



A
Project Report
on
Fake News Detection Using Machine Learning
submitted as partial fulfillment for the award of
BACHELOR OF TECHNOLOGY
DEGREE

SESSION 2022-23
in
Computer Science & Engineering

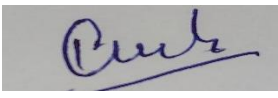
By
Naresh Singh (2000290109012)
Suraj Dhoundiyal (2000290109018)
Gunjan Kardam (1900290100061)

Under the supervision of
Prof. Bharti
KIET Group of Institutions, Ghaziabad

Affiliated to
Dr. A.P.J. Abdul Kalam Technical University, Lucknow
(Formerly UPTU)
May 2023

DECLARATION

We hereby declare that this submission is our own work and that, to the best of our knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgment has been made in the text.

Signature: 

Name: Naresh Singh

Roll No.: (2000290109012)

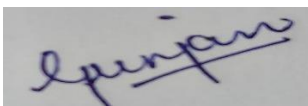
Date: 31/05/2023

Signature: 

Name: Suraj Dhoundiyal

Roll No.: (2000290109018)

Date: 31/05/2023

Signature: 

Name: Gunjan Kardam

Roll No.: (1900290100061)

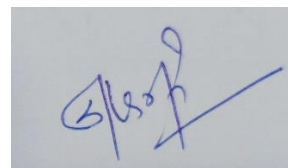
Date: 31/05/2023

CERTIFICATE

This is to certify that Project Report entitled “**Fake News Detection Using Machine Learning**” which is submitted by **Naresh Singh (2000290109012)**, **Suraj Dhoundiyal (2000290109012)** & **Gunjan Kardam (1900290100061)** in partial fulfillment of the requirement for the award of degree B. Tech. in Department of Computer Science & Engineering of Dr. A.P.J. Abdul Kalam Technical University, Lucknow is a record of the candidates own work carried out by them under my supervision. The matter embodied in this report is original and has not been submitted for the award of any other degree.

.

.



Date: 31/05/2023

Prof. Bharti

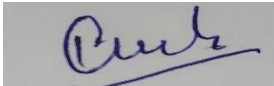
(Asst. Professor)

ACKNOWLEDGEMENT

It gives us a great sense of pleasure to present the report of the B. Tech Project undertaken during B. Tech. Final Year. We owe special debt of gratitude to **Prof Bharti**, Department of Computer Science & Engineering, KIET, Ghaziabad, for her constant support and guidance throughout the course of our work. Her sincerity, thoroughness and perseverance have been a constant source of inspiration for us. It is only his cognizant efforts that our endeavors have seen light of the day.

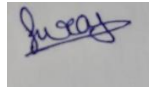
We also take the opportunity to acknowledge the contribution of **Dr. Vineet Sharma**, Head of the Department of Computer Science & Engineering, KIET, Ghaziabad, for his full support and assistance during the development of the project. We also do not like to miss the opportunity to acknowledge the contribution of all the faculty members of the department for their kind assistance and cooperation during the development of our project.

We also do not like to miss the opportunity to acknowledge the contribution of all faculty members, especially faculty/industry person/any person, of the department for their kind assistance and cooperation during the development of our project. Last but not the least, we acknowledge our friends for their contribution in the completion of the project.

Signature: 

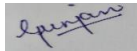
Name: Naresh Singh

Roll No.: (2000290109012)

Signature: 

Name: Suraj Dhoundiyal

Roll No.: (2000290109018)

Signature: 

Name: Gunjan Kardam

Roll No.: (1900290100061)

ABSTRACT

We aim to provide the user with the ability to classify the news as fake or real and check the authenticity of the website publishing the news. In recent years, due to the booming development of online social networks, fake news for various commercial and political purposes has been appearing in large numbers and widespread in the online world. With deceptive words, online social network users can get infected by this online fake news easily, which has brought about tremendous effects on the offline society already. An important goal in improving the trustworthiness of information in online social networks is to identify the fake news timely. This paper aims at investigating the principles, methodologies, and algorithms for detecting fake news articles, creators and subjects from online social networks and evaluating the corresponding performance. Information preciseness on Internet, especially on social media, is an increasingly important concern, but web-scale data hampers, ability to identify, evaluate and correct such data, or so called "fake news," present in these platforms. In this paper, we propose a method for "fake news" detection and ways to apply it on Facebook, one of the most popular online social media platforms. This method uses Naive Bayes classification model to predict whether a post on Facebook will be labeled as real or fake. The results may be improved by applying several techniques that are discussed in the paper. Received results suggest, that fake news detection problem can be addressed with machine learning methods.

TABLE OF CONTENTS

	Page No.
DECLARATION.....	ii
CERTIFICATE.....	iii
ACKNOWLEDGEMENTS.....	iv
ABSTRACT.....	v
LIST OF FIGURES.....	ix
LIST OF TABLES.....	x
LIST OF ABBREVIATION.....	xi
 CHAPTER 1 (INTRODUCTION)	
1.1. Introduction.....	1
1.2. Project Description.....	1
1.3. Motivation.....	2
1.4. Objective.....	2
1.5. Overview of Project.....	3
 CHAPTER 2 (LITERATURE REVIEW)	
2.1. Media Rich Fake news detection.....	4
2.1.1. Weekly Supervised Learning for fake news detection on twitter.....	4
2.2. Fake News Detection in social media.....	5
2.3. The Spread of fake news by social Bots.....	5
2.4. Misleading online content.....	5
2.5. Detecting fake news stories via multimodal analysis.....	6
2.6 Detecting and filtering rumor in social media using news media event.....	6

CHAPTER 3 (PROPOSED METHODOLOGY)

3.1. Existing Methodology.....	8
3.2. Proposed Methodology.....	8
3.3. System Architecture.....	9
3.4. Modules.....	12
3.4.1. Data Use.....	13
3.4.2. preprocessing.....	13
3.4.3. Feature Extraction.....	13
3.4.4. Training the Classifier.....	14
3.5. Algorithms.....	14
3.5.1. Naive Bayes.....	14
3.5.2. SVM.....	15
3.5.3. Logistic Regression.....	16
3.5.4. Decision Tree.....	17
3.5.5. Random Forest.....	18

CHAPTER 4 (RESULTS AND DISCUSSION)

4.1. Requirement Analysis.....	20
4.2. Functional Requirements.....	20
4.3. Non-Functional Requirements.....	21
4.4. System Requirements.....	21
4.4.1 Hardware Requirements.....	22
4.4.2 Software Requirements.....	22
4.5. System Design and Testing Plan.....	22

4.5.1. Input Design.....	22
4.5.2. Output Design.....	22
4.6. Social Feasibility.....	23
4.7. Data Flow Diagram.....	23
4.8. Performance Analysis.....	24
4.9. Accuracy.....	25
CHAPTER 5 (CONCLUSIONS AND FUTURE SCOPE)	
5.1. Conclusion.....	26
5.2. Future Scope.....	26
REFERENCES.....	27
APPENDEX1.....	28
APPENDIX2.....	30
CERTIFICATES.....	39

LIST OF FIGURES

Figure No.	Description	Page No.
Figure.1	Architecture diagram	9
Figure.2	The Classification Algorithms	10
Figure.3	Fake Detector Model	11
Figure.4	Support vector Machine	16
Figure.5	Logistic Regression	17
Figure.6	Data Flow Diagram	24
Figure.7	Accuracy of the Modal	25

LIST OF TABLES

Table. No.	Description	Page No.
1	Comparison of classifiers	25

LIST OF ABBREVIATIONS

NLTK	Natural Language Toolkit
POS	Part of Speech
TF-IDF	Term Frequency - inverse Document Frequency

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

As an increasing amount of our lives is spent interacting online through social media platforms, more and more people tend to hunt out and consume news from social media instead of traditional news organizations. The explanations for this alteration in consumption behaviors are inherent within the nature of those social media platforms:

- (i) it's often more timely and fewer expensive to consume news on social media compared with traditional journalism, like newspapers or television; and
- (ii) it's easier to further share, discuss, and discuss the news with friends or other readers on social media. For instance, 62 percent of U.S. adults get news on social media in 2016, while in 2012; only 49 percent reported seeing news on social media

No doubt we have several advantages of this digital world but it also has its disadvantages as well. There are different issues in this digital world. One of them is fake news. Someone can easily spread a fake news. Fake news is spread to harm the reputation of a person or an organization. It can be a propaganda against someone that can be a political party or an organization. There are different online platforms where the person can spread the fake news. This includes the Facebook, Twitter etc. Machine learning is the part of artificial intelligence that helps in making the systems that can learn and perform different actions (Donepudi, 2019). A variety of machine learning algorithms are available that include the supervised, unsupervised, reinforcement machine learning algorithms. The algorithms first have to be trained with a data set called train data set. After the training, these algorithms can be used to perform different tasks. Machine learning is using in different sectors to perform different tasks. Most of the time machine learning algorithms are used for prediction purpose or to detect something that is hidden.

1.2 PROJECT DESCRIPTION

The term 'fake news' became common parlance for the issue, particularly to describe factually incorrect and misleading articles published mostly for the purpose of making money through page views. In this paper, it is seemed to produce a model that can accurately predict the likelihood that a given article is fake news. Facebook has been at the epicenter of much critique

following media attention. They have already implemented a feature to flag fake news on the site when a user sees it; they have also said publicly they are working on to distinguish these articles in an automated way. Certainly, it is not an easy task. A given algorithm must be politically unbiased – since fake news exists on both ends of the spectrum – and give equal balance to legitimate news sources on either end of the spectrum. In addition, the question of legitimacy is a difficult one. However, in order to solve this problem, it is necessary to have an understanding on what Fake News.

1.3 MOTIVATION

We will be training and testing the data, when we use supervised learning, it means we are labeling the data. By getting the testing and training data and labels we can perform different machine learning algorithms but before performing the predictions and accuracies, the data is need to be preprocessing i.e., the null values which are not readable are required to be removed from the data set and the data is required to be converted into vectors by normalizing and tokening the data so that it could be understood by the machine. Next step is by using this data, getting the visual reports, which we will get by using the Mat Plot Library of Python and Sickit Learn. This library helps us in getting the results in the form of histograms, pie charts or bar charts.

1.4 OBJECTIVE

The objective of this project is to examine the problems and possible significances related with the spread of fake news. We will be working on different fake news data set in which we will apply different machine learning algorithms to train the data and test it to find which news is the real news or which one is the fake news. As the fake news is a problem that is heavily affecting society and our perception of not only the media but also facts and opinions themselves. By using the artificial intelligence and the machine learning, the problem can be solved as we will be able to mine the patterns from the data to maximize well defined objectives. So, our focus is to find which machine learning algorithm is best suitable for what kind of text dataset. Also, which dataset is better for finding the accuracies as the accuracies directly depends on the type of data and the amount of data. The more the data, more are your chances of getting correct accuracy as you can test and train more data to find out your results.

1.5 OVERVIEW OF PROJECT

With the advancement of technology, digital news is more widely exposed to users globally and contributes to the increment of spreading and disinformation online. Fake news can be found through popular platforms such as social media and the Internet. There have been multiple solutions and efforts in the detection of fake news where it even works with tools. However, fake news intends to convince the reader to believe false information which deems these articles difficult to perceive. The rate of producing digital news is large and quick, running daily at every second, thus it is challenging for machine learning to effectively detect fake news.

CHAPTER 2

LITERATURE REVIEW

2.1 MEDIA RICH FAKE NEWS DETECTION

In general, the goal is profiting through clickbaits. Clickbaits lure users and entice curiosity with flashy headlines or designs to click links to increase advertisements revenues. This exposition analyzes the prevalence of fake news in light of the advances in communication made possible by the emergence of social networking sites. The purpose of the work is to come up with a solution that can be utilized by users to detect and filter out sites containing false and misleading information. We use simple and carefully selected features of the title and post to accurately identify fake posts. The experimental results show a 99.4% accuracy using logistic classifier.

2.1.1 WEAKLY SUPERVISED LEARNING FOR FAKE NEWS DETECTION ON TWITTER

The problem of automatic detection of fake news in social media, e.g., on Twitter, has recently drawn some attention. Although, from a technical perspective, it can be regarded as a straightforward, binary classification problem, the major challenge is the collection of large enough training corpora, since manual annotation of tweets as fake or non-fake news is an expensive and tedious endeavor. In this paper, we discuss a weakly supervised approach, which automatically collects a large-scale, 4 but very noisy training dataset comprising hundreds of thousands of tweets. During collection, we automatically label tweets by their source, i.e., trustworthy, untrustworthy source, and train a classifier on this dataset. We then use that classifier for a different classification target, i.e., the classification of fake and non-fake tweets.

2.2 FAKE NEWS DETECTION IN SOCIAL MEDIA

Fake news and hoaxes have been there since before the advent of the Internet. The widely accepted definition of Internet fake news is: fictitious articles deliberately fabricated to deceive readers”. Social media and news outlets publish fake news to increase readership or as part of psychological warfare. In general, the goal is profiting through Clickbaits lure users and entice curiosity with flashy headlines or designs to click links to increase advertisements revenues. This exposition analyzes the prevalence of fake news in light of the advances in communication made possible by the emergence of social networking sites. The purpose of the work is to come up with a solution that can be utilized by users to detect and filter out sites containing false and misleading information. We use simple and carefully selected features of the title and post to accurately identify fake posts. The experimental results show a 99.4% accuracy using logistic classifier.

2.3 THE SPREAD OF FAKE NEWS BY SOCIAL BOTS

The massive spread of fake news has been identified as a major global risk and has been alleged to influence elections and threaten democracies. Communication, cognitive, social, and computer scientists are engaged in efforts to study the complex causes for the viral diffusion of digital misinformation and to develop solutions, while search and social media platforms are beginning to deploy countermeasures. However, to date, these efforts have been mainly informed by anecdotal evidence rather than systematic data. Here we analyze 14 million messages spreading 400 6 thousand claims on Twitter during and following the 2016 U.S. presidential campaign and election. We find evidence that social bots play a key role in the spread of fake news. Accounts that actively spread misinformation are significantly more likely to be bots.

2.4 MISLEADING ONLINE CONTENT

Big Data Analytics and Deep Learning are two high-focus of data science. Big Data has become important as many organizations both public and private have been collecting massive amounts of domain-specific information, which can contain useful information about problems such as national intelligence, cyber security, fraud detection, marketing, and medical informatics.

Companies such as Google and Microsoft are analyzing large volumes of data for business analysis and decisions, impacting existing and future technology. Deep Learning algorithms extract high-level, complex abstractions as data representations through a hierarchical learning process.

2.5 Detecting fake news stories via multimodal analysis

In this paper, the Authors Describe - Online fake news is a specific type of digital misinformation that poses serious threats to democratic institutions, misguides the public, and can lead to radicalization and violence. Hence, fake news detection is an important problem for information science research. In the paper, they focus on the automatic identification of fake content in online news stories. They use automated methods for fake news detection. And consider three approaches for fake news detection one based only on textual features, one on visual features, and one combining both. The results for the accuracy obtained on the testing set (averaged over 100 iterations) using different algorithms and considering the three different models. They use Python Scikit-learn implementation of multiple well-known algorithms such as Logistic Regression. The results identify textual and visual features that are more likely to be associated with fake news. They also suggest that multimodal analysis can help improve the performance of purely textual or purely visual fake news detectors. These results pave the way for better understanding of the fake news phenomenon, including its psychological underpinnings, and aid further research on multimodal fake news detection.

2.6 Detecting and filtering rumor in social media using news media event

The Authors- Nithya Kandasamy and Krishnamoorthi Murugasamy while describing the main theme of the paper explain, fake news is unverified at the time of posting. Therefore, it is necessary to detect and remove the fake news before it spreads widely. Their proposed rumor detection method compares the social media content with news media and applies the support vector machine (SVM) as a binary classification technique. And the experimental results proved that the proposed method will be useful for avoiding social damages caused by rumors in social media. They collect the Data set from social media and news website and from the twitter.

Rumor detection model is trained with news events then tweets are given as input to the rumor detection model. Using mining model based on concept,²³ semantic structure of each tweet is identified. We have taken SVM, ICDM, Decision Tree, Naive Bayes, and K-NN classification techniques to classify class label as Rumor or Not Rumor. In this SVM achieves the highest accuracy of 82% than other machine learning models. Furthermore, input from multiple social networking sites may be gathered for identifying rumor detection based on that user preferences can be identified.

CHAPTER 3

PROPOSED METHODOLOGY

3.1 EXISTING METHODOLOGY

There exists a large body of research on the topic of machine learning methods for deception detection, most of it has been focusing on classifying online reviews and publicly available social media posts. Particularly since late 2016 during the American Presidential election, the question of determining 'fake news' has also been the subject of particular attention within the literature. Conroy, Rubin, and Chen outline several approaches that seem promising towards the aim of perfectly classify the misleading articles. They note that simple content-related n-grams and shallow parts-of-speech tagging have proven insufficient for the classification task, often failing to account for important context information. Rather, these methods have been shown useful only in tandem with more complex methods of analysis. Deep Syntax analysis using Probabilistic Context Free Grammars have been shown to be particularly valuable in combination with n-gram methods. Feng, Banerjee, and Choi can achieve 85%-91% accuracy in deception related classification tasks using online review corpora.

Since this problem is a kind of text classification, implementing a Naive Bayes classifier will be best as this is standard for text-based processing. The actual goal is in developing a model which was the text transformation (count vectorizer vs TFIDF vectorizer) and choosing which type of text to use (headlines vs full text). Now the next step is to extract the most optimal features for COUNTVECTORIZER or TFIDF-vectorizer, this is done by using a few the most used words, and/or phrases, lower casing or not, mainly removing the stop words which are common words such as “the”, “when”, and “there” and only using those words that appear at least a given number of times in a given text dataset.

3.2 PROPOSED METHODOLOGY

This presents the methodology used for the classification. Using this model, a tool is implemented for detecting the fake articles. In this method supervised machine learning is used for classifying the dataset. The first step in this classification problem is dataset collection

phase, followed by preprocessing, implementing features selection, then perform the training and testing of dataset and finally running the classifiers. describes the proposed system methodology. The methodology is based on conducting various experiments on dataset using the algorithms described in the previous section named Random Forest, SVM and Naïve Bayes, majority voting and other classifiers. The experiments are conducted individually on each algorithm, and on combination among them for the purpose of best accuracy and precision

3.3 SYSTEM ARCHITECTURE

Fig. 1 shows the methodology contains many steps which includes Data Collection, Data Pre-processing, Data Vectorization, Data Splitting, Model training and further the evaluation of the trained model. In data collection large dataset of real and fake articles from a variety of news sources are collected and labelled. Data Pre-processing of that labelled data is done in order to fill the null values, balancing of data, Tokenization, Stopword 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 vectored data is further split data in a ratio of 3:1 for train and test 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 our model is tested on various ML classifiers.

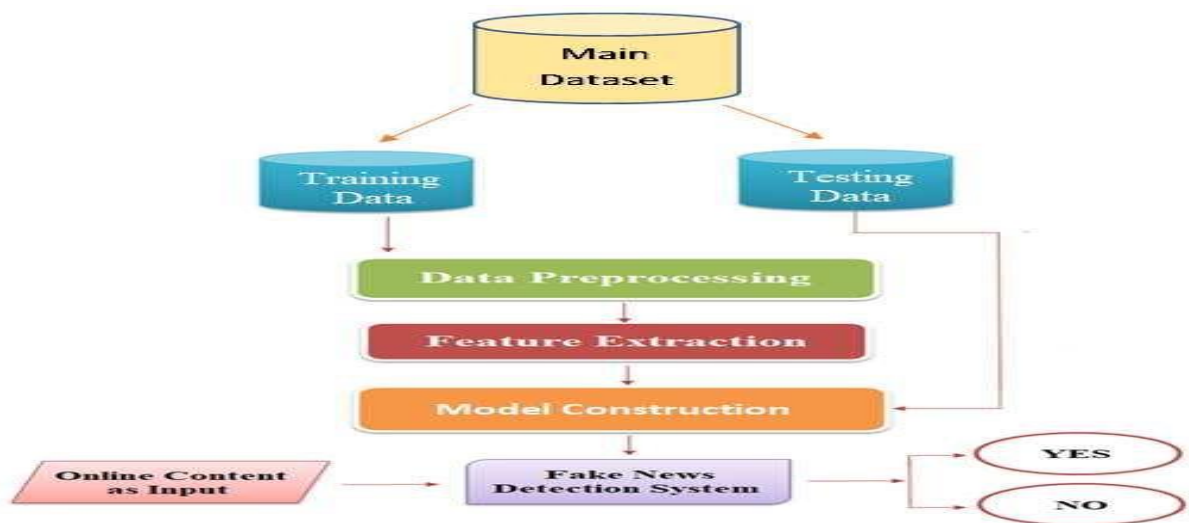


Figure 1. Architecture diagram

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.

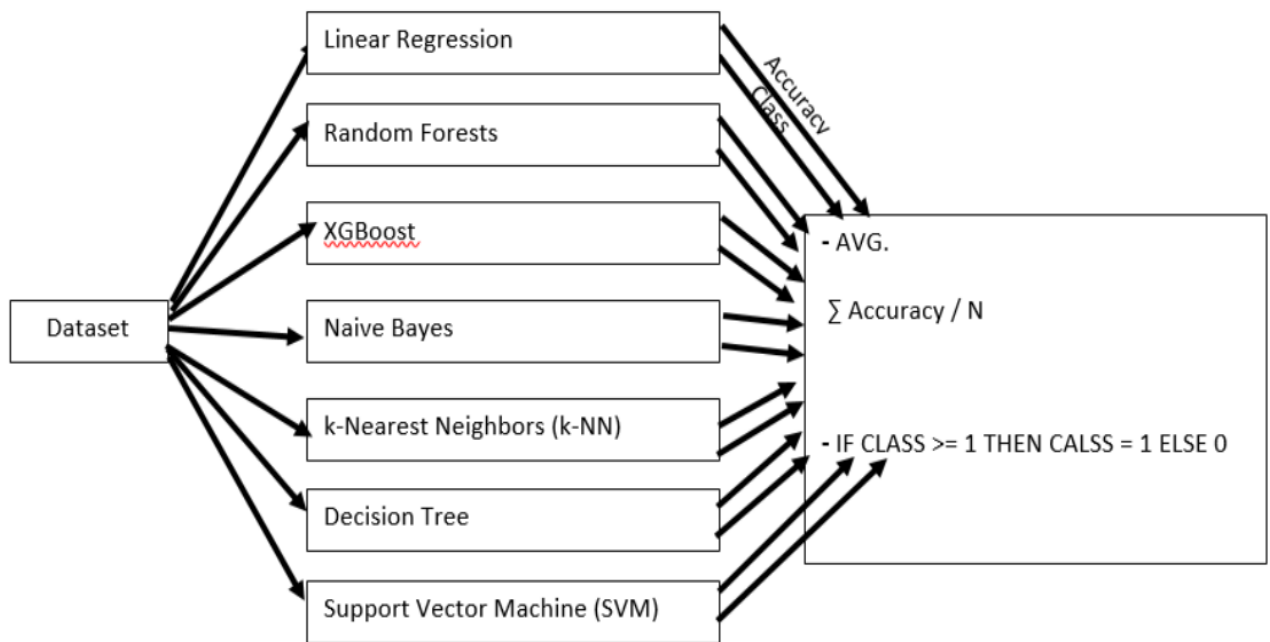


Figure 2. The Classification Algorithms

In the process of model creation, the approach to detecting political fake news is as follows: First step is collection political news dataset, (the Liar dataset is adopted for the model), perform preprocessing through rough noise removal, the next step is to apply the NLTK (Natural Language Toolkit) to perform POS and features are selected. Next perform the dataset splitting apply ML algorithms (Naïve bays and Random Forest) then create the proposed classifier model. The Fig 2 shows that after the NLTK is applied, the Dataset gets successfully preprocessed in the system, then a message is generated for applying algorithms on trained portion. The system response with N.B and Random Forest are applied, then the model is created with response message. Testing is performed on test dataset, and the results are verified, the next step is to monitor the precision for acceptance. The model is then applied on unseen data selected by user. Full dataset is created with half of the data being fake and half with real articles, thus making the model's reset accuracy 50%. Random selection of 80% data is done

from the fake and real dataset to be used in our complete dataset and leave the remaining 20% to be used as a testing set when our model is complete. Text data requires preprocessing before applying classifier on it, so we will clean noise, using Stanford NLP (Natural language processing) for POS (Part of Speech) processing and tokenization of words, then we must encode the resulted data as integers and floating-point values to be accepted as an input to ML algorithms. This process will result in feature extraction and vectorization; the research using python sk-learn library to perform tokenization and feature extraction of text data, because this library contains useful tools like Count Vectorizer and Tfidf Vectorizer. Data is viewed in graphical presentation with confusion matrix.

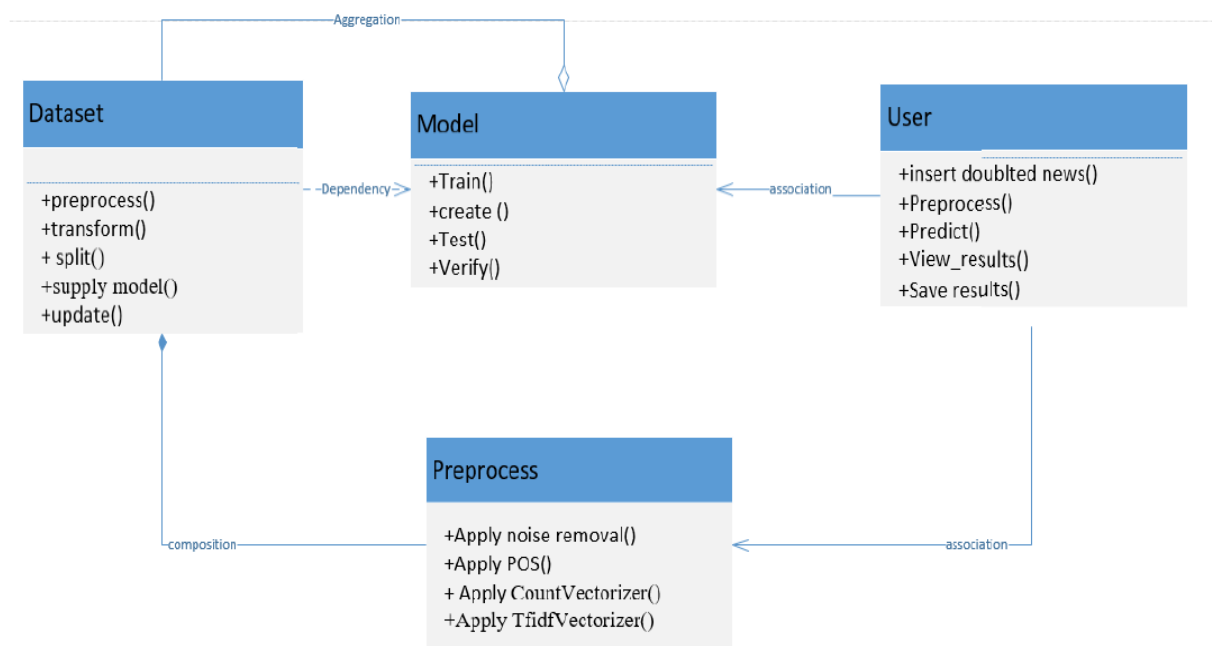


Figure 3. Fake Detector Model

This section discusses the chosen dataset, The LIAR-PLUS Master that has been used for cleaning and extracting the data set, and the algorithms are applied. This dataset has automatically extracted proof sentences from the full-text verdict article published in Politifact by journalists. As shown in following figure 4 we used the features of Truth values, in addition we applied part of speech on the statement to get another 4 features (nouns, verbs, preposition and sentences) and each record is labeled by class label as (0, 1, 2, 3) to be used in training the model. The following steps have been used to evaluate the precision of the news.

1. Liar-dataset is preprocessed 12.8K
2. The texts in multiple contexts is taken from multiple sites and is labelled manually. Then it is transformed from TSV format into CSV format using Python.
3. The next step is to clean the noise using NLP libraries and v2 library. The noise involves ids, dots, commas, quotations, and by stemming terms, delete the suffix. The next step is to use POS (Part of speech) which will turn the dataset into tokens and statistical values.
5. Extract unigram and bigram features by using TFIDF Vectorizer function of python sklearn. Feature extraction library to generate TF-IDF n-gram features.
6. Divide the dataset into 70% for train and 30% for test using python sklearn.
7. Produce classification model ipynb file after applying all the algorithms.
8. Test model precision on the test portion of dataset and produce confusion matrix.
9. Evaluate accuracy, precision, recall, and f1-score for fake and real news.
10. Design the interface to be used for testing unseen news by user.
4. Perform feature extraction by choosing lexical features, Such as word count, average word length, length of article, number count, number of sections of speech (adjective).

The data is divided it into two parts: The first section, which consists of 75% of the data, is a trained data, where the algorithm detects the real news and false news, then the data is labeled in the form of 0 and 1 where 0 is for false news and 1 for true news. After that, the rest of the data, which is 25% of it, will do a test on it, so that it is sure whether the news is nature or forged, and then return it in case it was right or wrong, and according to the percentage of right and wrong, the algorithm percentage will be formed.

3.4 MODULES

3.4.1 Data Use

So, in this project we are using different packages and to load and read the data set we are using pandas. By using pandas, we can read the .csv file and then we can display the shape of the dataset with that we can also display the dataset in the correct form. We will be training and

testing the data, when we use supervised learning, it means we are labeling the data. By getting the testing and training data and labels we can perform different machine learning algorithms but before performing the predictions and accuracies, the data is need to be preprocessing i.e., the null values which are not readable are required to be removed from the data set and the data is required to be converted into vectors by normalizing and tokening the data so that it could be understood by the machine. Next step is by using this data, getting the visual reports, which we will get by using the Mat Plot Library of Python and Sklearn. This library helps us in getting the results in the form of histograms, pie charts or bar charts.

3.4.2 Preprocessing

This presents the methodology used for the classification. Using this model, a tool is implemented for detecting the fake articles. In this method supervised machine learning is used for classifying the dataset. The first step in this classification problem is dataset collection phase, followed by preprocessing, implementing features selection, then perform the training and testing of dataset and finally running the classifiers. describes the proposed system methodology. The methodology is based on conducting various experiments on dataset using the algorithms described in the previous section named Random Forest, SVM and Naïve Bayes, majority voting and other classifiers. The experiments are conducted individually on each algorithm, and on combination among them for the purpose of best accuracy and precision

3.4.3 Feature Extraction

Feature extraction s the process of selecting a subset of relevant features for use in model construction. Feature extraction methods helps in to create an accurate predictive model. They help in selecting features that will give better accuracy. When the input data to an algorithm is too large to be handled and i ts supposed to be redundant then the input data will be transformed into a reduced illustration set of features also named feature vectors. Altering the input data to perform the desired task using this reduced representation instead of the full-size input. Feature extraction is performed on raw data prior to applying any machine learning algorithm, on the transformed data in feature space.

3.4.4 Training the Classifier

As In this project We are using Scikit-Learn Machine learning library for implementing the architecture. Scikit Learn is an open-source python Machine Learning library which comes bundled in 3rd distribution anaconda. This just needs importing the packages and you can compile the command as soon as you write it. If the command does not run, we can get the error at the same time. I am using 4 different algorithms and I have trained these 4 models i.e., Naïve Bayes, Support Vector Machine, K Nearest Neighbors and Logistic Regression which are very popular methods for document classification problem. Once the classifiers are trained, we can check the performance of the models on test-set. We can extract the word count vector for each mail in test-set and predict it class with the trained models.

3.5 Algorithms Used

3.5.1 Naive Bayes

Naive Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. It is mainly used in text classification that includes a high-dimensional training dataset.

Naive Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.

It is a probabilistic classifier, which means it predicts based on the probability of an object. Some popular examples of Naïve Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles.

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

The Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

- **Naive:** It is called Naive because it assumes that the occurrence of a certain feature is independent of the occurrence of other features. Such as if the fruit is identified on the basis of color, shape, and taste, then red, spherical, and sweet fruit is recognized as an apple. Hence each feature individually contributes to identify that it is an apple without depending on each other.

- **Bayes:** It is called Bayes because it depends on the principle of Bayes' Theorem.

3.5.2 Support Vector Machine- SVM

- SVMs are a set of supervised learning methods used for classification, and regression.
- Effective in high dimensional spaces.
 - Uses a subset of training points in the support vector, so it is also memory efficient.

The advantages of support vector machines are

- Still effective in cases where the number of dimensions is greater than the number of samples.
- Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.
- Versatile: different kernel functions can be specified for the decision function.

Common kernels are provided, but it is also possible to specify custom kernels.

The disadvantages of support vector machines include:

- If the number of features is much greater than the number of samples, avoid over-fitting in choosing Kernel functions and regularization term is crucial.

SVMs do not directly provide probability estimates, these are calculated using an expensive five-fold cross-validation.

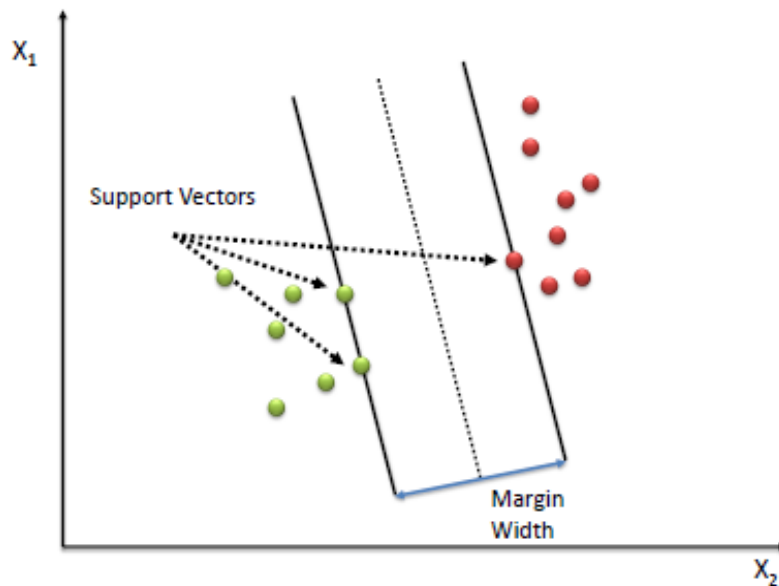


Figure 4 - SVM

3.5.3 Logistic Regression

- Linear model for classification rather than regression.
- The expected values of the response variable are modeled based on combination of values taken by the predictors.
- Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.
- Logistic Regression is much like the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas logistic regression is used for solving the classification problems.
- In Logistic regression, instead of fitting a regression line, we fit a "Shaped logistic function, which predicts two maximum values (0 or 1).
- The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc.
- Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.

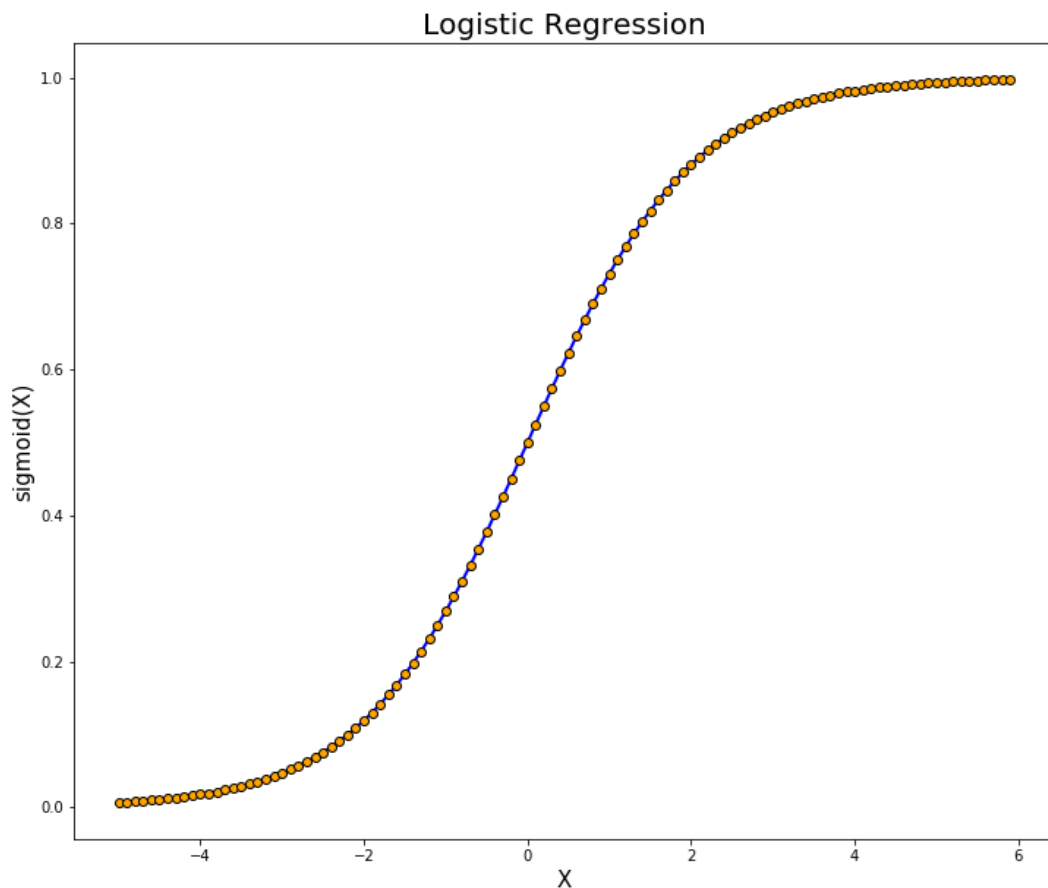


Figure.5 Logistic Regression

3.5.4 Decision Tree

It is a supervised learning model mainly used for classification problems. 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 trees are the tree models where the target variable can take a discrete set of values. Equations are used to calculate Entropy mathematically which measures the impurity in the 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 tests 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, like 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 CART

algorithm, which stands for Classification and Regression Tree algorithm. A Decision Tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees.

The Decision Tree Algorithm belongs to the family of supervised machine learning algorithms. It can be used for both a classification problem as well as for a regression problem.

The goal of this algorithm is to create a model that predicts the value of a target variable, for which the decision tree uses the tree representation to solve the problem in which the leaf node corresponds to a class label and attributes are represented on the internal node of the tree.

3.5.5 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 is used to calculate Gini Index to measure variance. It takes less training time as compared to other algorithms. It predicts output with high accuracy, even for the large dataset it runs efficiently. It can also maintain accuracy when a large proportion of data is missing.

- It can be used both for classification and regression. It is also the most flexible and easy to use algorithm. A forest consists of trees. It is said that the more trees it has, the more robust a forest is. Random Forests create Decision Trees on randomly selected data samples, get predictions from each tree and select the best solution by means of voting. It also provides a pretty good indicator of the feature importance.
- Random Forests have a variety of applications, such as recommendation engines, image classification and feature selection. It can be used to classify loyal loan applicants, identify fraudulent activity and predict diseases. It lies at the base of the Boruta algorithm, which selects important features in a dataset.
- Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. 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.

- As the name suggests, "Random Forest is a classifier that 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.

CHAPTER 4

RESULTS AND DISCUSSION

Algorithm's accuracy depends on the type and size of your dataset. More the data, more chances of getting correct accuracy. Machine learning depends on the variations and relations. Understanding what is predictable is as important as trying to predict it. While making algorithm choice, speed should be a consideration factor.

4.1 REQUIREMENT ANALYSIS

Requirement analysis, also called requirement engineering, is the process of determining user expectations for a new modified product. It encompasses the tasks that determine the need for analyzing, documenting, validating, and managing software or system requirements. The requirements should be documentable, actionable, measurable, testable and traceable related to identified business needs or opportunities and define to a level of detail, sufficient for system

4.2 FUNCTIONAL REQUIREMENTS

It is a technical specification requirement for the software products. It is the first step in the requirement analysis process which lists the requirements of software systems including functional, performance and security requirements. The function of the system depends mainly on the quality hardware used to run the software with given functionality. **Usability**- It specifies how easy the system must be use. It is easy to ask queries in any format which is short or long, porter stemming algorithm stimulates the desired response for user.

Robustness- It refers to a program that performs well not only under ordinary conditions but also under unusual conditions. It is the ability of the user to cope with errors for irrelevant queries during execution.

Security-The state of providing protected access to resource is security. The system provides good security and unauthorized users cannot access the system there by providing high security.

Reliability-It is the probability of how often the software fails. The measurement is often expressed in MTBF (Mean Time Between Failures). The requirement is needed in order to ensure that the processes work correctly and completely without being aborted. It can handle any load and survive and even capable of working around any failure.

Compatibility-It is supported by version above all web browsers. Using any web servers like localhost makes the system real-time experience.

Safety-Safety is a measure taken to prevent trouble. Every query is processed in a secured manner without letting others to know one's personal information.

4.3 NON- FUNCTIONAL REQUIREMENTS

Portability-It is the usability of the same software in different environments. The project can be run in any operating system.

Performance-These requirements determine the resources required, time interval, throughput and everything that deals with the performance of the system.

Accuracy-The result of the requesting query is very accurate and high speed of retrieving information. The degree of security provided by the system is high and effective.

Maintainability-Project is simple as further updates can be easily done without affecting its stability. Maintainability basically defines that how easy it is to maintain the system. It means how easy it is to maintain the system, analyze, change and test the application. The maintenance of this project is simple as further updates can easily be done without affecting its stability.

4.4 SYSTEM REQUIREMENTS

4.4.1 HARDWARE REQUIREMENTS:

Processor - Any updated processor

Hard disk - 100GB

RAM - MIN 4GB

4.4.2 SOFTWARE REQUIREMENTS:

Operating System - Windows

Coding language - PYTHON

4.5 SYSTEM DESIGN AND TESTING PLAN

4.5.1 INPUT DESIGN

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:

- [1] What data should be given as input?
- [2] How the data should be arranged or coded?
- [3] The dialog to guide the operating personnel in providing input.
- [4] Methods for preparing input validations and steps to follow when error occur.

4.5.2 OUTPUT DESIGN

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

The output form of an information system should accomplish one or more of the following objectives.

- Convey information about past activities, current status or projections of the
- Future.
- Signal important events, opportunities, problems, or warnings.
- Trigger an action.
- Confirm an action.

4.6 SOCIAL FEASIBILITY

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

4.7 DATA FLOW DIAGRAM

- The DFD is also called as bubble chart. It is a simple graphical formalism that can be used to represent a system in terms of input data to the system, various processing carried out on this data, and the output data is generated by this system.
- The data flow diagram (DFD) is one of the most important modeling tools. It is used to model the system components. These components are the system process, the data used by the process, an external entity that interacts with the system and the information flows in the system.
- DFD shows how the information moves through the system and how it is modified by a series of transformations. It is a graphical technique that depicts information flow and the transformations that are applied as data moves from input to output.

- DFD is also known as bubble chart. A DFD may be used to represent a system at any level of abstraction. DFD may be partitioned into levels that represent increasing information flow and functional detail.
- It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration.

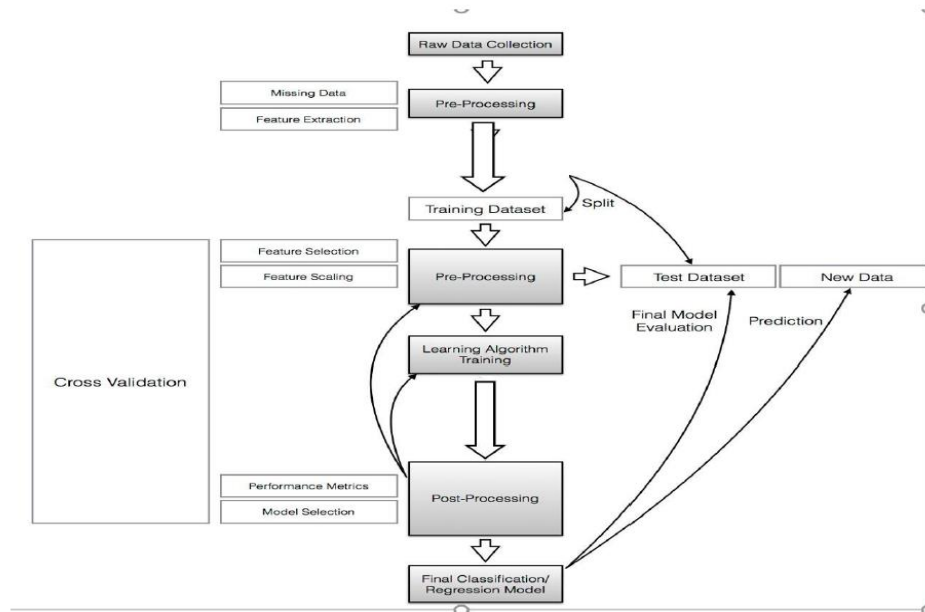


Figure.6 Data flow Diagram

4.8 Performance Analysis

The maximum accuracy of the Linear Regression Classifier.

Table -1 (Comparison of 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

4.9 Accuracy

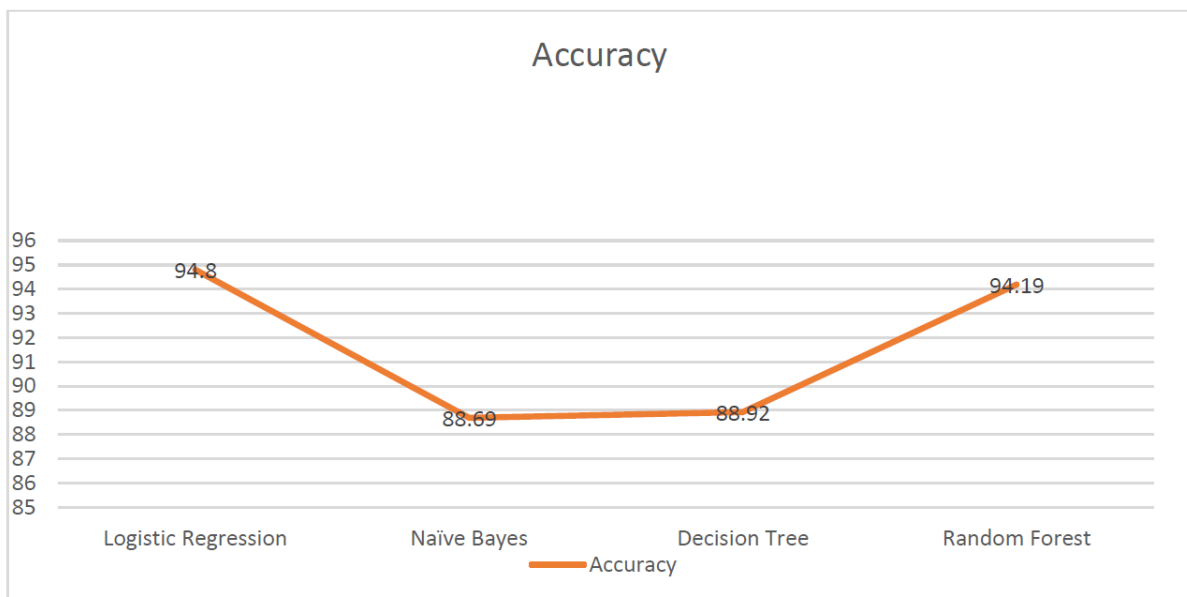


Figure.7 - Accuracy of the model

CHAPTER 5

CONCLUSION AND FUTURE SCOPE

7.1 CONCLUSION

Many people consume news from social media instead of traditional news media. However, social media has also been used to spread fake news, which has negative impacts on individual people and society. In this paper, an innovative model for fake news detection using machine learning algorithms has been presented. This model takes news events as an input and based on twitter reviews and classification algorithms it predicts the percentage of news being fake or real.

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential. This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus, the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

7.2 FUTURE SCOPE

This is based on the crowd sourcing dataset and the web covering dataset. These are the static datasets. Through these we can only test the data which is present in the predefined training data sets. The paper gives the appropriate result for the test data that is present in the training datasets. Thus, the future scope of the paper is connecting this methodology to the internet news which gives results even for the test data that is not present in the training data sets. We can even change to some other better classifier to classify the data other than naïve bayes and logistic regression.

REFERENCES

- [1]. Parikh, S. B., & Atrey, P. K. (2018, April). Media-Rich Fake News Detection: A Survey. In 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR) (pp. 436-441). IEEE.
- [2]. Conroy, N. J., Rubin, V. L., & Chen, Y. (2015, November). Automatic deception detection: Methods for finding fake news. In Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community (p. 82). American Society for Information Science.
- [3]. Helmstetter, S., & Paulheim, H. (2018, August). Weakly supervised learning for fake news detection on Twitter. In 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp. 274-277). IEEE. [4]. Stahl, K. (2018). Fake News Detection in social media.
- [5]. Della Vedova, M. L., Tacchini, E., Moret, S., Ballarin, G., DiPierro, M., & de Alfaro, L. (2018, May). Automatic Online Fake News Detection Combining Content and Social Signals. In 2018 22nd Conference of Open Innovations Association (FRUCT) (pp. 272-279). IEEE.
- [6] Tacchini, E., Ballarin, G., Della Vedova, M. L., Moret, S., & de Alfaro, L. (2017). Some like it hoax: Automated fake news detection in social networks. arXiv preprint arXiv:1704.07506.
- [7] Ahmad, Iftikhar, Muhammad Yousaf, Suhail Yousaf, and Muhammad Ovais Ahmad. "Fake news detection using machine learning ensemble methods." Complexity 2020 (2020).
- [8] 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).
- [9] 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.

APPENDIX 1

Python: Python is an interpreted, high-level, general purpose programming language created by Guido Van Rossum and first released in 1991, Python's design philosophy emphasizes code Readability with its notable use of significant White space. Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects. Python is dynamically typed and garbage collected. It supports multiple programming paradigms, including procedural, object-oriented, and functional programming.

Sklearn: Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering, and dimensionality reduction via a consistent interface in Python. This library, which is largely written in Python, is built upon NumPy, SciPy and Matplotlib.

NumPy: NumPy is a library for the python programming language, adding support for large, multi- dimensional arrays and matrices, along with a large collection of high level mathematical functions to operate on these arrays. The ancestor of NumPy, Numeric, was originally created by Jim with contributions from several other developers. In 2005, Travis created NumPy by incorporating features of the competing Numarray into Numeric, with extensive modifications. NumPy is open-source software and has many contributors.

Librosa: Librosa is a Python package for music and audio analysis. Librosa is basically used when we work with audio data like in music generation (using LSTMs), Automatic Speech Recognition.

It provides the building blocks necessary to create the music information retrieval systems. Librosa helps to visualize the audio signals and do the feature extractions in it using different signal processing techniques.

Seaborn: Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics. Seaborn is a library in Python predominantly used for making statistical graphics. Seaborn is a data visualization library built on top of matplotlib and closely integrated with pandas' data

structures in Python. Visualization is the central part of Seaborn which helps in exploration and understanding of data.

SciPy: SciPy contains modules for optimization, linear algebra, integration, interpolation, special functions, FFT, signal and image processing, ODE solvers and other tasks common in science and engineering. SciPy is also a family of conferences for users and developers of these tools: SciPy (in the United States), EuroSciPy (in Europe) and SciPy.in (in India). Enthought originated the SciPy conference in the United States and continues to sponsor many of the international conferences as well as host the SciPy website. SciPy is a scientific computation library that uses NumPy underneath. It provides more utility functions for optimization, stats, and signal processing.

APPENDIX 2

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 misuses 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]

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

(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

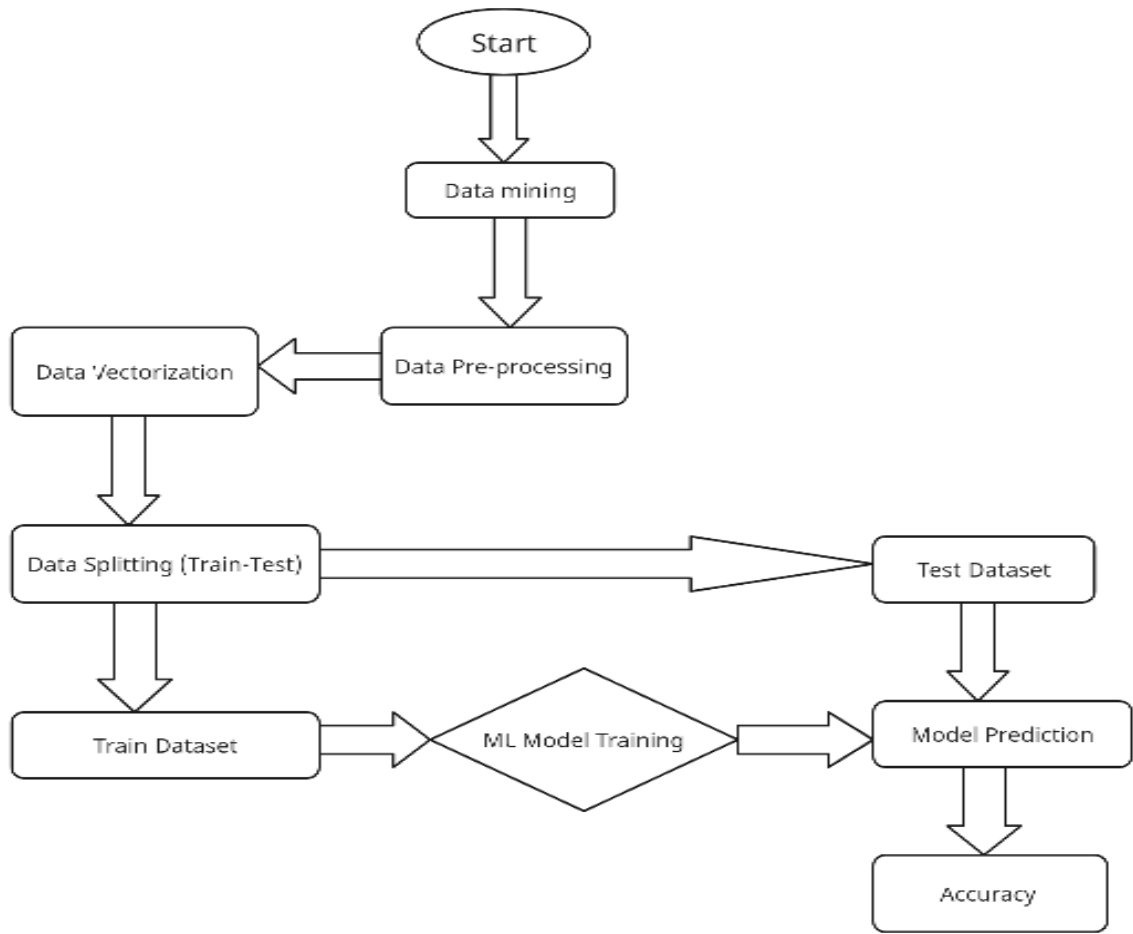


Figure 1: Flowchart of model

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.

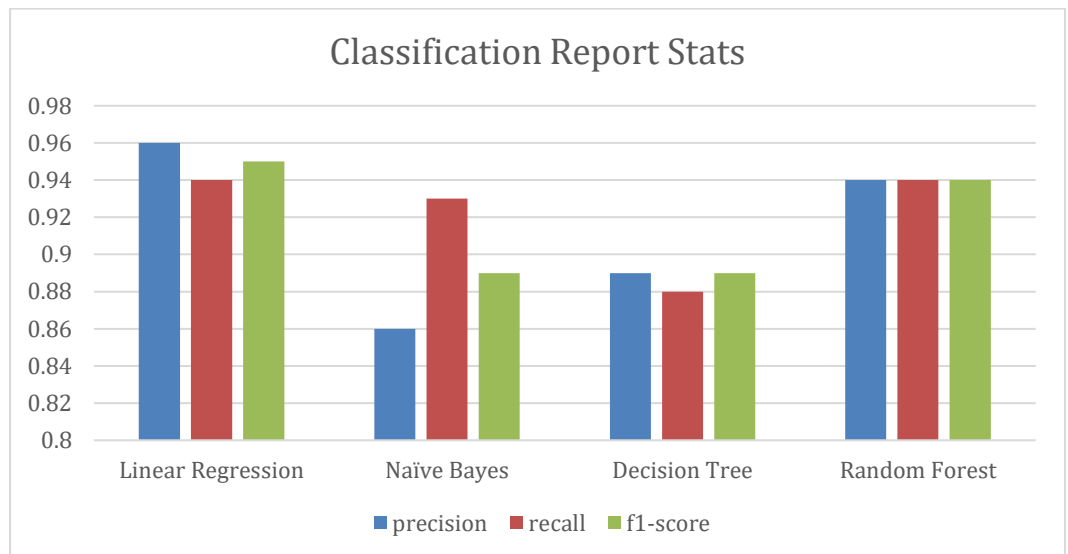
Once the model has been trained and evaluated, it can be deployed in a production environment, where it can be used to automatically classify new articles as they are true or fake. As the dataset keeps changing so does the model, so it is important to regularly update the model with new data and retrain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

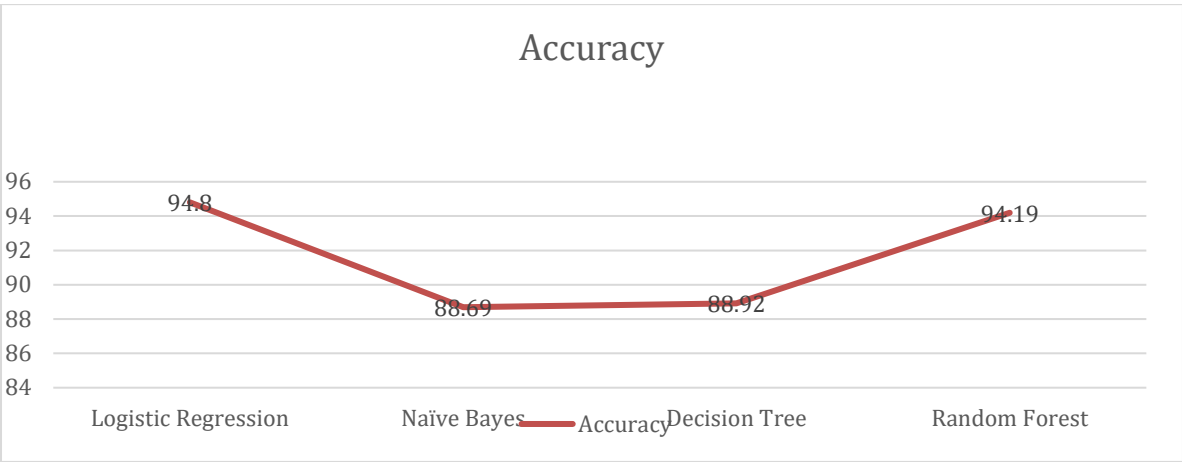


Figure 2: Accuracy of ML Models

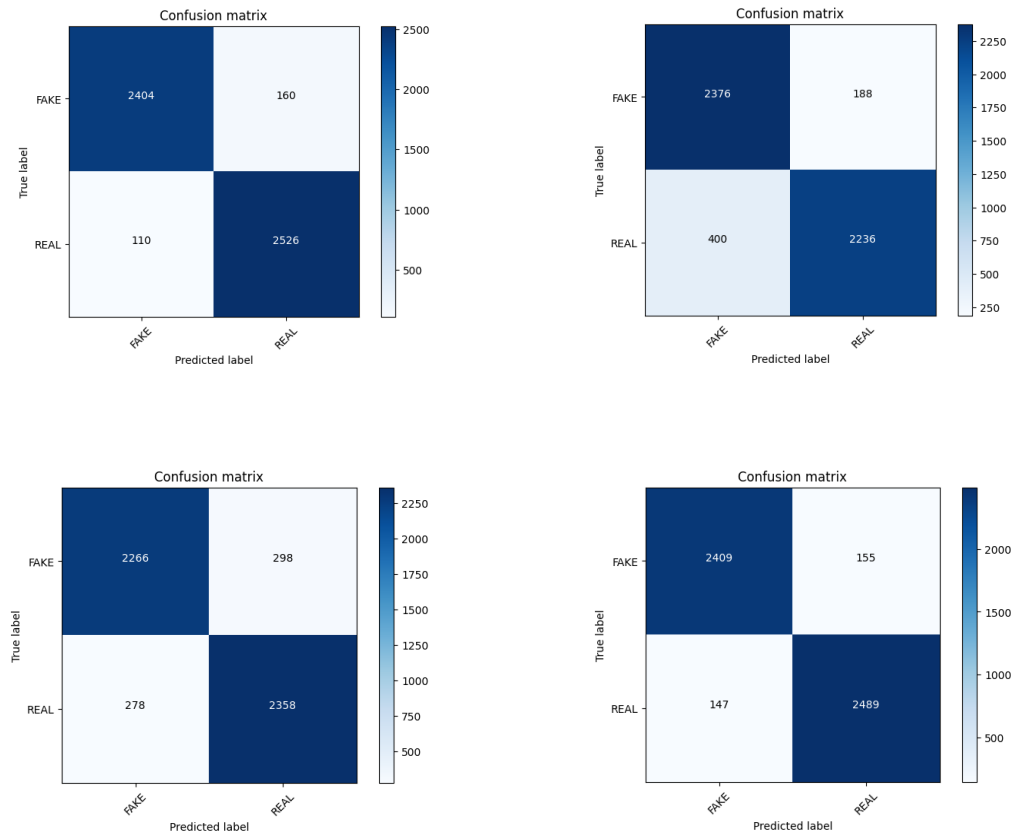


Figure 3: Confusion Matrix of ML Classifiers

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.

CERTIFICATES

Paper ID: IT-ICACSIT-DRGP-010423-1358



INSTITUTE FOR TECHNOLOGY AND RESEARCH
International Conference on
Advanced Computer Science and Information Technology

Certificate

*This is to certify that **Naresh Singh** has presented a paper entitled "**Fake News Detection using Machine Learning Algorithms**" at the International Conference on Advanced Computer Science and Information Technology (ICACSIT) held in Durgapur, India on 01st April, 2023.*



Chairman
**Institute for Technology
and Research**



INSTITUTE FOR TECHNOLOGY AND RESEARCH
International Conference on
Advanced Computer Science and Information Technology

Certificate

This is to certify that *Suraj Dhoundiyal* has presented
a paper entitled "*Fake News Detection using Machine
Learning Algorithms*" at the International Conference
on Advanced Computer Science and Information
Technology(ICACSIT) held in Durgapur, India
on 01st April, 2023.



Chairman
Institute for Technology
and Research



INSTITUTE FOR TECHNOLOGY AND RESEARCH
International Conference on
Advanced Computer Science and Information Technology

Certificate

This is to certify that *Gunjan Kardam* has presented a
paper entitled "*Fake News Detection using Machine
Learning Algorithms*" at the International Conference
on Advanced Computer Science and Information
Technology (ICACSIT) held in Durgapur, India
on 01st April, 2023.



Chairman
**Institute for Technology
and Research**