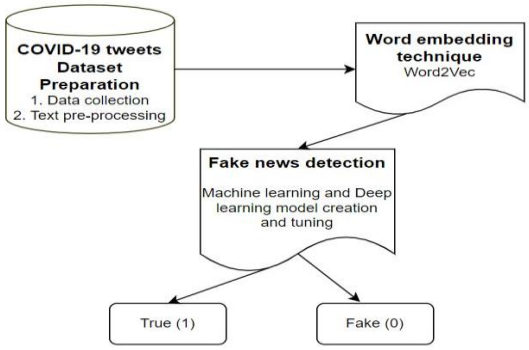

Fig. 1. Architecture of the proposed model

3.1 Dataset description

Fake news dataset is prepared by combining datasets from three different sources namely fakerealcovid-19, CoAID-master, COVID-19 twetter dataset which are obtained from public repository Kaggle, GitHub related to COVID-19 tweets, for two reasons. Firstly, they share a framework that consists of two categories: real news and fake news. Secondly, the outliers and constraints of each separate data collection are lessened when the data sets are combined. Few more records were generated using Gretel synthetic data generator tool by providing sample records. Three features are considered namely Tweet id, Tweet, Label from the dataset having total records 12427, out of which 6922 records are Fake and 5505 records are True. According to the four categories, including fake and true news, an even split of the available data is taken into consideration: The number of brief sentences that depict fake news is higher than that of true news, fake news content's comprehension is not as good as that of real news, fake news stories have a higher subjectivity than legitimate news articles, records of real news is more than false information records.

4 Results and Discussions

An intensive experiment was carried on Jupyter Notebook, Python 3.10 using data from GitHub and Kaggle to ensure the model's efficacy. Randomly, 80% of the dataset is divided into training groups and 20% into testing groups. After achieving the utmost level of accuracy, the results were recorded.

4.1 Data Collection and Cleaning

The dataset for identifying fake news is gathered from Kaggle and GitHub repositories connected to COVID-19 tweets, manually marked as 0 for fake and 1 for true by checking it through fact-checking websites (few), and then checked for NaN, na values, and missing values (if any). In order to prevent losing significant tweets, these values are handled by importing a panda Dataframe and using mean on the dataset that is produced inside of the fillna function. There are totally 12427, out of which 6922 records are Fake and 5505 records are True having the fields tweet id, tweet, label as 0 for fake and 1 for real.

4.2 Data Visualization

The visualization of data helps to understand the dataset's structure. As illustrated in Fig. 2, data is visualized using WordCloud images for both bogus and legitimate news.

Fig. 2. WordCloud for Fake News

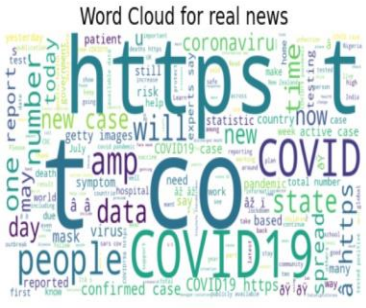

Fig. 3. WordCloud for Real News

4.3 Feature Extraction

It converts unprocessed information into a format that algorithms can comprehend and successfully analyze, making it a crucial step to be considered in machine learning or deep learning, one such reason is Numerical representation of Words using a word embedding technique Word2Vec. Gensim package is imported to use Word2Vec model and a sample of Words to vector conversion is done as shown in Fig. 4. by considering the word ‘coronavirus’ and the most similar words for the word ‘twitter’ in the matrix form.

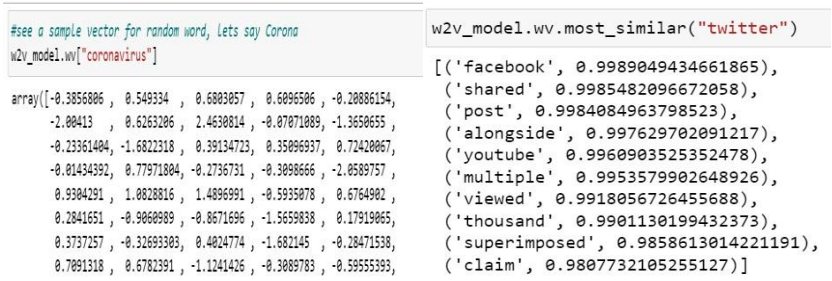

Fig. 4. Sample of Words to vector conversion for the word ‘coronavirus’ and ‘twitter’

4.4 Model selection

Based on provided datasets, machine learning models, including SVM, Gaussian NB, RF, and LR, are chosen after considering their limitations. Selecting tweet content as one of the features in feature engineering is done because it provides information about news, whether it is fake or real, also no data imbalances between fake and actual tweets, which might lead to biased model efficiency with a classifier favoring the majority class. When choosing deep learning models, like LSTM and Bi-LSTM, consideration is given to the model's limits based on prepared datasets. These datasets are chosen because they are not prone to overfitting, have acceptable data quality, and usage of sufficient amount of processing resources. A few more procedures are considered to solve these limits on the supplied dataset, including handling missing values, fixing class imbalance, and preprocessing and cleaning the data.

4.5 Performance Measure

Results for Accuracy, Precision, Recall, and F1score were computed to assess the effectiveness of proposed model. These metrics are the accepted performance measures for categorization issues. Confusion matrix compares predicted labels to actual labels in a table format to indicate how well a classification model performed

Table 2. Confusion matrix

Predicted/Actual	**Predicted Positive**	**Predicted Negative**
Actual Positive	True Positive	False Negative
Actual Negative	False Positive	True Negative

a=True Positive, b=True Negative, c=False Positive d=False Negative. Accuracy, Precision, Recall, F1-Score are calculated as follows

$$Accuracy=(a+b)/(a+b+c+d) \quad (1)$$

$$Precision=a/(a+b) \quad (2)$$

$$Recall=a/(a+d) \quad (3)$$

$$F1\text{-}Score=2*(Precsion*Recall)/(Precsion+Recall) \quad (4)$$

The performance metric outperforms other earlier studies on misinformation detection. The outcomes of the proposed framework are contrasted with those from other earlier researchers' framework in Table 3. where Precision, Recall and F1-score of Proposed RF is 87%,86%,86% outperformed as compared to RF of existing work. Table 4. details how well the models performed on the publicly accessible datasets with respective to Precision, Recall, F1-score achieving highest score with Bi-LSTM on fake and real news records.

Table 3. Comparative Analysis of Proposed Model with Existing Works

Reference	Model	Precision	Recall	F1-Score
[13]	SVM	0.62	0.62	0.61
	RF	0.60	0.60	0.59
Proposed	SVM	0.82	0.81	0.81
	RF	0.87	0.86	0.86

Table 4. Performance Metrics

Model	Label	Precision	Recall	F1-Score
SVM	Fake	0.81	0.86	0.83
	Real	0.82	0.76	0.79
RF	Fake	0.86	0.91	0.88
	Real	0.88	0.82	0.85
NB	Fake	0.71	0.80	0.75
	Real	0.71	0.62	0.66
LR	Fake	0.81	0.85	0.83
	Real	0.81	0.76	0.79
LSTM	Fake	0.86	0.89	0.87
	Real	0.86	0.82	0.84
Bi-LSTM	Fake	0.86	0.92	0.89
	Real	0.88	0.83	0.86

On the COVID-19 Twitter dataset, the Bi-LSTM model scored 87.3% accuracy. Fig.5 shows the comparison of ML and DL models on fake and real COVID -19 tweets data using performance metrics Accuracy.

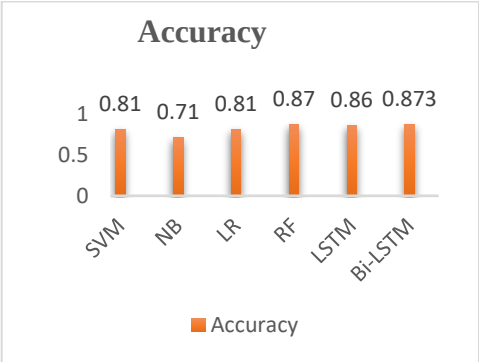

Fig. 5. Accuracy comparision

5 Conclusion and Future Scope

The goal of the proposed effort is to use three different sets from the Kaggle and GitHub repositories to categories whether the COVID-19 tweets posted in social media are fake or real. To increase the performance of the models, data pre-processing is done via tokenizing, deleting stopwords, and lemmatization. Word2Vec word embedding technology is used to convert words to vectors in a way that captures the semantic linkages and contextual meanings of the words. The data is then trained using DL classifiers like LSTM and Bi-LSTM as well as ML classifiers like SVM, RF, NB, and LR. Among all the models, Bi-LSTM model provides better results in terms of accuracy.

Other word embedding techniques, such as Glove, FastText, and Transformer-based BERT for fake news detection, could be used to further the work.

References

1. Verma, Pawan Kumar, Prateek Agrawal, Ivone Amorim, and Radu Prodan "WELFake: word embedding over linguistic features for fake news detection." IEEE Transactions on Computational Social Systems, VOL. 8, NO. 4, AUGUST 2021, pp: 881-893.
2. Al-Rakhami, Mabrook S., and Atif M. Al-Amri. "Lies kill, facts save: Detecting COVID-19 misinformation in twitter." IEEE Access, VOLUME 8, pp: 155961-155970.
3. Che, Hangjun, Baicheng Pan, Man-Fai Leung, Yuting Cao, and Zheng Yan, "Tensor Factorization with Sparse and Graph Regularization for Fake News Detection on Social Networks." IEEE Transactions on Computational Social Systems 2023.

4. Rastogi, Shubhangi, and Divya Bansal. "A review on fake news detection 3T's: Typology, time of detection, taxonomies." International Journal of Information Security 22.1 (2023), pp: 177-212.
5. Kondamudi, Medeswara Rao, Somya Ranjan Sahoo, Somya Ranjan Sahoo, Nandakishor Yadav, "A comprehensive survey of fake news in social networks: Attributes, features, and detection approaches." Journal of King Saud University-Computer and Information Sciences 35.6 (2023), pp: 101571
6. Mohawesh, Rami, Xiao Liu, Hilya Mudrika Arini, Yutao Wu, Hui Yin, "Semantic graph-based topic modelling framework for multilingual fake news detection." AI Open 4 (2023) Elsevier, pp:33–41.
7. Altheneyan, Alaa, and Aseel Alhadlaq. "Big data ML-based fake news detection using distributed learning." IEEE Access 11 (2023), pp: 29447-29463.
8. Jose, Xavier, SD Madhu Kumar, and Priya Chandran. "Characterization, classification and detection of fake news in online social media networks." 2021 IEEE Mysore Sub Section International Conference (MysuruCon). IEEE, 2021, pp: 759-765.
9. Gupta, Archit, Arnav Batla, Chaitanya Kumar, Dr. Goonjan Jain, "Comparative Analysis of Machine Learning Models for Fake News Classification." 2023 3rd International Conference on Intelligent Technologies (CONIT). IEEE, 2023.
10. Tian, Ziyan, and Sanjeev Baskiyar. "Fake News Detection using Machine Learning with Feature Selection." 2021 6th International Conference on Computing, Communication and Security (ICCCS). IEEE, 2021.
11. Ercan Canhasi, Rexhep Shijaku, and Erblin Berisha. 2022. Albanian Fake News Detection. ACM Trans. Asian Low-Resour. Lang. Inf. Process. 21, 5, Article 86 (November 2022), pp:1-24, https://doi.org/10.1145/3487288.
12. Shim, Jae-Seung, Yunju Lee, and Hyunchul Ahn. "A link2vec-based fake news detection model using web search results." Expert Systems with Applications 184 (2021), Elsevier, pp: 115491.
13. VasuAgarwal, H.ParveenSultana, SrijanMalhotra, AmitrajitSarkar, "Analysis of classifiers for fake news detection.", International Conference on Recent Trends in Advanced Computing 2019, Procedia Computer Science 165 (2019), Elsevier, pp: 377-383.
14. Faustini, Pedro Henrique Arruda, and Thiago Ferreira Covoes. "Fake news detection in multiple platforms and languages." Expert Systems with Applications 158 (2020), Elsevier, pp: 113503.
15. Mangal Deepak, and Dilip Kumar Sharma. "Fake news detection with integration of embedded text cues and image features." 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Amity University, Noida, India, IEEE, 2020.
16. Yingtong Dou, Kai Shu, Congying Xia, Philip S. Yu, Lichao Sun. 2021, "User Preference-aware Fake News Detection", In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21), July 11–15, 2021, Virtual Event, Canada. ACM, New York, NY, USA, https://doi.org/10.1145/3404835.3462990.
17. Essa, Ehab, Karima Omar, and Ali Alqahtani. "Fake news detection based on a hybrid BERT and LightGBM models", Complex & Intelligent Systems (2023), pp: 1-12.

18. Sastrawan, I. Kadek, I. P. A. Bayupati, and Dewa Made Sri Arsa. "Detection of fake news using deep learning CNN–RNN based methods." ICT Express 8.3 (2022), pp: 396-408.
19. Saleh Hager, Abdullah Alharbi, and Saeed Hamood Alsamhi. "OPCNN-FAKE: Optimized Convolutional Neural Network for Fake News Detection." IEEE Access 9 (2021), pp: 129471-129489.
20. Thilagam, P. Santhi. "Multi-layer perceptron based fake news classification using knowledge base triples.", Applied Intelligence 53.6 (2023): 6276-6287.
21. Sharma, Upasna, and Jaswinder Singh. "Review of Feature Extraction Techniques for Fake News Detection." Advances in Information Communication Technology and Computing: Proceedings of AICTC 2022. Singapore: Springer Nature Singapore, 2023, pp: 389-399.
22. Singh, Lovedeep. "Fake News Detection: a comparison between available Deep Learning techniques in vector space." 2020 IEEE 4th Conference on Information & Communication Technology (CICT). IEEE, 2020.
23. Sanjay Kumar, Akshi Kumar, Abhishek Mallik, and Rishi Ranjan Singh, "OptNet-Fake: Fake News Detection in Socio-Cyber Platforms Using Grasshopper Optimization and Deep Neural Network." IEEE Transactions on Computational Social Systems (2023).
24. Islam, Taminul, MD Alamin Hosen, Akhi Mony, MD Touhid Hasan, Israt Jahan, and Arindom Kundu. "A proposed Bi-LSTM method to fake news detection." In 2022 International Conference for Advancement in Technology (ICONAT), pp. 1-5. IEEE, 2022.