IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 6, June 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.165

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006115 |

Fake News Detection Using Machine Learning

Boobala Muralitharan D[1], Priyanka P[2], Reshma Banu L[3], Shifa Shujauddin[4], Vidhyabharathi V[5]

Professor, Department of Computer Science and Engineering, Saranathan College of Engineering, Trichy, India[1]

Student, Department of Computer Science and Engineering, Saranathan College of Engineering, Trichy, India[2]

Student, Department of Computer Science and Engineering, Saranathan College of Engineering, Trichy, India[3]

Student, Department of Computer Science and Engineering, Saranathan College of Engineering, Trichy, India[4]

Student, Department of Computer Science and Engineering, Saranathan College of Engineering, Trichy, India[5]

ABSTRACT: The easy access and exponential growth of the information available on social media networks has made it intricate to distinguish between false and true information. The easy dissemination of information by way of sharing has added to exponential growth of its falsification. The credibility of social media networks is also at stake where the spreading of fake information is prevalent. Thus the goal of this project is to create a tool for detecting the language patterns that characterize fake and real news through the use of machine learning and natural language processing techniques. Machine learning algorithms are used for identification of fake news. The following four classifiers are applied: Logistic Regression, Random Forest, Decision Tree and Gradient Boosting Algorithm. Simple classification is not completely correct in fake news detection because classification methods are not specialized for fake news. With the integration of machine learning and text-based processing, we can detect fake news and build classifiers that can classify the news data. Text classification mainly focuses on extracting various features of text and after that incorporating those features into classification. The big challenge in this area is the lack of an efficient way to differentiate between fake and non-fake due to the unavailability of corpora. Experimental analysis based on the existing dataset indicates a very encouraging and improved performance.

KEYWORDS: Machine Learning, Natural Language Processing, Logistic Regression, Decision Tree, Gradient Boosting, Random Forest.

I. INTRODUCTION

News consumption is a double-edged sword. On the one hand, its low cost, easy access, and rapid dissemination of information lead people to seek out and consume news. It enables the wide spread of “fake news”, i.e., low quality news with intentionally false information. The extensive spread of fake news has the potential for extremely negative impacts on individuals and society. Therefore, fake news detection has recently become an emerging research that is attracting tremendous attention. First, fake news is intentionally written to mislead readers to believe false information, which makes it difficult and nontrivial to detect based on news content.Therefore this project aims to develop a Fake News Detection system using Machine Learning and Natural Language Processing. Natural Language Processing(NLP) is an area of computer science and Artificial Intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program computers to fruitfully process large amounts of natural language data. It encompasses techniques that can utilize text, create models and produce predictions.Natural language processing (NLP) is a subfield of linguistics, computer science, information engineering, and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyse large amounts of natural language data. Its accuracy will be tested using machine learning algorithms. The algorithm or the machine learning models are used to study the data of past news reports and predict the chances of a news report being fake or not. These models must be able to detect fake news in a given scenario. This project is motivated by the widespread problem of fake news. It has become a greater issue because of the advancements in AI which brings along artificial bots thatmay be used to create and spread fake news.The situation is dire because many people believe anything they read on the internet and the ones who are amateur or are new to the digital technology may be easily fooled. A similar problem is fraud that may happen due to spam or malicious emails and messages. So, it is compelling enough acknowledge this problem take on this challenge to control the rates of crime, political unrest, grief, and thwart the attempts of spreading fake news.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006115 |

II. RELATED WORK

In [1] the research surveys the current state-of-the-art technologies that are instrumental in the adoption and development of fake news detection. "Fake news detection" is defined as the task of categorizing news along a continuum of veracity, with an associated measure of certainty. Veracity is compromised by the occurrence of intentional deceptions. The nature of online news publication has changed, such that traditional fact checking and vetting from potential deception is impossible against the flood arising from content generators, as well as various formats and genres. The paper provides a typology of several varieties of veracity assessment methods emerging from two major categories -- linguistic cue approaches (with machine learning), and network analysis approaches. We see promise in an innovative hybrid approach that combines linguistic cue and machine learning, with network-based behavioral data. Although designing a fake news detector is not a straightforward problem, we propose operational guidelines for a feasible fake news detecting system.

News currently spreads rapidly through the internet. Because fake news stories are designed to attract readers, they tend to spread faster. For most readers, detecting fake news can be challenging and such readers usually end up believing that the fake news story is fact. Because fake news can be socially problematic, a model that automatically detects such fake news is required. In [2], we focus on data-driven automatic fake news detection methods. We first apply the Bidirectional Encoder Representations from Transformers model (BERT) model to detect fake news by analyzing the relationship between the headline and the body text of news. To further improve performance, additional news data are gathered and used to pre-train this model. We determine that the deep-contextualizing nature of BERT is best suited for this task and improves the 0.14 F-score over older state-of-the-art models.

In [3], to detect fake news on social media, a data mining perspective is presented that includes the characterization of fake news in psychology and social theories. This article looks at two main factors responsible for the widespread acceptance of fake messages by the user which is naive realism and confirmatory bias. It proposes a general two-phase data mining framework that includes feature extraction and modeling, analyzing data sets, and confusion matrix for detecting fake news.
In [4], the study uses Artificial Intelligence, Natural Language Processing, and Machine Learning techniques to conduct binary categorization of diverse news items available online. We want to give users the ability to classify news as fake or real, as well as verify the legitimacy of the website that published it. This exploration proposed using one AI calculations (choice tree) to distinguish the phony news. In this paper, the full dataset size rises to 20,761 examples, while the testing test size approaches 4,345 examples. The preprocessing steps start with cleaning information by eliminating pointless unique characters, numbers, English letters, and void areas, lastly, eliminating stop words is carried out. From that point forward, the most famous element extraction technique (TF-IDF) is utilized prior to applying the two proposed characterization calculations. The outcomes show that the best exactness accomplished approaches 98.11% utilizing the choice tree model.

In [5], the author introduced the concept of the importance of NLP in stumbling across incorrect information. They have used time frequency-inverse document frequency (TF-IDF) of bigrams and probabilistic context-free grammar detection. Shloka Gilda introduced the concept of the importance of NLP in stumbling over incorrect information. They used Bi-Gram Count Vectorizer and Probabilistic Context-Free Grammar (PCFG) to detect deceptions. They examined the data set in more than one class of algorithms to find out a better model. The count vectorizer of bi-grams fed directly into a stochastic gradient descent model which identifies noncredible resources with an accuracy of 71.2%.

III. SYSTEM ARCHITECTURE

We have created fake news detection model based on text of the news article. The proposed system of Fake News Detection and Identification is solely based on the content of the news articles. Natural Language processing is used to derive information from the massive amount of text used as input. The data is cleaned and prepared for machine leaning models to be trained through Data Preprocessing. Exploratory Data Analysis is carried on the data, involving statistics and visualizations to analyze and identify trends in the data set. The pre-processed data is converted into numeric value in the form of vectors using two different vectorization techniques: Bag of Words (Count vectorizer) and Term Frequency-Inverse Document Frequency (TF-IDF Vectorizer). There are pre-training algorithms available in the NLP toolkit, which has been utilized in the project. It consists of four different classifiers, each having an accuracy level of over ninety percentage. The motive of this project is to increase the accuracy of detecting fake news more than the present results that are available.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006115 |

Input is collected from various sources such as newspapers , social media and stored in datasets. System will take input from datasets. The datasets undergo preprocessing and the unnecessary information is removed from it and the data types of the columns are changed if required. Jupyter notebook and python libraries are used in the above step. For fake news detection , we have to train the system using dataset. Before entering to the detection of fake news , entire dataset is divide into two datasets. 80% is used for training and 20% is used for testing. During training , the model using the train dataset. In testing , the test dataset is given as input and the output is predicted. After the testing time , The predicted output and the actual output are compared using confusion matrix obtained.

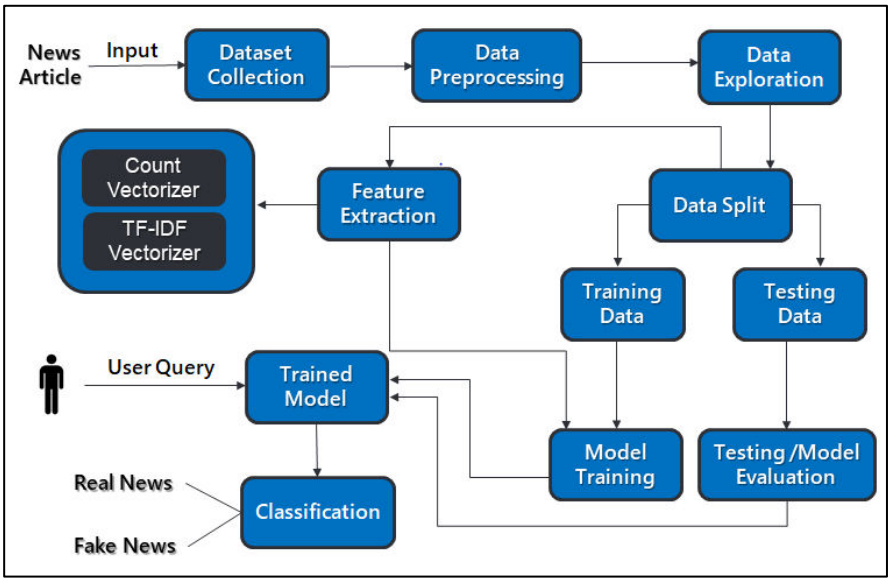

Fig.1 System Architecture Diagram

IV. SYSTEM IMPLEMENTATION

1. DATA COLLECTION

To work with Fake News Detection project using Machine Learning, a huge amount of Data is required to train the Machine learning models or classifiers. Collecting and preparing the dataset is one of the most crucial parts while creating a fake news detection model. The technology applied behind it cannot work properly if the dataset is not well prepared and pre-processed. The dataset for this project is taken from Kaggle. This data science site contains a diverse set of compelling, independently-contributed datasets for machine learning. The dataset that has been used is explained below:

The dataset used for this project are in csv format named fake.csv, true.csv.

It contains 5 columns viz

1- Title
2- Text
3- Subject
4- Date
5- Label (Fake/True)

2. PRE- PROCESSING

Data Social media data is highly unstructured – majority of them are informal communication with typos, slangs and bad-grammar etc. quest for increased performance and reliability has made it imperative to develop techniques for utilization of resources to make informed decisions. To achieve better insights, it is necessary to clean the data before it can be used for predictive modeling. For this purpose, basic pre-processing was done on the News training data. This step was comprised of data cleaning. While reading data, we get data in the structured or unstructured format. A structured format has a well-defined pattern whereas unstructured data has no proper structure. In between the 2 structures, we have a semi-structured format which is a comparably better structured than unstructured format.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006115 |

Cleaning (or pre-processing) the data typically consists of a number of steps:

Step 1: Data Integration
This project derives a large dataset by integrating two publicly available datasets – Fake and real news, each taken separately. There are 40836 labeled news articles in the dataset after integration, which contains the entries of Fake and True articles in roughly equal numbers.

Step 2: Dimensionality Reduction
Dimension Reduction techniques ensure the integrity of data while reducing the data. Data reduction is a process that reduces the volume of original data and represents it in a much smaller volume. Data reduction techniques are used to obtain a reduced representation of the dataset that is much smaller in volume by maintaining the integrity of the original data. By reducing the data, the efficiency of the data mining process is improved, which produces the same analytical results. The reduction of the data may be in terms of the number of rows (records) or terms of the number of columns (dimensions). All the columns which is not used for analysis and prediction, such as the "date" and "title" column is removed from the dataset.

Step 3: Punctuation Removal
Punctuation can provide grammatical context to a sentence which supports our understanding. But for our vectorizer which counts the number of words and not the context, it does not add value, so we remove all special characters.

Step 4: Tokenization
Tokenization is the process of breaking up the paragraph into smaller units such as sentences or words. Each unit is then considered as an individual token. The fundamental principle of Tokenization is to try to understand the meaning of the text by analyzing the smaller units or tokens that constitute the paragraph. To do this, the NLTK library is used. NLTK is the Natural Language Toolkit library in python that is used for Text Preprocessing.

Step 5: Stopwords Removal
Stop words are a collection of words that occur frequently in any language but do not add much meaning to the sentences. These are common words that are part of the grammar of any language. They don't tell us much about our data so we remove them. Every language has its own set of stop words. For example, some of the English stop words are "the", "he", "him" etc. To remove them, stopwords are imported from nltk.corpus. This helps in dimensionality reduction by eliminating unnecessary information.

Step 6: Stemming
Punctuation Stemming helps reduce a word to its stem form. It often makes sense to treat related words in the same way. It is the process of reduction of a word into its root or stem word. It removes suffices, like "ing", "ly", "s", etc. by a simple rule-based approach. It reduces the corpus of words but often the actual words get neglected. The word affixes are removed leaving behind only the root form or lemma. To do this first PorterStemmer is imported from nltk.stem and an object of the PorterStemmer class is created. After that using the PorterStemmer object the stem method is called to perform stemming on our wordlist. After implementing the stemming function all the words in our list have been reduced to their stem words or lemma.

Step 7: Lemmatization
Stemming does not always result in words that are part of the language vocabulary. It often results in words that have no meaning to the users. In order to overcome this drawback, the concept of Lemmatization is used. After the process of Lemmatization is applied we see that each word is converted into a meaningful parent word. Thus we observe that we can accurately specify the context of lemmatization by passing in the desired parts of speech in the parameter of the lemmatize method.

V. EXPLORATORY DATA ANALYSIS

The process of Data Exploration is unavoidable and one of the major step to fine-tune the given data set(s) in a different form of analysis to understand the insights of the key characteristics of various entities of the data set like columns, rows by applying Pandas, NumPy, Statistical Methods, and Data visualization packages. Exploratory Data Analysis is a data analytics process to understand the data in depth and learn the different data characteristics, often with visual means. This allows one to get a better feel of the data and find useful patterns in it. It is crucial to

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006115 |

understand it in depth before data analysis is performed and data is run through an algorithm. The patterns in the data need to be identified and determine which variables are important and which do not play a significant role in the output. Further, some variables may have correlations with other variables. You also need to recognize errors in your data. All of this can be done with Exploratory Data Analysis. It helps us gather insights and make better sense of the data, and removes irregularities and unnecessary values from data.

```
subject
Government News     1570
Middle-east          778
News                9050
US_News              783
left-news           4459
politics            6841
politicsNews       11272
worldnews          10145
Name: text, dtype: int64
```

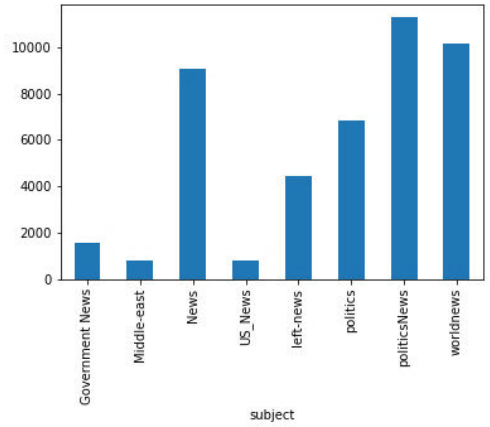

Fig.2Subject- wise Article Count

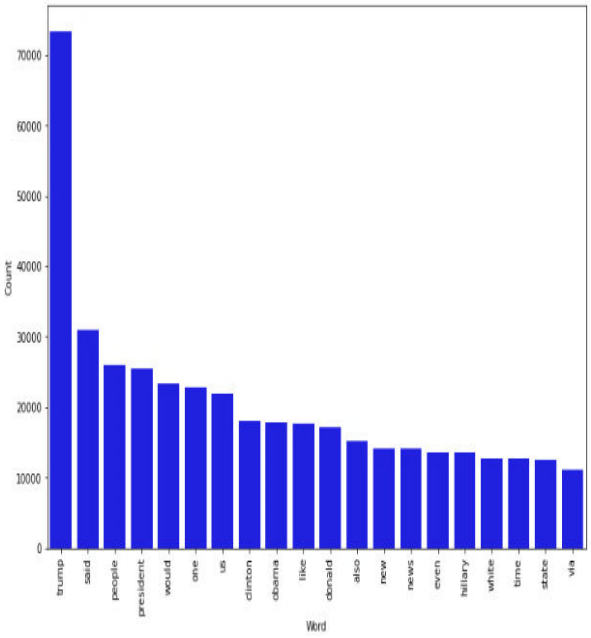

Fig.3 Most Frequent Words Count Graph

```
target
fake    23481
true    21417
Name: text, dtype: int64
```

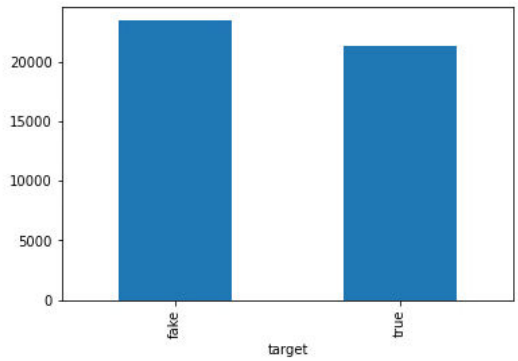

Fig.4Fake and True News Count

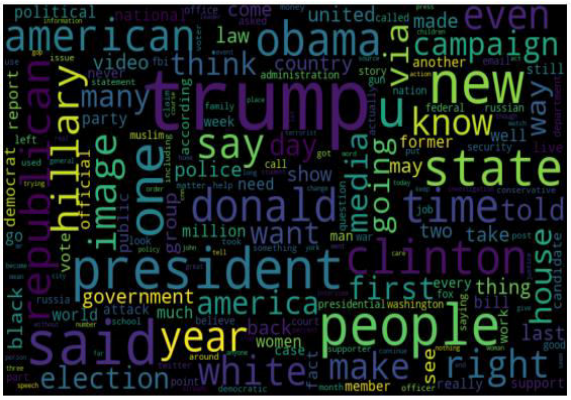

Fig.5 Word Cloud

VI. FEATURE EXTRACTION

A number of features can be generated using text data like word count, frequency of large words, frequency of unique words, n-grams etc. By creating a representation of words that capture their meanings, semantic relationships, and numerous types of context they are used in, we can enable computer to understand text and perform Clustering, Classification etc.

Vectorizing is the process of encoding text as integers i.e. numeric form to create feature vectors so that machine learning algorithms can understand our data. This project has implemented two types of vectorization techniques.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006115 |

i. Count Vectorization

Bag of Words (BoW) or CountVectorizer describes the presence of words within the text data. It gives a result of 1 if present in the sentence and 0 if not present. It, therefore, creates a bag of words with a document-matrix count in eachtext document.

	also	campaign	clinton	could	country	donald	donald trump	election	even	first	...
0	1	0	0	1	0	0	0	0	0	0	...
1	0	0	0	0	1	1	1	0	0	0	...
2	0	0	0	2	0	1	1	1	1	1	...
3	6	1	10	0	0	0	0	0	0	0	...
4	0	0	0	0	0	0	0	2	0	0	...
...	...	...	...	...	...	...	...	...	...	...	...
35913	0	0	0	0	0	0	0	0	0	0	...
35914	0	0	0	0	0	0	0	0	0	0	...
35915	1	0	0	0	0	3	3	1	0	0	...
35916	3	0	0	2	1	1	1	0	2	1	...
35917	0	0	0	0	0	0	0	0	0	0	...

Fig.6Document-Term Matrix for Count Vectorizer

ii. TF- IDF Vectorization

It computes “relative frequency” that a word appears in a document compared to its frequency across all documents TF-IDF weight represents the relative importance of a term in the document and entire corpus. TF stands for Term Frequency: It calculates how frequently a term appears in a document. Since, every document size varies, a term may appear more in a long sized document that a short one.Thus, the length of the document often divides Term frequency.
IDF stands for Inverse Document Frequency: A word is not of much use if it is present in all the documents. Certain terms appear many times in a document but are of little importance. IDF weighs down the importance of these terms and increase the importance of rare ones. The more the value of IDF, the more unique is the word. TF-IDF is applied on the body text, so the relative count of each word in the sentences is stored in the document matrix.

$$TF(t,d) = Number\ of\ times\ t\ occurs\ in\ document\ 'd' / Total\ word\ count\ of\ document\ 'd'$$
$$IDF(t,d) = Total\ number\ of\ documents / Number\ of\ documents\ with\ term\ t\ in\ it$$
$$TFIDF(t,d) = TF(t,d) * IDF(t)$$

	also	campaign	clinton	could	country	donald	donald trump	election	even	first	...
0	0.24	0.00	0.00	0.28	0.00	0.00	0.00	0.00	0.00	0.00	...
1	0.00	0.00	0.00	0.00	0.14	0.11	0.12	0.00	0.00	0.00	...
2	0.00	0.00	0.00	0.13	0.00	0.06	0.06	0.08	0.07	0.07	...
3	0.28	0.06	0.71	0.00	0.00	0.00	0.00	0.00	0.00	0.00	...
4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.56	0.00	0.00	...
...	...	...	...	...	...	...	...	...	...	...	...
35913	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	...
35914	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	...
35915	0.10	0.00	0.00	0.00	0.00	0.31	0.32	0.14	0.00	0.00	...
35916	0.13	0.00	0.00	0.10	0.05	0.04	0.04	0.00	0.11	0.05	...
35917	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	...

Fig.7 Document-Term Matrix for TF-IDF Vectorizer

VII. CLASSIFICATION ALGORITHMS

This section deals with training the classifier. Different classifiers were investigated to predict the class of the text. We explored specifically four different machine-learning algorithms: Logistic regression, Decision Tree, Random Forest

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006115 |

and Gradient Boosting Algorithm. The implementations of these classifiers were done using Python library Sci-Kit Learn.

i. Logistic Regression

Logistic regression is one of the most popular Machine Learning algorithms. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. In Fake News Detection the output is either Fake or True. Instead of giving the exact value as Fake or True, it gives the probabilistic values which lie between Fake and True. An 'S' shaped logistic function called the Sigmoid function, predicts two maximum values (Fake or True). The curve from the logistic function indicates the likelihood of the news being either Fake or True based on the article content.

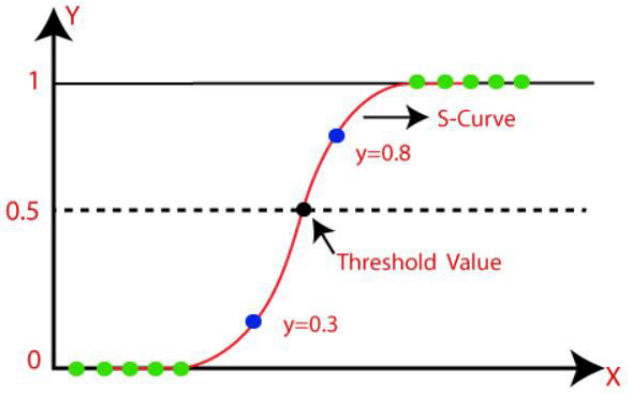

Fig.8 Logistic Regression Curve

The final equation for Logistic Regression:

$$log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

ii. Random Forest

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. It contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output as either Fake or True. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

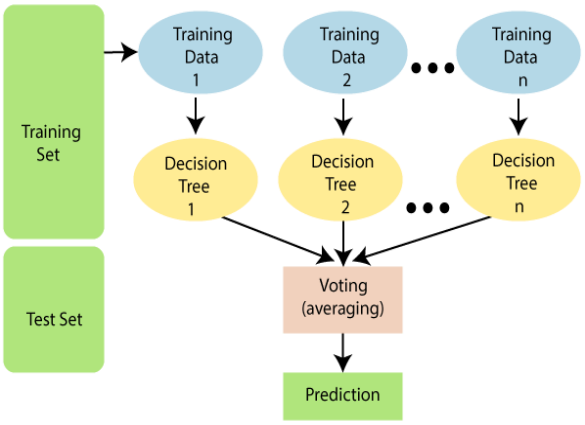

Fig.9Random Forest Classifier

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006115 |

iii. Decision Tree

Decision Tree is a Supervised learning technique. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. The decisions or the test are performed on the basis of features of the given dataset. It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions. It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure. In order to build a tree, we use the Classification and Regression Tree algorithm. A decision tree simply asks a yes or no question, and based on the answer , it further split the tree into subtrees.

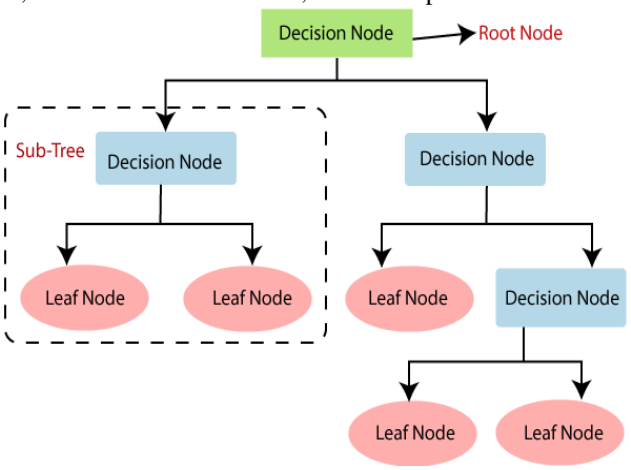

Fig.10Decision Tree Classifier

iv. Gradient Boosting

Gradient Boosting is one of the most popular machine learning algorithms for tabular datasets. It is powerful enough to find any nonlinear relationship between your model target and features and has great usability that can deal with missing values, outliers, and high cardinality categorical values on your features without any special treatment. A GBM combines distinct decision trees' predictions to bring out the final predictions. These dissimilar decision trees capture the different information from the data. The nodes in each decision tree take a distinct subset of the features for picking out the best split. This signifies that actually these decision trees aren't all identical and therefore they are able to capture distinct signals from the data. Moreover, every tree considers the errors made by the previous decision tree. Hence, every successor tree is made on the error of the previous tree.

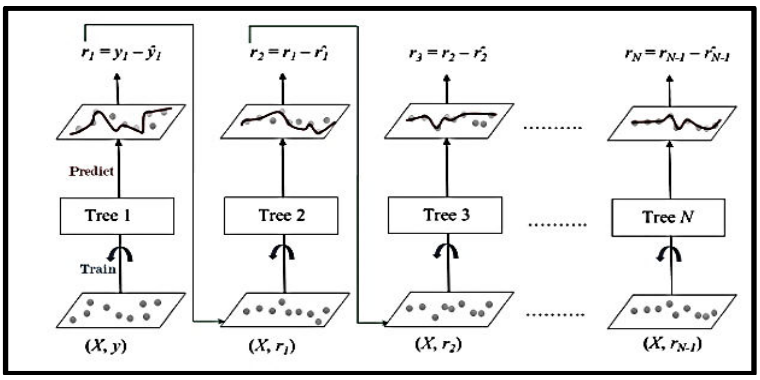

Fig.11 Gradient Boosting Classifier

VIII. MODEL EVALUATION

Model Evaluation is the process through which we quantify the quality of a system's predictions. We expect machine learning models to provide accurate and trustworthy predictions. It is important to assess how machine learning models generalize on test data. We must always test how a model generalizes on unseen data. To do this, we

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006115 |

measure the newly trained model performance on a new and independent dataset. This model will compare labeled data with it's own predictions. Many learning algorithms have been proposed. In order to understand the relative merits of these alternatives, it is necessary to evaluate them. The primary approaches to evaluation can be characterized as either theoretical or experimental. Theoretical evaluation uses formal methods to infer properties of the algorithm, such as its computational complexity and also employs the tools of computational learning theory to assess learning theoretic properties. Experimental evaluation applies the algorithm to learning tasks to study its performance in practice.

IX. RESULTS AND DISCUSSION

1. EVALUATION METRICS

Evaluate the performance of algorithms for fake news detection problem; various evaluation metrics have been used. In this subsection, we review the most widely used metrics for fake news detection. Most existing approaches consider the fake news problem as a classification problem that predicts whether a news article is fake or not:

• True Positive (TP): when predicted fake news pieces are actually classified as fake news;
• True Negative (TN): when predicted true news pieces are actually classified as true news;
• False Negative (FN): when predicted true news pieces are actually classified as fake news;
• False Positive (FP): when predicted fake news pieces are actually classified as true news.

2. CONFUSION MATRIX

Confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. It allows the visualization of the performance of an algorithm. It is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class. This is the key to the confusion matrix. The confusion matrix shows the ways in which your classification model is confused when it makes predictions. It gives us insight not only into the errors being made by a classifier but more importantly the types of errors that are being made.

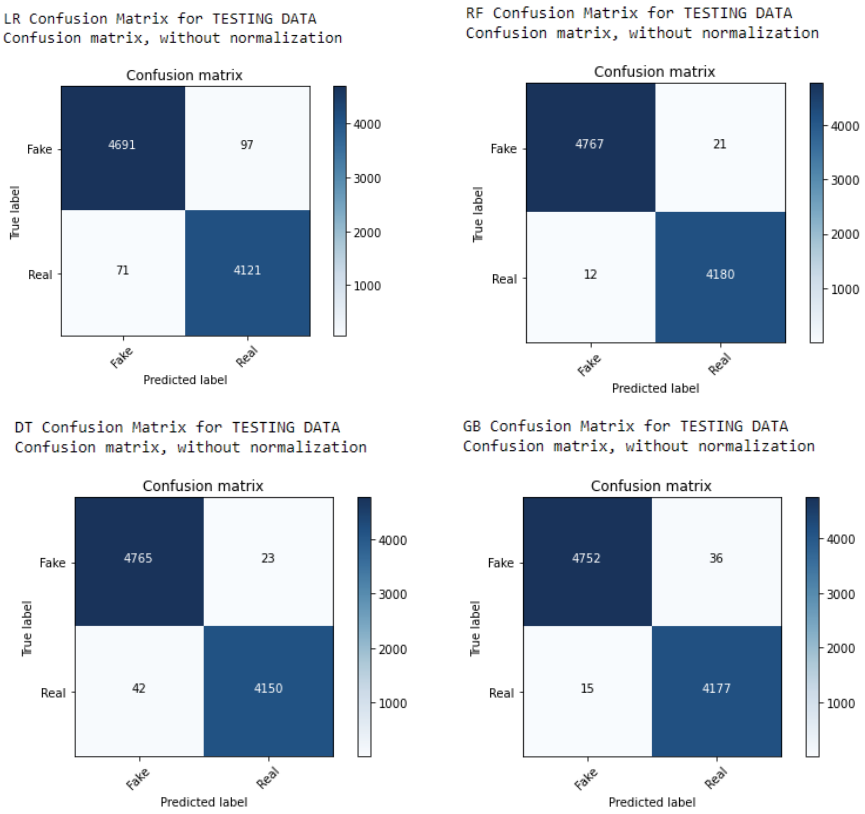

Fig.12 Confusion Matrices for each Algorithm

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006115 |

X. CLASSIFICATION REPORT

It is one of the performance evaluation metrics of a classification-based machine learning model. It displays your model's precision, recall, F1 score and support. It provides a better understanding of the overall performance of our trained model.True Positives, False Positives, True negatives and False Negatives are used to predict the metrics of a classification report To understand the classification report of a machine learning model, it is needed to know all of the metrics displayed in the report.

i. Precision

Precision is defined as the ratio of true positives to the sum of true and false positives.

$$\text{Precision} = \frac{TP}{TP + FP}$$

ii. Recall

Recall is defined as the ratio of true positives to the sum of true positives and false negatives.

$$\text{Recall} = \frac{TP}{TP + FN}$$

iii. F1 Score

The F1 is the weighted harmonic mean of precision and recall. The closer the value of the F1 score is to 1.0, the better the expected performance of the model is expected to be.

$$F1 = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

iv. Support

Support is the number of actual occurrences of the class in the dataset. It doesn't vary between models, it just diagnoses the performance evaluation process.

LR Classification Report for TESTING DATA

	precision	recall	f1-score	support
fake	0.99	0.98	0.98	4788
true	0.98	0.98	0.98	4192
accuracy			0.98	8980
macro avg	0.98	0.98	0.98	8980
weighted avg	0.98	0.98	0.98	8980

RF Classification Report for TESTING DATA

	precision	recall	f1-score	support
fake	1.00	1.00	1.00	4788
true	1.00	1.00	1.00	4192
accuracy			1.00	8980
macro avg	1.00	1.00	1.00	8980
weighted avg	1.00	1.00	1.00	8980

DT Classification Report for TESTING DATA

	precision	recall	f1-score	support
fake	0.99	1.00	0.99	4788
true	0.99	0.99	0.99	4192
accuracy			0.99	8980
macro avg	0.99	0.99	0.99	8980
weighted avg	0.99	0.99	0.99	8980

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006115 |

GB Classification Report for TESTING DATA

	precision	recall	f1-score	support
fake	1.00	0.99	0.99	4788
true	0.99	1.00	0.99	4192
accuracy			0.99	8980
macro avg	0.99	0.99	0.99	8980
weighted avg	0.99	0.99	0.99	8980

Fig.13 Classification Report for each Algorithm

XI. ACCURACY

Machine learning model accuracy is the measurement used to determine which model is best at identifying relationships and patterns between variables in a dataset based on the input, or training, data. The better a model can generalize to 'unseen' data, the better predictions and insights it can produce, which in turn deliver more business value. The cost of errors can be huge, but optimizing model accuracy mitigates that cost.

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + TN + FP}$$

Classifier	Accuracy	
	Training Data	Testing Data
Logistic Regression	99.99	99.63
Random Forest	99.42	99.12
Decision Tree	99.99	99.27
Gradient Boosting	99.66	99.43

Table.1 Accuracy of each Algorithm

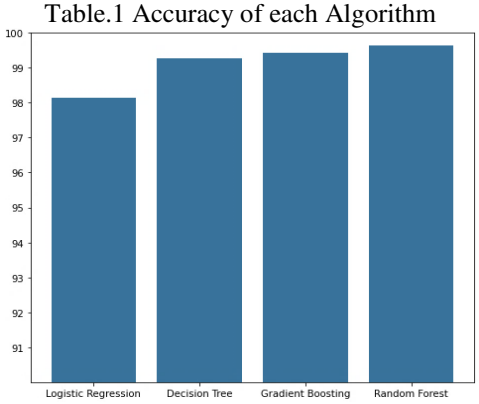

Fig.14 Comparison of the Accuracy of Models

XII. CONCLUSION AND FUTURE WORK

Counterfeit News is one of the most well-known peculiarities that significantly affect our public activity, particularly in the political space. These days, making counterfeit news turns out to be extremely simple in light of clients' far and wide utilizing the web and online entertainment. Accordingly, the recognition of slipperiness news is a urgent issue that should be significant essentially due to its difficulties like the restricted measure of the benchmark datasets and how much the distributed news consistently. This exploration proposed using one AI calculations to distinguish the phony news. In this paper, the full dataset size rises to 20,761 examples, while the testing test size approaches 4,345 examples. The preprocessing steps start with cleaning information by eliminating pointless unique characters, numbers, English letters, and void areas, lastly, eliminating stop words is carried out. From that point forward, the most famous element extraction technique (TF-IDF) is utilized prior to applying the two proposed characterization calculations. The

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006115 |

outcomes show that the best exactness accomplished approaches 98.11% utilizing the decision tree model.
To implement in real life, we can design a application or website, where users could enter the links of news or copy paste the news. We can add integrated feature in social media platforms. In the future, this model could be improved a lot using more features which could also tackle not only fake news articles but also tackle rumours spread by individuals.Fake news detection has many open issues that require attention of researchers. For instance, in order to reduce the spread of fake news, identifying key elements involved in the spread of news is an important step. Graph theory and machine learning techniques can be employed to identify the key sources involved in the spread of fake news. Likewise, real time fake news identification in videos can be another possible future direction.

REFERENCES

1. N. K. Conroy, V. L. Rubin, and Y. Chen, “Automatic deception detection: methods for finding fake news,” Proceedings of the Association for Information Science and Technology, vol. 52, no. 1, pp. 1–4, 2015.
2. H. Jwa, D. Oh, K. Park, J. M. Kang, and H. Lim, “exBAKE: automatic fake news detection model based on bidirectional encoder representations from transformers (BERT),” Applied Sciences, vol. 9, no. 19, 2019.
3. K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, “Fake news detection on social media,” ACM SIGKDD Explorations Newsletter, vol. 19, no. 1, pp. 22–36, 2017.
4. S. Vosoughi, D. Roy, and S. Aral, “The spread of true and false news online,” Science, vol. 359, no. 6380, pp. 1146–1151, 2018.
5. S. Gilda, "Notice of Violation of IEEE Publication Principles: Evaluating machine learning algorithms for fake news detection," 2017 IEEE 15th Student Conference on Research and Development (SCOReD), 2017, pp. 110-115, DOI: 10.1109/SCORED.2017.8305411.
6. Douglas, “News consumption and the new electronic media,” The International Journal of Press/Politics, vol. 11, no. 1, pp. 29–52, 2006.
7. J. Wong, “Almost all the traffic to fake news sites is from facebook, new data show,” 2016.
8. D. M. J. Lazer, M. A. Baum, Y. Benkler et al., “The science of fake news,” Science, vol. 359, no. 6380, pp. 1094–1096, 2018.
9. S. A. Garc´ıa, G. G. Garc´ıa, M. S. Prieto, A. J. M. Guerrero, and C. R. Jimenez, “The impact of term fake news on the scientific ´ community scientific performance and mapping in web of science,” Social Sciences, vol. 9, no. 5, 2020.
10. A. D. Holan, 2016 Lie of the Year: Fake News, PolitiFact, Washington, DC, USA, 2016.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.165

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details