Fake News Detection using Machine Learning Algorithms

Naresh Singh (naresh11113638@gmail.com), Suraj Dhoundiyal (suraj.dhoundiyal1999@gmail.com), Gunjan Kardam (gunjankardam123@gmail.com), Bharti(bharti.cse@kiet.edu)

Department of Computer Science and Engineering, KIET Group of Institutions, Ghaziabad, UP, India

Abstract

Earlier when there was no internet then life was not that easy for every human being. Every small work requires a lot of bookish research which people can use further in their works but now as technology has evolved people can get every piece of information in just a single click from their computer or mobile phone. In this new era when technology has evolved that much then definitely there would be some people or organizations who can misuse the technology. Nowadays one of the biggest misuse of technology is the spreading of fraudulent content on the internet or some other platforms by some people or organizations for the sake of their benefits and some political agenda which may benefit them but can sometimes create chaos among normal people. To stop the spreading of fake news there is a need to target those articles present on the internet and remove them. In this research, an ML model is trained which can detect fraud news and further actions can be taken on the news. Various ML models are used in this research and maximum accuracy of 95% is achieved by using Logistic Regression.

Keywords- Machine Learning, Fake news, models.

1. Introduction

In today's digital age, the spread of false or disinformation, usually called fake news, has become a serious problem. With the increase in the usage of social networking websites and applications, it has become increasingly difficult to distinguish between credible sources and misinformation. **[6]** This has led to the need for reliable and efficient methods to identify fake news. One promising approach is the use of machine learning techniques. This research explores the use of machine learning algorithms for the detection of fake news and evaluates their effectiveness. Additionally, the study also investigates the effect of fake news on society and the importance of developing robust systems to encounter fraudulent news. The challenges in catching fake news, such as the complexity of natural language and the dynamic nature of the information being shared are common. The effectiveness of using various ML models on this subject varies from technique to technique. Text classification techniques to identify fake news, including feature extraction, feature selection, and model evaluation are used in this study. The use of NLP techniques for extracting features such as sentiment analysis, named entity recognition, and text summarization that can be used to check fake news. It's important to consider the role of social media in the spread of fraudulent news. Social media platforms have become a primary source of information for many people, and false news spreads very fast on platforms like Facebook, Twitter, etc.

The use of the ML model can help to identify and flag fake news online, and can also be used to track the spread of misinformation and to take steps to limit its impact or to remove such types of fraud articles from social media applications. Another important aspect to consider is

the role of fact-checking in fake news detection. Fact-checking is verifying the accuracy of the information and is an important step in identifying and correcting false or misleading information. ML models can be used to automate the process of fact-checking, and to identify potentially false information before it is widely disseminated. It is also important to consider the potential biases that may be present in the data used to train ML models for fake news detection. Biases in the training data can lead to inaccurate or unfair results, and it's important to address these biases and to ensure that the models are fair and unbiased.

This research evaluates the conduct of various machine learning models using performance evaluation metrics such as accuracy, precision, recall, and F1-score. A discussion of how the results of the research can be applied to improve the accuracy of fake news detection systems and mitigate the impact of fake news on society. The ethical considerations of using ML for fake news detection, such as avoiding bias and protecting privacy. The research in the field of fake news detection using ML models is done by various researchers and there will be always a scope for future research.

1.1.Motivation

In this new era of technology, people have a very easy to access every piece of information available on the internet. Especially in India where people have very affordable data prices from the telecom operators, it becomes the second largest social media market in the world with 755 million active users in India only. Some people are not able to easily identify which data available on the internet is true or fraudulent and thus sometimes people get trapped in some sort of vicious circle which further leads to harming them sometimes financially or sometimes by any other means. In the research an ML model is trained which can easily identify fraudulent news in the market so that it can help people through social media applications and they do not easily get distracted from any fraudulent news thus internet can become a safer and more authentic place for every netizen.

1.2. Literature Review

In the year 2017, Mykhailo Granik and Volodymyr Mesyura published their research in Ukraine. **[12]** In their research, they used Naïve Bayes Classifier and achieved an accuracy of approximately 74% which is not very good but at that time this accuracy was very good because they lacked much training data. Also in the year 2018, Akshay Jain and Amey Kasbe published their research in India **[13]** which showed an accuracy of approx. 83%. In their study, they told several methods to apply on social media platforms like Facebook etc. which can easily identify fake news directly on that platform. Two scholars from Saudi Arabia, **[2]** conducted their research on various ML algorithms and found their highest accuracy of 99% with the Naïve Bayes algorithm. In this research, a brief study of other researchers is done and found that various researchers used supervised learning algorithms which led them to complete their research by finding the highest accuracy from any one of those algorithms.

2. Machine Learning Models

(a) Logistic regression is a predictive technique that utilizes a logistic function to model a dual outcome based on one or more independent variables. It is a statistical method used for classification problems, such as predicting whether a person has a certain disease based on certain measurements. The function generates an "S" curve, which allows for the modeling of probabilities of a certain event occurring. The model parameters are learned through maximum likelihood estimation and can be regularized to prevent overfitting.

The mathematical expression for the LR hypothesis function is Eq (1).

$$h_\theta(X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X)}} \quad \text{Eq (1)}$$

Using a sigmoid function, the output of logistic regression is converted to a value that is probable; the objective is to find the best probability by minimizing the cost function as shown by Eq (2) cost function calculation.**[1]**

$$cos\,t(h_\theta(x), y) = \begin{cases} log\big(h_\theta(x)\big), & y = 1 \\ -log\big(1 - h_\theta(x)\big), & y = 0 \end{cases} \quad \text{Eq (2)}$$

(b) Support Vector Machine locates a hyperplane with the greatest degree of data point separation between classes. This hyperplane is also referred to as the decision boundary in a two-class problem. Support vectors are the data points closest to the decision boundary, which is how the algorithm gets its name. SVMs are powerful and adaptable, but they can be affected by the kernel and input data scaling. Eq. (3) provides a mathematical representation **[1]** and definition of the SVM model's cost function.

$$J(\theta) = \frac{1}{2}\sum_{j=1}^{n} \theta_j^2 \quad \text{Eq (3)}$$

(c) Naive Bayes applies the popular Bayes theorem **[8]** methodology to identify the most promising hypothesis inside a given space. The dataset in this case serves as the space for the algorithm. According to the Bayes theorem:

$$p\left(\frac{h}{d}\right) = \frac{p\left(\frac{d}{h}\right) * p(h)}{p(d)} \quad \text{Eq (4)}$$

The best-fitting hypothesis for a given 'd' dataset is represented by 'h' in Eq (4). The "p(d/h)" calculates the likelihood of "h" about "d". The prediction is made via probabilistic calculations by quantifying the dataset into a likelihood table for a hypothesis

(d) Decision Tree is a supervised learning model mainly used for classification problems. **[10]** The internal nodes of the tree show the feature of the dataset and the leaf node shows the outcome and branches represent the decision rules. Classification tree are the tree models where the target variable can take a discrete set of values. Eq (5) and Eq (6) are to calculate Entropy mathematically which measures the impurity in the node.

$$E(S) = \sum_{i=1}^{c} -p_i\, log_2\, p_i \quad \text{Eq (5)}$$

$$E(T, X) = \sum_{c \in X} P(c) E(c) \quad \text{Eq (6)}$$

(e) Random Forest is an ensemble learning method for classification problems. It consists of several decision trees, each of which was trained using a distinct subset of the training data and set of characteristics. The average of all the trees' forecasts is used to determine the outcome. Random Forest is also known for its ability to handle high dimensional and correlated features, it can also be used to estimate feature importance. They are very adaptable and can handle a lot of input variables without having to delete any of them. Eq (7) is used to calculate Gini Index to measure variance.**[1]**

$$G_{ind} = 1 - \sum_{i=1}^{c} (P_i)^2 \quad \text{Eq (7)}$$

3. Preliminaries

3.1.Data Mining

Data mining plays a crucial role in detecting fake news by identifying patterns and knowledge in large amounts of dataset. It involves utilizing machine learning, statistics, and database systems, among other methods, to extract useful information from data sets. To train the model more efficiently there should be a large dataset with good quality information. **[5]**

Successful research and the training of machine learning models depend heavily on datasets. The dataset used here is taken from Kaggle for both training and testing of the model. The dataset contains a total of 20k news which includes both fake as well as true news.**[11]**

3.2. Data Pre-processing

Preprocessing of data is done by using multiple techniques in which the raw data is converted to such a form that is usable to train the machine learning model. Various preprocessing techniques are applied to make the algorithm more efficient so that the accuracy of the model provides great results. Because it can have a significant impact on the outcomes of the data mining process, data preprocessing is a crucial step. At very first null values are checked and they are filled **[2]**. Further, duplicate data entry is checked in the dataset and after that, whether the data is balanced or not is analyzed. Data balancing is the technique where it is checked that the number of true data is equal to the number of false data. The dataset was already balanced so there was no need to balance the data in this study. Further, Natural Language Processing (NLP) techniques are used to process the data. There are three steps of NLP pre-processing they are Tokenization, stopword removal, and stemming. Breaking down a sentence into small words called tokens in the process of Tokenization. After this stopwords were removed these words can create problems in the process of model training **[4]**. The last step in preprocessing using NLP techniques is Stemming, which is the process of transforming a word into its root form. Here, porter-stemmer is used for the stemming of the data which is available in the nltk (Natural Language toolkit library) **[4]** in python. Now the data is well organized and now the data is ready for further tasks. Now the data will go for the vectorization task.

3.3.Data Vectorization:

Machine learning models cannot be directly trained with raw text data. There is a need to convert text data into the form of vectors to train the model. Text documents generally have different sizes but the vector generated using the vectorizer has the same length. Each vector represents a specific feature. There are various vectorizers in the NLP technique like Count Vectorizer, TF-IDF Vectorizer, Word2vec, etc. In this research TF-IDF vectorizer is used as it is more suitable for the research. The TF-IDF Vectorizer calculates the TF and IDF values for each word in the document collection and then combines them to create a numerical representation of the text data.

3.4.Data Splitting

The dataset is divided into two parts one is used for the training of the model and the other one is used for the testing of the trained model. Data is generally split into 75% for training and 25% for testing but the ratio can be decided by the programmer itself. This research also trains model in a 75%-25% format for training and testing respectively. It should be ensured that data is properly balanced before it gets split as it may result in bias training of the model and further can give poor accuracy. After the splitting of the dataset, research proceeds for model training.

3.5.Performance Evaluation Metrics

Model prediction is a statistical technique in which probability is applied to an unknown event to predict outcomes. It analyzes past and present data of the predicted outcomes and scrutinizes them. After the model training, the model prediction test is done using testing data on multiple classifiers and the classification report is generated which contains precision, recall, and f1-score **[2]** on which we calculated the accuracy of model further Confusion Matrix is also generated for all the classifiers shown in figure 2. By using the formulas of Eq (8), Eq (9), and Eq (10), precision, recall, and f1-score can be calculated mathematically. **[3]**

$$precision = \frac{TP}{TP+FP} \quad \text{Eq (8)}$$

$$recall = \frac{TP}{TP+FN} \quad \text{Eq (9)}$$

$$f1 - score = 2 * \frac{precision*recall}{precision+recall} \quad \text{Eq (10)}$$

4. Proposed Methodology

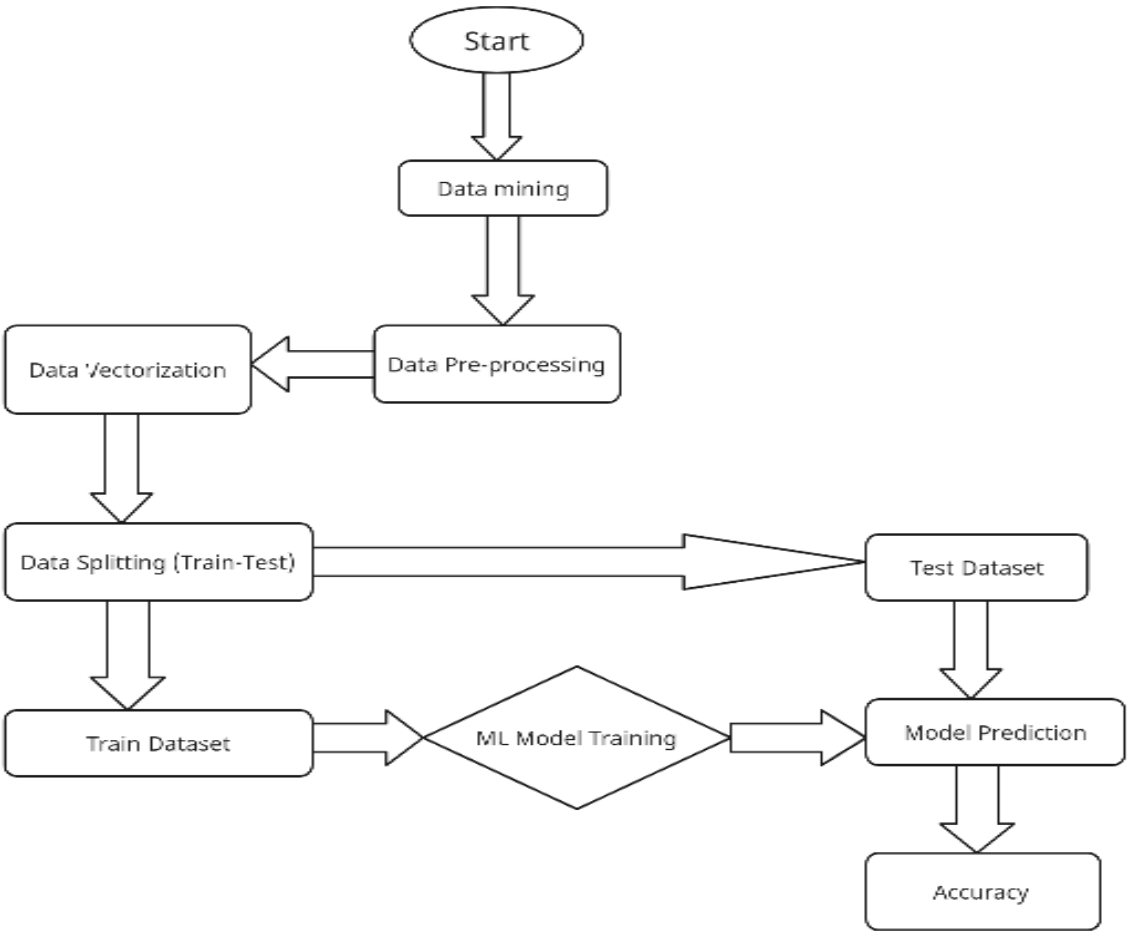


Figure 1: Flowchart of model

Figure 1 shows the methodology contains many steps which include Data Collection, Data Pre-processing, Data Vectorization, Data Splitting, Model training, and further evaluation of the trained model. In data collection, large datasets of real and fake articles from a variety of news sources are collected and labelled. Data Pre-processing of that labelled data is done to fill the null values, balancing of data, Tokenization, Stopwords removal, and stemming. Now the vectorization of pre-processed data is done so that the text data of the dataset gets converted into vectors which is understandable by the machine. The vector data is further split data in a ratio of 3:1 for training and testing purpose. The 75% training data goes for model training and thus the final trained model is evaluated on the 25% of testing data and thus the accuracy of this model is tested on various ML classifiers.

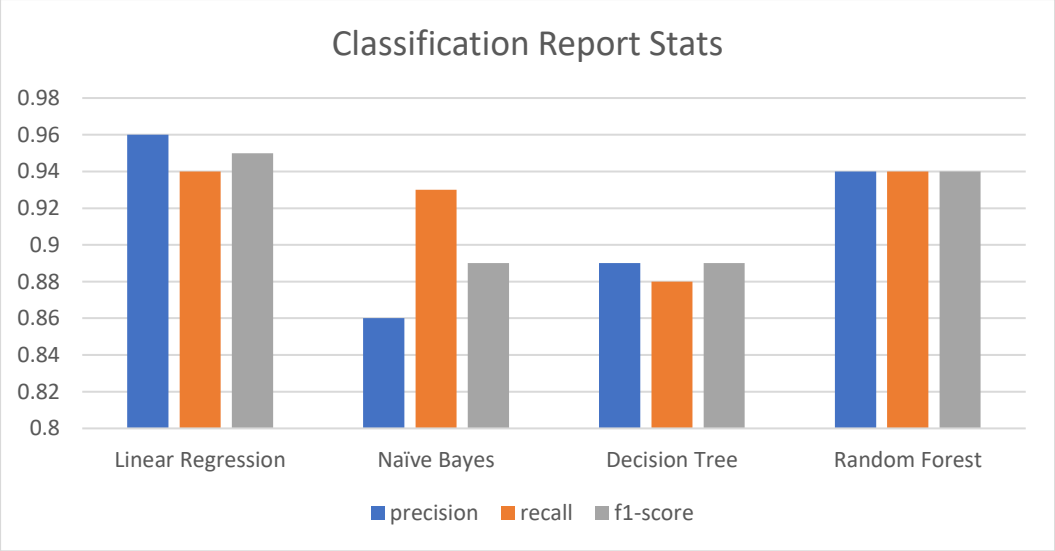
Once the model has been trained and evaluated, it can be deployed in a production environment, where it can be used to automatically classify new articles as they are true or fake. As the dataset keeps changing so does the model, so it is important to regularly update the model with new data and retrain the model if necessary to maintain its performance.

It's worth noting that this is an approach which is followed in this study, different studies may have slight variations in their pipeline or may use different models.

5. Results and Discussion

Table 1: Comparison of various ML Classifiers

ML Classifier	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)
Logistic Regression	96	94	95	94.80
Naïve Bayes	86	93	89	88.69
Decision Tree	89	88	89	88.92
Random Forest	94	94	94	94.19



The line chart below clearly shows the accuracy of all classifiers and here it can be seen the maximum accuracy of the Linear Regression Classifier.

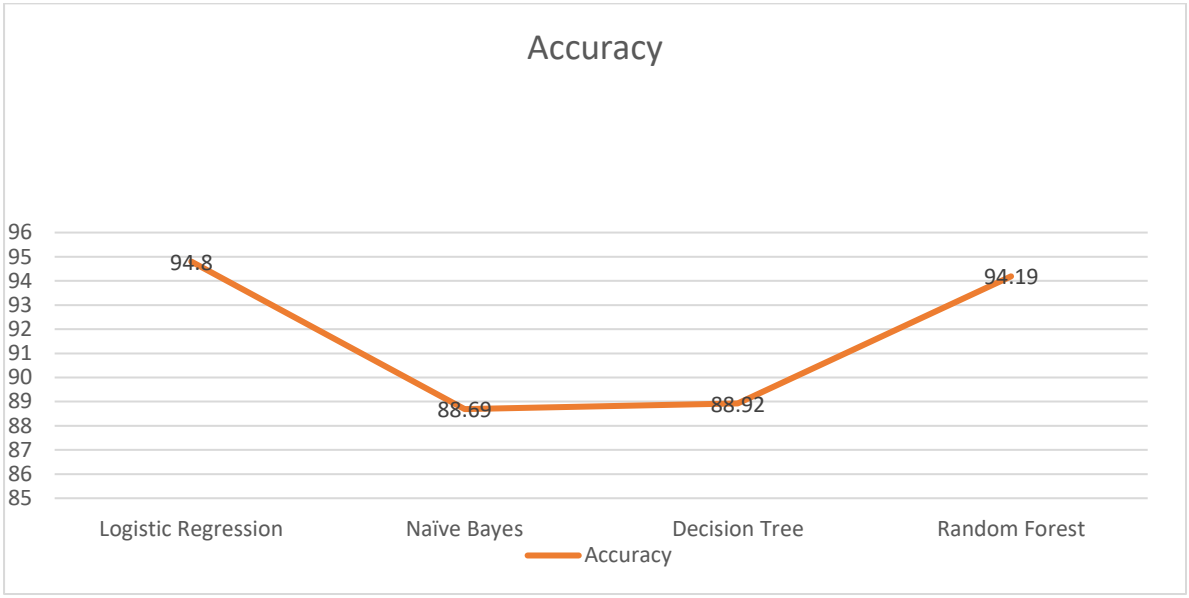


Figure 2: Accuracy of ML Models

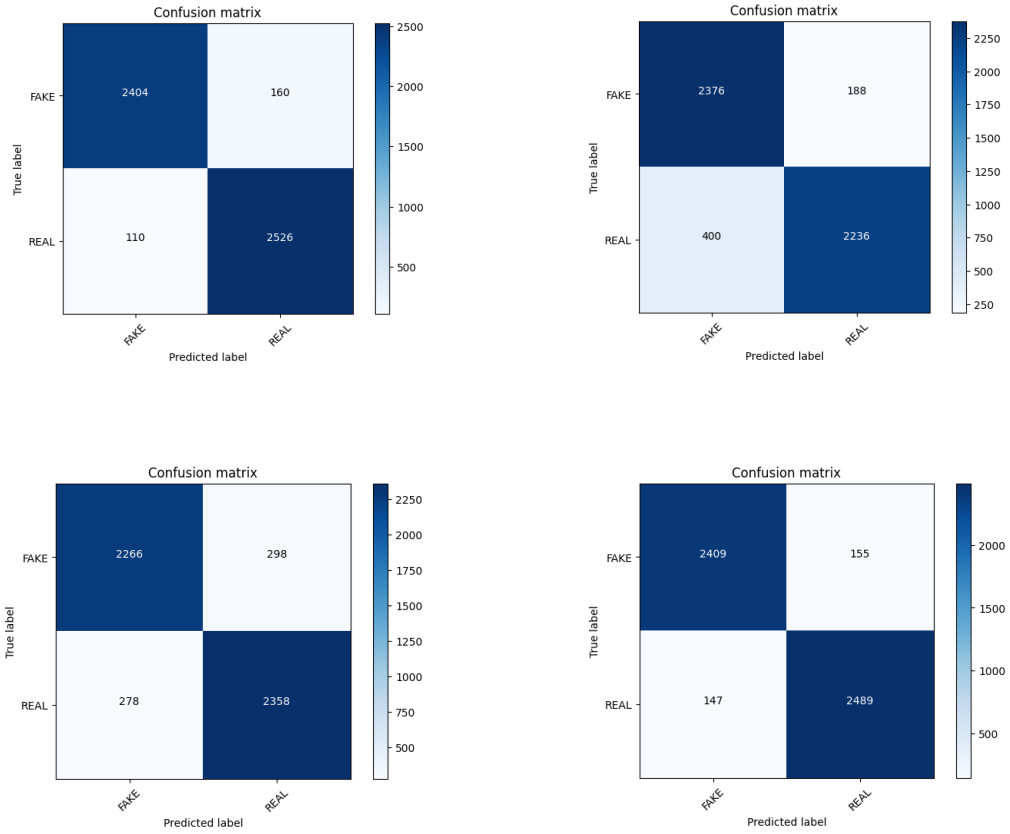


Figure 3: Confusion Matrix of ML Classifiers

6. Conclusion

Fake news detection is a very dynamic topic as every day millions of events occur, and numbers of fake news are spread in the market. In this research a single model is not enough to detect fake news and achieve high accuracy. In this research various classifiers were used from which, Logistic Regression results out as the best classifier which gives an accuracy of 94.8 percent and is very close to it, we see Random Forest with 94.19 percent accuracy and further Decision Tree and Naïve Bayes with 88.92 and 88.69 percent respectively. Logistic Regression gives the best result on our dataset that does not mean it will give the same accuracy on other datasets as well. As it is already said that fake news is a dynamic subject and models need to be regularly trained to maintain their efficiency.

7. References

1) Ahmad, Iftikhar, Muhammad Yousaf, Suhail Yousaf, and Muhammad Ovais Ahmad. "Fake news detection using machine learning ensemble methods." *Complexity* 2020 (2020).
2) Albahr, Abdulaziz, and Marwan Albahar. "An empirical comparison of fake news detection using different machine learning algorithms." International Journal of Advanced Computer Science and Applications 11, no. 9 (2020).
3) Ahmad, Tahir, Muhammad Shahzad Faisal, Atif Rizwan, Reem Alkanhel, Prince Waqas Khan, and Ammar Muthanna. "Efficient Fake News Detection Mechanism Using Enhanced Deep Learning Model." Applied Sciences 12, no. 3 (2022): 1743.
4) Chauhan, Tavishee, and Hemant Palivela. "Optimization and improvement of fake news detection using deep learning approaches for societal benefit." International Journal of Information Management Data Insights 1, no. 2 (2021): 100051.
5) Meesad, Phayung. "Thai fake news detection based on information retrieval, natural language processing and machine learning." SN Computer Science 2, no. 6 (2021): 1-17.
6) Manzoor, Syed Ishfaq, and Jimmy Singla. "Fake news detection using machine learning approaches: A systematic review." In 2019 3rd international conference on trends in electronics and informatics (ICOEI), pp. 230-234. IEEE, 2019.
7) Hiramath, Chaitra K., and G. C. Deshpande. "Fake news detection using deep learning techniques." In 2019 1st International Conference on Advances in Information Technology (ICAIT), pp. 411-415. IEEE, 2019.
8) Jain, Anjali, Avinash Shakya, Harsh Khatter, and Amit Kumar Gupta. "A smart system for fake news detection using machine learning." In 2019 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), vol. 1, pp. 1-4. IEEE, 2019.
9) Khanam, Z., B. N. Alwasel, H. Sirafi, and M. Rashid. "Fake news detection using machine learning approaches." In IOP Conference Series: Materials Science and Engineering, vol. 1099, no. 1, p. 012040. IOP Publishing, 2021.
10) https://www.kaggle.com/c/fake-news/data?select=train.csv
11) Granik, Mykhailo, and Volodymyr Mesyura. "Fake news detection using naive Bayes classifier." In *2017 IEEE first Ukraine conference on electrical and computer engineering (UKRCON)*, pp. 900-903. IEEE, 2017.
12) Jain, Akshay, and Amey Kasbe. "Fake news detection." In *2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*, pp. 1-5. IEEE, 2018.