BANGLA FAKE NEWS DETECTION USING MACHINE LEARNING

Submitted By

Khandakar Tazinur Rahman
182-35-2557
Department of Software Engineering
Daffodil International University

Supervised By

Professor Dr. Engr. Abdul Kadar Muhammad Masum
Professor
Department of Software Engineering
Daffodil International University

A thesis submitted in partial fulfillment of the requirement for the degree of Bachelor of Science in Software Engineering

Spring 2024

Approval of the Thesis

APPROVAL

This thesis titled on "**Detection of Bangla Fake News using Machine learning Technique**", submitted by **Khandakar Tazinur Rahman (ID: 182-35-2557)** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approval as to its style and contents.

BOARD OF EXAMINERS

---------- **Chairman**

Dr. Engr. AKM Masum
Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

---------- **Internal Examiner 1**

Dr. Marzia Ahmed
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

---------- **Internal Examiner 2**

Md. Rajib Mia
Lecturer
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

---------- **External Examiner**

Dr. Md. Manowarul Islam
Associate Professor
Dept. of Computer Science and Engineering
Jagannath University

Declaration

DECLARATION

I announce that I am rendering this study document under **Dr. Engr. AKM Masum**, Professor, Department of Software Engineering, Daffodil International University. I therefore, state that this work or any portion of it was not proposed here therefore for Bachelor's degree or any graduation.

Supervised By

Dr. Engr. AKM Masum,

Professor,

Department of Software Engineering,

Daffodil International University

Submitted By

Khandakar Tazinur Rahman

182-35-2557

Department of Software Engineering

Daffodil International University

Abstract

Knowledge on the other hand is very much potent and the effect of sharing knowledge is highly significant but the sad thing is that sharing knowledge is now associated with share falsehood which has far-reaching consequences. Due to social media people are continuously exposed to fake news in the digital age even if they are not directly a consumer of the news. Social media and news sites are standard means of getting updated information, hence it is all too simple to spread fake news. Such news is packaged as credible but has wrong information on it, or is entirely fabricated. Secondly, the increase in the cases of lynching has been witnessed in recent time because of prevalence of fake news. Furthermore, all COVID-19 falsehoods have led to massive confusion and panic across the world. Still, there is a possibility to automate the fake news' detection, but the major part of such systems is designed for English only. As the babies of the speak bengali around the world, that paper puts forward a novel specifically for the Bangla language to identify fake news. For the pre-processing and feature extraction part on our dataset we have employed several methods. Our insights derived from the performance analysis of the Passive Aggressive Classifier and Support Vector Machine both of which have 93% accuracy reveal that. 8% and 93. specifically, the method achieves a recognition accuracy of over 90%, and false negative and false positive rates of 5%, respectively, higher than other machine learning classifiers.

Dedication

I would like to Dedicate this work to those who were affected by fake news, and encouraged me to focus on this topic. I dedicate this work to our family and faculty, for their help let me get this far. And more so to my supervisor who encouraged and helped me throughout the process.

Acknowledgement

As such, I think that it is fitting that I get to express my initial words of thank to Almighty Allah for granting my this thesis a hitch-free production. Secondly, the author's concern is also to acknowledge them for encouraging them throughout this research endeavor, their supervisor **Dr. Engineer Abdul Kader Masum**. It never felt like he was forcing himself to do it for me and if ever i asked for him to assist me or to do something for me he would readily do it with much of zest. And the last thing I would like to thank my family for their patience and support during the preparation of this report. Here precisely it is evident, and beyond reasonable doubt that without the help of these people, it is out rightly impossible to carry out this research work not to mention prosecuting this Undergraduate Degree.

Table of Contents

List of Figures

List of Tables

Nomenclature

The following set of symbols and acronyms will be subsequently utilized within the body of the paper.

LR=Model Logistic Regression

ML=Machine Learning

MNB=Model of Multinomial Naive Bayes

NLP=Natural Language Processing

PAC=Classifier of Passive-Aggressive

SVM=Model Support Vector Machine

TF-IDF=Term Frequency-Inverse Document Frequency

Chapter I

Introduction

False and incorrect rumor that are intended to fool people into believing it to be real is known as fake news. The phrase "fake news" gained widespread recognition in 2016 during the US presidential campaign when Donald Trump was delaying whatever he was being accused of by utilizing this word. Fake news is spread for political or commercial purposes in an effort to sway public opinion or generate revenue. Fake news influences people's thoughts and persuades them to adopt a certain viewpoint. More often than not, fake news is a concoction of facts and falsehoods. "A false news can travel halfway across the globe before the truth starts to wear trousers. " is a well-known proverb. Fake news distributes much more fast and to a larger audience on social media and digital technology, where information is shared freely and rapidly. We often notice bogus news on internet news sources. The purpose of this erroneous and misleading rumor deceive the watcher. These days, soc-media overflowing by stories of that. The authors this false rumor are not concerned with the topics they cover; instead, they are only looking to profit from it. For example, they may get readers to visit their websites via Facebook and then monetize those visits through advertising. Because the news is intended to be "Clickbait," the writer makes money on the amount of people who click on it. It is quite hard to tell the real news from this phony news. When fake news spreads from the internet to the offline world, it may have detrimental effects. In 2018, a crowd in Mexico killed two individuals who they believed to be child abductors after spreading false information about child abduction over WhatsApp [1]. The mob did not verify the veracity of the information before taking their lives. Similar incidents had occurred in Myanmar and India. Additionally, false information on Facebook and WhatsApp in Sri Lanka, Myanmar, and India led to tragic violence [1]. False information propagates hate and has psychological effects on individuals. According to a nationwide study conducted Bangladesh's incidence of seeing false news is greatest in rural areas (67%), followed by urban areas (63%), according the Management and Resources Development Initiative. and lowest in metropolitan areas (52.5%). Furthermore, just half of those who seek for news online make an effort to determine whether the content they encounter is factual or subjective [2]. Therefore, our suggested method for creating this kind of mechanism is to apply ML methods by the identification of bogus news.

1.1 Motivation

One of the biggest issues facing any nation these days is the rise in lynchings and other violent crimes as a result of the dissemination of false information. Social media and the internet environment have an impact on people. As a result, they spread the word about any story they come across with an alluring title even when they are unsure of its veracity. With the aid of social media, anybody with a keyboard can generate fake news, distribute it quickly, and profit from it. It doesn't matter whether the false information incites violence or not. Some individuals don't care if incorrect information hurts other people because they are fixated by the number of Reactions and Shares Received Following sharing anything on soc-media. In Bangladesh, a significant portion of internet users lack adequate digital literacy. Bangladeshi internet users searched social media for health-related information during the COVID-19 epidemic. The first false material about COVID-19 to surface online in this nation was religious in nature, stating that consuming Thankuni leaves (Indian pennywort) and routinely reciting Bismillah (in the name of Allah) would guard against infection [3]. Furthermore, there was a report on WhatsApp and Facebook that individuals were attempting to abduct children because human sacrifices had to be performed in order to construct the Padma Bridge. The crowds murdered a number of individuals on the street because they thought they were kidnappers [4]. A Facebook account in Ramu disseminated information about a Quran being desecrated in 2012 [5]. Twelve temples and fifty homes were destroyed by a group of around twenty-five thousand people. The Buddhist who was charged with disseminating this was not guilty. A few days ago, a lady accused her husband of domestic abuse while she and her kid were live on Facebook. Nothing similar was discovered by the cops when they got there. A media report later refuted this. Lately, a great deal of Facebook accounts and organizations have been posting requests for assistance from ill people or groups that don't exist. Our proposal is to create a model that employs several machine learning classifiers to more precisely identify bogus news, hence halting its dissemination.

1.2 Problem Statement

Fake news is driving lynching like wildfire over Bangladesh and many other countries. Conversely, the spread of misleading data is aggravating the COVID-19 epidemic scenario. The Management and Resources Development Initiative (MRDI), funded by Unicef, carried a national poll in Bangladesh and found that 63.6% of respondents had come across false news after assuming what they had seen on social media or the internet to be accurate. Moreover, almost two thirds of people never or hardly review the news source that first published it [2]. As much as we can, we wish to fight it as it has grown to be a major concern for Bangladesh. Though they are user-based services, fact checking sites like act-Check, Jachais, and Rumour-Scann do exist; people submit inaccurate information on their own when they find it and by reporting false information, a biassed group of people may often change the context. An autonomous mechanism would be thus increasingly open and tidy.

more precisely, present technologies cannot help one to examine Bangla words and characters. mostly that practically every model created so can work using English words and letters. One needs a model that can understand Bangla characters and words. Someone cannot know whether fake news they came across on Viber, Imo, WhatsApp, or any other messaging platform. Therefore, we need a system that can handle every problem if we are to solve all these flaws in the present models.

1.3 Objective and Contributions

A civilization can be brought under the grip of fake news. For years, scientists have been trying to develop a computerised model capable of identifying false news such that it could enable readers may see it right away and stay away from manipulation. There are many sophisticated algorithms available for identifying rumors only able to identify sentences. There aren't many distinguish bogus rumor. Our key goal create a sophisticated algorithm that can recognise content written in Bangla that is false or fraudulent. To build this model, we'll use a range of state-of-the-art technologies, such as NLP and machine learning model. Using the automated process, the enthusiast to ascertain it's a rumor. This idea applies to the article as well as the headline.

Detect rumor, you need examine each sentence. Using a lot of unnecessary words and punctuation makes it harder to spot fake news. To get rid of these unnecessary words and punctuation, we employed a number of methods of preprocessing unique to not many pre-processing technologies available to bengali. That method would be very beneficial to any upcoming investigator hoping to study the Bengali. Other than that, that idea in its entirety might be a very useful tool for our country and everybody else who reads or speaks Bangla around the world. This essay has the power to stop widespread violence in a civilization and help all of humanity.

1.4 Thesis Structure

There are several chapters in our thesis paper. Every single piece has been broken down into its own sections. Background information, including an overview of the literature and algorithms, is included in Chapter 2. The summary of the several articles that we have read, examined, and assessed for easier comprehension and comparison is included in the literature review. Despite the large number of research articles we have evaluated, we have included around eight studies that are relevant to our study. It includes several classifiers in algorithms that we have used in our work. These classifiers are the Random Forest classifier, Decision Tree classifier, PAC, MNB, SVM, and LR. We presented our suggested model in Chapter 3. The description of the model and the data are included in this part. Data processing and feature extraction are included in data description. The primary focus of data description is the source of our dataset and the methods used to prepare it for training and testing in order to get the desired outcome. Experiments are covered in Chapter 4, where we also talk about how we arrived at the conclusion. We have examined the outcomes from the various classifiers in Chapter 5. We wrapped off the report in Chapter 6 by providing a summary and talking about potential future research.

Chapter 2

Background Study

The harm that fake news may do is immense. We have witnessed several instances of conflict stemming from the dissemination of false information. These days, having a smartphone, the internet, and other gadgets makes things much simpler than they were in the past. These days, internet and smartphone usage is widespread even in rural places. Therefore, it is simpler to quickly disseminate any information around the nation in a short amount of time. However, for many years, Bangladesh did not implement any successful measures to avert this issue. However, academics have been attempting to stop it in recent years by using ML, NLP, deep learning, and other clever techniques. The following are a some of the studies that Bangladeshi scholars have completed:

In order to develop a system, linguistic characteristics and neural network-based techniques have been investigated in [6]. The only aim of this study is to identify bogus news in Bangla. The sixth most popular language is Bengali [7]. The detection of false news in Bangla has seen very little study in the field of natural language processing. Various techniques have been used to extract features. Lexical, syntactic, semantic, metadata, and punctuation aspects are examples of classical linguistic features. However, in the neural network section, they have used convolutional neural networkls, Long Short-Term Memory, and developed language models from past times. Put in use Support Vector Machine, Random Forest, and Logistic Regressions approaches, another the corresponding F1-score accuracy 47%, 56%, and 54%. It is clear from analyzing the POS tag's F1-score that this method cannot identify false news. Furthermore, no development over the arbitrary baseline was seen. morever F1-score using the SVM classifier was 91% once all language data were included. RF outperformed the other two in every instance. They will focus on the neural network models' character-level properties. They want to use them in next studies. However, they will keep adding to their collection of data. With 8.5k annotations, they want to reach 50k. For the experiment, they utilized LR, RF, and SVM models; SVM outperformed the others. On the other hand, the result's correctness was not discussed.

A person developed using Support Vector Machine and Multinomial Naive Bayes Classifier to determine whether a Bangla news source is authentic or not. The dataset consists of gathered by various Bengali rumor items, with 60% of the data being legitimate rumor and 41% being fraudulent rumor, in order train the model to identify fake news in Bangla. Pre-processing the data is essential to achieving greater accuracy. Preprocessing is completed by eliminating punctuation, unique characters, numerical figures, and particular symbolism. Before text sent into the classification algorithms,

the features are extracted using Count vectorizer and TF-IDF. The data was divided into train and test sets in a 70:30 ratio for classification purposes.

The dataset is classified using MNB and SVM classifiers, with classification accuracy of 93% and 96.64%, respectively.

A approach for identifying bogus news in Bengali was developed by Shafayat Bin Shabbir Mugdha et al., and it is capable of accurately determining whether or not the news is genuine based just on the headlines [9]. In order to achieve the goal, a fresh The dataset included created by GNV approach in order for running this model. As part of the data preparation, stopping words were eliminated, tokenisation was completed, and derived was done. Stemming is process of eliminating diacritical marks and inflected words from words. Features are extracted using TF-IDF and Extra Trees Classifier. Even though it is a classifier, the Extra Trees Classifier is used in this model in feature selection method choose the finest elements, and the findings are afterwards used in classifiers to enhance performance and outcomes. Several classifiers were used for classification, including GNV, RF Classifier, Support Vector Machine or Logistic Regression. Greatest precision, on the other hand, was 87.42 percent for Gaussian Naive Bayes.

2.1 Review of the Literary Work

[10] presents few ML methods in their datasets including analysis metrics, NBM, Support Vector Machine, RC, and DT. These models represent the text in the TF-IDF matrix and are used for categorization of it. For assessment liar liar data-set employs NBM, Support Vector Machine, RC, and DT model. Conversely, a uses technique collection uneven. Choosing the closest neighbours under the minority class aids this approach. Techniques, Naive-Bayes, Linear SVM, and Ridge classifier have average accuracy yet, NBM poor on spotting bogus classifies and news majority of the texts as credible. Linear models perform the best for the corpus of bogus news. Decision tree performs 89.4%, Naive-Bayes obtains 85.3%; ridge classifier 93.98% and linear SVM has an accuracy of 94.7%. Not enough for many situations is fake news identification limited to text's supervised algorithms. Additional information should be taken under consideration as the author's information in order to get a suitable answer. This work proposes an automated derived from, which data validate material. [11] many methods of assessing false news are used. Based on the news the article is analysing, it adopts many strategies. This work features LIWC software as a psycholinguistic tool. This work computed subjectivity to evaluate the toxicity of a news story. This work used many classifiers with relation to AUC and F1 score. Among these classifiers are XGBoost, Random Forests, failed in the AUC(72%) and F1(75%) six scores. Conversely, with both Random Forests and XGBoost projected better than the other classifiers. Although identified bogus news effectively, their accuracy still lacked some

degree; so, applying the most recent and more classifier approaches can help to raise the accuracy much more. Predicting authenticity depends on tracking its origins using deep learning and neural networks as merely may employ. Some characteristics for the classifier are trained, occurred, emotion identification. Two Naive Bayes classifier and a Random Forest one. Three datasets in all, including more than 40000 papers, were employed. The classifiers were trained and tested partly using one of the datasets. We just tested the classifiers using the remaining datasets. We applied an original dataset, FakeNewsNet, and ISOT. They selected the optimum accuracy for identification of false news by testing many characteristics on several classifiers. Various characteristics were retrieved, and datasets were used. With had best. Second, with regard to dataset obtained maximum. Outperformance, using ISOT dataset, the lacks the 84.96%. Using FakeNewsNet data, accuracy fell short One NB. One may utilise other characteristics in tandem with this one to boost accuracy. The VADER function produced the least of results. The accuracy was rather poor, hence VADER is not a useful tool for the false news categorization. With its accuracy rates, the ISOT dataset topped the other datasets according to researchers. Features ER, PoS, and VADER brought a poor accuracy rate. More research can help to raise the accuracy rates. Furthermore one may perform a mix of the characteristics. Furthermore not done was the categorization of actual, satirical, and false news—which should be included going forward. In [13] bogus news is found using a novel technique of attitude identification. Stance detection links two separate works of writing. Using the words "agree," "disagree," "discuss," and "unrelated," they noted the posture taken in the news story and title. It developed a method capable of spotting fake news by precisely forecasting the link between news events and headlines.Their approach drew on a Fake News Challenge (FNC-1) dataset. Representation, was the section utilised in data processing. Using, they trained it. Still, their best model used concatenated preprocessed data to forecast the desired posture using dense neural network architecture. Their approach was able to depict, both locally and internationally, the relative importance of a term in article-headline pairings. To updates and casesThey computed the cosine similarity between TF-IDF pairings to make sure of the similarity between headline-article pairs. finest by using. This is a better result than others. In order to build on these findings and develop an autonomous false news detection software, they want to conduct a related study on more datasets. They described a graph-based, semi-supervised approach to spotting fake news in their work [14]. The use of content-based detection algorithms served as their primary weapon. Their solution consists of three parts: creating an article similarity network, inferring missing labels using graph learning techniques, and embedding articles in Euclidean space. The work's main contributions were the use of word embeddings to create latent representations of news items in a lower-dimensional Euclidean space and the capture of contextual similarities across news items using a graph-based representation scheme. The cast of fictitious occupations thus became a difficulty thanks to Graph Neural Network architectures that can function effectively on minimal tagged input. They indicated a range of 2% to 20%. They discovered that there is significant similarity between the results of 84% labeled data and 20% labelled data.

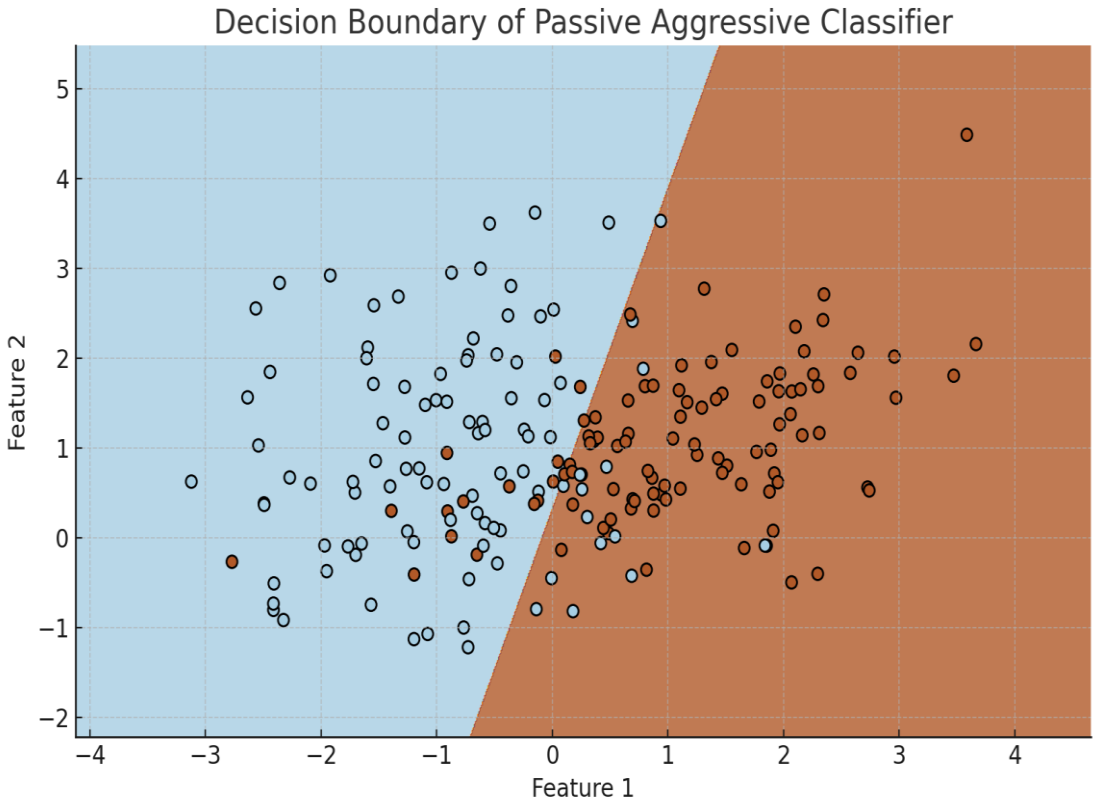
Nevertheless, they discovered that their AGNN and GCN models performed better than the others when they assessed a recent paper.lassifies

2.2 Algorithms for Machine Learning

In our suggested approach, we have used six classifiers to identify bogus news in Bangla.

2.2.1 The Classifier of Passive-Aggressive

When a large data stream exists, an online method known as a passive-aggressive algorithm is applied. This is essentially generates a training sample, absorbs knowledge from it, and modifies the classifier after that discards. Crammer et al [15] first put out this method. Correct forecasts stay passive; on the other hand, erroneous guesses cause response that is aggressive.



This method generates a forecast by means of weighted vectors multiplied with normalized data, therefore determining. It predicts Then notes the actual class of that document. Positive papers have a true class value of +1; negative documents have a value of -1. The method upgrades the weight vector using these true class values may always higher.

Weight vectors' with the document. Considered as a loss. The prediction erroneous if the score of the document. And should score of the document be on the correct side of the decision border, two techniques— Passive and Aggressive—are used.

Passive algorithms look to see whether the dot product exceeds +1. Greater than +1 indicates proper classification of the document. The method so maintains. Conversely, this method investigates number falls less. Should be less, that method determines using. That reveals its distance from the decision boundary.

This method recalues the weight vector once the loss value is known. The new dot product will therefore precisely be +1.

2.2.2 Multinomial Naive Bayes

A document's word count is multinomially distributed in the Multinomial Naive Bayes model [16]. Contains immutable collection of MPs for identification, and the number of classes is fixed as well. It begins by calculating the prior probabilities conditional probabilities each within likelihood equal to sum all files that belong to that class divided by the total number of documents. A word's probability of belonging to a class is calculated by adding its frequency of occurrence in that class to the overall frequency of occurrence in that class, divided by the vocabulary size, and then adding 1.

$$\text{Probability of a class, } P(c) = \frac{N_c}{N}$$

$$\text{Likelihood of a word given a class, } P(w \mid c) = \frac{\text{count}(w, c) + 1}{\text{count}(c) + |V|}$$

Figure 2.2: Equations of Naive Bayes

It will calculate the 'argmax' among the classes by calculating the the probabilities. If you give this model a test document, it can tell you which category it belongs in.

2.2.3 SVM

A SVM is a kind of SML technique solving two-group identification problems by use of classification techniques. An SVM model given labelled training data for every category will classify new text into corresponding categories.

The Support Vector Machine (SVM) algorithm is proficient in performing both classification and regression tasks.The potential segregate 2 categories.The goal of the programme is to find an aircraft with the maximum margin.

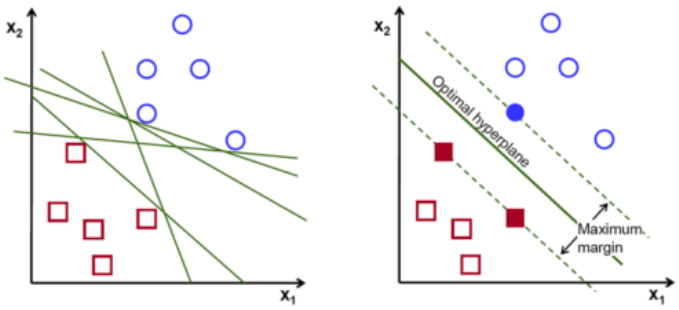


Figure 2.3: Possible Hyperplanes

The function c(x, y, f(x) has as its definition:

One minus y * f(x), or 0 if y * f(x) is less than or equal to 1, and one otherwise defines the function c(x, y, f(x).

Equation (2.1) shows SVM algorithm cost function. .

Eq. depicts the Support Vectors method. Cost is zero when expected and real values have the same sign. The technique calculates the loss value regardless of zero cost.

$$c(x, y, f(x)) = \begin{cases} 0, & \text{if } y * f(x) \geq 1 \\ 1 - y * f(x), & \text{else} \end{cases}$$

(2.1)

A regularisation parameter is added to balance minimising loss with maximising the margin. The value of w is updated by subtracting the product of α and 2λw from w. The value of w is updated by adding the product of α and the difference between yi multiplied by xi and 2λw. The value is 2.3.

$$w = w - \alpha \cdot (2\lambda w) \tag{2.2}$$

$$w = w + \alpha \cdot (y_i \cdot x_i - 2\lambda w) \tag{2.3}$$

The functions computing the gradient are equations (2.2) and (2.3). While Eq. (2.3) is employed when the model produces erroneous predictions, Equation (2.2) is used in the absence of any misclassification.

2.2.4 LR

A supervised machine learning approach in NLP uses logistic regression as its foundation for categorization. It is a probabilistic classifier with strong neural network connections. Training and Testing are the two stages of logistic regression. During the training phase, the cross-entropy loss techniques and stochastic gradient descent are used to train. Additionally, it computes p(y—x) for each test case x during the test phase and returns. Each xi is first multiplied by its weight wi by the classifier. Next, it adds the bias term b to the total of the weighted characteristics. Z indicates the total of the weighted characteristics. The weight wi indicates the input feature's significance to the classification choice. Furthermore, the bias term b, usually referred to as the intercept, is a real value. To generate probability, the classifier uses the sigmoid function on z. Next, using a loss function that quantifies the degree of correspondence, it computes the cross-entropy loss.

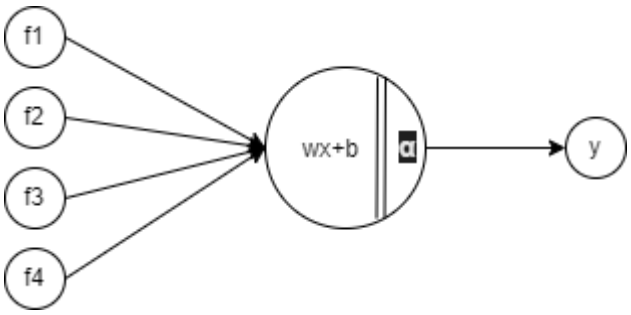
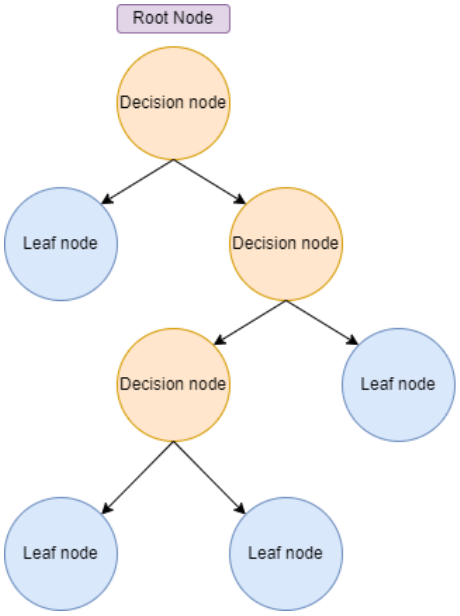


Figure 2.4: Logistic Regression Model

After that, a stochastic gradient descent technique is used to update the weights in order to iteratively reduce this loss function.

2.2.5 DTC

A DTC divides a dataset until it locates nodes that contain only one kind of class of data, or pure leaf nodes. A supervised learning method called decision trees is mostly used to address categorization issues. This classifier has a tree structure, with core nodes representing dataset attributes, leaf nodes representing outcomes, and branches representing decision rules. The Decision Node and Leaf Node are the two nodes that make up the decision tree. Leaf nodes assist in determining the class of a new data point, whereas decision nodes provide a condition to divide the data.

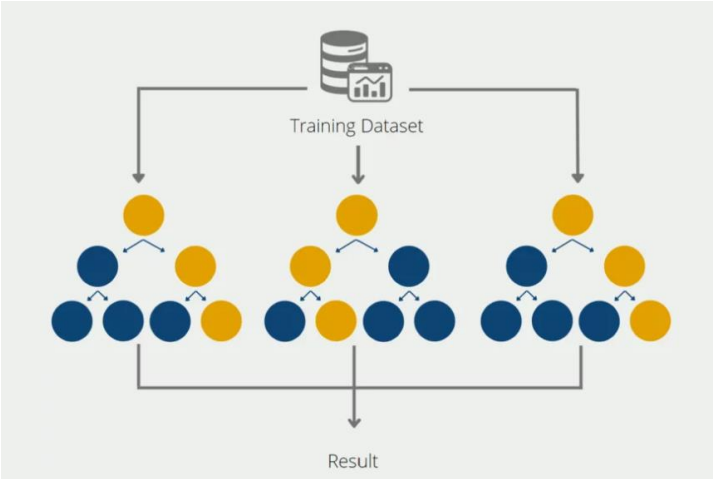


DTC uses Gini Index and Information Gain to forecast the class of the provided dataset. Finding the pure leaf nodes is the primary goal. The root node is divided first by the algorithm. There will be two states in each split, and the method will calculate each state's entropy. The amount of information in a state is measured by its entropy. It evaluates each potential split and selects the one with the lowest entropy since least entropy optimizes the acquisition of information. In order to find the ideal feature and matching threshold, the model iterates over each potential feature and feature value. It doesn't stop doing this until it locates the tree's pure leaf nodes. Although its calculation is simpler, the Gini Index performs the exact same function as Information Gain.

2.2.6 Random Forest

A classification method called random forests uses several decision trees. Using feature randomization and bagging approaches, it creates decision trees.

This technique seeks to produce a forest composed of uncorrelated trees by building each tree separately. In this classification procedure, every single tree produces a forecast for a class; the class with the greatest number of votes becomes the model's prediction in the end.



Creating bootstrap datasets and carrying out aggregation are two processes that are used to create the composite approach known as bagging.

Creating new datasets from the original dataset is the first step in building a random forest. In order to ensure that every dataset, method uses random sampling with replacement, which permits instances to possibly exist multiple times. Aggregation is the findings of many DT. For every dataset, features are selected at random via the

feature randomness method. It does weaken bond of the feature randomization additionally bootstrapping, DT finds a unique forecast.

Chapter III

Suggested Framework

3.1.1 Overview of Dataset

Referred to inputs. In essence, the predictors serve as the input. A dataset is a collection of data files. One or more database tables may be found in tabular datasets. In a table, each row represents an instance of a given variable, which is indicated by each column. Multiple documents or files make up a dataset [21].

Bangla Fake News Detection's main obstacle has been gathering the input data. The Bangla language has received very little investigation on this subject. Obtaining an appropriate dataset that aligns with our study is akin to requesting the moon. On the other hand, as fake news detection algorithms are now being developed extensively, That substantial quantity accessible for many prominent language. Acquiring a good dataset was challenging for a language with limited resources like Bangla. The information was compiled from several news outlets and websites.

Table 3.1: Contents of Dataset

File Name	Authentic48K.csv	Fake-1K.csv
Number of contents(rows)	48k	1.3k
Columns	7	7
Features	Article Id, Date, Domain, Headline, Content, Label	Article Id, Date, Domain, Headline, Content, Label

Two of the four Microsoft Excel.csv files—Authentic48K.csv and Fake-1K.csv—were utilized in our study. About 48k of the labeled dataset's 14 news are real, while the remaining 1.3k are false. We gathered an additional dataset created by GUB.

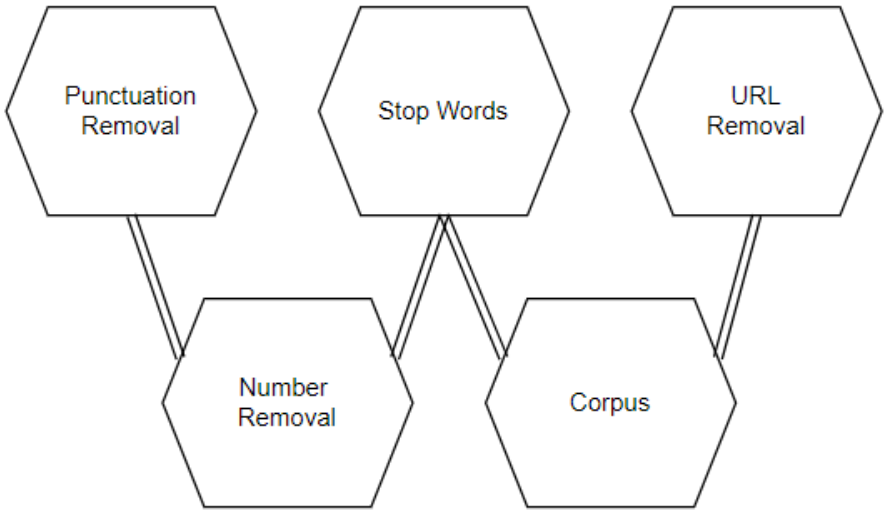
ArticleID	Serial number of the news
Domain/Source	Website address from where the news is collected
Date	Published Date
Category	Type of news
Headline/Title	Headline of the news
Content	Body of the news
Label	Information that if the news is fake or not

That tagged, however in order identifying Bangla Fake rumor, it must undergo further processing.

3.1.1 Preparing the Data

Information preparation converting unprocessed info into a type that is well-suited for specific requirements. Real-world data may lack completeness and consistency. The statistics exhibit inaccuracies and deficiencies patterns. Info preparation has the capability to address these concerns. Data preprocessing is the process of appropriately preparing raw data for later processing. Data preparation in Machine Learning (ML) involves transforming the dataset into a format that can be readily understood and analyzed by the algorithm [22].

Data undergoes a series of sequential processes throughout the preparation phase. During the first stage info cleansing, eliminated numbers, punctuation, null values, duplicate values, and other unwanted elements from the original dataset. Next, we eliminated the stop words. Eliminating stop words is the process of filtering away words that has little semantic significance.



In order to facilitate further analysis, we organize the data into two distinct forms to ensure cleanliness and standardization. They are -

• **Corpus**: A corpus is a compilation of written or spoken materials. In order find information, using PY(python) framework known as PANDAS.Pandas. Then stored in a DataFrame, which essentially functions..

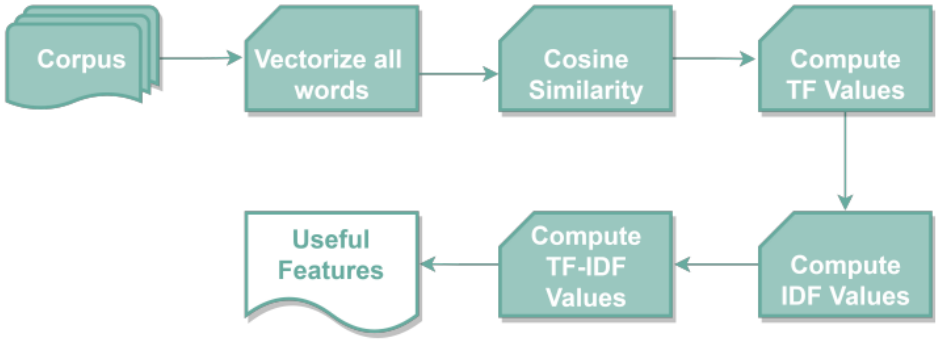
• **LINK REMOVE**: Link's relate to hyperlinks elements inside the material. These links do not provide any relevant information. As part of the preparation stage, we eliminated the link's.

The corpus is now prepared.

The process of turning raw data into a format that meets our needs precisely is known as data preparation. Real-world data may not be consistent or comprehensive. There are flaws and errors in certain patterns or trends in the data. These issues may be resolved by data preparation. Preparing raw data for further processing in a suitable manner is known as data preparation. Making the dataset in a format that the algorithm can easily comprehend and evaluate is known as data preparation in machine learning (ML) [22].

3.1.2 Feature Extraction

The process of extracting data from a corpus that indicates the significance of a certain word or phrase is known as feature extraction. Our use of term frequencyOne of the finest feature extraction methods in our sample is Inverse Document Frequency (TF-IDF). TF-IDF transforms the corpus into meaningful features as classification models do not allow us to transmit the corpus directly. The number of times a word appears in a sentence relative to the total number of words in the phrase is determined using the Term Frequency (TF) approach. Additionally, the Inverse Document Frequency (IDF) approach calculates a word's frequency of occurrence in all sentences.

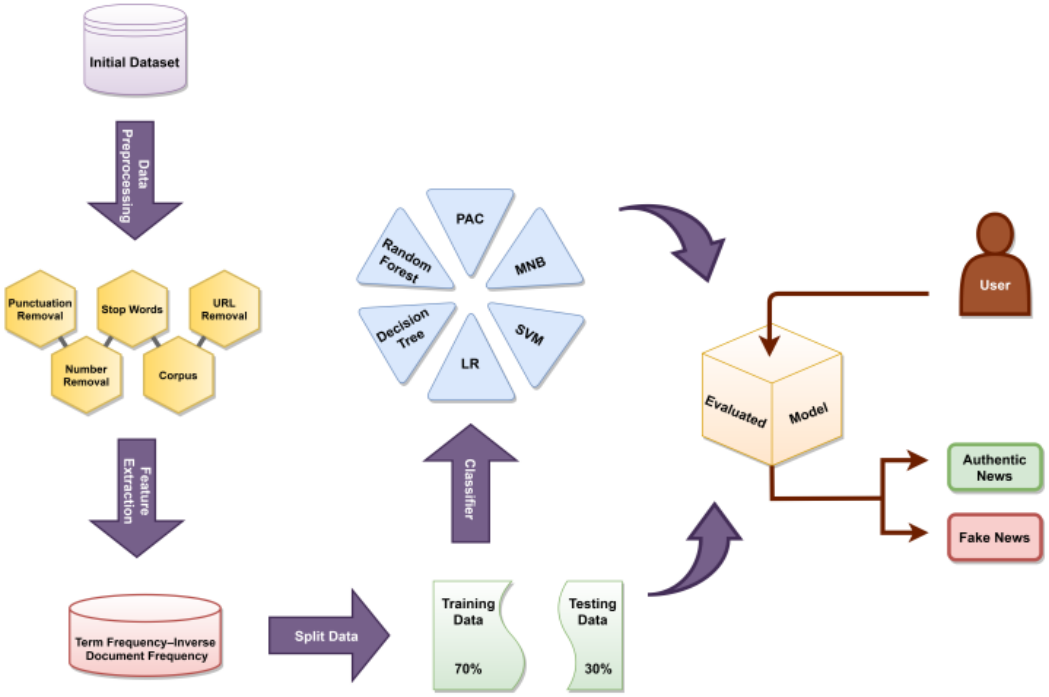


The value of the T F-I DF. A word or phrase is considered rarer the higher its TF-IDF score. We have loaded the "TfidfVectorizer" from the scikit-learn package in order to Apply TF-IDF. The formula does not need to be manually written. Following the library's import, we built a "TfidfVectorizer" object. With the use of the thing method an array to which our corpus has been supplied.

Initially, this function turns each word into a vector. Next, it uses the Cosine Similarity function to determine how similar two or more vectors are to one another. Finally, it offers the whole corpus's TF-IDF values in numerical form.

3.2 Model Description

We have attempted to choose the best strategy for the model in order to achieve our objective of identifying bogus news. Figure 3.3 shows the suggested model. In order to feed the machine learning classifiers and extract features, the data had to be processed



after the original dataset was collected. This stage is essential for solving various string classification issues and aids in obtaining accurate data from the dataset. Many hours of effort were needed for the preparation phase, which included removing punctuation marks, stop words, URLs, numbers, case conversion, named entity recognition, stemming, lemmatizing, and other things. Due to the lack of appropriate equipment, we were forced to disregard some procedures since we are dealing with the Bangla language. Therefore, in order to clean the dataset, we eliminated digits, stop words, and punctuation. A technique called TF-IDF was used to extract features. Following the first two stages, the data was divided into training and testing segments for the machine learning classifier using a 70:30 ratio. We have experimented with many classifier

types, including Random Forest, Multinomial Naive Bayes (MNB), Support Vector Machine (SVM), Passive-Aggressive classifier (PAC), and Logistic Regression. Subsequently, the model determines which classifier is the best at identifying false information. Our plan is to provide an online platform where users may manually enter material and verify its legitimacy after the model has been deployed.

Chapter IV

Trial and error

4. 1 The Classification Dataset

With 49.3k data, we have a sizable dataset for categorization. Nevertheless, of these, 48,000 are real data, while the remaining 1.3K are false reports. It is clear that the difference between fictitious and authentic datasets is enormous. We would have gotten skewed results in favor of real news if we had utilized the complete dataset. As a result, the model would not be the best choice for categorizing any data.

4. 2 Elimination of Punctuation

To apply our proposed model we had to go through a bunch of steps. After collecting the dataset, the main challenge was to prepare it for the machine learning classifiers. In the preprocessing part we removed punctuation marks (! , . ? : ; {} [ ] ( ) ‘ “ / \ etc).

4. 3 Elimination of Stop Words

This kind stop words arise in those that do not supply us any required meaning. We must thus exclude such terms from the corpus. Us dealing with the Bengali linguistic,

and not much has been done in that regard., languages stop word libraries are generally accessible for the mainstream and we had to manually enter Bangla stopwords.

Input: আমি ভাত খাই। তোমাদের দেশের নাম কি? সবার জন্য আইন সমান।

Output: ভাত খাই। দেশের নাম। আইন সমান।

4.4 Stemming

In natural language processing, stemming has become quite popular as every letter boiled down to its base form. Text inflection is reduced by stemming.. We aimed to use stemming on our dataset during pre-processing. Sadly, however, we did not locate any instruments capable of exact stemming on Bangla text.

4.5 Extracting Features

Using TF-IDF vectorizer for feature extraction will help us to calculate the descriptiveness of a term and get the word relevance between our synthetic and real texts.

	333	334	335	336	337	338	339	340	341	342	343
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.139566
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000
...	...	...	...	...	...	...	...	...	...	...	...
4017	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000
4018	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000
4019	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000
4020	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000
4021	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000

4. 6 Combining the Headline and Content

For headline and text, we set up two separate columns. At first, we discovered two unique findings on content and headline. We discovered later that concatenating the two columns produces better results. The outcome revealed poor result. That occurred with 8–9 words, are only a sentence. The models so could not be readily taught.

Table 4.2: Before Merging ‘Headline’ and ‘Content’ Columns

	headline	content	class
0	মুরগির হামলায় শেয়াল নিহত	বাংলায় একটা প্রবাদ আছে, শেয়ালের কাছে মুরগী বর্গা দেওয়া। প্রবাদটা এজন্য বলা হয় যে, শিয়ালের কাজই হলো মুরগী খেয়ে ফেলা। শেয়ালের কাছ থেকে মুরগী কখনো নিজেকে রক্ষা করতে পারেনা।তবে ফ্রান্সের এক বিশ্ববিদ্যালয়ের শিক্ষক-শিক্ষার্থীরা জানাচ্ছে, এবার এক আশ্চর্য্যজনক ঘটনা ঘটছে। মুরগীর হামলায় শিয়াল মারা গেছে।ঘটনাটি গত সপ্তাহের, ফ্রান্সের উত্তর-পূর্বাঞ্চলের বুতানিয়া এলাকার একটি কৃষিবিষয়ক বিদ্যালয়ে। ওই বিদ্যালয়ে একটি মুরগির খামার রয়েছে। মুরগির ঘরের এক কোনায় সকালে মৃত একটি শিয়াল পড়ে থাকতে দেখে শিক্ষার্থীরা।৬ হাজার মুরগি রয়েছে খামারটিতে। শিয়ালটির মৃতদেহ যেখানে পাওয়া গেছে, সেখানে তিন হাজার মুরগি ছিল। মুরগিগুলো সারা দিন বাইরে চরে বেড়ায়। সন্ধ্যা হলে নিজ থেকে ঘরে উঠে আসে। সূর্য ডুবে গেলেই দরজা বন্ধ হয়ে যায় স্বয়ংক্রিয়ভাবে।স্থানীয় বন্য প্রাণী বিশেষজ্ঞরা জানিয়েছেন, এই ঘটনায় তাঁরা অবাক। শিয়ালটি বাচ্চা ও অনভিজ্ঞ ছিল। এতগুলো মুরগির সামনে পড়ে সে সম্ভবত ভড়কে গিয়েছিল।Source- BBC Bangla	0
1	বিটিভিতে যেবার আমি ইন্টারভিউ দিতে গেলাম	BTV থেকে লোকজন আসছে, ইন্টারভিউ নিবে।চারজনের টিম, এদের মধ্যে স্যুট, টাই পড়া বস একজন, দুইজন ক্যামেরাম্যান, আরেকজন উঠতি বয়সের, কড়া মেকাপ দেওয়া নারী মডেল।বস যিনি, আমার সাথে হাত মিলিয়ে পরিচয় হবার পর, পাশে দাঁড়ানো এক সিকিউরিটি গার্ডকে বললেন, '... তুমি এক কাজ করো, নুরুল ভাই কফির কাপ হাতে নিয়ে তোমার সামনে দিয়ে হেঁটে যাবে, তুমি আকাশ বাতাস কাঁপাইয়া ঠাস করে একটা সেলুট দিবা, এইটা ভিডিও করে রাখি, বাঙ্গালি পাবলিকের ভুং ভাং পছন্দ।'... এইটা না করলে হয় না, কেমন যেনা বান্যানা লাগে – ভদ্রলোকের আইডিয়া শুনে আমি রীতিমতো বিব্রত।'... আরে স্যার কী বলেন? আপনার ফ্যাক্টরিতে আদব কায়দা আছে, এইটা তোফায়েল সাহেব দেখতে চাইবে।'... তোফায়েল সাহেব কে?''কমার্স মিনিস্টার তোফায়েল সাহেব। নেন– কফির কাপে চুমুক দিতে দিতে হাঁটা শুরু করেন।'হাঁটাহাঁটি ভিডিও হবার পর, আমার অফিস রুমে বসে ইন্টারভিউ শুরু হলো।'স্যার, ক্যামেরা অন হবার পর আপনি বলা শুরু করবেন।'কী বলবো?''বাংলাদেশের আর্থ সামাজিক প্রেক্ষাপটে হোম টেক্সটাইলসের অবদান আর তাতে ইউনিলায়েন্স টেক্সটাইলসের ভূমিকা।'আমি কিছুক্ষণ তবদা মেরে থাকলাম। 'যেই লাইন বললেন, এইটার বাংলা কী?''এইটাই বাংলা, একটু গুছাইয়া এইটা ঐটা বইল্যা দিয়েন, আর শেষে বর্তমান সরকারের ইতিবাচক ভূমিকা ক্লিয়ার করে দিবেন, এতটুকুই, পারবেন না? স্যার, আপনার পকেটে চিরুনি আছে?''না তো, কেন?''আমার চুলটা একটু আচড়াইয়া নিতাম, চুল এলোমেলো থাকলে আপনার ভাবী আবার রাগ করে, আমার ক্যামেরা ফেস ভালো, খালি এলোমেলো চুলটাই সমস্যা করে। এইটা নিয়া ডাইরেক্টর স্যার একদিন ডাক দিয়া বলছেন, আজিজ, তোমার এইটা চুল না বা" এইটাই বুঝলাম না, উনি লোক ভালো, কিন্তু মুখ খারাপ।	0
2	বিদেশ থেকে উন্নতমানের বিরোধীদল আমদানি করার পরামর্শ দিলেন অনলাইন রাজনীতি বিশেষজ্ঞরা	অদ্ভুত বিরোধীদলহীনতায় ভুগছে সরকার। এ এক অন্যরকম শূন্যতা। বাবা-মায়ের সাথে থেকেও আজকালকার ছেলেমেয়েরা যেমন একাকীত্বে ভোগে, তেমনি সবকিছু ছাড়া বেষ্টিত থাকলেও বর্তমান সরকারের পাশে 'কী যেন নেই! কী যেন নেই!'প্রচন্ড অনিচ্ছায় জাতীয়পার্টি বিরোধীদলের চেয়ারে বসলেও, তারা ঠিক যেন বিরোধীদল নয়। অনেকটা বিরোধীদলের রেপ্লিকার মত। এমতাবস্থায় সরকারকে বিরোধীদল আমদানির পরামর্শ দিয়েছেন অনলাইন রাজনীতি বিশেষজ্ঞরা।নাম প্রকাশে অনিচ্ছুক জনৈক বিরোধীদল বিশেষজ্ঞ আমাদের বলেন, 'বর্তমান সরকারের এখন একটা জিনিসেরই অভাব, আর সেটা হল উন্নতমানের বিরোধীদল। অতিসত্বর এ ব্যাপারে পদক্ষেপ নেয়া উচিত। দেশে না পাওয়া গেলে বিদেশ থেকে উন্নতমানের বিরোধীদল আনাতে হবে।''আলীবাবা' অথবা 'অ্যামাজন' থেকে অর্ডার করে বিরোধীদল আনানো ঠিক হবে কিনা এমন প্রশ্নে ওই বিশেষজ্ঞ চোখ কপালে তুলে বলেন, 'এটা মোটেই ঠিক হবে না! অনলাইনে অর্ডার করলে কখনোই ভাল মাল পাওয়া যায় না।' জাতীয়পার্টিকে অনলাইনে অর্ডার করে আনানো হয়েছে, এমন ধারণাও তিনি পোষণ করেন।'বর্তমান বিরোধীদলকে 'মেড ইন চায়না' উল্লেখ করে তিনি আরো বলেন, 'চায়না হ্যান্ডসেটের মত এদেরকেও স্পিকারে সাউন্ড বেশি। তবে সরকারের উচিত 'মেড ইন জাপান' অথবা 'মেড ইন ভিয়েতনাম' হ্যান্ডসেট ব্যবহার করা'। এদিকে বিরোধীদল উৎপাদন করে এমন দেশ Google-এ না পেয়ে আমরা সরকারের উচ্চপর্যায়ে কথা বলতে যাই। সেখানে গিয়ে আমরা প্রচন্ড ধমধমে পরিবেশ দেখতে পাই। যেন পুরো অফিসেরই 'মন খারাপ'। অফিসের এক কোণায় স্পিকারে প্লো সাউন্ডে গান বাজতে শুনি- 'তোমরা কেউ কি দিতে পারো বিরোধীদলের ভালবাসা?'	0
3	অবসর নেয়ার ঘোষণা দিলেন মেসি !	রাশিয়া বিশ্বকাপ নকআউট পর্বে ফ্রান্সের সাথে ৪-৩ গোলে পরাজয়ের গ্লানিতে ফুটবল থেকে অবসরের ঘোষণা দিলেন মেসি। খেলা শেষে মাঠ ছাড়তে ছাড়তে বিড়বিড় করে এমনটাই বলছিলেন আর্জেন্টাইন এই সুপারস্টার !ভালোই চলছিলো ম্যাচ। শুরুতে ফ্রান্স গোল দিলে মাঠ কাঁপালেও পরবর্তীতে দুটি গোল যুক্ত হয় আর্জেন্টিনার ঝুলিতে। বুক চেতিয়ে খেলা শুরু করেন মেসি। হঠাৎই বদলে যায় ম্যাচের দৃশ্যপট। একে একে ফ্রান্স আরো তিনটি গোল ঢুকিয়ে দেয় আর্জেন্টিনার জালে। এরই সাথে বিশ্বকাপ থেকে বিদায় ঘন্টা বেজে যায় আর্জেন্টিনার!মাঠ ছাড়ার সময় গেটের চিপায় থাকা এক সাংবাদিক দেখতে পান, ক্লান্তি অবসাদপূর্ণ শরীরে মাঠ ছাড়তে ছাড়তে মেসি মাথা নিচু করে কি যেনো বলছেন। ওই সাংবাদিক মেসির লিপ রিডিং করে বুঝতে পারেন, মেসি বলছে, "আমি আর ফুটবল খেলবোনা। আমি আর ফুটবল খেলবোনা।"ব্যাতাসের বেগে কথাটি স্টেডিয়ামজুড়ে ছড়িয়ে পরে। পরবর্তীতে মেসির অবসর নেয়ার ব্যাপারে দ্রুত এ খবরটি স্টেডিয়াম প্রাঙ্গনে ভাইরাল হয়ে যায়। এবং অনেকেই তা বিশ্বাস করে বসেন। যদিও আনুষ্ঠানিকভাবে মেসি এখনো কিছু জানাননি !এখন শুধুই এ কিংবদন্তীর মুখ থেকে অবসর নেয়ার ঘোষণাটি শোনার পালা...	0

	headline & content	class
0	মুরগির হামলায় শেয়াল নিহত। বাংলায় একটা প্রবাদ আছে, শেয়ালের কাছে মুরগী বর্গা দেওয়া। প্রবাদটা এজন্য বলা হয় যে, শিয়ালের কাজই হলো মুরগী খেয়ে ফেলা। শেয়ালের কাছ থেকে মুরগী কখনো নিজেকে রক্ষা করতে পারেনা।তবে ফ্রান্সের এক বিশ্ববিদ্যালয়ের শিক্ষক-শিক্ষার্থীরা জানাচ্ছে, এবার এক আশ্চর্য্যজনক ঘটনা ঘটছে। মুরগীর হামলায় শিয়াল মারা গেছে। ঘটনাটি গত সপ্তাহের, ফ্রান্সের উত্তর-পূর্বাঞ্চলের বুতানিয়া এলাকার একটি কৃষিবিষয়ক বিদ্যালয়ে। ওই বিদ্যালয়ে একটি মুরগির খামার রয়েছে। মুরগির ঘরের এক কোনায় সকালে মৃত একটি শিয়াল পড়ে থাকতে দেখে শিক্ষার্থীরা।৬ হাজার মুরগি রয়েছে খামারটিতে। শিয়ালটির মৃতদেহ যেখানে পাওয়া গেছে, সেখানে তিন হাজার মুরগি ছিল। মুরগিগুলো সারা দিন বাইরে চরে বেড়ায়। সন্ধ্যা হলে নিজ থেকে ঘরে উঠে আসে। সূর্য ডুবে গেলেই দরজা বন্ধ হয়ে যায় স্বয়ংক্রিয়ভাবে।স্থানীয় বন্য প্রাণী বিশেষজ্ঞরা জানিয়েছেন, এই ঘটনায় তাঁরা অবাক। শিয়ালটি বাচ্চা ও অনভিজ্ঞ ছিল। এতগুলো মুরগির সামনে পড়ে সে সম্ভবত ভড়কে গিয়েছিল।Source- BBC Bangla	0
1	বিটিভিতে যেবার আমি ইন্টারভিউ দিতে গেলাম। BTV থেকে লোকজন আসছে, ইন্টারভিউ নিবে।চারজনের টিম, এদের মধ্যে স্যুট, টাই পড়া বস একজন, দুইজন ক্যামেরাম্যান, আরেকজন উঠতি বয়সের, কড়া মেকাপ দেওয়া নারী মডেল।বস যিনি, আমার সাথে হাত মিলিয়ে পরিচয় হবার পর, পাশে দাঁড়ানো এক সিকিউরিটি গার্ডকে বললেন, '... তুমি এক কাজ করো, নুরুল ভাই কফির কাপ হাতে নিয়ে তোমার সামনে দিয়ে হেঁটে যাবে, তুমি আকাশ বাতাস কাঁপাইয়া ঠাস করে একটা সেলুট দিবা, এইটা ভিডিও করে রাখি, বাঙ্গালি পাবলিকের ভুং ভাং পছন্দ।'... এইটা না করলে হয় না, কেমন যেনা বান্যানা লাগে – ভদ্রলোকের আইডিয়া শুনে আমি রীতিমতো বিব্রত।'... আরে স্যার কী বলেন? আপনার ফ্যাক্টরিতে আদব কায়দা আছে, এইটা তোফায়েল সাহেব দেখতে চাইবে।'... তোফায়েল সাহেব কে?''কমার্স মিনিস্টার তোফায়েল সাহেব। নেন– কফির কাপে চুমুক দিতে দিতে হাঁটা শুরু করেন।'হাঁটাহাঁটি ভিডিও হবার পর, আমার অফিস রুমে বসে ইন্টারভিউ শুরু হলো।'স্যার, ক্যামেরা অন হবার পর আপনি বলা শুরু করবেন।'কী বলবো?''বাংলাদেশের আর্থ সামাজিক প্রেক্ষাপটে হোম টেক্সটাইলসের অবদান আর তাতে ইউনিলায়েন্স টেক্সটাইলসের ভূমিকা।'আমি কিছুক্ষণ তবদা মেরে থাকলাম। 'যেই লাইন বললেন, এইটার বাংলা কী?''এইটাই বাংলা, একটু গুছাইয়া এইটা ঐটা বইল্যা দিয়েন, আর শেষে বর্তমান সরকারের ইতিবাচক ভূমিকা ক্লিয়ার করে দিবেন, এতটুকুই, পারবেন না? স্যার, আপনার পকেটে চিরুনি আছে?''না তো, কেন?''আমার চুলটা একটু আচড়াইয়া নিতাম, চুল এলোমেলো থাকলে আপনার ভাবী আবার রাগ করে, আমার ক্যামেরা ফেস ভালো, খালি এলোমেলো চুলটাই সমস্যা করে। এইটা নিয়া ডাইরেক্টর স্যার একদিন ডাক দিয়া বলছেন, আজিজ, তোমার এইটা চুল না বা" এইটাই বুঝলাম না, উনি লোক ভালো, কিন্তু মুখ খারাপ।	0
2	বিদেশ থেকে উন্নতমানের বিরোধীদল আমদানি করার পরামর্শ দিলেন অনলাইন রাজনীতি বিশেষজ্ঞরা। অদ্ভুত বিরোধীদলহীনতায় ভুগছে সরকার। এ এক অন্যরকম শূন্যতা। বাবা-মায়ের সাথে থেকেও আজকালকার ছেলেমেয়েরা যেমন একাকীত্বে ভোগে, তেমনি সবকিছু ছাড়া বেষ্টিত থাকলেও বর্তমান সরকারের পাশে 'কী যেন নেই! কী যেন নেই!'প্রচন্ড অনিচ্ছায় জাতীয়পার্টি বিরোধীদলের চেয়ারে বসলেও, তারা ঠিক যেন বিরোধীদল নয়। অনেকটা বিরোধীদলের রেপ্লিকার মত। এমতাবস্থায় সরকারকে বিরোধীদল আমদানির পরামর্শ দিয়েছেন অনলাইন রাজনীতি বিশেষজ্ঞরা।নাম প্রকাশে অনিচ্ছুক জনৈক বিরোধীদল বিশেষজ্ঞ আমাদের বলেন, 'বর্তমান সরকারের এখন একটা জিনিসেরই অভাব, আর সেটা হল উন্নতমানের বিরোধীদল। অতিসত্বর এ ব্যাপারে পদক্ষেপ নেয়া উচিত। দেশে না পাওয়া গেলে বিদেশ থেকে উন্নতমানের বিরোধীদল আনাতে হবে।''আলীবাবা' অথবা 'অ্যামাজন' থেকে অর্ডার করে বিরোধীদল আনানো ঠিক হবে কিনা এমন প্রশ্নে ওই বিশেষজ্ঞ চোখ কপালে তুলে বলেন, 'এটা মোটেই ঠিক হবে না! অনলাইনে অর্ডার করলে কখনোই ভাল মাল পাওয়া যায় না।' জাতীয়পার্টিকে অনলাইনে অর্ডার করে আনানো হয়েছে, এমন ধারণাও তিনি পোষণ করেন।'বর্তমান বিরোধীদলকে 'মেড ইন চায়না' উল্লেখ করে তিনি আরো বলেন, 'চায়না হ্যান্ডসেটের মত এদেরকেও স্পিকারে সাউন্ড বেশি। তবে সরকারের উচিত 'মেড ইন জাপান' অথবা 'মেড ইন ভিয়েতনাম' হ্যান্ডসেট ব্যবহার করা'। এদিকে বিরোধীদল উৎপাদন করে এমন দেশ Google-এ না পেয়ে আমরা সরকারের উচ্চপর্যায়ে কথা বলতে যাই। সেখানে গিয়ে আমরা প্রচন্ড ধমধমে পরিবেশ দেখতে পাই। যেন পুরো অফিসেরই 'মন খারাপ'। অফিসের এক কোণায় স্পিকারে প্লো সাউন্ডে গান বাজতে শুনি- 'তোমরা কেউ কি দিতে পারো বিরোধীদলের ভালবাসা?'	0
3	অবসর নেয়ার ঘোষণা দিলেন মেসি ।। রাশিয়া বিশ্বকাপ নকআউট পর্বে ফ্রান্সের সাথে ৪-৩ গোলে পরাজয়ের গ্লানিতে ফুটবল থেকে অবসরের ঘোষণা দিলেন মেসি। খেলা শেষে মাঠ ছাড়তে ছাড়তে বিড়বিড় করে এমনটাই বলছিলেন আর্জেন্টাইন এই সুপারস্টার !ভালোই চলছিলো ম্যাচ। শুরুতে ফ্রান্স গোল দিলে মাঠ কাঁপালেও পরবর্তীতে দুটি গোল যুক্ত হয় আর্জেন্টিনার ঝুলিতে। বুক চেতিয়ে খেলা শুরু করেন মেসি। হঠাৎই বদলে যায় ম্যাচের দৃশ্যপট। একে একে ফ্রান্স আরো তিনটি গোল ঢুকিয়ে দেয় আর্জেন্টিনার জালে। এরই সাথে বিশ্বকাপ থেকে বিদায় ঘন্টা বেজে যায় আর্জেন্টিনার!মাঠ ছাড়ার সময় গেটের চিপায় থাকা এক সাংবাদিক দেখতে পান, ক্লান্তি অবসাদপূর্ণ শরীরে মাঠ ছাড়তে ছাড়তে মেসি মাথা নিচু করে কি যেনো বলছেন। ওই সাংবাদিক মেসির লিপ রিডিং করে বুঝতে পারেন, মেসি বলছে, "আমি আর ফুটবল খেলবোনা। আমি আর ফুটবল খেলবোনা।"ব্যাতাসের বেগে কথাটি স্টেডিয়ামজুড়ে ছড়িয়ে পরে। পরবর্তীতে মেসির অবসর নেয়ার ব্যাপারে দ্রুত এ খবরটি স্টেডিয়াম প্রাঙ্গনে ভাইরাল হয়ে যায়। এবং অনেকেই তা বিশ্বাস করে বসেন। যদিও আনুষ্ঠানিকভাবে মেসি এখনো কিছু জানাননি !এখন শুধুই এ কিংবদন্তীর মুখ থেকে অবসর নেয়ার ঘোষণাটি শোনার পালা...	0

name of model	accuracy	precision	recall	f1-score
Passive Aggressive Classifier	0.863956	0.896979	0.942540	0.919195
Multinomial Naive Bayes	0.861350	0.859813	0.993016	0.921626
Support Vector Machine	0.874902	0.875211	0.988571	0.928444
Logistic Regression	0.868908	0.868560	0.990159	0.925382
Decision Tree Classifier	0.858223	0.901912	0.928254	0.914894
Random Forest Classifier	0.878030	0.893255	0.966984	0.928659

	name of model	accuracy	precision	recall	f1-score
0	Passive Aggressive Classifier	0.919401	0.931003	0.939647	0.935305
1	Multinomial Naive Bayes	0.867012	0.830469	0.987001	0.901994
2	Support Vector Machine	0.928613	0.932004	0.954503	0.943119
3	Logistic Regression	0.910190	0.911528	0.947075	0.928962
4	Decision Tree Classifier	0.859528	0.891080	0.881151	0.886088
5	Random Forest Classifier	0.911341	0.892099	0.974930	0.931677

4. 7 Findings Using Various Data Split Ratios

Two guidelines exist for selecting the ideal. We investigated our model to see how it is performing in line with them. Though technically legitimate news shows more on our feeds, false news is quite common in real life. These are some findings using various types.

name of model	accuracy	precision	recall	f1-score
Passive Aggressive Classifier	0.922033	0.925291	0.949461	0.937220
Multinomial Naive Bayes	0.848242	0.805159	0.992618	0.889115
Support Vector Machine	0.925165	0.917838	0.964225	0.940460
Logistic Regression	0.918552	0.908945	0.963657	0.935502
Decision Tree Classifier	0.851027	0.877620	0.879614	0.878616
Random Forest Classifier	0.910198	0.881666	0.985804	0.930831

	name of model	accuracy	precision	recall	f1-score
0	Passive Aggressive Classifier	0.919530	0.929715	0.940845	0.935247
1	Multinomial Naive Bayes	0.853849	0.811853	0.993662	0.893604
2	Support Vector Machine	0.919530	0.919212	0.953521	0.936053
3	Logistic Regression	0.906481	0.899801	0.954930	0.926546
4	Decision Tree Classifier	0.862549	0.877049	0.904225	0.890430
5	Random Forest Classifier	0.914311	0.890236	0.982394	0.934048

	name of model	accuracy	precision	recall	f1-score
0	Passive Aggressive Classifier	0.933043	0.938953	0.948605	0.943755
1	Multinomial Naive Bayes	0.863478	0.815663	0.994126	0.896095
2	Support Vector Machine	0.926957	0.919831	0.960352	0.939655
3	Logistic Regression	0.906087	0.893004	0.955947	0.923404
4	Decision Tree Classifier	0.865217	0.891369	0.879589	0.885440
5	Random Forest Classifier	0.926957	0.903924	0.980910	0.940845

name of model	accuracy	precision	recall	f1-score
Passive Aggressive Classifier	0.937935	0.944129	0.954067	0.949072
Multinomial Naive Bayes	0.869490	0.826954	0.992344	0.902131
Support Vector Machine	0.935035	0.933148	0.961722	0.947220
Logistic Regression	0.925174	0.918647	0.961722	0.939691
Decision Tree Classifier	0.868329	0.894788	0.887081	0.890918
Random Forest Classifier	0.932135	0.907733	0.988517	0.946404

In all cases, find support vector machines (SVM) and passive-aggressive classifier (PAC) performed satisfactorially. While various result, in them 70/30 produced the greatest performance.

4. 8 The Algorithm of Gradients Boosting

A GB kind of method. That provides findings based forthcoming technique. That method lowers prediction results fault. Used that method another hand discovered current method is doing great.

Gradient Boosting Classifier Acuuracy Score -> 0.9064810787298826

	precision	recall	f1-score	support
0	0.93	0.81	0.87	879
1	0.89	0.96	0.93	1420
accuracy			0.91	2299
macro avg	0.91	0.89	0.90	2299
weighted avg	0.91	0.91	0.91	2299

4. 9 Manually Verifying News

We have created a manual testing model where a user may enter the news and several classifiers would predict if the news is false or not, thereby allowing us to experiment with our model.

```
news = str(input())
manual_testing(news)
```

এবার রাজধানীতে ঘটলো আরেকটি চাঞ্চল্যকর হত্যাকান্ডের ঘটনা। ধারের টাকা পরিশোধ না করায় প্রচন্ড গরমের মধ্যে সোয়েটার পড়িয়ে মেরে ফেলা হয়েছে খায়রুল আলম নামে এক যুবককে। রাজধানীর মানিকনগর পাওয়ারহাউস এলাকায় এই হত্যাকান্ডের ঘটনাটি ঘটেছে।\n\nজানা যায়, দীর্ঘদিন যাবৎ খায়রুলের কাছে টাকা পায় হারুন নামের এক প্রতিবেশী। অনেকবার টাকা দিবো দিচ্ছি বলেও খায়রুল ধারের টাকাগুলো পরিশোধ করেনি।\n\nএক পর্যায়ে বিরক্ত হয়ে যায় পাওনাদার হারুন। অতঃপর বুধবার দুপুর ২টায় খায়রুলকে বাসা থেকে ডেকে নিয়ে জোরপূর্বক সোয়েটার পড়িয়ে দেয় হারুন সহ এলাকার আরো কিছু ব খাটে যুবকেরা। ফলে অতিরিক্ত ডিহাইড্রেশনের কারণে মারা যায় খায়রুল।\n\nখায়রুলের গা থেকে সোয়েটার খুলে তার মরদেহটি ঢাকা মেডিকেল মর্গের নিকট হস্তান্তর করা হয়েছে।\n\nধারণা করা হচ্ছে, চাঞ্চল্যকর এ খুনের দায়ে অভিযুক্ত হারুন ও তার দলবল ইতিমধ্যেই ঢাকার বাইরে গিয়ে গা ঢাকা দিয়েছে। খুনিদের যতদ্রুত সম্ভব আটক করা হবে বলে জানিয়েছেন মানিকনগর থানা কর্মকর্তা এসআই জামসেদ। - প্রথম আলু

```
Logistic Regression Prediction: Fake News
Decision Tree Prediction:  Fake News
Random Forest Prediction: Fake News
Passive Aggressive Prediction: Fake News
Multinomial Naive Bayes Prediction: Not A Fake News
Support Vector Machine Prediction: Fake News
```

That makes clear that the model requires us to provide fictitious data. Apart from Multinomial Naive Bayes, all classifiers have been able to identify bogus news; different classifiers have projected the model.

Chapter 5

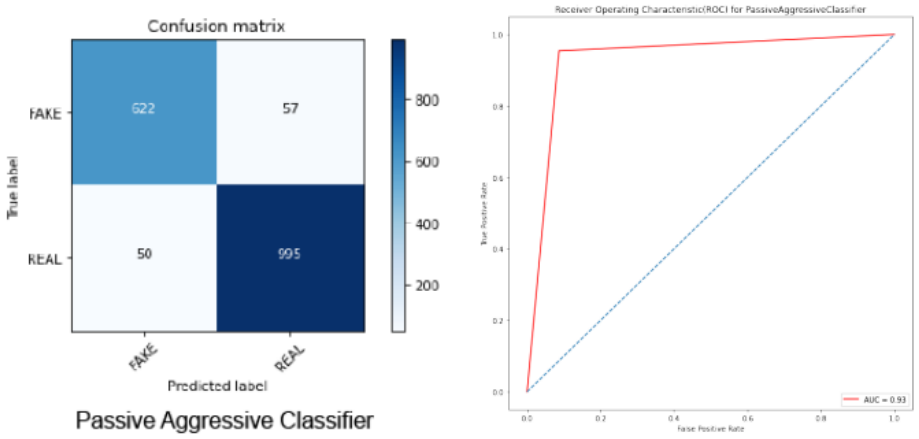
Result Analysis

We'll provide result of every classifier to evaluate our model. Our training set included 70% of the dataset and our testing set consisted of 30%.

Here we provide the classifier performance in every sense-

1. PAC

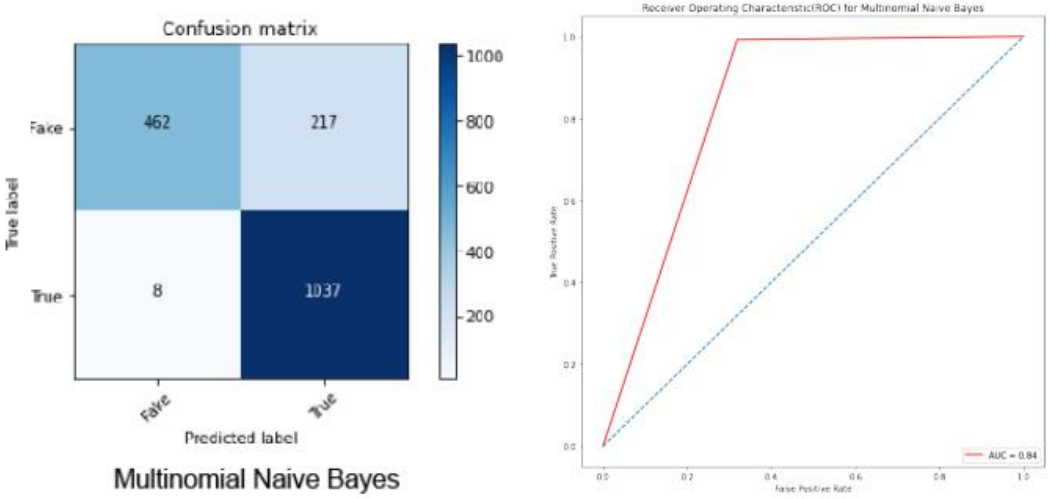
Accuracys - 94.1%; Precisions - 94.6%; Recalls - 95.6%; F1 scores – 95.0%



Discovered that 995 accurate news were accurately recognized as real news and 622 false news were successfully identified as such. With an Area Under the ROC curve (AUC) of 0.93, forecasts were quite excellent. Moreover, the classifier has validity shown by its recall and F-1 score of 95.4% and 94.9%.

2. MNV

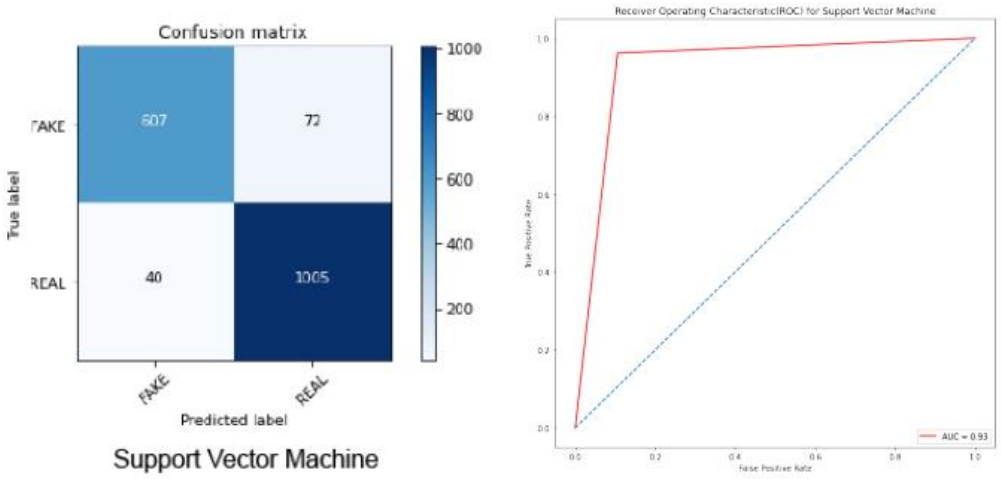
Accuracy- 94.5%; Precisions - 91.6%; Recalls - 93.6%; F1 scores – 97.5%



.

It shows rather 86.2%. Precisions awarded an 82.6% mark. We discovered that 1037 real news items were accurately recognized as legitimate news and 462 bogus news. With 0.81, forecasts turned out to be really excellent. Moreover, the classifier has validity shown by its recall and F-1 score of 99.2% and 90.2%.

3. SVM



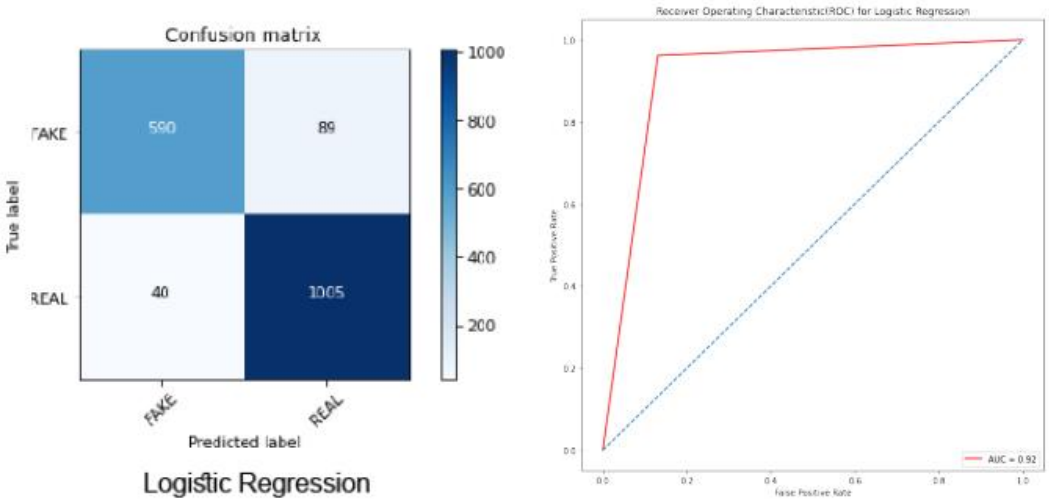
Accuracy- 91.5%; Precisions - 81.6%; Recalls - 91.7%; F1 scores – 89.5%

It shows 91.5%. Precisions earned 93.3% rating. We discovered that 1005 accurate news items were properly classified as real news and 607 bogus news items were accurately labeled as such. With an Area Under the ROC curve (AUC) of 0.93, forecasts turned out to be excellent. Moreover, the classifier displays validity with a recall and F-1 score of 96.2% and 94.7%.

4. LR

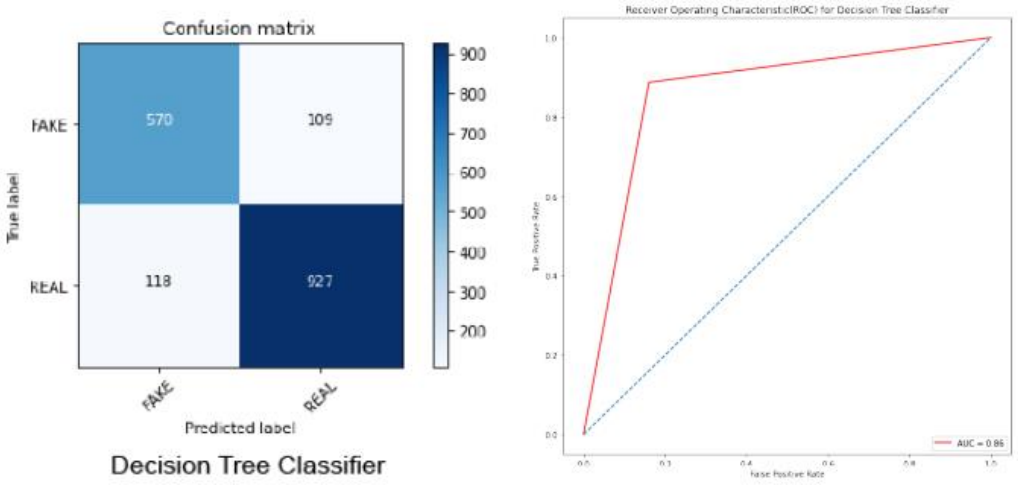
Accuracy- 90.5%; Precisions - 92.6%; Recalls - 91.9%; F1 scores – 9=87.1%

It shows that accuracy is rather 92.5%. Precision earned a 91.9% rating. We discovered that 1005 real news items were accurately recognized as legitimate news and 590 bogus news items were accurately identified as such. With a ROC curve (AUC) value of 0.92, forecasts turned out to be really excellent. Moreover, the classifier has validity shown by its recall and F-1 score of 96.2% and 87.1%.



5. DTC

Accuracy- 98.3%; Precisions - 81.6%; Recalls - 83.3%; F1 scores – 77.5%

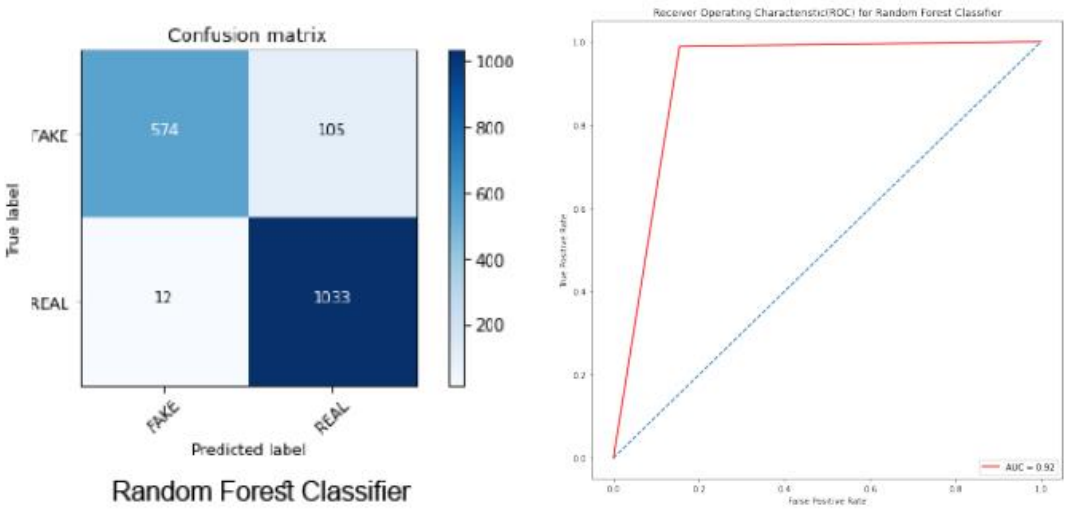


It shows rather 86.8%. Precision received an 89.5% rating. We discovered that 927 real news items were properly recognized as legitimate news and 570 false news items were properly classified as phony. With an Area Under the ROC curve (AUC) of 0.86, forecasts proved to be really excellent.

Moreover, the classifier displays validity with a recall and F-1 score of 88.7% and 89.1%.

6. RFC

Accuracy- 89.2%; Precisions - 89.6%; Recalls - 91.9%; F1 scores – 91.4%

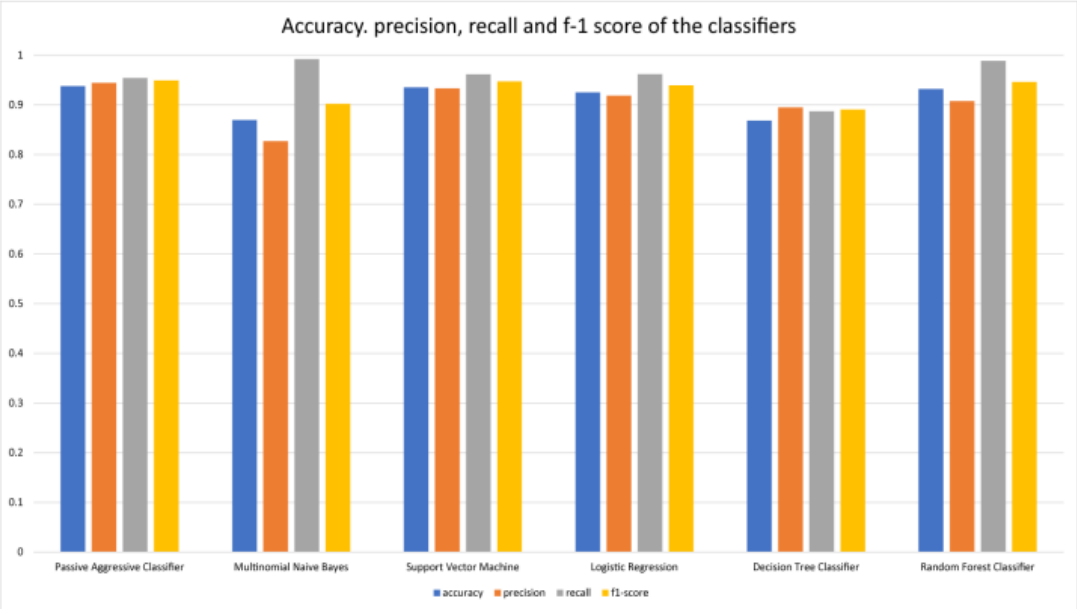


It shows rather 93.2%. Precision earned 90.8% mark. We discovered that 1033 real news items were successfully classified as legitimate news and 574 bogus news items were appropriately labeled as such. With of 0.90, quite excellent. Moreover, classifier has validity shown by its recall and F-1 score of 98.9% and 94.6%.

-

name of model	accuracy	precision	recall	f1-score
Passive Aggressive Classifier	0.937935035	0.944128788	0.954066986	0.949071871
Multinomial Naive Bayes	0.869489559	0.826953748	0.992344498	0.902131361
Support Vector Machine	0.935034803	0.933147632	0.961722488	0.947219604
Logistic Regression	0.925174014	0.918647166	0.961863788	0.939691445
Decision Tree Classifier	0.868329466	0.894787645	0.88708134	0.890917828
Random Forest Classifier	0.932134571	0.907732865	0.988516746	0.946404031

It shows PAC had best accuracy of 92.7%. Furthermore displaying exceptional accuracy of 93.5% and 93.2% respectively were SVM and RFC. The lowest of 86.8% shown by the Decision Tree Classifier also resulted from the kind of the algorithm. The approach is greedy in character and deterministic. They also show a propensity to overfit. Furthermore shown is poor accuracy of Multinomial Naive Bayes (MNNB) compared to other classifiers—86.9%. This happened because MNB assumes the characteristics to be independent of one another.



Chapter IV

In summary

By means of investigation properly categorized bengali false rumor using many ML algorithms. Forward gathering false and real info. We also pre-processed the data eliminating stopwords, numerals, punctuation, etc. We next combined the datasets' contents and headlines. Following that, we included TF-IDF into feature extraction. We obtained our answer at last using the classifiers. It attained result ninety percent. Support Vector Machine also Passive Aggressive Classifier performed better than other ones. Furthermore, many classifiers have been used to provide a whole picture of their performance on Bangla false news. If researchers choose to investigate in this area going forward, they may reasonably estimate makes clear that TF-IDF helps the classifiers recognize Bangla false news more effectively. Though they perform better with large datasets, other methods as "Word2Vec" also can extract features. We could not investigate further feature extraction methods without a synthetic dataset. Moreover, not many tools operate with Bangla natural language processing. We think that our model will work better if we can resolve this problem. As much as feasible, we want to grow our phony dataset. We also want to try certain techniques that have been helpful in other widely spoken languages and try to modify develop a plugin showing the outcome right away should someone enter false or accurate news. At last, Bangla would like to employ as is presently considered as one of the greatest transformers and has great possibilities. We are sure that these next initiatives will greatly advance our study.