A Novel Three-Level Voting Model for Detecting Misleading Information on COVID-19

Shovan Bhowmik[1,*[0000−0001−7218−9875]], Priyo Ranjan Kundu Prosun[2[0000−0003−1512−7680]], and Kazi Saeed Alam[2[0000−0001−6863−1309]]

[1] Bangladesh Army International University of Science and Technology, Cumilla Cantonment, Bangladesh
Corresponding Author: bhowmik.sshovon5795@gmail.com
[2] Khulna University of Engineering & Technology, Khulna, Bangladesh
priyo.prosun1997@gmail.com
saeed.alam@cse.kuet.ac.bd

Abstract. Fake or misleading information detection is attracting researchers from all around the world in recent years as both society and the political world are greatly influenced by it. Moreover, various popular social media sites such as Twitter, Facebook, Instagram, etc. have accelerated the increase in the dissemination of rumors, false cures, conspiracy theories in the forms of posts, articles, videos, URLs during the COVID-19 pandemic. Thus there is an extensive need to find new techniques to verify or check the reliability of the online contents, which has inspired us to conduct this research to automatically detect misleading information. Our main aim is to create a three-level voting model to categorize the information into two classes: 'real' or 'misleading'. Five conventional mining algorithms and three ensemble models have been deployed with two distinct feature extraction techniques accompanying multiple sets of n-gram profiles on a benchmark dataset. Our research outcome shows that the Linear Support Vector Machine algorithm and Bagging ensemble model classifier have carried out significantly in recognizing misleading information which has been surpassed by our proposed novel three-level voting scheme. Our proposed model yields the best performance using TF-IDF for feature extraction with 96% accuracy.

Keywords: Coronavirus · Covid-19 · Fake News · Text Classification · Misleading Information · Three-level Voting · Machine Learning Models

1 Introduction

'Fake News' can be delineated as deceptive information which conveys through various social media sites and news portal to manipulate human opinion. With the growth of social media, in particular the Facebook News Feed and Twitter, the incidents of fake news has been escalated.[1]
In china, initially guessed as acute Pneumonia was further declared as COVID-19 by World Health Organization (WHO) which was spread by 'Sars-CoV-2' virus [2]. COVID-19 is a worldwide health issue which needs strong awareness,

thorough maintenance of individual and public hygienic practices, and tidiness of all locations. Social media platforms greatly influence the distribution of misleading information relevant to COVID-19 in all facets of our everyday lives. The existence of a Coronavirus, it's attributes and specifics of which are not yet well recognized. Moreover, due to the dissemination of misleading information about this virus, general people are getting frightened. Deceiving information might be designed to threaten the economies of a country, weaken people's trust in their governments, or allow a single commodity to make massive profits, which has already occurred with COVID-19 [3]. As a result, there is an immediate need to provide the public with a method to validate the authenticity of knowledge relating to COVID-19.

Motivated by the need for automatic recognition of false news, several Mining Algorithms and Natural Language Processing features have been used [15][11][5]. In this article, we have implemented a novel three-level voting architecture which is combined by text analysis based on NLP features and various ML classification methods. The performance of our three-level voting model has surpassed the performance of all five distinct classifying methods, to be specific, Multinomial Naive Bayes (MNB), Linear Support Vector Classifier (LSVC), Decision Tree Model (DT), Logistic Regression (LR), K-Nearest Neighbor (KNN) and three different ensemble classifiers namely Bagging, Gradient Boosting(GBoost) and AdaBoost (AdB). Using a benchmark dataset constructed of both real and fake news, experimental evaluation has been performed and very promising results have been obtained. In order to confirm the standard of the data, four distinct cross-validation methods were conducted and performance was measured for five metrics.

2 Related Works

Various strategies have been put forward for the identification of misleading news. This section summarizes the similar research work that has been conducted to determine misinformation on diverse platforms. Rubin et al. brought forward a 'Satire Detection' model in which they employed a Support Vector Machine (SVM) based algorithm augmented by 5 predictive attributes such as absurdity, humor, grammar, negative-affect and punctuation [4]. Their combinations were further tried out on 360 news articles. Elhadad et al. introduced a model for false news identification on social media sites [5]. For detection purpose, they created a feature vector to extract hybrid attributes from the metadata of news documents. The efficiency of their approach was evaluated using nine ML classification algorithms on three different datasets.

Wang et al. [6] also presented LIAR, a new sate of data that can be used to spot misleading information automatically. While LIAR is significantly greater in size, unlike other data sets, this data set does not include entire articles, it consists of 12800 pre-classified short statements from politicalFact.com. Rubin et al. employed both Vector Space Modeling (VSM) and Rhetorical Structure Theory (RST) to distinguish between inaccurate and accurate news [7]. RST

captures the coherence of the news article in terms of the practical interaction among the useful text units. The hierarchical formation of each news article can also be depicted by RST. By contrast, VSM is used to determine the connections among rhetorical structures. Ahmed et al. proposed an architecture for detecting misleading news that uses both n-gram and Term Frequency - Inverse Document Frequency (TF-IDF) to extract features [8].

Nowadays, two of the best feature extraction methods which researchers use are Bag of words [9] and Tf-Idf. Thota et al. [10] proposed a fake news identification system with the help of several Deep Learning architectures. Their approach of calculating Tf-Idf vectors on the basis of unigrams and bigrams turned out to be very successful, which outperformed some existing models to some extent. Al Asaad et al. presented an architecture that merges various Machine Learning classification techniques [11] such as Multinomial Naive Bayes [16] and Lagrangian SVM algorithm, for text categorization [12]. The credibility of their proposed system was tested upon a real/fake news dataset. Ibrishimova and Li [13] researched different deceptive news concepts and introduced a definition focused on absolute truthfulness and the source's comparative validity. In addition, they proposed a false news classification system, that uses automated and non-automated testing of information as well as stylistic features.

In a nutshell, a number of contemporary works focus on the evolution of identification systems that can recognize misleading information. For the detection process, the majority of them depend on manually labeled data. In this article, we present a three-level voting (TLV) structure, by using ground-truth accumulated from dependable and neutral information sources. We are mainly concentrating on information relevant to the COVID-19 pandemic.

3 Proposed News Classification Model

In our suggested paradigm, several well known machine learning (ML) models and ensemble techniques have been applied. We have used a publicly available large-scale dataset to justify our three-level voting system. Figure 1 illustrates the comprehensive methodology of our work.

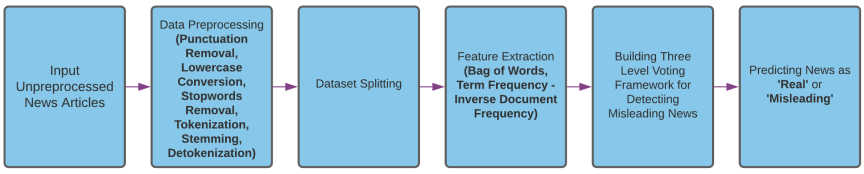

Fig. 1. Overall Misleading Information Detection Process.

3.1 Dataset Description

Most datasets for fake news detection research have contained political events. For being a new domain, there was no such dataset established related to COVID-19 in the initial stage of our work. Thus, we have used a newly created benchmark dataset named CoAID [14] that is publicly accessible. The dataset is diversified by various COVID-19 related healthcare data such as fake news articles on websites and claims of one or two sentences on social media platforms around the timeline from December 1, 2019 to November 30, 2020. Various reliable media outlets such as: Healthline, ScienceDaily, NIH, MNT, Mayo Clinic, Cleveland Clinic, WebMD, WHO and CDC have served as the source for real news whereas misleading news is retrieved from several fact-checking sites named LeadStories, PolitiFact, etc. Also, posts related to COVID-19 from many famous social platforms particularly Facebook, Twitter, Instagram, etc. are also included in the dataset. The collected dataset contents are labeled as 'Real' or 'Misleading' but it required cleaning as it was noisy. After going through some certain preprocessing steps, multiple features have been extracted to apply our proposed voting model. The final clean data constituted 3,749 articles(2,035 news and 1714 social media posts) in total among which 2,192 articles were labeled 'real' and 1,557 articles were labeled 'misleading'.

3.2 Data Preprocessing

Since we have collected raw data from the dataset, we need to perform dataset cleaning to get the refined data so that we can use them for the implementation of our work. For this reason, we have used several preprocessing methods to refine our raw data. Firstly, we have removed all the punctuations from the dataset. Then the whole document is converted to lower case so that consistency is maintained after which we have performed tokenization. The next step is to get rid of all the most frequently occurring words known as 'stop words'. Stop words are mainly insignificant words which if not are removed will create noise in text categorization. Following this, to reduce the type of words in our dataset, we have performed stemming in all the words to transform them into their root form. In the final step, detokenization is performed to get clean, refined and suitable data so that they can be used in the next stage of our work. Figure 2. Illustrates the overall preprocessing steps.

3.3 Dataset Dividation

The refined dataset has been divided into 80%-20% train-test ratio. After implementation of the splitting work, features were fitted to the train set. As our dataset was not perfectly balanced, we justified our architecture with a number of cross-validation methods namely, K-Fold, Stratified K-Fold, Shuffle Split and Stratified Shuffle Split. The value of 'K' was set to 10 folds while applying these cross-validation techniques.

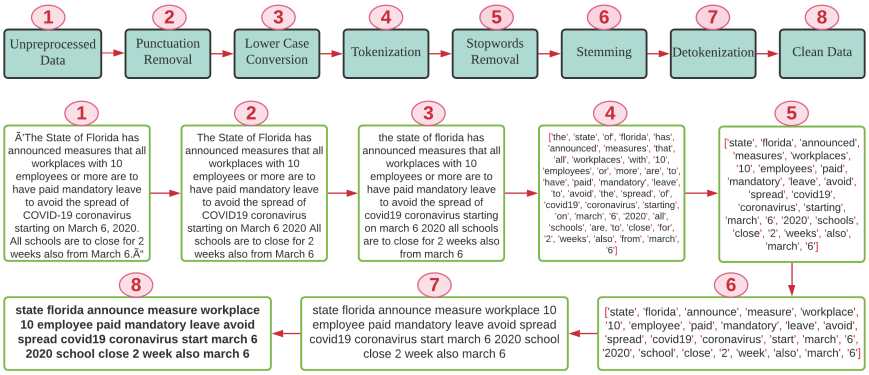

Fig. 2. Preprocessing Step.

3.4 Feature Extraction

It is really difficult to learn from a large amount of data for classifying text as the total number of words, terminologies and clauses are high. This makes the entire model arithmetically costly. Moreover, immaterial and redundant features hinder the classification model's accuracy and efficiency. Therefore, a feasible solution is to obtain important attributes to minimize an enormous amount of data and get rid of large feature space. We have considered two important features of natural language processing, 'BoW' and 'TF-IDF' in our research. The 'Bag of Words' (BoW) model computes the number of appearances of a single word in a document by taking input of documents, thereby representing numbers analogous to the text which can be regarded as fixed-length vector characteristics. Raw words are thus encrypted as key-value mapping to hold the recurrence of words in every document. Additionally, for the acquisition of information, we have utilized the 'Term Frequency - Inverse Document Frequency' (TF-IDF) model in our work. TF-IDF is a statistical technique that adds numeric loads to textual data used during the mining process. By calculating the frequency of terms in documents, it measures the relative importance of a term within the overall dataset. The key feature of this model is that the inverse frequency of documents negates the term frequency. Alternatively, it weighs down the influence of similar words in the document by scaling up the infrequent words.

3.5 Classification Process

To classify the news content, some base level ML models have been utilized in [15]. A single voting model with only mining algorithms has been designed in [5] to classify misleading data as well. In this paper, we have devised a novel TLV model consisting of machine learning algorithms as well as several ensemble techniques. Figure 3 portrays our proposed framework.

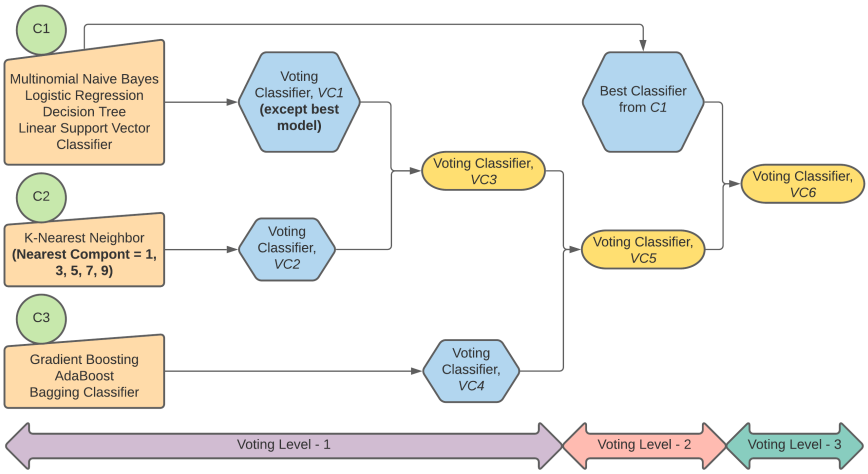

Fig. 3. The proposed Three Level Voting Scheme.

In our model, we have grouped different classifiers into three groups. $C1$: MNB, LR, DT, LSVC; $C2$: KNN with nearest component = 1, 3, 5, 7, 9 and $C3$: GBoost, AdB, Bagging

Three voting classifiers have been created from $C1$, $C2$ and $C3$ based on accuracy of the algorithms. $VC1$: Three classifiers (except the one which has highest accuracy) from $C1$ have been associated; $VC2$: $C2$ with all the variations have been associated; $VC4$: $C3$ have been associated; The three levels have been constructed as follows:
$1stLevel$: $VC3$ has been formed from $VC1$ and $VC2$
$2ndLevel$: $VC5$ has been formed from $VC3$ and $VC4$
$3rdLevel$: $VC6$ has been formed from $VC5$ and the classifier from $C1$ with the highest accuracy
Hard voting has been applied in our proposed model on all the results obtained from the ML algorithms and ensemble methods. Because of hard voting, if two of the three algorithms in a classifier group anticipate a piece of news as 'misleading', the whole voting classifier will assess the article as 'misleading'.

The articles contained redundant vocabularies. We have invoked 'BoW' and 'TF-IDF' features from the news articles to predict whether the considered article is 'real' or 'misleading'. In the TLV framework, these extracted features have been employed for recognition and different evaluation standards have been calculated to determine our TLV model success.

Table 1. Performance Evaluation Metrics Results for Selected Traditional Classifiers.

Metric	Feature Generation		MNB	LR	DT	LSVC	GBoost	AdB	Bagging	TLV
Accuracy	BOW		0.93	0.92	0.86	0.92	0.87	0.77	0.92	**0.94**
	TF-IDF	Word	0.84	0.90	0.87	0.93	0.88	0.84	0.93	0.93
		Character	0.85	0.90	0.85	0.92	0.90	0.86	0.92	0.93
		Unigram	0.82	0.89	0.88	0.94	0.88	0.82	0.94	**0.96**
		Bi-gram	0.80	0.79	0.79	0.81	0.75	0.74	0.81	0.75
		Tri-gram	0.78	0.77	0.76	0.78	0.77	0.77	0.78	0.79
		N-gram(2:3)	0.82	0.80	0.80	0.83	0.78	0.78	0.83	0.70
Precision	BOW		0.93	0.92	0.86	0.92	0.90	0.98	0.92	0.94
	TF-IDF	Word	0.94	0.93	0.88	0.93	0.91	0.92	0.93	0.92
		Character	0.94	0.92	0.86	0.92	0.92	0.93	0.92	0.93
		Unigram	0.93	0.93	0.89	0.94	0.92	0.93	0.94	0.95
		Bi-gram	0.93	0.93	0.82	0.85	0.93	0.96	0.85	0.76
		Tri-gram	0.96	0.97	0.91	0.94	0.97	0.98	0.94	0.89
		N-gram(2:3)	0.92	0.92	0.83	0.86	0.93	0.96	0.86	0.69
Recall	BOW		0.92	0.92	0.86	0.92	0.87	0.77	0.92	0.94
	TF-IDF	Word	0.84	0.90	0.87	0.93	0.88	0.84	0.93	0.92
		Character	0.85	0.90	0.85	0.92	0.90	0.86	0.92	0.93
		Unigram	0.82	0.89	0.88	0.94	0.88	0.82	0.94	0.95
		Bi-gram	0.79	0.79	0.79	0.81	0.75	0.74	0.81	0.74
		Tri-gram	0.78	0.77	0.76	0.78	0.77	0.77	0.78	0.76
		N-gram(2:3)	0.82	0.80	0.80	0.83	0.78	0.78	0.83	0.70
F1-Score	BOW		0.92	0.92	0.86	0.92	0.88	0.85	0.92	0.94
	TF-IDF	Word	0.87	0.91	0.87	0.93	0.89	0.86	0.93	0.92
		Character	0.88	0.91	0.85	0.92	0.90	0.88	0.92	0.93
		Unigram	0.85	0.90	0.88	0.94	0.89	0.85	0.94	0.95
		Bi-gram	0.84	0.83	0.80	0.82	0.81	0.82	0.82	0.73
		Tri-gram	0.85	0.85	0.82	0.84	0.85	0.86	0.84	0.81
		N-gram(2:3)	0.85	0.84	0.81	0.84	0.84	0.85	0.84	0.68
AUC	BOW		0.91	0.90	0.82	0.89	0.87	0.86	0.90	0.90
	TF-IDF	Word	0.90	0.92	0.83	0.92	0.88	0.84	0.92	0.90
		Character	0.91	0.90	0.80	0.90	0.90	0.90	0.90	0.90
		Unigram	0.88	0.92	0.86	0.93	0.91	0.88	0.93	0.92
		Bi-gram	0.85	0.84	0.75	0.78	0.73	0.76	0.78	0.72
		Tri-gram	0.79	0.72	0.61	0.75	0.71	0.75	0.75	0.70
		N-gram(2:3)	0.84	0.8	0.74	0.78	0.77	0.84	0.78	0.64

4 Experimental Results

A three level voting model has been designed to speculate the veracity of articles. Key features have been materialized from clean data using 'BoW' and 'TF-IDF' and fit our proposed model. Five popular classifiers (MNB, DT, LR, K-NN and LSVC) including three ensemble based approaches (GBoost, AdB and Bagging) have been selected for creating this model. We have assessed various performance appraisal measurements (Accuracy, Precision, Recall, F1-Score and AuC) to determine the efficiency of our proposed framework. The table 1 has substantiated the performance comparison of our model from other mining models. We have

Table 2. Comparative Analysis of Cross Validation Technique Results for TLV Model.

Cross-Validation	Feature Generation		Accuracy	Precision	Recall	F1-score	Auc
K-fold	BOW		0.85	0.89	0.88	0.89	0.82
	TF-IDF	Word	0.84	0.85	0.90	0.88	0.82
		Character	0.86	0.89	0.94	0.91	0.90
		Uni-gram	0.84	0.84	0.91	0.87	0.90
		Bi gram	0.42	0.79	0.37	0.50	0.44
		Tri gram	0.50	0.58	0.65	0.51	0.47
		N-gram(2:3)	0.52	0.68	0.53	0.60	0.52
Stratified K-fold	BOW		**0.95**	0.96	0.90	0.93	0.92
	TF-IDF	Word	0.92	0.94	0.95	0.95	0.96
		Character	0.92	0.95	0.95	0.95	0.95
		Uni gram	0.91	0.94	0.93	0.94	0.94
		Bi gram	0.52	0.90	0.42	0.57	0.43
		Tri gram	0.78	0.79	0.96	0.87	0.87
		N-gram(2:3)	0.63	0.90	0.58	0.71	0.65
Shuffle Split	BOW		0.93	0.97	0.93	0.94	0.92
	TF-IDF	Word	0.94	0.97	0.96	0.97	0.97
		Character	0.93	0.96	0.97	0.97	0.96
		Uni gram	0.93	0.95	0.94	0.95	0.91
		Bi gram	0.65	0.91	0.64	0.75	0.66
		Tri gram	0.78	0.82	0.80	0.81	0.82
		N-gram(2:3)	0.71	0.89	0.72	0.80	0.73
Stratified Shuffle Split	BOW		**0.95**	0.97	0.92	0.95	0.94
	TF-IDF	Word	**0.96**	0.95	0.96	0.96	0.96
		Character	**0.96**	0.95	0.96	0.96	0.96
		Uni gram	0.94	0.95	0.94	0.95	0.96
		Bi gram	0.69	0.89	0.68	0.78	0.72
		Tri gram	0.74	0.83	0.84	0.84	0.79
		N-gram(2:3)	0.71	0.89	0.74	0.81	0.76

got 94% accuracy for TLV considering 'BoW' as extracted features. TLV accuracy has excelled other models' accuracy. Consideration of TF-IDF (Unigram) for feature retrieval has provided 96% accuracy. The notable result has been obtained for other features as well. Moreover, adequate results have been reached

for precision, recall, F1-Score and AUC. Adequate marks for TF-IDF (Bi-gram, Tri-gram and N-gram(2:3)) have not been achieved since our data was collected from heterogeneous sources where N-Gram features are not able to find the semantic match between the words, characters, syllables or bytes. Besides, when taking N-grams as a feature value, we set 5000 as the highest number of features. Previously, similar results have been obtained in [8].

Our model has been substantiated by applying various cross-validation strategies to prevent our model from being skewed. The overall performance of our proposed model for cross-validation strategies is provided in table 2.

We have reached 95% accuracy for 'BoW' vector analysis when both stratified K-Fold and Stratified Shuffle Split processes have been leveraged. When Stratified Shuffle Split has been applied for TF-IDF (word, character), the accuracy has been excelled to 96%. Satisfactory results have been attained as performance evaluation metrics for other attributes as well.

We have incorporated similar types of algorithms in the same group. It has helped us to preserve the neutrality of the models' decision. The results can vary according to different features. Therefore, the best classifier from $C1$ has been separated for further use in 3rd level voting. Multiple voting levels have helped to predict the label of the contents accurately as in each level voting is applied for label identification. For example, if a label is predicted as 'real' in the first level and 'misleading' in the second level, there is an option to decide the label in the third level where the result has been compared with the best classifier of $C1$. Thus our TLV has increased the performance than individually applied ones for the majority of the features.

5 Conclusion & Future Works

Fake News on Covid-19 is continuously affecting the daily life of mankind either intentionally or unintentionally. In this work, a voting model has been designed for predicting real or misleading news. Our proposed TLV model has surpassed general mining techniques by voting for a single label in three stages. We have attained the maximum 96% accuracy for the CoAid dataset which may be increased more if we extend the amount of data in our dataset. In the further study, we will try to validate our model by applying more diversified public and private datasets consisting of articles, tweets, social media posts, etc. Instead of ensembling three algorithms for a single voting classifier, we will try to add more algorithms to measure the performance our system. We have also planned to apply Artificial Neural Network and Recurrent Neural Network for more stability of our work. Our suggested TLV framework can be applied in other related works in the field of NLP for e.g., Spam classification, Cyberbullying, Opinion Mining etc. for further study. Our proposed model can be easily implemented in software based application in the state of art.

References

1. Higdon, N. (2020). The Anatomy of Fake News: A Critical News Literacy Education. University of California Press.
2. World Health Organization Official Website. Accessed: Mar. 21, 2020.[Online]. Available: https://www.who.int.
3. Ling, T., Hoh, G., Ho, C., & Mee, C. (2020). Effects of the coronavirus (COVID-19) pandemic on social behaviours: From a social dilemma perspective. Technium Soc. Sci. J., 7, 312.
4. Rubin, V. L., Chen, Y., & Conroy, N. K. (2015). Deception detection for news: three types of fakes. Proceedings of the Association for Information Science and Technology, 52(1), 1-4.
5. Elhadad, M. K., Li, K. F., & Gebali, F. (2020). Detecting Misleading Information on COVID-19. Ieee Access, 8, 165201-165215.
6. Wang, W. Y. (2017). " liar, liar pants on fire": A new benchmark dataset for fake news detection. arXiv preprint arXiv:1705.00648.
7. Rubin, V. L., Conroy, N. J., & Chen, Y. (2015, January). Towards news verification: Deception detection methods for news discourse. In Hawaii International Conference on System Sciences (pp. 5-8).
8. Ahmed, H., Traore, I., & Saad, S. (2017, October). 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 (pp. 127-138). Springer, Cham.
9. Pimpalkar, A. P., & Raj, R. J. R. (2020). Influence of Pre-Processing Strategies on the Performance of ML Classifiers Exploiting TF-IDF and BOW Features. ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal, 9(2), 49-68.
10. Thota, A., Tilak, P., Ahluwalia, S., & Lohia, N. (2018). Fake news detection: A deep learning approach. SMU Data Science Review, 1(3), 10.
11. Smitha, N., & Bharath, R. (2020, July). Performance Comparison of Machine Learning Classifiers for Fake News Detection. In 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA) (pp. 696-700). IEEE.
12. Al Asaad, B., & Erascu, M. (2018, September). A tool for fake news detection. In 2018 20th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC) (pp. 379-386). IEEE.
13. Ibrishimova, M. D., & Li, K. F. (2019, September). A machine learning approach to fake news detection using knowledge verification and natural language processing. In International Conference on Intelligent Networking and Collaborative Systems (pp. 223-234). Springer, Cham.
14. Cui, L., & Lee, D. (2020). Coaid: Covid-19 healthcare misinformation dataset. arXiv preprint arXiv:2006.00885.
15. Reddy,P.S., Roy, D., Manoj, P., Keerthana, M., & Tijare, P. (2019). A Study on Fake News Detection Using Naïve Bayes, SVM, Neural Networks and LSTM.
16. Granik, M., & Mesyura, V. (2017, May). Fake news detection using naive Bayes classifier. In 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON) (pp. 900-903). IEEE.