Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 08 May 2026
Revised 19 June 2026
Accepted 03 July 2026

Natural Resources for Human Health 2026, Vol. 6 (11s): 1408–1423

KEYWORDS:
Natural Language Processing, BERT, RoBERTa, Machine learning, Deep learning, Text classification, Fake News Detection and Misinformation detection.

Health Misinformation Detection Using BERT and RoBERTa: A Comparative Study of Machine Learning and Deep Learning Approaches

V. Karunakaran [1]*, Rajat Goel *[2], Rajasekar Velswamy 3, Jeyakrishna Venugopal *4, Sreedevi.V 5

[1]Department of CSE (Artificial Intelligence and Machine Learning), Sri Eshwar College of Engineering, Coimbatore
[2,4,5] Manipal University Jaipur, Jaipur, India
[3]Department of Computer Science and Engineering, SRM institute of Science and Technology, Vadapalani Campus, Chennai
*Corresponding Authors: rajat.goel@jaipur.manipal.edu , karunasel@gmail.com, Jeyakrishnan.v@jaipur.manipal.edu

ABSTRACT: This expansion of fake news over the internet platforms is currently a significant challenge in the digital age that has led to misinformation in the political arena, society, and in the health sector. Deep learning and machine learning methods of automatic detection of fake news have limited research interest over the past few years. In this research paper, a comparative analysis of six classification models is presented to identify a fake news and they are Logistic Regression, Random Forest Tree, Support Vector Machine (SVM), Bidirectional Long Short-Term Memory (BiLSTM), Robustly Optimized BERT Pretraining Approach (RoBERTa) and Bidirectional Encoder Representations from Transformers (BERT). The Fake news and Real news Dataset (44, 898 news articles) is experimented on. Preprocessing functions are applied to the dataset thus lowering the alphabets, stop words removal, punctuation removal and the URL filtering; after which the classification process is done using the dataset. All the six models are trained and tested in the same experimental conditions and with performance measures of accuracy, recall, F1-score and precision. The experimental results indicate that the BERT model achieves the highest classification rate of 99.95 percent and precision, F1-score and recall of 1.00, which is higher than other models. RoBERTa and Random Forest score 99.80 then SVM at 99.52, then BiLSTM at 99.00 and lastly, Logistic Regression at 98.95. The comparative analysis reveals that the transformer models are superior to machine learning and deep learning models in detecting fake news, and can be applied in future investigation of detecting misinformation.

1. INTRODUCTION

In the digital world age, internet and social networking sites have highly transformed the way information is acquired and disseminated all over the world. These sites have not only ensured that news and information reaches people in a very short span of time, but they have also become a readily available tool to spread misinformation. Fake news refers to false or misleading news which is deliberately made and propagated to mislead people and shape their perception and make the society perplexed [1]. The propagation of fake news will lead to radical impacts within various industries including politics, healthcare and social cohesion.The invention of social media, such as Instagram, Facebook, YouTube and Twitter, have contributed a lot to spreading misinformation.. Research has revealed that counterfeit news disseminates six times quicker than honest news in social media

outlets [2]. The rapid spread of fake news in case of emergencies such as elections, pandemics, and natural disasters has also caused panic in people, influenced their voting behaviour and eroded the credibility of official news. An example of the unparalleled increase in misinformation in health that happened during the COVID-19 pandemic is the WHO declaration of the infodemic and the pandemic. [3].

Detecting fake news manually is a very difficult one and a time-consuming process because of the huge/large amount of content produced on online platforms on a secondly basis. Therefore, deep learning and Machine learning based automatic fake news detectors have become a significant research topic during the past few years. The most popular classic machine learning models such as the Logistic Regression, Random Forest Tree and Support Vector Machine (SVM) have been widely implemented to the text classification problem of detecting fake news [4]. These models rely on the manually developed features such as TF-IDF representation to classify news pieces as true or fake. Particularly, recurrent neural network-based models such as the Long Short-Term Memory (LSTM) and Bidirectional Long Short Term Memory (BiLSTM) have been proved to be more realistic when it comes to tasks related to Natural Language Processing (due to their capability of capturing sequential relationships among the textual data [5]. Models based on transformers (such as Bidirectional Encoder Representations from Transformers (BERT) and Robustly Optimized BERT Pretraining Approach (RoBERTa)) have more recently achieved state-of-the-art performance in various natural language processing tasks, such as text classification, sentiment analysis, and fake news detection [6].

These models rely on self attention processes and the already existing language representations to get the contextual meaning of words in a sentence and thus are highly applicable in fake news detection. Although multiple approaches to fake news detection exist using deep learning, machine learning, and transformer-based models, little literature exists to compare all of the model types under the same experiment. Both the traditional machine learning models and the deep learning models are explained in the existing literature, but not in detail in comparison with both the models. To resolve this drawback, this research paper has examined six different classification models: Logistic Regression, SVM, Random Forest, BiLSTM, BERT and RoBERTa to classify fake news using the Real and Fake News Dataset comprising 44,898 news articles..

The primary outputs of this paper are the following: • A comparative study of six models of classification covering conventional Machine Learning (ML), Deep Learning (DL), and Transformer-based models to detect fake news.

- A strong text preprocessing pipeline that includes the removal of stopwords, punctuations, URLs filtering, and lowercasing to improve the performance of the model.
- Extensive experimental analysis built on traditional measures of performance such as accuracy, F1-score, recall, and precision with the same experimental conditions.
- Empirical evidence that the fine-tuned BERT model is the most accurate in classification with 99.95% accuracy, it is higher than any other models in detecting fake news.

The rest of this paper will be structured as follows. Section 2 provides the corresponding literature on fake news detection. The dataset and preprocessing are discussed in Section 3. The methodology and models are described in section 4. The Section 5 includes the discussion and results of the experiment. In Section 6, future research direction is provided at the end of the paper.

2. LITERATURE REVIEW

The threat of fake news has been increasing and has brought a lot of research interest in various fields. The spectrum of methods of automatic fake news detection that have been investigated by researchers includes traditional machine learning approaches, deep learning, transformer-based methods, and hybrid architectures. In this section, the literature available is reviewed in four broad categories: (1) foundational and survey studies, (2) traditional machine learning methods, (3) deep learning methods, and (4) transformer-based and advanced methods.

2.1 Foundational & Survey Studies

The first survey on automatic detection of fake news, by Conroy et al., proposed a typology of veracity assessment methods that arise as a result of two broad categories: linguistic cues approaches, which involve machine learning, and network analysis approaches. The authors identified the potential of a hybrid method that incorporates linguistic signals along with network-

based behavioural information and suggested practical principles of developing a workable fake news detection framework. Their contribution focused on the idea that linguistic processing ought to go across different levels such as lexical processing, up to discourse-level processing in order to achieve the best performance [7].

D'ulizia et al. reviewed twenty-seven popular fake news detectors datasets systematically, giving a detailed description of each dataset in eleven characteristics that include news domain, use purpose, language, size, and media platform. The review has shown that the available data is rather limited and most of them are in English and are centred on political news. The authors detected the main issues of creating high-quality benchmark datasets and requested the creation of multimedia, multi-lingual, and large-scale fake news datasets, especially regarding COVID-19 misinformation [8].

Alghamdi et al. have provided a comprehensive survey of the machine learning and deep learning approaches that can be used to detect fake news. This review includes content, context and hybrid feature sets. This study has discussed the advanced transformer-based models and their efficiency in fake news describing; the issues of rapid evolution of fake news generation techniques, and a need for advanced robust and efficient detection mechanisms have also been identified. From this study, it can be seen that transformer based model (BERT and its variants) performed best as compared to conventional machine learning and deep learning model. [9].

2.2 Conventional Machine Learning Methods.

Perez-Rosas et al. introduced two new fake news detection datasets (with seven news areas, respectively), and learning experiments to build accurate fake news detectors using linguistic features. The authors were able to explore the differences between the language used in the fake news content and legitimate news content, and they demonstrated how computational linguistics can be applied to automatically identify fake news with high accuracy of up to 76%. The study reached a conclusion that future research should include linguistic features, meta-features, in addition to multimodality, in order to improve the detection performance [10].

The detection tool suggested by Al Asaad and Erascu is based on supervised machine learning techniques and representation of text using Bag-of-Words, TF-IDF, and Bi-gram Frequency. Probabilistic classifiers and linear classifiers were used to classify the type of news based on the title and content, and it was concluded that the linear classification using the TF-IDF technique was most effective in the classification of news content. This research emphasizes that simple feature extraction methods cannot be used and data mining methods should be employed in future. [11].

Khanam et al. studied the applications of traditional machine learning models in the classification of fake news using the scikit-learn library of Python and investigated the tokenization, feature extraction, feature selection approaches to optimize classification accuracy. The authors suggested adding part-of-speech textual analysis as a supplementary quantitative characteristic to the already available methods and suggested using Random Forest to enhance accuracy. The research established that the majority of previous studies that applied Naive Bayes algorithms had an accuracy of between 70 and 76 per cent, indicating that further developed methods are required [12].

Shaikh and Patil applied SVM, Passive Aggressive and Naïve Bayes Classifier along with TDIF feature extraction technique for the fake news detection system. The experimental findings also showed that using TF-IDF and SVM as a classifier with an accuracy of 95.05 is better than using Naive Bayes and Passive Aggressive methods. The research reaffirmed that SVM with TF-IDF is a well-performing baseline model in the task of text-based fake news classification [13].

Ahmad et al. suggested a ML ensemble method of automatic classification of news articles as misinformation or disinformation. The authors tested various textual features to differentiate fake content and real content and trained sets of machine learning algorithms with multiple ensemble methods on four data sets that were real. The experimental assessment helped verify that ensemble learner strategy is more efficient than single classifiers, which proves the viability of multiple model integration to enhance detection rates [14].

Reis et al. have performed an exploratory analysis with hundreds of thousands of models with a large variety of features in order to find out the true discriminatory value of the features for detecting fake news. It has been found that only a very small share of models were able to achieve the level of detection performance equal to 0.85 AUC, which means that the task is quite challenging indeed. The researchers have managed to discover that certain kinds of fake news tend to be associated with some specific feature combinations used in models, which means that modeling can become a fruitful area for further studies. [15].

2.3 Deep Learning Approaches

Thota et al. presented an approach for detection of fake news with the help of neural network models to detect the stance of two headlines and article body. Proposed TF-IDF Dense Neural Network model showed 94.21% accuracy on the test data, which was 2.5% more than the previously existing models. It was shown that the latent space representation technique based on Word2Vec always yielded less accurate results compared to plain TF-IDF or Bag-of-Words approaches, particularly on long news articles, therefore some other feature engineering techniques should be used for social media datasets [16].

Kaliyar et al. proposed deep Convolutional Neural Network model, named FNDNet, that could discriminate real from fake news automatically using hidden layers and without any need to craft features manually. The experiment conducted with this deep neural network model showed high accuracy results on benchmark dataset: this classifier provided state-of-the-art accuracy equal to 98.36 on the test data according to several criteria like precision, recall, F1 score, Wilcoxon statistic etc. According to the findings, the deep models based on CNN performed considerably better than other models [17].

Bahad et al. suggested a fake news detector framework based on Bidirectional LSTM-Recurrent Neural Network with the GloVe word embeddings to quantify the correlation between news article headline and body text. The two publicly available datasets experimental data proved the BiLSTM model to be the best choice in fake news detection as compared to CNN, vanilla RNN, and unidirectional LSTM models. The authors concluded that BiLSTM-RNN is considerably more efficient than unidirectional models, and that adaptive learning rate algorithm contributes considerably to the vanishing gradient issue in sequential models [18].

By employing 1,356 fake news events collected from Twitter and PolitiFact, Kumar et al. studied different deep learning techniques to detect fake news. In their research paper, they used CNNs, LSTMs, ensembles, and attention mechanisms for comparison and established that using a combination of CNN and Bidirectional LSTM with attention mechanism was the best approach for detecting fake news, which achieved an accuracy of 88.78, compared to other baseline models. Their findings indicated that CNN combined with LSTM and the attention mechanism could be useful in fake news detection [19].

In the proposed fake news detection system of this paper, different classifiers including logistic regression, naive Bayes, support vector machine (SVM), random forest and deep neural network (DNN) were compared. The authors used a combination of feature extraction technique and the effectiveness of this technique when combined with the classifier was tested with experimental analysis. The results of the experiment proved the importance of selecting the proper combination of classification algorithm and feature representation for fake news detection [20].

In the fake news classification, Nasir et al. proposed a hybrid deep learning model, which is based on Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). The model was tested on two synthetic news collections (ISOT and FA-KES) and the results of detection are significantly better than non-hybrid baseline models. Generalization on different data sets was also attempted by the authors who have reported promising results and added that complex architectures such as CNN and RNN are superior to single models to identify fake news [21].

The authors Monti et al. created a new automatic fake news detector model, which is based on a geometric deep learning algorithm, through which the graph convolutional neural networks are used to construct a heterogeneous information set, including news text, user profile, social network, and propagation behavior. This model was trained and tested using Twitter news articles that had been fact-researched by professional fact-checking organizations and had a very high ROC AUC of 92.7%. It has been demonstrated in the paper that fake news is reliable to be detected early on when it has already spread over a few hours and that propagation based approaches can be thought of as an alternative or as a complement to content based approaches [22].

Yuan et al. believed that the existing literature and data were lacking for detection of fake news in new news domains and they proposed Domain-Adversarial and Graph-Attention Neural Network (DAGA-NN) model for the same purpose. Extensive experiments conducted on large scale Twitter and Weibo multimedia data sets yielded that the proposed model was highly effective in the task of fake news detection on events and domain samples even if limited training data on the domain samples is provided. The article has highlighted the shortcomings of the traditional machine learning in cross-domain fake news detection and proposed a solution to it, which has potential in real-life implementation [23].

2.4 Advanced and Transformer-Based

Kaliyar et al. suggested FakeBERT, a BERT-based deep learning model, integrating various parallel blocks of single-layer deep Convolutional Neural Networks with BERT to address the ambiguity of natural language understanding. The suggested model exploited the two-way training potential of the BERT to extract semantic and long-distance relationship within the sentences, with a classification accuracy of 98.90 percent in the fake news detection tasks. The findings showed that FakeBERT is more effective than current models and showed that a combination of transformer-based language representations with convolutional architectures is highly beneficial in detecting fake news [24].

Raza and Ding suggested a fake news detection system based on transformers that uses the data in news, as well as in the social context. The proposed model utilized a Transformer architecture that had an encoder to learn useful representations and a decoder to predict future behavior using past observations. The model also included a successful labeling method to overcome the problem of label shortage and proved that the model can find fake news with increased accuracy several minutes after spreading, which is especially helpful in the case of early detection [25].

Segura-Bedmar and Alonso-Bartolome have done a fine-grained categorisation of fake news in both unimodal and multimodal in the Fakeddit dataset. Experiments implemented showed that the multimodal model, based on CNN architecture, which consists of the text and image data, achieved the best scores and BERT was the most successful model when using unimodal models with text data only and was able to achieve an accuracy of 78%. The study revealed that the combination of text and image information significantly boosts the fake news detection performance, and the use of visual information significantly boosts the performance for manipulated content and satire types of fake news [26].

2.5 Summary and Research Gap

The literature survey has revealed that fake news detection has been addressed in various ways ranging from classical machine learning models to deep learning models, and transformer based models. The first methods used were manual feature based methods such as TF-IDF, bag-of-words with classifiers such as Logistic Regression, SVM and Naive Bayes which gave a moderate accuracy. Afterwards, deep learning networks such as CNN, LSTM, and BiLSTM have shown a higher performance as they learn discriminatory features of text data on their own. More recently, models based on transformers like BERT have provided state-of-the-art performance using pre-trained contextual representations of language. However, most of the existing literature compare either of the traditional machine learning or the deep learning models and no thorough and comprehensive study of all the model types under same experimental conditions is performed. To fill this gap, the current study presents a comparative analysis of six classification models including LR, SVM, RF, BiLSTM, BERT and RoBERTa, under the same experimental conditions and the same data set, and gives trustworthy benchmarking information for researchers and practitioners in the field of fake news detection.

3. DATASET DESCRIPTION

3.1 Dataset Overview

The dataset considered in this paper is the Fake and Real News Dataset, which is openly accessible on the Kaggle repository. The initial set was compiled and maintained by Cllement Bisaillon and has been extensively utilized in fake news detection studies. It is made up of two separate CSV files: one containing fake news articles (Fake.csv) and one containing real news articles (True.csv). The 44,898 news articles (with 23,481 fake news articles and 21,417 real news articles) give rise to a dataset almost equally distributed among the different classes for binary classification tasks.

3.2 Dataset Attributes

The news articles dataset contains four attributes in each news article: title, text, subject and date. The title attribute is a reflection of the news article's heading, and the text attribute is a reflection of the body of the news article. The subject attribute describes the topic category of news article and the date attribute contains the date of publishing of news article. To collect the actual news articles, the articles on Reuters.com, a credible and reputable news source was collected, while the fake news articles were collected on several unreliable websites identified as such by Politifact, a fact checking organization in the United States, and Wikipedia. The sample includes news stories published in 2015-2018, with the majority of the news on the topic of political and world news.

3.3 Dataset Statistics

Table 1 presents a detailed summary of the dataset statistics used in this study.

Table 1: Dataset Statistics

Attribute	Details
Dataset Name	Fake and Real News Dataset
Source	Kaggle Repository
Total Articles	44,898
Fake News Articles	23,481 (52.3%)
Real News Articles	21,417 (47.7%)
Number of Attributes	4 (title, text, subject, date)
Real News Source	Reuters.com
Fake News Source	Politifact & Wikipedia flagged sites
Date Range	2015 – 2018
Primary Topics	Politics & World News
File Format	CSV

3.4 Data Preprocessing

The raw text data of the dataset is highly noisy and can negatively impact the performance of the classification models. Thus, a set of text preprocessing actions were performed to clean and standardize the text data and then extract features and train models. The steps involved in the preprocessing pipeline were the following:

Step 1: Text Combination: The title and text attributes of every news article were joined in one content field to offer a more resourceful textual expression to be categorized. This mixture makes sure that the headline and the body of the article involve in the decision of the classification.

Step 2: Lowercasing: All the text has been made lowercase in order to have consistency and to avoid the model taking the same word in various cases as separate tokens. As an example, lowercasing of words like Trump and trump is considered the same token.

Step 3: URL Removal: Regular expressions were used to remove the hyperlinks and URLs contained within the text because they lack any meaningful semantic value in terms of classification.

Step 4: Removal of Punctuation: Punctuation marks such as commas, full stop, exclamation marks and special characters were eliminated in the text in order to minimise noise and simplify the vocabulary.

Step 5: Number Removal: Numbers and alpha numerical tokens were eliminated in the text because they do not typically represent important discriminating information to classify fake news.

Step 6: Removal of Stop words: A stop words list of the Natural Language Toolkit (NLTK) was utilized to remove common English stop words, including the, is, at, which, and on. Stop words are the commonly appearing words that do not have much semantic meaning and their elimination assists the model to concentrate on words that are more significant and discriminative.

Step 7: Whitespace Normalization: The whitespace characters and newline characters were eliminated and the text was normalized to single spaces between words. The cleaned content field was then the input to the feature extraction and model

training after all the preprocessing steps were applied. Table 2 shows a sample comparison of text prior to and after preprocessing.

Table 2: Sample Text Before and After Preprocessing

	Sample Text
Before	"Modi FAKE News!! Visit https://example.com for more. 123 people said this is the best!!"
After	"Modi's fake news people said best"

3.5 Data Splitting

A conventional 80:20 split ratio was used to split the preprocessed dataset into training and testing sets. In this regard, 35,918 articles were applied in training the classification models and 8,980 articles were left to test and assess the performance of the models. The division of data was done in a fixed random state of 42 to make results same in all the experiments. Table 3 shows the data splitting statistics.

Table 3: Data Splitting Statistics

Split	Articles	Percentage
Training Set	35,918	80%
Testing Set	8,980	20%
Total	44,898	100%

4. METHODOLOGY

4.1 System Overview

This paper proposes a comparative model of automatic fake news detection, which compares six models of traditional machine learning, deep learning and transformer-based methods. The overall research system is illustrated in Figure 1, which contains seven major steps, namely, data collection, text preprocessing, feature extraction, train/test splitting, model training, evaluation, and result comparison. In the first stage, the articles are collected in the Kaggle repository and the Fake and Real News Dataset comprises 44,898 articles. The second stage is the stage of preprocessing the raw text data to eliminate noise and standardize the content. The third step is to apply the feature extraction methods depending upon the type of model. The fourth stage is to partition the data into training and testing data with 80:20 ratio. On the preprocessed data, six classification models are trained and evaluated in the fifth stage. Standard measures of performance are used to test all models at the sixth stage. The last step will be the comparison of the outcomes of all six models to find the best model to detect fake news.

4.2 Logistic Regression

Logistic Regression: Is a statistical classification algorithm which estimates the likelihood of binary outcome as a logistic function. Despite its simplicity, Logistic Regression has been shown to be quite effective in text classification as it is able to handle high dimensional features space. The baseline model in this research is Logistic Regression with TF-IDF (Term Frequency-Inverse Document Frequency) vectors of the preprocessed text data which are represented as TF-IDF vectors. TF-IDF vectorizer was also used with a maximum number of features set to 5000, which will represent the most important terms in the corpus. A solver, Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) with a limit of 1,000 iterations was used to train the model until convergence. Logistic Regression estimates the likelihood of an article to be fake as shown in Equation (1):

$$P(y=1|x) = 1 / (1 + e^{(-z)}) \ldots (1)$$

where $z = w^T x + b$, x is a TF-IDF feature vector, w is the weight vector, and b is the bias term. The model predicts the article as fake if $P(y=1|x) > 0.5$, and real otherwise.

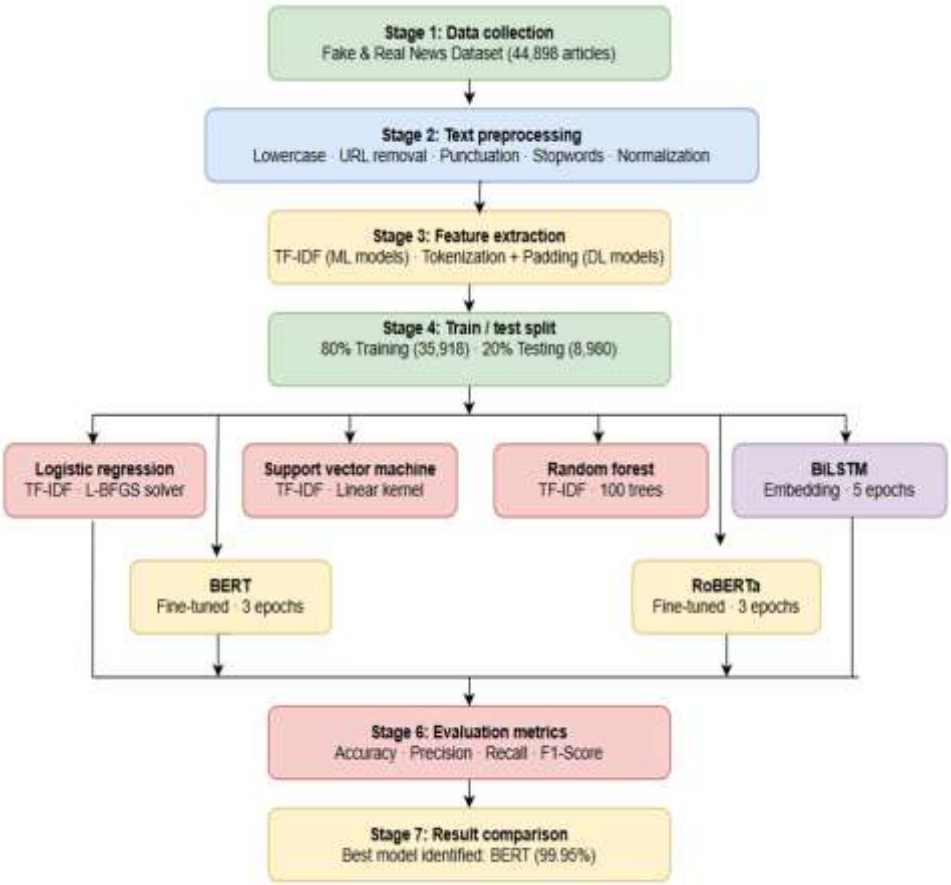


Fig. 1. Overall Architecture for fake news detection

4.3 Support Vector Machine (SVM)

Support Vector Machine (SVM) is one of the powerful supervised learning algorithm that identifies the best possible hyperplane that will maximally separate the classes in the feature space. Another advantage of SVM is that it can be used for high dimensional text classification, where feature representation is sparse. The TF-IDF features were used in this research with a maximum feature size of 5,000 and used with a Linear SVM classifier. Linear SVM was selected over kernel based SVM as it is computation efficient for a large scale text data set. The decision function of the is: Linear SVM:

$$f(x) = sign\,(w^T x + b) \quad \ldots (2)$$

where w is the weight vector learned during training, x is the TF-IDF feature vector of the input article, and b is the bias term. The model classifies an article as fake if $f(x) = +1$ and real if $f(x) = -1$.

4.4 Random Forest

Random Forest is an ensemble learning algorithm which constructs multiple decision trees in the training step and outputs the mode of the results of each tree. Each decision tree of the forest is trained using a random subset of the training data and a random subset of the features, which reduces the risk of overfitting and improves the ability to generalize. The random forest classifier with 100 decision trees and features of TF-IDF representing maximum feature number of 5000 has been used in this paper. The majority of the all 100 trees gives the final classification decision as indicated by Equation (3):

$$y\_pred = mode\ \{ T_1(x), T_2(x), \ldots, T_{100}(x) \} \quad \ldots (3)$$

where $T_i(x)$ represents the prediction of the i-th decision tree for input article x. The use of multiple trees provides robust and stable predictions and reduces the variance associated with individual decision trees.

4.5 Bidirectional Long Short-Term Memory (BiLSTM)

Long Short-Term Memory (LSTM) is a particular form of Recurrent Neural Network (RNN) that is developed to learn long-term dependencies in sequential data through the use of gating mechanisms to regulate information flow. BiLSTM is the same as LSTM but performs the processing of the sequence in both forward and backward direction, allowing the model to learn the context of both backward and forward tokens in the sequence. In this study, text data was tokenized and padded using 10,000 most frequent words and the limit of sequence lengths was set to 300. The diagram of the BiLSTM model contains the following layers:

Table 4: BiLSTM Model Architecture

Layer	Configuration	Output
Embedding	10,000 vocab, 128 dims	(batch, 300, 128)
BiLSTM (1st)	64 units, return seq.	(batch, 300, 128)
BiLSTM (2nd)	32 units	(batch, 64)
Dense	64 units, ReLU	(batch, 64)
Dropout	Rate = 0.5	(batch, 64)
Output Dense	1 unit, Sigmoid	(batch, 1)

The model was compiled using the Adam optimizer with binary cross-entropy loss and trained for 5 epochs with a batch size of 64. The output of the BiLSTM model is computed as given by Equation (4):

$$h\_t = [h\_forward\_t ; h\_backward\_t] \quad \ldots (4)$$

where h_forward_t and h_backward_t represent the hidden states from the forward and backward LSTM layers respectively at time step t, and the semicolon denotes vector concatenation.

4.6 Bidirectional Encoder Representations from Transformers (BERT)

BERT is a language model that uses a pre-trained transformer network and is proposed by Devlin et al. that uses the mechanism of bidirectional self-attention to learn deep contextual representations of text. Compared to traditional sequential models, BERT learns the entire input sequence at once in both directions and therefore allows it to learn the rich contextual dependencies between tokens. This paper used the Bert for Sequence Classification as the fine-tuning of the pre-trained bert-base-uncased model to binary fake news classification. The tokenizer applied to the input text was the BERT tokenizer, whose maximum sequence length was 128 tokens, the input was padded and truncated to have an equal input size. The dataset was randomly sampled to 10,000 articles which were then fine-tuned because of the computational limitations of the Google Colab environment. The optimizer employed in the fine-tuning was AdamW and a learning rate of 2e-5 was applied in 3 epochs with a batch size of 16. BERT model yields classification output as shown in Equation (5):

$$y = softmax(W \cdot h\_[CLS] + b) \quad \ldots (5)$$

where h[CLS] is the hidden state of the special [CLS] token from the final transformer layer, W is the weight matrix of the classification head, and b is the bias vector. The [CLS] token aggregates the contextual information of the entire input sequence and serves as the sentence-level representation for classification.

4.7 Robustly Optimized BERT Pretraining Approach (RoBERTa)

RoBERTa is an optimized version of BERT which was proposed by Liu et al. which enhances the pretraining strategy of BERT with a number of key alterations such as training on larger datasets with longer sequences, eliminating the Next Sentence Prediction (NSP) task, and dynamic masking during training. These changes lead to better language representations and better downstream task performance as opposed to the original BERT model. The architecture was a fine-tuning of the pre-trained roberta-base model on the Roberta for Sequence Classification architecture that was used to detect fake news in this study. BERT was then replicated with the same experimental setup, except that RoBERTa tokenizer was used, the maximum sequence length was 128 tokens, 10,000 articles were sampled, AdamW was used with a learning rate of 2e-5, epochs were 3, and a batch size was 16. RoBERTa model classification output is formulated in the same way as BERT in Equation (5), replacing the encoder used in BERT with RoBERTa encoder to produce contextual token representations.

4.8 Evaluation Metrics

All six classification models were evaluated using the four most commonly used evaluation metrics in the field of text classification and fake news detectors: Accuracy, Precision, Recall and F1-Score. All the measurements are calculated using the values of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) gained in the confusion matrix as in Equations (6) to (9):

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad \ldots (6)$$

$$Precision = TP / (TP + FP) \quad \ldots (7)$$

$$Recall = TP / (TP + FN) \quad \ldots (8)$$

$$F1\text{-}Score = 2 \times (Precision \times Recall) / (Precision + Recall) \quad \ldots (9)$$

Accuracy is a measure of the general proportion of the correctly classified news articles of all the articles in the test set. Precision measures the number of articles predicted to be fake that actually are fake – it provides an idea of the model's reliability in positive predictions. Recall is the proportion of actual fake articles that can be recognized by the model, the ability of the model to recognize all fake cases. F1-Score It is a harmonic mean of Precision and Recall and provides a balanced evaluation of the model accuracy, helpful when the distribution between the classes is not perfectly even. A summary of the experimental settings of all the six models of this study is presented in Table 5.

Table 5: Summary of Experimental Settings

Model	Category	Features	Key Parameters
Logistic Regression	Traditional ML	TF-IDF (5,000)	L-BFGS, max_iter=1000
SVM	Traditional ML	TF-IDF (5,000)	Linear kernel
Random Forest	Traditional ML	TF-IDF (5,000)	100 trees, random_state=42
BiLSTM	Deep Learning	Tokenization + Padding	Adam, 5 epochs, batch=64
BERT	Transformer	BERT Tokenizer (128)	AdamW, lr=2e-5, 3 epochs
RoBERTa	Transformer	RoBERTa Tokenizer (128)	AdamW, lr=2e-5, 3 epochs

5. RESULTS AND DISCUSSION

5.1 Experimental Setup

All the experiments were run in the Google Colaboratory (Colab) platform with a T4 free of charge Google-provided graphics card accelerator. All experiments were done on the Fake and Real News Dataset which consists of 44,898 articles. The dataset was preprocessed using the pipeline developed in Section III and split into a training set and a testing set with 80:20 ratio. For traditional ML models, the features were TF-IDF with the max number of features being 5000. The maximum length of the sequences in the case of BiLSTM model was 300, which were padded before being tokenized. In the case of the BERT and RoBerta models, fine-tuning was done on a subset of 10,000 randomly sampled articles because of computational limits, and the maximum token length was 128. All of the models were run with Python 3 and scikit-learn, TensorFlow, Keras, and HuggingFace Transformers libraries.

5.2 Experimental Results

The results of the classification tests of all six models were provided in the detail in Table 6 with the test set. The findings are presented in the Accuracy, Precision, Recall and F1-Score of the Real and Fake news categories.

Table 6: Detailed Classification Results of All Six Models

Model	Class	Precision	Recall	F1-Score	Support	Accuracy
Logistic Regression	Real	0.99	0.99	0.99	4247	**98.95%**
	Fake	0.99	0.99	0.99	4733	
SVM	Real	0.99	1.00	0.99	4247	**99.52%**
	Fake	1.00	1.00	1.00	4733	
Random Forest	Real	1.00	1.00	1.00	4247	**99.80%**
	Fake	1.00	1.00	1.00	4733	
BiLSTM	Real	0.99	0.99	0.99	4247	**99.00%**
	Fake	0.99	0.99	0.99	4733	
BERT (Best)	**Real**	**1.00**	**1.00**	**1.00**	**961**	**99.95%**
	Fake	**1.00**	**1.00**	**1.00**	**1039**	
RoBERTa	Real	1.00	1.00	1.00	961	**99.80%**
	Fake	1.00	1.00	1.00	1039	

Table 7: Comparative Summary of All Models

Model	Accuracy (%)	Precision	Recall	F1-Score	Rank
Logistic Regression	98.95	0.99	0.99	0.99	6
BiLSTM	99	0.99	0.99	0.99	5
SVM	99.52	1	1	1	4
Random Forest	99.8	1	1	1	3
RoBERTa	99.8	1	1	1	2
BERT	**99.95**	**1**	**1**	**1**	**1**

5.3 Result Discussion

Table 6 and Table 7 details the accuracy of all six classification models on Fake and Real News Dataset, respectively, which is found to perform well with the accuracy ranging from 98.95% to 99.95%. This denotes that both the conventional machine learning and deep learning models can be successfully used to distinguish between fake and real news articles when used on this dataset. But there are greater disparities on performance amongst model categories that are explained in the subsequent sub-sections.

5.3.1 Performance of Traditional Machine Learning Models

Random Forest returned the highest accuracy of 99.80% as compared to SVM and Logistic Regression at 99.52 and 98.95 respectively. When the TF-IDF feature representation is used, all the three models have obtained a precision, recall and F1-score of 0.99-1.00 for both classes, which suggests that the models are very good at doing classification. The reason behind the good performance of the Random Forest is the ensemble nature of the method which involves the joining of 100 decision trees to give a strong and stable classification decision. SVM was effective because it was able to identify optimal decision boundaries in high-dimensional feature space or TF-IDF space. Logistic Regression, the simplest of models, had a competitive accuracy of 98.95, which is affirmation that it is an effective model to use as a strong baseline of text classification tasks. Nevertheless, such classical models fully use the TF-IDF features that only reflect the frequency statistics of words in the text but do not reflect the semantic meaning of words and their contextual interaction with other words

5.3.2 Performance of Deep Learning Model

BiLSTM model attained the accuracy of 99.00 percent with the precision, recall and F1-scores of 0.99 of both classes. Although this performance is competitive to the traditional machine learning models, it was not a significant improvement on this dataset. BiLSTM model sequentially processes the text both forward and backward and therefore it is capable of capturing contextual dependencies between words. It is however to be expected that the relatively small difference between traditional models can be explained by the nature of the dataset, with lexical and statistical features represented by TF-IDF already being highly discriminative between fake and real news articles. However, that the BiLSTM model can be trained on raw sequences of tokens without manually designed features is an important benefit when it comes to extrapolation to new data.

5.3.3 Performance of Transformer-Based Models

The overall performance of the transformer-based models was the best in this study. BERT achieved the highest accuracy (99.95%), and the highest precision, recall and F1-score of both classes (1.00), and was the best model of all six methods measured. RoBERTa achieved the accuracy of 99.80% while the accuracy of BERT was 99.82%. BERT's success might stem from its ability to learn rich representations of words in context through pre-training on large-scale text corpora in a deep, bidirectional way, instead of relying on basic word frequency features. The fine-tuned BERT model makes use of the [CLS] token representation to encode the semantic content of the whole article and uses a classification head to differentiate between fake and real news with high accuracy. These findings validate the concept that transformer-based models, particularly BERT,

are the most effective models to use for fake news detection tasks and that they greatly surpass traditional machine learning models in terms of understanding the context of language.

5.4 Graphs and Visualizations

The accuracy comparison bar chart in Figure 2 shows how all six models have improved the accuracy of their performance in a progressive way starting with the traditional machine learning models through to the transformer-based models. The confusion matrices of each of the six models are shown in figure 3 which shows the number of true positive, true negative, false positive and false negative of each model. Figure 4 represents the training accuracy curve and the validation accuracy curve of the BiLSTM model in 5 training epochs, thus the model doesn't overfit. All these visualizations lead to the overall view of the performance features of every model used in this study.

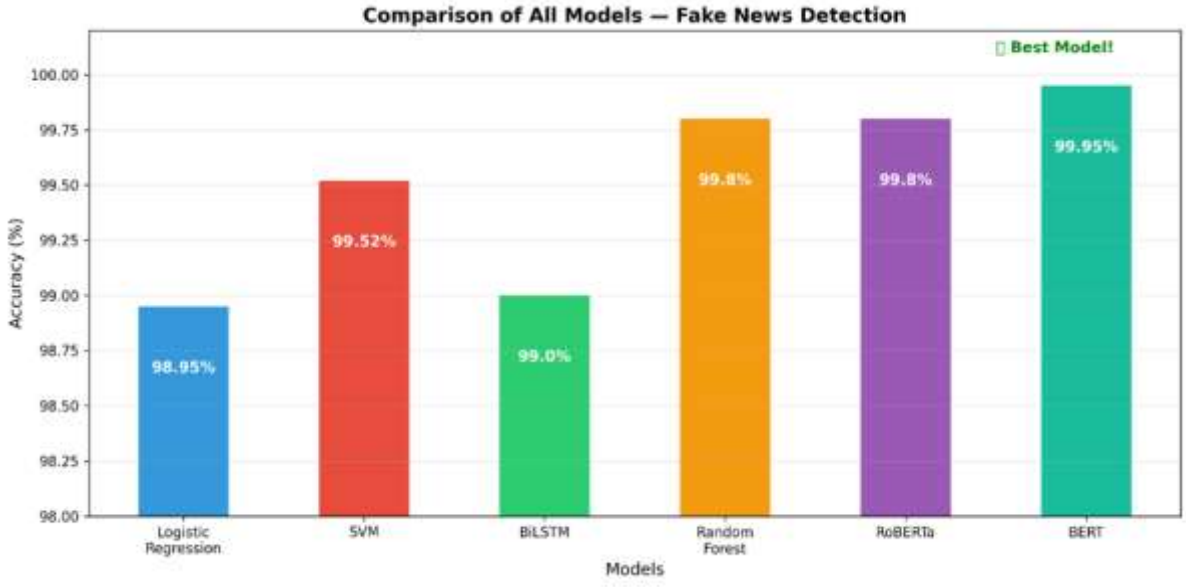


Fig. 2. Accuracy comparison of all six models

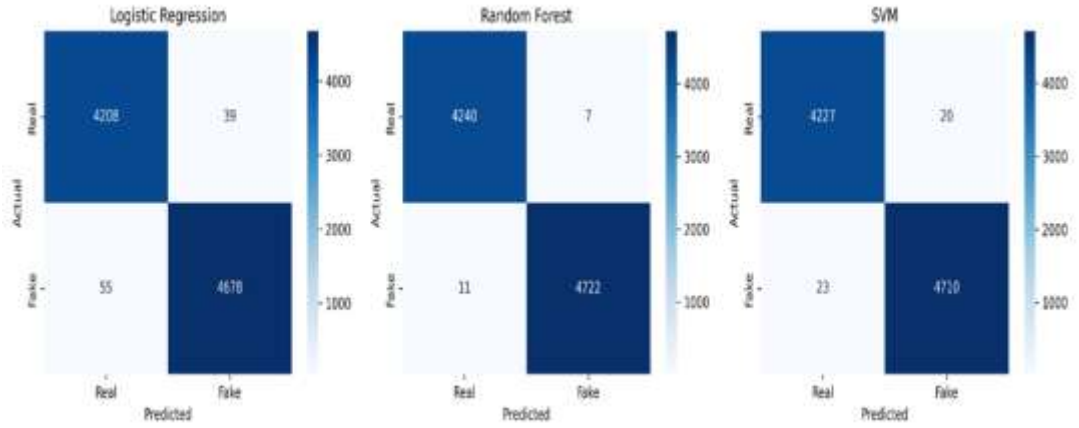


Fig. 3. Confusion matrices of all six models

Natural Resources for Human Health
eISSN: 2583-1194

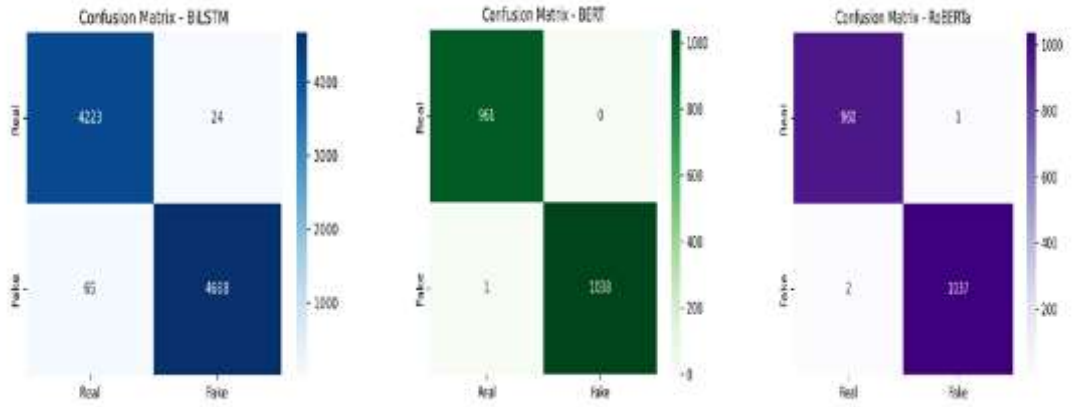
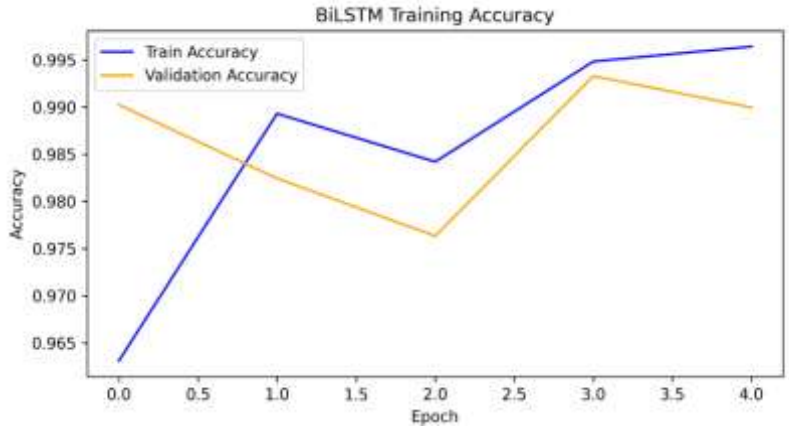


Fig. 4. BiLSTM training and validation accuracy curves



5.5 Summary of Findings

Based on the analysis of the experimental results and discussion made in this section, the following is a summary of the results of this experiment. First, the six classification models were found to be very accurate with the average value of over 98 percent accuracy on Fake and Real News Dataset which illustrated the possibility of automated fake news detection using traditional and advanced machine learning models. Secondly, the fine-tuned BERT has the highest accuracy of classification which is 99.95%, while the precision, recall and F1-score are all 1.00, indicating that a transformer-based model is the best method of fake news detection. Third, Random Forest is outperforming SVM and Logistic Regression with an accuracy of 99.80 which is competitive with the transformer based RoBERTa model. Fourth, the BiLSTM deep learning model achieved 99.00 per cent accuracy, demonstrating that deep learning models can also be a contender for performance without the need for hand-designed features. Fifth, the overall comparative analysis carried out in this paper under the same experimental conditions provides a good benchmarking information, which can be useful for researchers and practitioners to select the appropriate model to use in the real world application of fake news detection.

6. CONCLUSION AND FUTURE WORK

In the current paper, six different classification models of automatic fake news detection have been detailed and compared using Fake and Real News Dataset of 44,898 news articles. The models under consideration fall into three categories, namely, traditional machine learning models such as Logistic Regression, Support Vector Machine, and Random Forest, a deep learning model that is built on Bidirectional Long Short-Term Memory (BiLSTM), and transformer-based models such as BERT and RoBERTa. All the models were trained and tested on the same experimental conditions with a uniform preprocessing pipeline. The experimental results demonstrated that that each of the six models achieved a high level of accuracy in the classification of fake news with a value of more than 98% which consequently showed the feasibility of fake news automated detection. The

most effective model for detecting fake news is the fine-tuned BERT model which achieved 99.95% accuracy, and 1.00 Recall and F1-Score. The scores of both RoBERTa and Random Forest were 99.80, followed by SVM with 99.52, BiLSTM with 99.00 and Logistic Regression with 98.95. These results support the fact that transformer-based models and especially BERT have a better contextual language comprehension and are more effective in fake news detection tasks, compared to traditional machine learning and deep learning models.

In spite of the high outcomes obtained, this research has some limitations that can be used to guide further research. Since fine-tuning of BERT and RoBERTa models was limited by computational resources, future research can consider fine-tuning on the entire dataset with access to high-performance computing systems to enhance accuracy further. Moreover, the dataset of this study is restricted to the English-language political news, and the future research may consider the analysis to the multilingual and multi-domain fake news datasets, such as the regional Indian languages, namely Tamil, Telugu, and Hindi. Moreover, the use of multimodal fake news detection methods that involve the use of images, patterns of social network propagation and user engagement data along with the textual information might be explored to enhance the detection rates in the actual social media setting. Also necessary would be the creation of explainable AI techniques that can give explainable predictions, which would increase the transparency and reliability of automated fake news detection systems to end users and practitioners.

ACKNOWLEDGEMENTS

AUTHOR CONTRIBUTIONS

V. Karunakaran: Conceptualization, methodology, data collection, investigation, and writing—original draft preparation.

Rajat Goel: Data analysis, machine learning implementation, validation, visualization, and manuscript review and editing.

Rajasekar Velswamy: Methodology development, experimental design, implementation, and critical review of the manuscript.

Jeyakrishna Venugopal: Supervision, conceptualization, research design, interpretation of results, manuscript review and editing, and project administration.

Sreedevi V.: Literature review, data validation, manuscript review and editing, and overall coordination.

All authors contributed to the development of the study, reviewed the manuscript critically, approved the final version, and agreed to be accountable for the work.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

ETHICS STATEMENT

Not applicable.

DATA AVAILABILITY STATEMENT

HTTPS://WWW.KAGGLE.COM/DATASETS/CLMENTBISAILLON/FAKE-AND-REAL-NEWS-DATASET

REFERENCES

[1] Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter, 19(1), 22–36.
[2] Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science, 359(6380), 1146–1151.
[3] World Health Organization. (2020). Infodemic management: A key component of the COVID-19 global response. World Health Organization. https://www.who.int/teams/risk-communication/infodemic-management
[4] Ahmed, H., Traore, I., & Saad, S. (2018). Detecting opinion spams and fake news using text classification. Security and Privacy, 1(1), e9.
[5] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780.
[6] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT 2019, 4171–4186.

[7] Conroy, N. K., Rubin, V. L., & Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. Proceedings of the Association for Information Science and Technology, 52(1), 1–4.
[8] D'ulizia, A., Caschera, M. C., Ferri, F., & Griffoni, P. (2021). Fake news detection: A survey of evaluation datasets. PeerJ Computer Science, 7, e518.
[9] Alghamdi, J., Luo, S., & Lin, Y. (2024). A comprehensive survey on machine learning approaches for fake news detection. Multimedia Tools and Applications, 83(17), 51009–51067.
[10] Pérez-Rosas, V., Kleinberg, B., Lefevre, A., & Mihalcea, R. (2018). Automatic detection of fake news. In Proceedings of the 27th International Conference on Computational Linguistics (pp. 3391–3401).
[11] Al Asaad, B., & Erascu, M. (2018). A tool for fake news detection. In 2018 20th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC) (pp. 379–386).
[12] Khanam, Z., Alwasel, B. N., Sirafi, H., & Rashid, M. (2021). Fake news detection using machine learning approaches. IOP Conference Series: Materials Science and Engineering, 1099(1), 012040.
[13] Shaikh, J., & Patil, R. (2020). Fake news detection using machine learning. In 2020 IEEE International Symposium on Sustainable Energy, Signal Processing and Cyber Security (iSSSC) (pp. 1–5).
[14] Ahmad, I., Yousaf, M., Yousaf, S., & Ahmad, M. O. (2020). Fake news detection using machine learning ensemble methods. Complexity, 2020(1), 8885861.
[15] Reis, J. C., Correia, A., Murai, F., Veloso, A., & Benevenuto, F. (2019). Explainable machine learning for fake news detection. In Proceedings of the 10th ACM Conference on Web Science (pp. 17–26).
[16] Thota, A., Tilak, P., Ahluwalia, S., & Lohia, N. (2018). Fake news detection: A deep learning approach. SMU Data Science Review, 1(3), 10.
[17] Kaliyar, R. K., Goswami, A., Narang, P., & Sinha, S. (2020). FNDNet — A deep convolutional neural network for fake news detection. Cognitive Systems Research, 61, 32–44.
[18] Bahad, P., Saxena, P., & Kamal, R. (2019). Fake news detection using bi-directional LSTM-recurrent neural network. Procedia Computer Science, 165, 74–82.
[19] Kumar, S., Asthana, R., Upadhyay, S., Upreti, N., & Akbar, M. (2020). Fake news detection using deep learning models: A novel approach. Transactions on Emerging Telecommunications Technologies, 31(2), e3767.
[20] Hiramath, C. K., & Deshpande, G. C. (2019). Fake news detection using deep learning techniques. In 2019 1st International Conference on Advances in Information Technology (ICAIT) (pp. 411–415).
[21] Nasir, J. A., Khan, O. S., & Varlamis, I. (2021). Fake news detection: A hybrid CNN-RNN based deep learning approach. International Journal of Information Management Data Insights, 1(1), 100007.
[22] Monti, F., Frasca, F., Eynard, D., Mannion, D., & Bronstein, M. M. (2019). Fake news detection on social media using geometric deep learning. arXiv preprint arXiv:1902.06673.
[23] Yuan, H., Zheng, J., Ye, Q., Qian, Y., & Zhang, Y. (2021). Improving fake news detection with domain-adversarial and graph-attention neural network. Decision Support Systems, 151, 113633.
[24] Kaliyar, R. K., Goswami, A., & Narang, P. (2021). FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. Multimedia Tools and Applications, 80(8), 11765–11788.
[25] Raza, S., & Ding, C. (2022). Fake news detection based on news content and social contexts: A transformer-based approach. International Journal of Data Science and Analytics, 13(4), 335–362.
[26] Segura-Bedmar, I., & Alonso-Bartolome, S. (2022). Multimodal fake news detection. Information, 13(6), 284.