Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 6, June : 2023

AN APPROACH FOR FAKE NEWS DETECTION USING MACHINE LEARNING THROUGH NLP

[1] Smita Rani Sahu, Assistant Professor, Department of Information Technology,
Aditya Institute of Technology & Management, Tekkali, India.
smitharanisahu.it@adityatekkali.edu.in
[2*] Dr. B.V. Ramana, Professor, Information Technology,
Aditya Institute of Technology & Management, Tekkali, India,
ramana.bendi@gmail.com.
[3] Dr. Kavitha, Assistant Professor, Information Technology,
Aditya Institute of Technology & Management, Tekkali, India,
kavitha.akki@adityatekkali.edu.in
[4] B. Manideep, Computer Science and Engineering,
Aditya Institute of Technology & Management, Tekkali, India,
bendimanideep@gmail.com
[5] Ramya Nuka, Department of Information Technology,
Aditya Institute of Technology & Management, Tekkali, India,
ramyanuka17@gmail.com

ABSTRACT

The time period of faux information elude to reports, images, and movies which can be shared to purposefully unfold fake information. Consumers regularly create or exchange facts on social networking sites, a number of which might be incorrect and do not have any impact. Information that is deemed to be fake news is inaccurate or deceptive that appears as real news. A lot of research is already targeted at detecting it. Fake news needs to be detected and it ought to be stopped earlier than it causes additional harm to the country. Fake news is rapidly increasing on social media within a short period of time. It is difficult to use the algorithms to classify reported works as misleading or false. Even experts in this field must consider a number of factors before determining an object's accuracy. The solution to the fake news problem has been revealed by enforcing a fake news discovery strategy that makes use of numerous categorization techniques. We employed classification algorithms in this model such as Random Forest, support vector machine (SVM), and Logistic Regression. For the purpose of training the advanced system, we advise including a dataset of real and fraudulent news. In accordance with the findings of the confusion matrix, we will put into practise feature selection strategies to assess and select the best feature in order to achieve the highest level of accuracy.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 6, June : 2023

INTRODUCTION

In recent years, fake news is a very common term that we have commonly experienced in our livelihood and mainly in social networks and it can be easily broadcasted with different procedures and purposes like politics, education, and economical problems The spread of fake news is rapidly growing in our society. So, at this time people are spending their too much time for connecting with social media platforms. People are in tough situations to confirm whether they got the information may be correct or may not be correct. It's a difficult factor to know the things which are right and which are not right. With the elements of fake news, fake application links, and fake calls, others were concentrated on obtaining personal information by clicking on fake application links, making it simple to obtain the personal information in this open environment. Some people are primarily concerned with finding people's interest in order to develop their own applications and quickly have their websites promoted by people clicking more often.For example, we are facing the problem with fake calls like they will offer you a huge amount that you got as a lottery or any kind of vehicles that you got as a lucky draw once we accept the link or we said any kind of OTP to them then with a span of seconds they will loot our data with their browsers and we will face the further consequences like report hacked by someone or personal data may be got shared, etc. Fake news means articles or other materials that are intentionally false and intended to manipulate viewers. Social media and the internet suffer from fake accounts, fake calls, fake posts, and fake news. The main goal of fake news is frequently to mislead readers and/or make them believe something that isn't always real. It spreads easily, shares, and discusses issues with familiarity and other users. Thousands of articles containing political and economic interest, a large amount of misinformation or intentional misinformation, it was created online as the most purposed and very fast online news services.So, it helps us to reduce the negative impact of inaccurate information. We must develop tools that can detect the propagation of incorrect information on social media platforms automatically. This work describes how to use supervised machine learning techniques on a labelled dataset that has been manually categorised and guaranteed to build a model that can assess whether an article is fraudulent based on its words, sentences, and sentences. The classification algorithms will make the most accurate predictions after we have applied them to the dataset. They make a distinction between authentic and fraudulent news.We suggested using machine learning techniques like Logistic Regression, Random Forestand Support Vector Machine (SVM) to build the

models.Which will predict the truth and false of an article. We have used datasets containing real or fake messages and got the best results.

RELATED WORK

The main target of our work is to describe the most efficient classification algorithm to identifying dummy news and to evaluate its accuracy. We tested various classification algorithms and used Logistic Regression, Support Vector Machine and Random Forest in our model. One of these three, the Random Forest achieves the highestperformance, but the moment for SVM is high compared to the Logistic Regression. A quick study of literature on the identification of bogus news was done. On the subject of detecting fake news, several studies have been conducted. For the purpose of detecting fake news, various strategies have been used, with varying degrees of success.

Khanam, Z., B. N. Alwasel, H. Sirafi, and M. Rashid[3].The six algorithms for detection are as follows: XGboost, Random Forest, Naïve Bayes, K-Nearest Neighbors, Decision Tree, and Support Vector Machine. XGBoost is depicting with an accuracy of above 75%, followed by Support Vector Machine and Random Forest with an accuracy of over 73%.

Baykara, Muhammet, and AwfAbdulrahman[2].They used algorithms are Random Forest, KNN, Linear SVM, Logistic Regression, AdaBoost, and XGBoost. These were the best methods we used to provide high-quality results, with an average accuracy rating of 91.23%.

Baarir, Nihel Fatima, and Abdelhamid Djeffal[1]. They used Linear Support Vector Machine classifier that gives an accuracy of 92%.

Shaikh, Jasmine, and Rupali Patil[4]. They used the classifiers are Passive Aggressive Classifier, Naïve Bayes Classifier and SVM compare to other algorithms SVM gives the highest accuracy of 95.05%.

Sharma, Uma, SidarthSaran[5]. They used the algorithms are Naïve Bayes, Random Forest, Logistic Regression, Passive Aggressive Classifier. It is concluded that PAC is the best among them with 93% accuracy.

Jain, Anjali, Avinash Shaky[7]. They employed SVM and Naive Bayes methods. The proposed method is effective and can accurately define if an outcome is correct up to 93.6% of the time.

Manzoor, Syed Ishfaq, and Jimmy Singla[6]. They used algorithms are Naïve Bayes, Decision Tree, SVM, Random Forest, and XG Boost.

Khan, JunaedYounus[9]. They used Machine learning models are Naïve Bayes, KNN, SVM, LR, and Random Forest, and Neural Network-Based and Deep Learning Models are Conventional Neural Network, Long Short TermMemory, BiLSTM, and C-LSTM. This paper shows Naïve Bayes has the 95% of accuracy.

Dutta, Pinky Saikia[8]. They compared the outcomes of both classifiers using methods like the Naive Bayes classifier and the Logistic Regression classifier.

Kaliyar, Rohit Kumar[10]. KNN, Naive Bayes, Random Forest, and CNN&LSTM are the algorithms they utilised; when compared to other algorithms, CNN&LSTM had the greatest accuracy.

METHODOLOGY AND IMPLEMENTATION

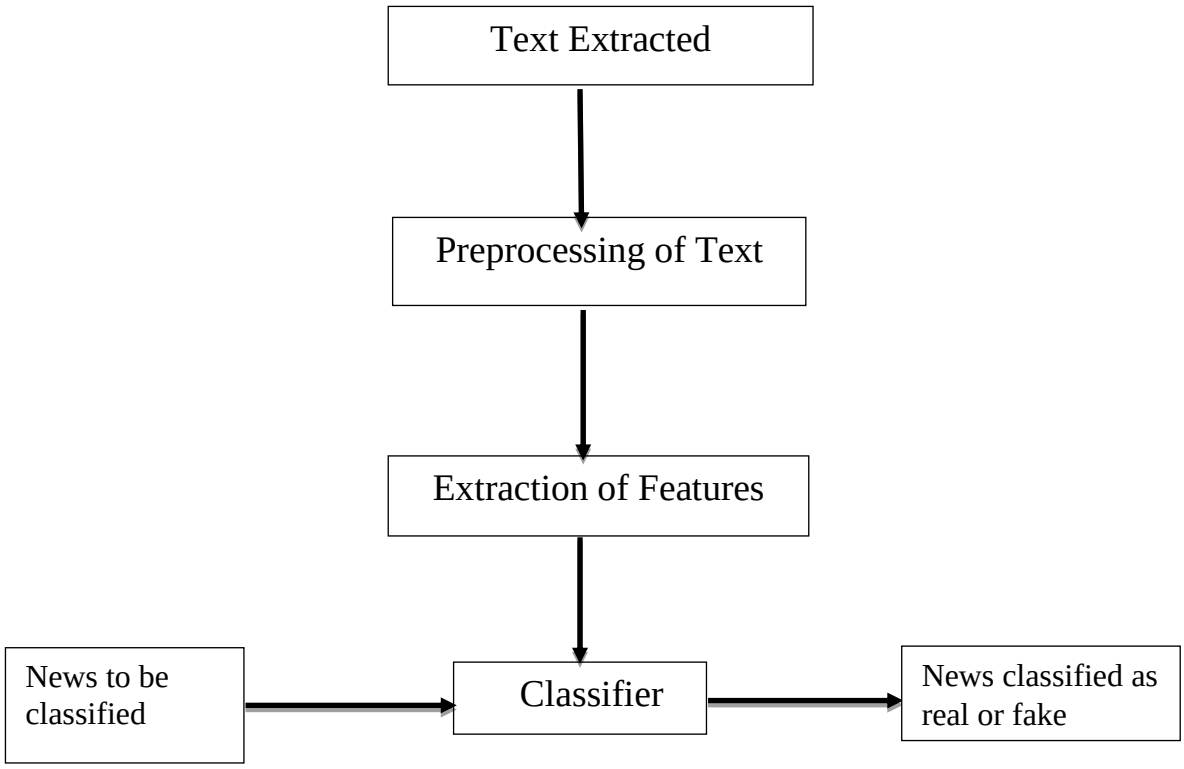

Fig 1. Flowchart for fake news detection

In this section, we have to present the system employed for categorization. A tool is used to identify fake articles using this paradigm. With this method, the dataset is classified using supervised machine learning and converting Textual Information and we have used for natural language processing (NLTK).Text categorization comes first in this classification procedure, then text preparation. This process is used to identify punctuations, news headlines, prepositions, stemming, and Stop Words Removal. Some of the stop words are, an the, a, is, for, etc. Stop Word Removal is an important technique in NLP, it can be processed and filtered. Stemming helps us to reduce the group of words which has similar meanings and works based on rules, such as removing "ing" if words end with "ing". We considered two various feature selection methods termed as Term Frequency Inverted Document Frequency and Term Frequency. TF describes about the significance of a word based on its circumstance in the document. It defines the fashionability of a word. As a result, a vocabulary of phrases is used to describe the paper. In other terms, IDF (Inverse Document Frequency) determines a word's rarity or describes how uncommon a word is. There is an approach for determining the importance of words in a document that is Term Frequency-Inverted Document Frequency(TD-IDF). To start, we clean out any unnecessary or pointless words or characters from the text data. Next is feature extraction using Term Frequency-Inverse Document Frequency. These three distinct algorithms—Support Vector Machine (SVM), Random Forest, and Logistic Regression—were considered. The Natural Language Toolkit for Machines is what we use to put these classifiers into action (NTLK). There were training and test sets created from the dataset. The remaining 80% of the dataset is utilised for training, while just 20% is used for testing.The information we used for Natural Language Processing and the other.

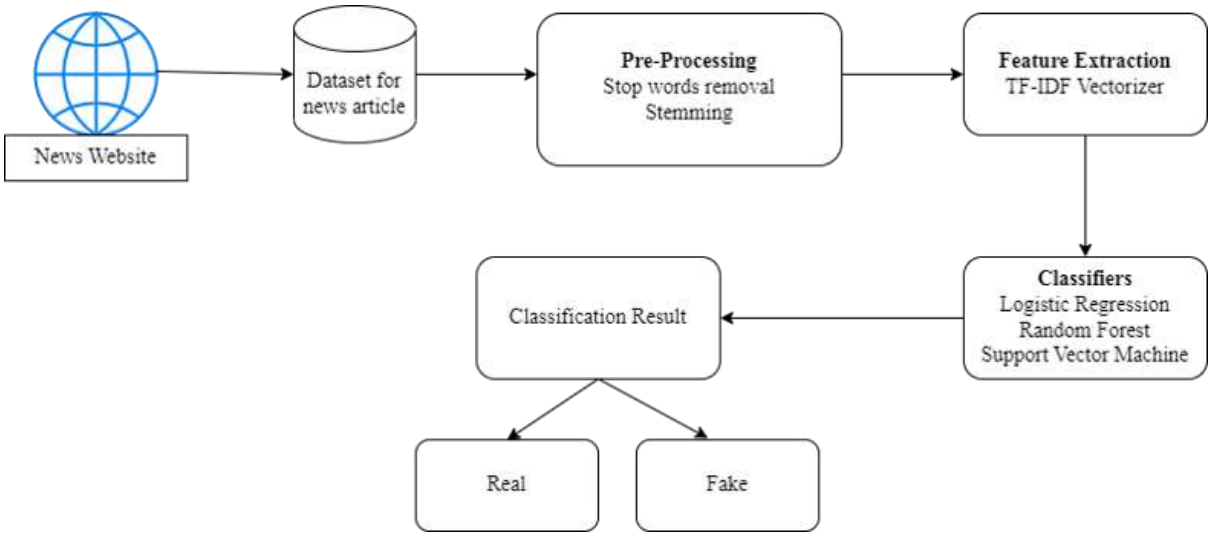

Fig-2:-An Architectural diagram

Dataset and collection:

• A dataset is a collection of data that generally consists of a database table and a single statistical data matrix, where each row represents a dataset member, and each column represents a variable.

• The dataset includes a list of values for each variable, including title, id, author, text, and label.
• We frequently gather various datasets from https://www.kaggle.com.
• The dataset, where the row and column values are (20800*5), contains both genuine and fake news.

TF-IDF: The TF-IDF gauges how frequently a term appears throughout the entire manuscript. In an effort to quantify the occurrence of that phrase, it attempts to provide a metric value. In text mining, this is widely and regularly used. This weight serves as a measurable indicator to assess a word's significance inside a corpus or collection of reports.

$$TF(t) = \frac{Number\ of\ terms\ t\ appears\ in\ the\ document}{Total\ no\ of\ terms\ in\ the\ document}$$

$$IDF(t) = log\left(\frac{Total\ no\ of\ documents}{Number\ of\ documents\ with\ term\ t\ in\ it}\right)$$

$$Thus,\ TD - IDF_{score} = TF * IDF$$

A. Support Vector Machine (SVM):

A support vector machine is one of the controlled learning algorithms (SVM). As a result, after training, the model is produced. Support Vector Machine main objective is to categorise recently collected data. This is the hyperplane or decision boundary that separates the dataset into the two classifications. For the class being considered, a point is chosen that is near to the class being opposed.A line touches the point that is parallel to the hyperplane. The largest margin is considered when designing the hyperplane. Less datasets yield better results for SVMs. The lengthy training process for SVM on large datasets is a disadvantage.

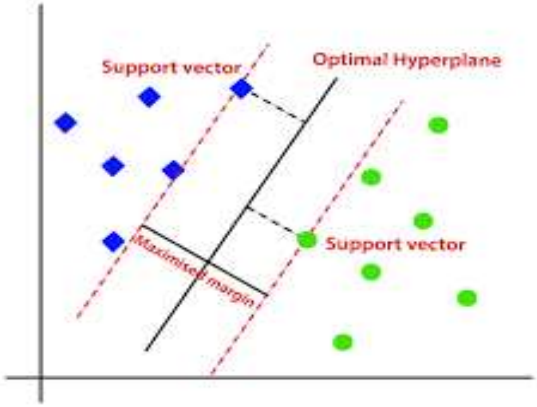

***Fig.2*. employing a hyperplane to divide two categories into subgroups.**

B. Random Forest:

Random Forest is a part of the supervised learning strategy. More trees lead to greater accuracy and fewer replanting worries in a forest. Random Forest is a classifier that, as its name implies, averages the predictions of many decision trees on various subsets of a given data set to increase the predicted accuracy of that data set. Instead of using only one prediction from one decision tree, a random forest forecasts the result based on the prediction

that has the highest support. Artificial intelligence classification and regression issues can be resolved using it.

C. Logistic Regression:

Instead of being a regression algorithm, it is a classification algorithm. It is used to estimate discrete values (binary values like 0/1, yes/no, and true/false) based on a set of independent variables (s). In layman's terms, it establishes the probability of an event by fitting data to a logit function. As a result, it is also known as logit regression. Since it predicts probability, the range of its output value is from zero to one.
The mathematical representation of the result's log probability is a linear combination of the predictor components.

Odds = p/(1-p) = probability of event occurrence / probability of not event occurrence

ln(odds) = ln(p/(1-p))

logit(p)=ln(p/(1-p)) = y0+y1X1+y2X2+y3X3....+ ykXk

Confusion Matrix:

A table known as a confusion matrix is often used to describe how well a classification model performs on a set of test data for which the true results are called. It makes it possible to visualise how well an algorithm performs. The prediction outcomes from the classification task are summarised in a confusion matrix. The number of accurate and wrong predictions for each class is expressed using count values. This is the secret of the confusion matrix.The confusion matrix describes how your classification model produces predictions when it is confused.

Table 1: Confusion Matrix

The following metrics can be defined by classifying this as a classification problem**:**

Total	**label 1 (Predicted)**	**Label 2 (Predicted)**
label 1 (Actual)	*TP*	*FN*
label 2 (Actual)	*FP*	*TN*

1. Precision = $\frac{|TP|}{|TP|+|FP|}$
2. Recall = $\frac{|TP|}{|TP|+|FN|}$
3. F1 Score = $2*\frac{Precision*Recall}{Precision+Recall}$

4. Accuracy = $\frac{|TP|+|TN|}{|TP|+|TN|+|FP|+|FN|}$

Confusion Matrices:

The confusion matrix, which displays the actual set and predicted sets, is shown below after applying several extracted features (TF-IDF) to three distinct classifiers (SVM, Random Forest and Logistic Regression).

Table 2: Support Vector Machine (SVM) confusion matrix table using TF-IDF features:

Total = 20800	**Support Vector Machine (SVM)**	
	Predicted Yes	***Predicted No***
Actual Yes	5120	74
Actual No	38	5168

Table 3:Using TF-IDF features, the confusion matrix table for the Random Forest algorithm –

Total = 20800	**Random Forest**	
	Predicted Yes	***Predicted No***
Actual Yes	5122	72
Actual No	19	5187

Table 4: Using the TF-IDF, the confusion matrix table for logistic regression features –

Total = 20800	**Logistic Regression**	
	Predicted Yes	***Predicted No***
Actual Yes	4899	295
Actual No	44	5162

Table 5: Precision, Recall, F1-score and AccuracyComparison Table for each of the three classifiers:

Classifiers	Precision	Recall	F1-Score	Accuracy
SVM	99%	99%	99%	98.92%
Random Forest	99%	99%	99%	99.125%
Logistic Regression	97%	96%	97%	96.74%

CONCLUSION

Online news articles and apps like Facebook and Twitter are progressively taking the role of newspapers that were once preferred as hardcopies. The WhatsApp forwards are one such excellent resource. Fake news is a growing concern that makes issues more difficult and alters or discourages individuals from using digital technologies. Our Fake News Detection system, which accepts user input and categorises it as real or false, was created to stop the problem. To do this, a variety of NLP and machine learning approaches must be used. The classification algorithms predict the ratio of fake news to real news. After the result is known, the algorithm will be able to identify whether or not an article is genuine.When compared to Logistic Regression, Support Vector Machine and Random Forest are the most accurate techniques.

REFERENCES

[1]Baarir, Nihel Fatima, and Abdelhamid Djeffal. "Fake news detection using machine learning." In 2020 2nd International Workshop on Human-Centric Smart Environments for Health and Well-being (IHSH), pp. 125-130. IEEE, 2021.

[2]Abdulrahman, Awf, and MuhammetBaykara. "Fake news detection using machine learning and deep learning algorithms." In 2020 International Conference on Advanced Science and Engineering (ICOASE), pp. 18-23. IEEE, 2020.

[3]Khanam, Z., B. N. Alwasel, H. Sirafi, and Mamoon 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.

[4]Shaikh, Jasmine, and Rupali Patil. "Fake news detection using machine learning." In 2020 IEEE International Symposium on Sustainable Energy, Signal Processing and Cyber Security (iSSSC), pp. 1-5. IEEE, 2020.

[5]Sharma, Uma, Sidarth Saran, and Shankar M. Patil. "Fake news detection using machine learning algorithms." International Journal of Creative Research Thoughts (IJCRT) 8, no. 6 (2020): 509-518.

[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]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.

[8]Dutta, Pinky Saikia, Meghasmita Das, Sumedha Biswas, Mriganka Bora, and Sankar Swami Saikia. "Fake news prediction: a survey." International Journal of Scientific Engineering and Science 3, no. 3 (2019): 1-3.

[9]Khan, Junaed Younus, Md Khondaker, Tawkat Islam, Anindya Iqbal, and Sadia Afroz. "A benchmark study on machine learning methods for fake news detection." arXiv preprint arXiv:1905.04749 (2019): 1-14.

[10]Kaliyar, Rohit Kumar. "Fake news detection using a deep neural network." In 2018 4th International Conference on Computing Communication and Automation (ICCCA), pp. 1-7. IEEE, 2018.