

FakeDetect: Bangla Fake News Detection Model based on Different Machine Learning Classifiers

by

Tasnuba Sraboni

17301017

Md. Rifat Uddin

17101016

Fahim Shahriar

21141084

Ruhit Ahmed Rizon

17101050

Shakib Ibna Shameem Polock

17101435

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
Brac University
June 2021

© 2021. Brac University
All rights reserved.

Declaration

It is hereby declared that

1. The thesis submitted is our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. The thesis acknowledges all main sources of help.

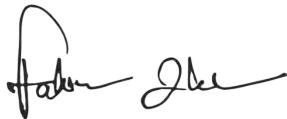
Students' Full Name & Signature:



Tasnuba Sraboni
17301017



Md. Rifat Uddin
17101016



Fahim Shahriar
21141084



Ruhit Ahmed Rizon
17101050



Shakib Ibna Shameem Polock
17101435

Approval

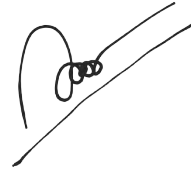
The thesis titled “FakeDetect: Bangla Fake News Detection Model based on Different Machine Learning Classifiers” submitted by

1. Tasnuba Sraboni (17301017)
2. Md. Rifat Uddin (17101016)
3. Fahim Shahriar (21141084)
4. Ruhit Ahmed Rizon (17101050)
5. Shakib Ibna Shameem Polock (17101435)

of Spring, 2021 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science and Engineering on June 06, 2021.

Examining Committee:

Supervisor:
(Member)




Muhammad Iqbal Hossain, PhD
Assistant Professor
Department of Computer Science and Engineering
Brac University

Thesis Coordinator:
(Member)

Md. Golam Rabiul Alam, PhD
Associate Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)



Sadia Hamid Kazi, PhD
Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

Information is power although fake information can have severe consequences when it gets viral. Living in the era of social media is like always getting influenced by the news of the online world even though it is fake. Moreover, online news portals and social media are becoming standardized for consuming information. It is effortless to spread fake news using these mediums. Fake news is represented as authentic news with the wrapping of inaccurate information. In recent times, the rate of lynching has increased because of the spread of fake news. Besides, COVID-19 related false information is affecting people by creating chaos and spreading panic worldwide. Some fake news automation systems exist to tackle this problem. However, they are largely developed for English. There are hundreds of millions of people who speak Bangla worldwide. In this work, we propose a model that can favorably detect fake news in Bangla. We have applied some pre-processing and feature extraction techniques to our dataset. Experimental analysis on real-world data demonstrates that Passive Aggressive Classifier and Support Vector Machine achieves 93.8% and 93.5% accuracy respectively which are higher than the other Machine Learning classifiers.

Keywords: Fake News; Bangla Fake News; Machine Learning; NLP; Tf-IDF; Passive Aggressive Classifier

Dedication

We would like to dedicate our work to all the victims of fake news who motivated us to work on this topic. We would also like to dedicate our work to our parents and teachers, without whom we would never have progressed as far as we have. And a special Thanks to our supervisor who supported us wholeheartedly.

Acknowledgement

First and foremost, we give thanks to the Almighty Allah for allowing us to finish our thesis without any serious setbacks.

Secondly, we would like to express our gratitude to our supervisor, Dr. Muhammad Iqbal Hossain, for his unwavering support and encouragement throughout our research. He was always willing to assist us anytime we needed it.

And, finally, we want to express our gratitude to our families. Without their continuous support, it would not be possible to pursue this research work as well as this Undergraduate Degree.

Table of Contents

Declaration	i
Approval	ii
Abstract	iii
Dedication	iv
Acknowledgement	v
Table of Contents	vi
List of Figures	viii
List of Tables	ix
Nomenclature	x
1 Introduction	1
1.1 Motivation	2
1.2 Problem Statement	3
1.3 Objective and Contributions	3
1.4 Thesis Structure	4
2 Background Study	5
2.1 Literature Review	6
2.2 Machine Learning Algorithms	9
2.2.1 Passive-Aggressive Classifier	9
2.2.2 Multinomial Naive Bayes	10
2.2.3 Support Vector Machine	10
2.2.4 Logistic Regression	11
2.2.5 Decision Tree Classifier	12
2.2.6 Random Forest	13
3 Proposed Model	14
3.1 Dataset Description	14
3.1.1 Data Preprocessing	15
3.1.2 Feature Extraction	16
3.2 Model Description	17

4	Experimentation	18
4.1	Dataset for Classification	18
4.2	Punctuation Removal	18
4.3	Stop Words Removal	19
4.4	Stemming	19
4.5	Extracting Features	20
4.6	Concatenation of Headline and Content	20
4.7	Results with Different Data Split Ratios	23
4.8	Gradient Boosting Algorithm	24
4.9	Manual Testing of News	25
5	Result Analysis	26
6	Conclusion	32
	Bibliography	34

List of Figures

2.1	Decision Boundary of Passive Aggressive Algorithm	9
2.2	Equations of Naive Bayes	10
2.3	Possible Hyperplanes	10
2.4	Logistic Regression Model	11
2.5	Decision Tree	12
2.6	Random Forest Model Making a Prediction	13
3.1	Data Flow of Data Preprocessing	15
3.2	Feature Extraction with TF-IDF	16
3.3	Proposed Model for Bangla Fake News Detection	17
5.1	Confusion Matrix and ROC Curve for PAC	26
5.2	Confusion Matrix and ROC Curve for MNB	27
5.3	Confusion Matrix and ROC Curve for SVM	28
5.4	Confusion Matrix and ROC Curve for LR	28
5.5	Confusion Matrix and ROC Curve for Decision Tree Classifier	29
5.6	Confusion Matrix and ROC Curve for Random Forest Classifier	30
5.7	Result Histogram	31

List of Tables

3.1	Contents of Dataset	14
3.2	Description of Column Title	15
4.1	Before Removing Punctuation Marks	18
4.2	After Removing Punctuation Marks	18
4.3	Before Removing Stop Words	19
4.4	After Removing Stop Words	19
4.5	Before and After Stemming on ‘Headline’ Column	20
4.6	TF-IDF Metrics	20
4.7	Before Merging ‘Headline’ and ‘Content’ Columns	21
4.8	After Merging ‘Headline’ and ‘Content’ Columns	21
4.9	Results Obtained from ‘Headline’ Column for Different Classifiers	22
4.10	Results Obtained from ‘Content’ Column for Different Classifiers	22
4.11	50:50 Train Test Result	23
4.12	60:40 Train Test Result	23
4.13	80:20 Train Test Result	24
4.14	70:30 Train Test Result	24
4.15	Classification Report of Gradient Boosting Algorithm	25
4.16	Manual Test of News	25
5.1	Result with different classifiers	30

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

LR Logistic Regression

ML Machine Learning

MNB Multinomial Naive Bayes

NLP Natural Language Processing

PAC Passive-Aggressive Classifier

SVM Support Vector Machine

TF-IDF Term Frequency-Inverse Document Frequency

Chapter 1

Introduction

Fake news is the news that is false and inaccurate but is designed to make people think that they are true and accurate. People, in general, became familiar with the term ‘Fake news’ in 2016 during the election campaign in the U.S. when Donald Trump was using this term to put off anything he was being accused of. The reason for fake news is either commercial or political to make money or change opinion. Fake news manipulates people’s minds to think a certain way and makes people support a particular opinion. Fake news is not always just lies; more often it is a mixture of lies and truths. There is a famous saying that goes: “A lie can get halfway around the world before the truth has got its pants on”. As information spreads very quickly and freely on digital technologies and social media, fake news also spread much more quickly and to many more people. Often on online news sites, we see fake news. This false and inaccurate news is designed to mislead the readers. Social media is flooded with this kind of news these days. People who write this fake news are not interested in the issues that they are writing; they are interested in making some money out of it. For instance, through Facebook, they can draw readers into their websites and then make money off of advertising. The writer profits on the number of people who click on the news as the news are designed as ‘Clickbait’. It is very difficult to distinguish this fake news from genuine ones. Fake news can have serious consequences when they come out of the online world and enter into the offline world. In 2018, fake news of child abduction went viral through WhatsApp in Mexico and then a mob burned two men, who were suspected to be child abductors, to death before checking if the information was correct or not [1]. A similar thing happened in India and Myanmar. Also, fatal violence was instigated because of fake news on Facebook and WhatsApp in India, Myanmar, and Sri Lanka [1]. Fake news affects people psychologically and spreads hatred. A national survey carried out by the Management and Resources Development Initiative (MRDI), found that the rate of Fake news experience is high in rural areas (66%), followed by urban areas (62.3%), while it is the lowest in metropolitan areas (52.5%) in Bangladesh. And when searching for news online, half of the people don’t try to find out whether the information is fact or opinion [2]. So, to develop a system like this, our proposed approach is to use machine learning techniques for fake news detection.

1.1 Motivation

The increasing occurrence of lynching and violence as consequences of spreading fake news has become one of the main problems in every country. People are influenced by social media and their online world. So, when they see any news having a luring headline they share it with others without knowing the authenticity of that news. Everyone with a keyboard can make up fake news and spread it within a minimum amount of time with the help of social media and can pocket the profit coming from their websites; they do not care if that piece of false information creates any violence or not. Some people are obsessed with the count of Reacts and Shares they get after posting something on social media; so they also do not care if any false information harms others. A huge share of internet users lack proper digital literacy in Bangladesh. During this COVID-19 pandemic, the netizens of Bangladesh looked up health-related information on social media. In this country, The first online misinformation related to COVID-19 was a religious one that claimed that protection from COVID-19 infection would be eating Thankuni leaves (Indian pennywort) saying Bismillah (in the name of Allah) regularly [3]. Moreover, a rumor spread through Facebook and WhatsApp that to build the Padma Bridge human sacrifices need to be made as offerings and that is why people are trying to kidnap children. Several people were killed on the street by the mobs after suspecting them as kidnappers [4]. In Ramu, there was news spread from a Facebook account of the desecration of a Quran in 2012 [5]. Almost 25 thousand people were involved in destroying 12 temples and 50 houses. The Buddhist who was accused to spread this was innocent. A few days back a woman and her daughter went live on Facebook and accused her husband of domestic violence. When the police arrived they found nothing like this. Later a media report cleared this. Recently, we have seen so many Facebook groups and accounts posting for the help of a sick person or a group which doesn't exist. To resist this spread of fake news, we are proposing a model using different machine learning classifiers which will detect fake news more accurately.

1.2 Problem Statement

Lynching occurring like a wave in many countries as well as in Bangladesh as a consequence of fake news. On the other hand, the COVID-19 pandemic situation is getting worse due to the spread of misinformation. With the support of Unicef, a national survey carried out by the Management and Resources Development Initiative (MRDI), found that 63.6 percent of people in Bangladesh have experienced that the news they get from social media or online and took it as authentic news, later they found out that this news was fake. And nearly two-thirds of the people never or some of the time look at which news source published it [2]. In Bangladesh, it has become a severe problem and we want to resist it as much as possible. Although there exist fact-checking websites (such as BD FactCheck, Jachai and Rumor Scanner), these fact-checking platforms are user-based services, people find something fake and report there on their own and often a biased group of people can change the context by false reporting. So, an automated system would be more transparent and neater. Moreover, the existing systems are not suitable for analyzing Bangla letters and words. Because almost all the models created so far can work on English letters and words. A model that can process Bangla letters and words is very essential. If anyone receives any fake news through WhatsApp, Imo, Viber, or any other messaging apps they cannot find out whether the news is fake or not. So, to tackle all these limitations of the existing models we need a system that can solve all these problems.

1.3 Objective and Contributions

Fake news has the power to manipulate an entire community. For years researchers have spent much time building an automated model that can detect fake news so that the reading can know it right away without being manipulated. There are many sophisticated models that can detect fake news but they only work on English words and phrases. There are only a few models that can detect fake news in Bangla. The main objective of our research is to build a sophisticated model that can detect fake or bogus content written in the Bangla language. To build this model we will use different advanced tools like machine learning, natural language processing. Using the automation system, reader will be able to identify whether the news they are reading is fake or not. This model will work for not only the article but also the headline.

To detect fake news we have to analyze every word and phrase. There are many unnecessary words or punctuation marks that do not play any role to detect fake news. To remove these unnecessary words and punctuation marks we have used different pre-processing techniques for the Bangla language. There are close to no tools for pre-processing in Bangla. This tool can be very handy for any future researcher who wants to work with Bangla Language. Apart from that, the entire model can be a huge asset for our country or anyone who speaks or reads Bangla content anywhere in the world. This paper can contribute to the entire humankind and can save a community from mass violence.

1.4 Thesis Structure

Our entire thesis paper is divided into several chapters. We have described each and every work in separate portions. **Chapter 2** contains background study which includes literature review and Algorithms. The literature review contains the summary of different papers that we have reviewed and analyzed for better understanding and comparison. Though we have reviewed a lot of research papers, we have added around eight papers that are related to our work. In algorithms, it contains different classifiers that we have used in our work. PAC, MNB, SVM, LR, Decision Tree classifier, Random Forest classifier are those classifiers. In **Chapter 3**, we have introduced our proposed model. This section includes data description, model description. Data description includes data processing and feature extraction. Data description mainly refers to where we got our dataset from and how we processed our data for testing and training to get the final result. **Chapter 4** contains experimentation where we have discussed how we have obtained the result. In **Chapter 5**, we have analyzed the results that we obtained from different classifiers. In **Chapter 6**, we have concluded the paper by summarizing it and discussing about future work.

Chapter 2

Background Study

Fake news can cause a lot of damage. We have seen many examples of conflict that was caused by spreading fake news. Nowadays the availability of the internet, smartphone, and other devices make these things even easier than before. Now people in rural areas also use smartphones and the internet. So it's easier to spread any information fast all over the country within a short period of time. However, there were no effective approaches taken to prevent this problem in Bangladesh for many years. But in recent years researchers are trying to prevent it in some smart ways using ML, NLP, deep learning etc. Some of the works which have been done by the researchers of Bangladesh are given below:

Linguistic features and neural network-based methods have been explored to create a system in [6]. This paper focuses solely on detecting Bangla fake news. Bangla is the sixth most used language [7]. Very little Natural Language Processing research has been done for detecting fake news in Bangla. Different methods have been used to extract features, for traditional linguistic features, they have used lexical features, syntactic features, semantic features, metadata, and punctuation. On the other hand, they have used CNN, LSTM, pre-trained language models in the neural network part. The methodologies they have used are SVM, RF, LR and their accuracy of F1-score is 46%, 55%, and 53% respectively. By evaluating the F1-score of the POS tag, we can conclude that this technique is unable to determine fake news. Moreover, there was no improvement over the random baseline. However, after the incorporation of all the linguistic features, F1-score achieved 91% using the SVM classifier. In all the cases RF performed worse than the other two. They will work on the character level features in neural network models. They will include them in future experiments. On the other hand, they will continue to expand their data set. They have annotated 8.5k and are hoping to make it 50k. They have used SVM, LR, RF models for the experiment where SVM performed better than others. However, there was nothing mentioned about the accuracy of the result.

To detect whether a Bangla News is fake or not Md Gulzar Hussain et al. [8] proposed a model using Multinomial Naive Bayes Classifier and Support Vector Machine Classifier. The model is being made to detect Bangla fake news so that the data set is being collected from different Bangla news articles where 60% of the data is real news and 40% data is fake news. To get higher accuracy pre-processing the data is a must. By removing punctuation marks, special characters, numerical

values, special symbols the preprocessing is done. Count vectorizer and TF-IDF are being used to extract the features before feeding the text into the classification algorithms. For classification, the data was split into 70:30 ratio for train and test. MNB and SVM Classifiers are used to classify the dataset where accuracy is 93% and 96.64% respectively.

Shafayat Bin Shabbir Mugdha et al created a methodology for Bengali false news identification that can successfully assess if the news is authentic or not based on the news headlines [9]. To run this model, a new data set was constructed using the Gaussian Naive Bayes technique to attain the target. Stop words were removed, tokenization was done, and stemming was done as part of the data preparation. By removing inflected words and diacritic markings from words, stemming has been accomplished. TF-IDF and Extra Trees Classifier are used to extract features. Although the Extra Trees Classifier is a classifier, it is utilized as a feature selection approach in this model to choose the most suited features and then apply the results in the classifiers to improve outcomes and performance. SVM, Logistic Regression, Random Forest Classifier, Gaussian Naive Bayes and many other classifiers were employed for classification. Gaussian Naive Bayes, on the other hand, had the best accuracy of 87.42 percent.

2.1 Literature Review

A couple of machine learning models on their datasets which include evaluation metrics, Naive-Bayes model, linear SVM, Ridge classifier, and decision tree are presented in [10]. These models are used for the classification of the text and representing them in the TF-IDF matrix. The liar-liar data set uses linear SVM, decision tree, Ridge classifier and max feature number models for evaluation. On the other hand, a fake corpus dataset applies SMOTE (Synthetic Minority Over-sampling Technique) as this data set is unbalanced. This model helps by choosing the nearest neighbors in the minority class. For the results, on the liar-liar corpus, there is an average accuracy of Naive-Bayes, Linear SVM and Ridge classifier but as at the recall par class, it shows that Naive-Bayes is bad at detecting fake news and classifies most of the texts as reliable. For the fake news corpus, linear models give the best performance. With linear SVM reaching an accuracy of 94.7%, ridge classifiers 93.98% and decision tree outperforms Naive-Bayes with an accuracy of 89.4%, Naive-Bayes gets 85.3%. Fake news detection only on the supervised models of text is not sufficient for all cases. To get a proper solution additional information should be taken to the light as the author's information. This paper suggests an automatic fact-checking model compiled with a knowledge base and the model will extract information from the text and verify the information with the database.

Various techniques to evaluate fake news are implemented in [11]. This paper takes different types of approaches depending on the news it is processing. LIWC software is used in this paper as a psycholinguistic feature. To measure the toxicity of a news article, this paper used Google's API, and TextBlob's API is used to compute subjectivity. This paper used different classifiers w.r.t AUC and F1 score. These classifiers include KNN, Naive Bayes, Random Forests, SVM with RBF kernel and XGBoost. Naive Bayes did not perform well in both the AUC(72%) and F1(75%)

scores. On the other hand, both Random Forests and XGBoost predicted better than the other classifiers, scoring 86% accuracy in XGBoost and 85% accuracy in Random Forests classifiers w.r.t AUC. Random Forests and XGBoost predicted fake news well but still lacked some accuracy, which can be improved, by using the latest and more classifier techniques to improve the accuracy even more. As only textual analysis might not provide the most accurate result we can use deep learning and neural networks to trace its source which will be crucial to predicting authenticity.

To train the machine some features for the classifier are used in [12] like- word count, gram count(counts of the number of times they appeared), the sentiment of the news analysis, lemmatization, named entity recognition. 2 classifiers were used- Random Forest classifier and a Naive Bayes classifier. A total of 3 datasets was used which contained more than 40000 articles. One of the datasets was used to partially train and test the classifiers. The remaining datasets were used to purely test the classifiers. ISOT, FakeNewsNet and an original dataset were used. They tested different features on different classifiers and chose the best accuracy for fake news detection. Different features were extracted and the datasets were applied. The ISOT dataset got the highest accuracy of 97.42% RF Count-word, 97.60% RF Count-ngram, 94.74% NB Count-word N, and 97.91% NB Count-ngram. Secondly, On the TF-IDF features, ISOT dataset got the highest accuracy too- 98.51% RF TF-IDF-word, 98.46% RF TF-IDF-ngram, 93.31% NB TF-IDF-word and 95.68% NB TF-IDF-ngram. The Random Forest classifier outperformed Naive Bayes classifier. Thirdly, the ER feature, which is not that of a good feature, got an accuracy of 85.71% RF and 74.93% NB with the ISOT dataset. On the PoS feature, the accuracy was not to the mark using the FakeNewsNet data. Accuracy of only 50.23% RF and 44.79% NB was achieved. This feature can be used with other features to increase accuracy. The VADER feature showed the worst result. The accuracy was very low and it implements that VADER is not a good feature for the classification of fake news. Researchers found that the ISOT dataset dominated the other datasets with its accuracy rates. The accuracy rate was low with features ER, PoS, and VADER. Further testing can be done so that accuracy rates increase. Moreover, a combination of the features can be done. Also, the classification of real, satire and fake news was not done which can be added in the future.

In [13], a unique approach of attitude identification is used to detect fake news. Stance detection establishes the link between two pieces of text. They identified the stance by utilizing the terms ‘agree’, ‘disagree’, ‘discuss’ and ‘unrelated’ in the news piece and title. It created an approach that can accurately predict the relationship between news items and headlines, allowing it to identify bogus news. Fake News Challenge (FNC-1) dataset was used in their model. The data processing part used Stop Word Removal, Punctuation Removal, Stemming, Word Vector Representation, Bag of Words, TF-IDF vectorizer and Sampling Techniques. To train their model, they employed TF-IDF with neural networks, BoW with dense neural networks and Pre-trained word embeddings with neural networks. However, their best model employed dense neural network architecture to predict the goal posture utilizing concatenated inputs of TF-IDF vector representations of words and pre-processed engineering features. Their method was able to capture the relative relevance of a word in article-headline pairs both locally and globally. They computed the

cosine similarity between headline-article TF-IDF pairs to find out the similarity between headline-article pairings. With cosine similarity input into a deep neural network, TF-IDF obtained the greatest performance of 94.31 percent on unigrams and bigrams. This result is better than the others. Finally, they plan to build on their findings by doing a similar study on different dataset in order to get closer to establishing an autonomous false news detection tool.

In their publication [14], they provided a system to detect fake news which works on the basis of graph and semi supervised. The use of content-based detection algorithms was their key tactic. Article embedding in Euclidean space, article similarity graph generation and inference of missing labels using graph learning techniques are the three aspects of their methodology. The paper's main advances were the use of word embeddings to create latent representations of news articles in a lower-dimensional Euclidean space and the capture of contextual similarities among articles using a graph-based representation scheme. The cast of fake news detection tasks is then transformed into a semi-supervised graph learning task utilizing Graph Neural Network designs that can perform well on limited labeled data. They used varying amounts of labeled data, from 2% to 20%. They noticed that the results of 20% labeled data and 84 percent labeled data are quite similar. They observed, however, that their AGNN and GCN models beat the competition when they assessed a new article.

2.2 Machine Learning Algorithms

We have used 6 classifiers in our proposed model to detect Bangla fake news.

2.2.1 Passive-Aggressive Classifier

Passive-Aggressive Algorithm is an on-line algorithm that is used when there is a big stream of data. It's an algorithm that basically gets a training example, learns from that example and updates the classifier, and then throws it away. This algorithm was proposed by Crammer et al [15]. For correct predictions it remains passive; on the other hand, for incorrect predictions, it responds aggressively. What this algorithm does is that it makes a prediction by multiplying normalized data with weight vectors and seeing if the document is positive or negative. For the dot product greater than zero, it predicts positive. After that, it observes the true class of that document. The true class is $+1$ for positive documents and -1 for negative documents. The algorithm uses these true class values to upgrade the weight vector so that the dot product for the positive prediction can be always greater than $+1$. The dot product is the similarity between the document and the weight vectors. A positive but smaller than $+1$ dot product value is considered as a loss.

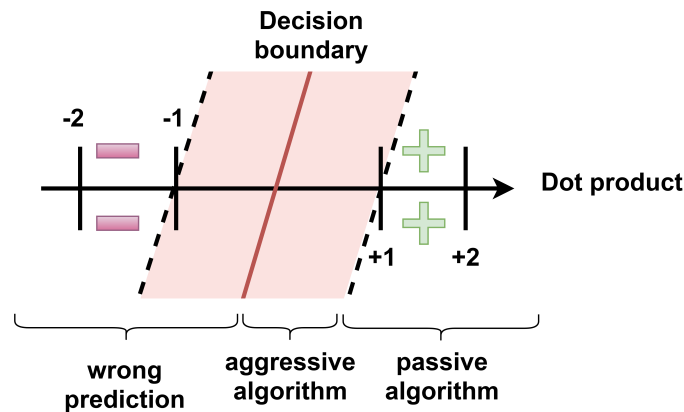


Figure 2.1: Decision Boundary of Passive Aggressive Algorithm

If the document's score falls on the left side of the decision boundary (Figure 2.1), it means the prediction is wrong. And if the document's score falls on the right side of the decision boundary, it uses two algorithms which are Passive and Aggressive. Passive algorithm checks if the dot product is bigger than $+1$. A bigger value than $+1$ means the document is classified correctly. So, the algorithm keeps the weight vector as it is. On the other hand, the Aggressive algorithm checks if the positive value is smaller than $+1$. If it is smaller than $+1$, this algorithm calculates a loss by the loss function. This loss value indicates how far it is from the decision boundary. After getting the loss value, this algorithm recomputes the weight vector. This will make the new dot product exactly $+1$.

2.2.2 Multinomial Naive Bayes

In the Multinomial naive bayes model, the words of a document are distributed as a multinomial [16]. There are a fixed number of classes and each class has an unchangeable set of multinomial parameters for classification. At first, it computes the probability of a class (priors) and the likelihood of a word given in that class (conditional probabilities). The probability of a class is the total count of documents belonging to a class over the total count of documents. And the likelihood of a word belonging in a class is the addition of 1 and how many times the word occurs in that class over the total number of all words that belong to that class and vocabulary size. By taking the dot product of prior and conditional probabilities, it will compute the ‘argmax’ among the classes. This model can predict under which class a test document will fall.

$$\begin{aligned} \text{Probability of a class, } P(c) &= \frac{N_c}{N} \\ \text{Likelihood of a word given a class, } P(w|c) &= \frac{\text{count}(w, c) + 1}{\text{count}(c) + |V|} \end{aligned}$$

Figure 2.2: Equations of Naive Bayes

2.2.3 Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problems. When for each category an SVM model is given sets of labeled training data, it categorizes new text. SVM is capable of doing both classification and regression.

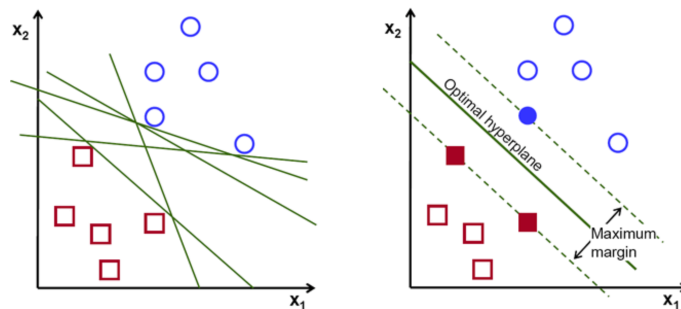


Figure 2.3: Possible Hyperplanes

The possible hyperplanes are used to separate the two classes of data points (Figure 2.3) [17]. The algorithm searches for a plane with the maximum margin. It is the maximum distance between data points of both classes. To provide some reinforcement so that future data can be classified with more accurately, margin distance is maximized.

$$c(x, y, f(x)) = \begin{cases} 0, & \text{if } y * f(x) \geq 1 \\ 1 - y * f(x), & \text{else} \end{cases} \quad (2.1)$$

Equation (2.1) is the cost function of the SVM algorithm. If the predicted value and the actual value are of the same sign, it generates the cost value 0. If the cost is not 0, then the algorithm calculates the loss value. To balance the margin maximization and loss, it adds a regularization parameter.

$$w = w - \alpha \cdot (2\lambda w) \quad (2.2)$$

$$w = w + \alpha \cdot (y_i \cdot x_i - 2\lambda w) \quad (2.3)$$

Equations (2.2) and (2.3) are the gradient functions. Equation (2.2) is used when there is no misclassification and Eq. (2.3) is used when the model predicts wrong. The algorithm updates the weights by using the gradients.

2.2.4 Logistic Regression

In NLP, logistic regression is the basis of a supervised machine learning algorithm for classification. It is a probabilistic classifier that is also very closely related to neural networks. Logistic regression has two phases – Training and Test. In the training phase, it trains a vector of weights w and a bias term b by applying stochastic gradient descent and the cross-entropy loss algorithms. And in the test phase, for each test example x , it computes $p(y=x)$ and gives back $y = 1$ or $y = 0$; here y is the higher probability label [18].

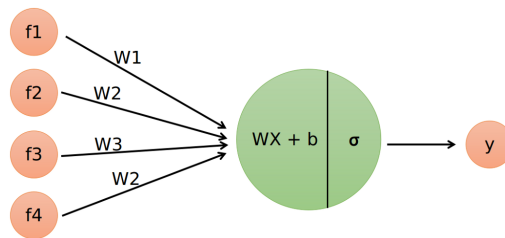


Figure 2.4: Logistic Regression Model

At first, the classifier multiplies each x_i by its weight w_i . Then it sums up all the weighted features and adds the bias term b with it. This sum of the weighted features is expressed by z . The weight w_i represents how important that input feature is to the classification decision. And the bias term b is a real number which is also known as intercept. The classifier applies the sigmoid function on z to create probability. Then it computes the cross-entropy loss by a loss function that expresses

the closeness of the classifier's output and the correct output (y , which is 0 or 1). After that, to update the weights so that it minimizes this loss function iteratively, it uses a stochastic gradient descent algorithm.

2.2.5 Decision Tree Classifier

Decision Tree is a binary tree that recursively keeps splitting the dataset until it finds pure leaf nodes which means the nodes that contain data with only one type of class. Decision Tree is a Supervised learning technique mostly used for solving classification problems. This classifier is tree-structured and the features of a dataset are represented by internal nodes, the outcome is represented by each leaf node and the decision rules are represented by branches. Two nodes of the decision tree are called Decision Node and Leaf Node. Decision nodes contain a condition to split the data and leaf nodes help to decide the class of a new data point [19].

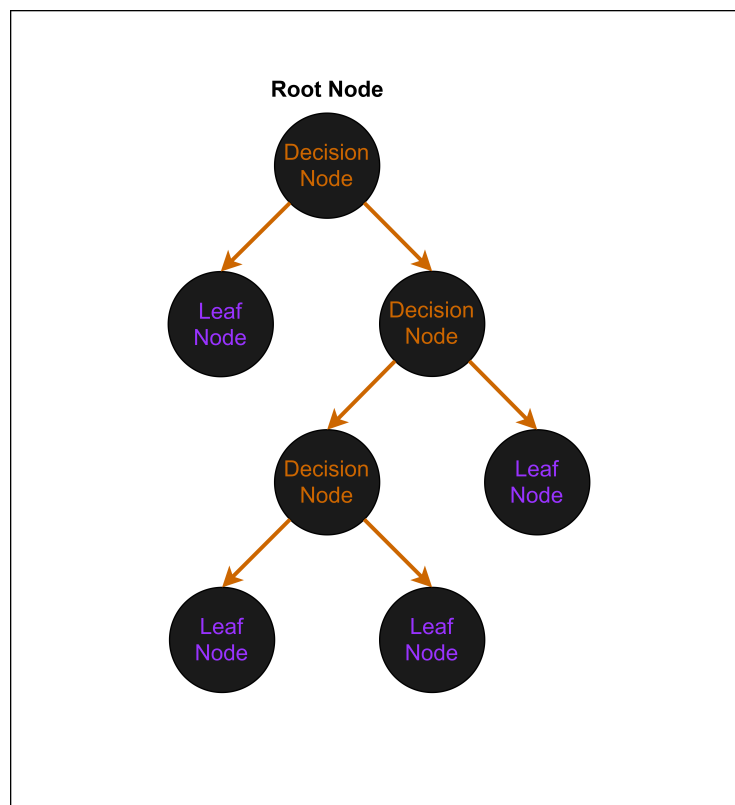


Figure 2.5: Decision Tree

The decision tree predicts the class of the given dataset by calculating Information Gain and Gini Index. The main objective is to find the pure leaf nodes. The algorithm at first splits the root node. Every split will have two states and the algorithm will compute the Entropy of every state. Entropy is the measure of information contained in a state. It compares every possible split and takes the one that gives the minimum entropy; because minimum entropy maximizes information gain. The model traverses through every possible feature and feature value to search for the best feature and the corresponding threshold. It keeps continuing the process until it finds the pure leaf nodes of the tree. Gini Index does exactly the same as Information Gain but its formula is simpler.

2.2.6 Random Forest

The random forest is a classification algorithm that contains many decision trees. It creates every decision tree by applying bagging and feature randomness processes. When building each individual tree, this algorithm tries to create a forest of trees that are not correlated. Every particular tree in this type of classification algorithm provides a prediction for a class and the specific class which receives the most votes eventually ends up becoming the prediction done by the model [20].

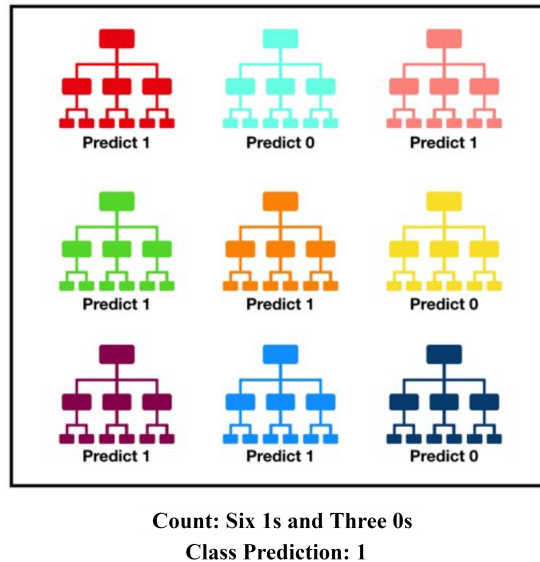


Figure 2.6: Random Forest Model Making a Prediction

Bagging is a combination of two processes – Bootstrap dataset and Aggregation. The first step of building a random forest is creating new datasets from the original one. The algorithm performs random sampling with a replacement which ensures every dataset will contain the same number of rows as the original one but the instances can be found more than once in a dataset. This process of creating new datasets is called bootstrapping. And the process of combining results from multiple decision trees is called Aggregation. Feature randomness selects features randomly for every dataset. It helps to reduce the correlation between the trees. Because of bootstrapping and feature randomness processes, every decision tree generates a different prediction.

Chapter 3

Proposed Model

3.1 Dataset Description

Inputs are the independent variables also known as features. Basically, the input is the predictors. A dataset is a collection of data. In tabular datasets, there are one or more database tables. Every column of a table denotes a particular variable and each row is an instance. Datasets consist of more than one document or file [21].

Collecting the input data has been the major challenge for Bangla Fake News Detection. On this topic, there is very little research on the Bangla language. A proper dataset that fits our research work is pretty much like asking for the moon. On the other hand, there is a large amount of dataset available for English and other mainstream languages (Spanish, French, Chinese, etc) as fake news detection models are vastly being developed at present. For a language like Bangla which has a scarcity of resources, it was quite difficult to have a decent dataset. We collected an annotated dataset built by the researcher from SUST, Bangladesh [6]. The data are gathered together from a lot of news websites and portals.

Content list -

Table 3.1: Contents of Dataset

File Name	Authentic-48K.csv	Fake-1K.csv	FinalData1.xlsx
Number of Contents (Rows)	48k	1.3k	2.5k
Columns	7	7	6
Features	ArticleID, Domain, Date, Category, Headline, Content, Label	ArticleID, Domain, Date, Category, Headline, Content, Label	Title, Statement, Category, Source, Date, Class

In our paper, we used 2 of the 4 Microsoft Excel .csv files named as- Authentic-48K.csv, Fake-1K.csv. This is a labeled dataset where around 48k news is authentic

and around 1.3k news is fake. We collected another dataset built by the researcher from Green University of Bangladesh [8]. This dataset contains 2.5k news and the title of the file is FinalData1.xlsx.

Table 3.2: Description of Column Title

Column Title	Description
ArticleID	Serial number of the news
Domain / Source	Website address from which the news is extracted
Date	Published Date
Category	Type of news
Headline / Title	Headline of the news
Content / Statement	Body of the news
Label / Class	Information that if the news is fake or not

Though this dataset is labeled, it needs further processing so that it can be used to train the model for detecting Bangla Fake News.

3.1.1 Data Preprocessing

Data preprocessing is a process of transforming raw data into a format that is quite aligned to our tasks. Real-world data is not always complete and consistent. These data contain errors and lack in certain behaviors or trends. Data preprocessing has the power to resolve these issues. Raw data are prepared properly for further processing by Data preprocessing. In Machine Learning (ML) processes, data preprocessing transforms the dataset in a form that could be easily interpreted and parsed by the algorithm [22].

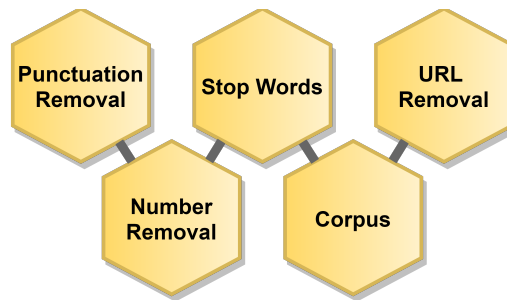


Figure 3.1: Data Flow of Data Preprocessing

Data goes through multiple steps during preprocessing. In the initial step of data cleaning, we removed punctuation, numbers, null values, duplicate values, etc. from the raw dataset. Then we removed stop words. Removing stop words is filtering out the words that have very little meaning. To get the data in a clean and standard format for further analysis we put all the data in two different formats. Which are-

- *Corpus*: Corpus is a collection of texts. To get the dataset in corpus format we used a python library for data analysis called Pandas and we put the dataset into a DataFrame which is basically a table.
- *URL Removal*: Urls are reference links or metadata and HTML tags (if there are any) in the content. Those links provide no valuable information. As a part of preprocessing, we removed the URLs from the dataset.

At this point the corpus is ready.

3.1.2 Feature Extraction

Feature extraction is a technique to extract information that represents the importance of a specific word or phrase in a corpus. We have used **Term Frequency-Inverse Document Frequency (TF-IDF)** in our corpus which is one of the best feature extraction techniques. As we cannot directly pass the corpus in classification models, TF-IDF converts it into useful features. By the Term Frequency (TF) method, it finds out the number of repetitions of a word in a sentence against total words in the sentence. And by Inverse Document Frequency (IDF) method, measures how common or rare the word is across all sentences. The value of the TF-IDF is the product of TF and IDF. The larger the TF-IDF value, the rare the word or phrase is.

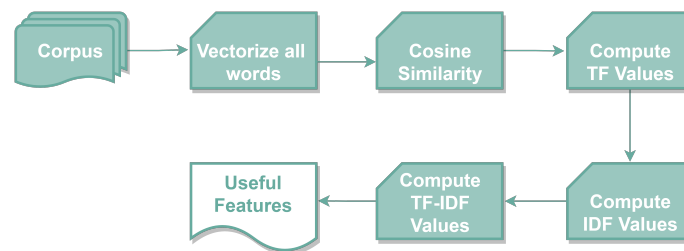


Figure 3.2: Feature Extraction with TF-IDF

From the scikit-learn library we have imported the ‘TfidfVectorizer’ in order to use TF-IDF. We do not have to write the formula manually. After importing the library we have created an object of ‘TfidfVectorizer’. Using that object we called the `fit_transform(corpus)` function in an array where we have passed our corpus. At first, this function converts every word into vectors. Then it applies Cosine Similarity which measures the similarity between two or more vectors. Lastly, it provides TF-IDF values of the entire corpus in numeric form.

3.2 Model Description

For the model, we've tried to choose the most effective approach to attain our goal of detecting fake news. We can see the proposed model from Figure 3.3, after collecting the initial dataset we needed to process the data to feed the machine learning classifiers and feature extraction. This step is necessary for all sorts of string classification problems and it helps to extract proper information from the dataset.

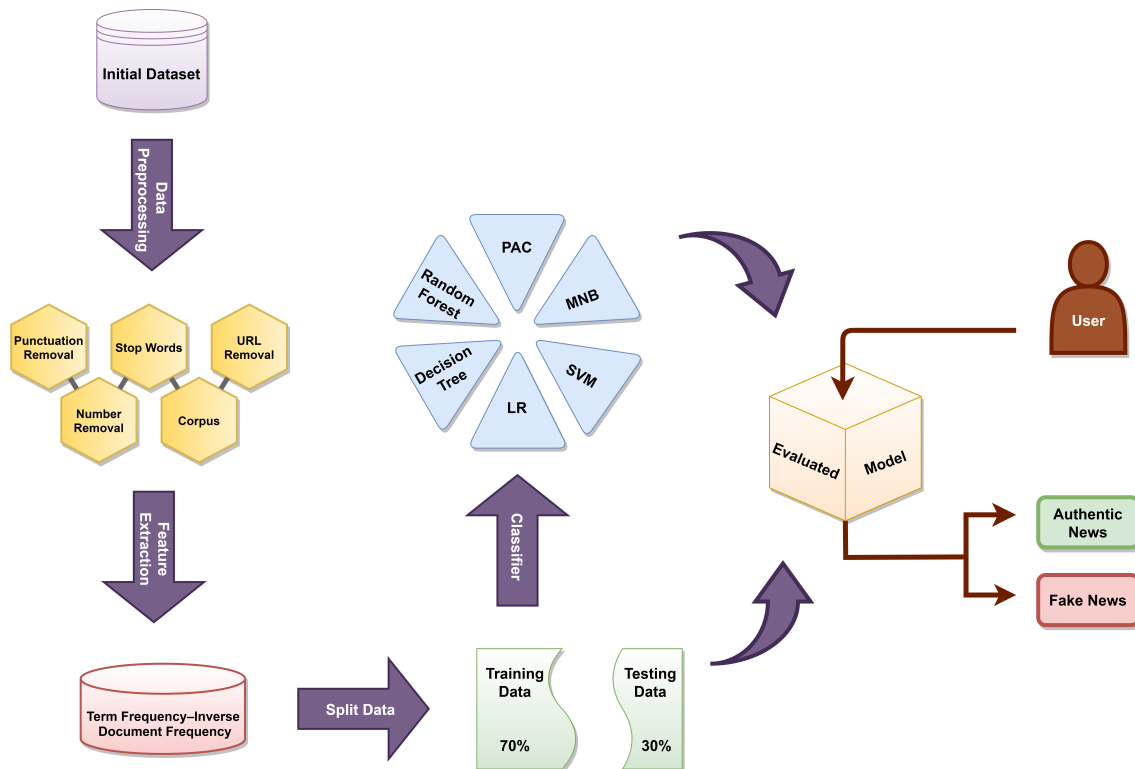


Figure 3.3: Proposed Model for Bangla Fake News Detection

The preprocessing part required a bunch of work which includes punctuation marks removal, stop-word removal, URL removal, number removal, case conversion, named entity recognition, stemming, lemmatizing etc. As we are working with the Bangla language, we had to overlook some of the techniques due to the unavailability of proper tools. So, we removed punctuation marks, stop words and numbers to clean the dataset. TF-IDF has been used for the feature extraction. After completing the first two steps we have split the data by 70:30 ratio for training and testing for the machine learning classifier. For classifiers, we have experimented with different types of classifiers such as Passive-aggressive classifier (PAC), Random Forest, Support Vector Machine (SVM), Multinomial Naive Bayes (MNB), and Logistic Regression. Then the model finalizes the best performing classifier for detecting fake news.

We have a thought in mind that after deploying the model we will build an online platform where a user can manually input content and check the authenticity of the content.

Chapter 4

Experimentation

4.1 Dataset for Classification

We have quite a big dataset for classification consisting of 51.8k data. However, among them, 49.5k of them are authentic data and the rest of the 2.3k is fake news. It is evident that there is a huge disproportion between fake and real datasets. If we used this entire dataset the result would have been biased towards authentic news. Hence, the model would not be an ideal model to classify any data. Keeping that in mind, we took 3.5k authentic data and 2.3k fake data from the dataset so that the model doesn't give any biased results.

4.2 Punctuation Removal

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).

Table 4.1: Before Removing Punctuation Marks

3	ব্যস্ত রাস্তায় গাড়ি থামিয়ে মোতানো হল নেতাকে। নারায়ণগঞ্জের কাঁচপুর এলাকায় চলমান গাড়ি থামিয়ে দিনে দুপুরে মোতানো হল নেতাকে। নেতার সাথে থাকাকালীন এক সহমত ভাই জানান, "বাচ্চিলামবিয়ার দাওয়াত খাতি। নেতারে কলাম ভাই অল্প পথ আরেকটু চাইপে রাখেন। অবস্থা বেগতিক দেখি রাস্তায় গাড়ি থামায় মোতলাম। কিন্তু ভাই মুইতে ভাসায় দিছে।" কিন্তু ব্যস্ত রাস্তায় এভাবে গাড়ি থামিয়ে রাস্তায় জ্যাম সৃষ্টি করে সাধারণ মানুষের দুর্ভোগ ভেঙে আনার ব্যপারে জানতে চাইলে সহমত ভাই জানান, "ভাই এর ক্ষমতাজানেন..."	0
4	জাস্টিন বিবারের কন্সার্টে উত্তেজনায় তিন তরুণীর মূত্র ত্যাগ। বুধবার রাতে মুম্বাই মাতাল জাস্টিন বিবার। মুম্বাইয়ের ডি ওয়াই পটেল স্টেডিয়াম রাত ৮টা থেকে ১০টা পর্যন্ত ছিল শুধুই বিবারময়। ২৩ বছরের পপ সেনসেশন স্টেজে তুর্কেন আতসবাজির মধ্যে দিয়ে। আর সঙ্গে সঙ্গেই দুলে উঠে ৪৫ হাজার দর্শকেরা সা গোটা স্টেডিয়াম। প্রিয় তারকাকে এত কাছ থেকে দেখতে পেয়ে হাজার হাজার তরুণী হাত দিয়ে ব্যস্ত ছিলেন তালি বাজানোর কাজে। উত্তেজনায় অনেকেই চিৎকার করে কালা করেছিলেন। অজ্ঞান হয়ে গিয়েছিলেন অনেকে...	0

Table 4.2: After Removing Punctuation Marks

3	ব্যস্ত রাস্তায় গাড়ি থামিয়ে মোতানো হল নেতাকে। নারায়ণগঞ্জের কাঁচপুর এলাকায় চলমান গাড়ি থামিয়ে দিনে দুপুরে মোতানো হল নেতাকে। নেতার সাথে থাকাকালীন এক সহমত ভাই জানান, "বাচ্চিলামবিয়ার দাওয়াত খাতি। নেতারে কলাম ভাই অল্প পথ আরেকটু চাইপে রাখেন। অবস্থা বেগতিক দেখি রাস্তায় গাড়ি থামায় মোতলাম। কিন্তু ভাই মুইতে ভাসায় দিছে। কিন্তু ব্যস্ত রাস্তায় এভাবে গাড়ি থামিয়ে রাস্তায় জ্যাম সৃষ্টি করে সাধারণ মানুষের দুর্ভোগ ভেঙে আনার ব্যপারে জানতে চাইলে সহমত ভাই জানান, "ভাই এর ক্ষমতাজানেন মিয়া হুদাই কথা কন কেন ভাই এর খুশি। ভাই যেইখা..."	0
4	জাস্টিন বিবারের কন্সার্টে উত্তেজনায় তিন তরুণীর মূত্র ত্যাগ। বুধবার রাতে মুম্বাই মাতাল জাস্টিন বিবার। মুম্বাইয়ের ডি ওয়াই পটেল স্টেডিয়াম রাত ৮ থেকে ১০ পর্যন্ত ছিল শুধুই বিবারময়। ২৩ বছরের পপ সেনসেশন স্টেজে তুর্কেন আতসবাজির মধ্যে দিয়ে। আর সঙ্গে সঙ্গেই দুলে উঠে ৪৫ হাজার দর্শকেরা সা গোটা স্টেডিয়াম। প্রিয় তারকাকে এত কাছ থেকে দেখতে পেয়ে হাজার হাজার তরুণী হাত দিয়ে ব্যস্ত ছিলেন তালি বাজানোর কাজে। উত্তেজনায় অনেকেই চিৎকার করে কালা করেছিলেন। অজ্ঞান হয়ে গিয়েছিলেন অনেকেই। তবে সবচেয়ে আলোচনায় এসেছে যে বিষয়টি তা হল চরম উ...	0

4.3 Stop Words Removal

Stop words occur more than any other types of words in a language which do not provide us any necessary information. So, we need to remove those words from the corpus. For the mainstream, languages stop word libraries are widely available since we are working with the Bangla language and there's not much work has been done with it so we had to do it manually by inputting Bangla stopwords.

Example for Bangla stop words:

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

Output: ভাত খাই। দেশের নাম। আইন সমান।

Table 4.3: Before Removing Stop Words

3	বাস্তব রাস্তায় গাড়ি থামিয়ে মোতানো হল নেতাকে। নারায়ণগঞ্জের কাঁচপুরএলাকায় চলমান গাড়ি থামিয়ে দিনে দুপুরে মোতানো হল নেতাকে। নেতার সাথে থাকাকে সহমত ভাই জানান যাচ্ছিলামবিয়ার দাওয়াত খাতি। নেতারে কলাম ভাই অল্প পথ আরেকটু চাইপে রাখেন। অবস্থা বেগতিকদেখি রাস্তায় গাড়ি থামায় মোতালাম। ভাই মুইতে ভাসায় দিছে। কিন্তু বাস্তব রাস্তায় এভাবে গাড়ি থামিয়েরাস্তায় জ্যাম সৃষ্টি করে সাধারণ মানুষের দুর্ভোগ ডেকে আনার ব্যপারে জানতে চাইলেসহমত ভাই জানান ভাই এর ক্ষমতাজানেন মিয়া হুদাই কথা কন কেন ভাই এর খুশি। ভাই যেইখা...
4	জাস্টিন বিবারের কনসার্টে উত্তেজনায় তিন তরুণীর মূত্র ত্যাগ। বুধবার রাতে মুম্বাই মাতাল জাস্টিন বিবার। মুম্বাইয়েরডি ওয়াই পটেল স্টেডিয়াম রাত া থেকে া পর্যন্ত ছিল শুধুই বিবারময়। বছরের পপসেনসেশন স্টেজে ঢুকেন আতসবাজির মধ্যে দিয়ে। আর সঙ্গে সঙ্গেই দুলে উঠে াজার দর্শকেতাসা গোটা স্টেডিয়াম। প্রিয় তারকাকে এত কাছ থেকে দেখতেপেয়ে হাজার হাজার তরুণী হাত দিয়ে ব্যস্ত ছিলেন তালি বাজানোর কাজে। উত্তেজনায়অনেকেই চিৎকার করে কান্না করেছিলেন। অজ্ঞান হয়ে গিয়েছিলেন অনেকেই। তবে সবচেয়েআলোচনায় এসেছে যে বিষয়টি তা হল চরম উ...

Table 4.4: After Removing Stop Words

3	বাস্তব রাস্তায় গাড়ি থামিয়ে মোতানো নেতাকে। নারায়ণগঞ্জের কাঁচপুরএলাকায় চলমান গাড়ি থামিয়ে দিনে দুপুরে মোতানো নেতাকে। নেতার থাকাকে সহমত ভাই জানান যাচ্ছিলামবিয়ার দাওয়াত খাতি। নেতারে কলাম ভাই অল্প পথ আরেকটু চাইপে রাখেন। অবস্থা বেগতিকদেখি রাস্তায় গাড়ি থামায় মোতালাম। ভাই মুইতে ভাসায় দিছে। বাস্তব রাস্তায় এভাবে গাড়ি থামিয়েরাস্তায় জ্যাম সৃষ্টি দুর্ভোগ ডেকে আনার ব্যপারে চাইলেসহমত ভাই জানান ভাই ক্ষমতাজানেন মিয়া হুদাই কন ভাই খুশি। ভাই যেইখানে মন চাইব মুতপে।
4	জাস্টিন বিবারের কনসার্টে উত্তেজনায় তিন তরুণীর মূত্র ত্যাগ। বুধবার রাতে মুম্বাই মাতাল জাস্টিন বিবার। মুম্বাইয়েরডি ওয়াই পটেল স্টেডিয়াম রাত া া শুধুই বিবারময়। বছরের পপসেনসেশন স্টেজে ঢুকেন আতসবাজির দিয়ে। সঙ্গেই দুলে উঠে াজার দর্শকেতাসা স্টেডিয়াম। প্রিয় তারকাকে দেখতেপেয়ে তরুণী হাত ব্যস্ত তালি বাজানোর কাজে। উত্তেজনায়অনেকেই চিৎকার কান্না করেছিলেন। অজ্ঞান গিয়েছিলেন অনেকেই। সবচেয়েআলোচনায় এসেছে চরম উত্তেজনায় তিন তরুণীর মূত্র ত্যাগ। জাস্টিন বিবারের কনসার্টে গান শুরুর পরই দর্শকরাবুঝতে পেরেছেন প্রতারিত...

4.4 Stemming

Stemming has been in use quite often in NLP where each word is reduced to its root form.. Stemming is used to decrease inflection from the texts. As a part of pre-processing, we tried to use stemming on our dataset. But, unfortunately, we could not find any tools that can do stemming on Bangla text precisely.

Here, we show the stemming we did on our headline column by using a stemmer function from BNLTK(Bangla Natural Language Processing Toolkit) -

Table 4.5: Before and After Stemming on ‘Headline’ Column

	headline		headline
400	কবরীর বাসায় চুরি	400	কবরীর বাসা চুরি
401	আবাসিকে শিল্পসহ খাতে বাড়ছে গ্যাসের দাম	401	আবাসি শিল্পসহ খা বাড় গ্যাস দাম
402	রেকর্ড বাড়তি অনুভূতি মার্শরাফী	402	রেকর্ড বাড়তি অনুভূতি মার্শরাফী
403	নির্ধারিত সময়ে শেষ পায়রা সেতুর নির্মাণ	403	নির্ধারিত সম শেষ পায় সেতুর নির্মাণ
404	সিনেমায় অভিশেক কোহলির	404	সিনেমা অভিশেক কোহলির
...
7197	তালকের স্বশুরবাড়িতে প্রবাসীর মরদেহ	7197	তালক স্বশুরবাড়ি প্রবাসীর মরদেহ
7198	বান্দরবানে পালিত মধু পূর্ণিমা	7198	বান্দরবান পালিত মধু পূর্ণিমা
7199	মাজেদাকে টাকা পুরস্কার	7199	মাজেদা টাকা পুরস্কা
7200	মন্ত্রিসভায় চূড়ান্ত অনুমোদনের অপেক্ষায় টেলিযোগা...	7200	মন্ত্রিসভা চূড়ান্ত অনুমোদন অপেক্ষা টেলিযোগাযো...
7201	ফুল ঝরে	7201	ফুল ঝর

Before Stemming After Stemming

From Table 4.5, we can see that after stemming, some of the words are wrongly reduced to a short form which can reduce the training accuracy of our models.

4.5 Extracting Features

For the feature extraction, as we’ve used TF-IDF vectorizer, to get the word relevance between our fake and authentic documents and also to compute the descriptiveness of a term.

Table 4.6: TF-IDF Metrics

	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 Concatenation of Headline and Content

We had 2 different columns for headline and content. Initially, we found 2 individual results on headline and content. Later, we found that by concatenating the 2 columns, we can get a better result. The result found from the headline column

had low accuracy. This happened because the headlines are just a sentence with 8-9 words in them. So, the models could not be easily trained.

Here in Table 4.7 and 4.8, we show our dataset before and after the concatenation of ‘Headline’ and ‘Content’ respectively-

Table 4.7: Before Merging ‘Headline’ and ‘Content’ Columns

headline	content	class
0	<p>বাংলায় একটি 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 আশ্চর্যজনক ঘটনা ঘটছে। মুরগীর হামলায় শিয়াল মারা গেছে। ঘটনাটি গত সপ্তাহের, ফ্রান্সের উত্তর-পূর্বাঞ্চলের বুতানিয়া এলাকার একটি কৃষিবিশ্বক বিদ্যালয়ে। ওই বিদ্যালয়ে একটি মুরগীর খামার রয়েছে। মুরগীর ঘরের এক কোনার সকালে মৃত একটি শিয়াল পড়ে থাকতে দেখে শিক্ষার্থীরা। ৬ হাজার মুরগী রয়েছে খামারটিতে। শিয়ালটির মৃতদেহ যেখানে পাওয়া গেছে, সেখানে তিন হাজার মুরগী 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 ঘটনার তারা অবাক। শিয়ালটি বাচ্চা ও অনভিজ্ঞ ছিল। এতগুলো মুরগীর সামনে পড়ে সে সন্তবত তড়কে গিয়েছিল।(Source- BBC Bangla</p>	0
1	<p>BTV থেকে লোকজন আসছে, ইন্টারভিউ নিবে। চারজনের টিম, এদের মধ্যে স্যুট, টাই পড়া বস একজন, দুইজন ক্যামেরাম্যান, আরেকজন উত্তীর্ণ 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 তাই কফির কাপ হাতে নিয়ে তোমার সামনে দিয়ে হেঁটে যাবে, তুমি আকাশ বাতাস কাঁপাইয়া ঠাস করে একটা সেলুট দিবা, এইটা ভিডিও করে রাখি, বাঙ্গালি পাবলিকের দুঃ ভাং পছন্দ!... এইটা না করলে হয় না, কেমন যেনো বানানো লাগে- ভুললোকের আইডিয়া শুনে আমি রীতিমতো বিব্রত!... আরে স্যার কী বলেন? আপনার ফ্যাঙ্কিরিতে আদব কায়া আছে, এইটা তোফায়েল সাহেব দেখতে চাইবে।... তোফায়েল সাহেব কে? কমার্স মিনিস্টার তোফায়েল সাহেব। নেন- কফির কাপ চুমুক দিতে দিতে হাঁটা শুরু করেন। হাঁটাখাটি 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 ইউনিলয়েন্স টেক্সটাইলসের ভূমিকা। আমি কিছুক্ষণ তবদা মেয়ে থাকলাম। ‘যেই লাইন বললেন, এইটার বাংলা কী?’ এইটাই বাংলা, একটু গুছাইয়া এইটা এঁটা বইল্যা এইটো, আর শেষে বর্তমান সরকারের ইতিবাচক ভূমিকা ক্রিয়াকর দিবেন, এতটুকুই, পারবেন না? স্যার, আপনার পরকটে চিক্রনি আছে? না তো, কেন? আমার চুলটা একটু আচড়াইয়া নিতাম, চুল এনেদেলে থাকলে আপনার ভাবী আবার রাগ করে, আমার ক্যামেরা ফেস ভালো, খালি এলেমেলো 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!</p>	0
2	<p>অদভূত বিরোধীদলহীনতায় ভুগছে সরকার। এ এক অন্যরকম শূন্যতা। বাবা-মায়ের সাথে থেকে আজকালকার ছেলেমেয়েরা যেমন একাকীত্বে ভোগে, তেমনি সর্বাঙ্কু ঘারা বেষ্টিত থাকলেও বর্তমান সরকারের পাশে কী যেন নেই। কী যেন নেই! প্রচলিত অনিচ্ছায় জাতীয়পাটি বিরোধীদের চেয়ারে বসলেও, তারা ঠিক যেন বিরোধীদল নয়। অনেকটা বিরোধীদের রেক্রিটার 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। অভিসম্বল 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। অনলাইন অর্ডার করলে কখনোই ভাল মাল পাওয়া যায় না। জাতীয়পাটিকে 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 বাই। সেখানে গিয়ে আমরা প্রচলিত ধর্মমত পরিবেশ দেখতে পাই। যেন পুরো অফিসেরই মন খারাপ। অফিসের এক কোণায় স্পিকারে গ্লো সাউন্ডে গান বাজতে শুনি- ‘তোমরা কেউ কি দিতে পারো বিরোধীদের ভালবাসা?’</p>	0
3	<p>রাশিয়া বিশ্বকাপ নকআউট পর্ব ফ্রান্সের সাথে ৪-৩ গোলে পরাজয়ের প্রান্তিতে ফুটবল থেকে অবসরের ঘোষণা দিলেন মেসি। খেলা শেষে মাঠ ছাড়তে ছাড়তে বিভ্রাট করে এমনটাই বলছিলেন আর্জেন্টাইন এই সুপারস্টার। তাহলেই চলছিলো ম্যাচ। শুরুতে ফ্রান্স গোল দিলে মাঠ কাঁপলেও পরবর্তীতে দুটি গোল মুক্ত হয় আর্জেন্টাইনের খুলিতা। বুক চেতিয়ে খেলা শুরু করেন মেসি। হঠাৎই বললে যায় ম্যাচের দৃশ্যপট। একে একে ফ্রান্স আরো তিনটি গোল চুকিয়ে দেয় আর্জেন্টাইনরা জালে। এরই সাথে বিশ্বকাপ থেকে বিদায় ঘন্টা বেজে যায় আর্জেন্টাইনরা 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 লিপি রিডিং করে বুঝতে পারেন, মেসি বলছে, ‘আমি আর ফুটবল খেলবোনা। আমি আর ফুটবল খেলবোনা।’ বাতাসের বেগে কথটি স্টেডিয়ামজুড়ে ছড়িয়ে পরে। পরবর্তীতে মেসির অবসর নেয়ার ব্যাপারে দ্রুত এ খবরটি স্টেডিয়াম প্রান্তনে ভাইরাল হয়ে যায়। এবং অনেকেরই তা বিশ্বাস করে বলেন। যদিও আনুষ্ঠানিকভাবে মেসি এখনো কিছু জানাননি। এখন শুধুই এ কিংবদন্তীর মুখ থেকে অবসর নেয়ার ঘোষণাটি শোনার পালা...</p>	0

Table 4.8: After Merging ‘Headline’ and ‘Content’ Columns

headline & content	class
<p>0</p> <p>বাংলায় একটি 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 আশ্চর্যজনক ঘটনা ঘটছে। মুরগীর হামলায় শিয়াল মারা গেছে। ঘটনাটি গত সপ্তাহের, ফ্রান্সের উত্তর-পূর্বাঞ্চলের বুতানিয়া এলাকার একটি কৃষিবিশ্বক বিদ্যালয়ে। ওই বিদ্যালয়ে একটি মুরগীর খামার রয়েছে। মুরগীর ঘরের এক কোনার সকালে মৃত একটি শিয়াল পড়ে থাকতে দেখে শিক্ষার্থীরা। ৬ হাজার মুরগী রয়েছে খামারটিতে। শিয়ালটির মৃতদেহ যেখানে পাওয়া গেছে, সেখানে তিন হাজার মুরগী 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 ঘটনার তারা অবাক। শিয়ালটি বাচ্চা ও অনভিজ্ঞ ছিল। এতগুলো মুরগীর সামনে পড়ে সে সন্তবত তড়কে গিয়েছিল।(Source- BBC Bangla</p>	0
<p>1</p> <p>BTV থেকে লোকজন আসছে, ইন্টারভিউ নিবে। চারজনের টিম, এদের মধ্যে স্যুট, টাই পড়া বস একজন, দুইজন ক্যামেরাম্যান, আরেকজন উত্তীর্ণ 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 তাই কফির কাপ হাতে নিয়ে তোমার সামনে দিয়ে হেঁটে যাবে, তুমি আকাশ বাতাস কাঁপাইয়া ঠাস করে একটা সেলুট দিবা, এইটা ভিডিও করে রাখি, বাঙ্গালি পাবলিকের দুঃ ভাং পছন্দ!... এইটা না করলে হয় না, কেমন যেনো বানানো লাগে- ভুললোকের আইডিয়া শুনে আমি রীতিমতো বিব্রত!... আরে স্যার কী বলেন? আপনার ফ্যাঙ্কিরিতে আদব কায়া আছে, এইটা তোফায়েল সাহেব দেখতে চাইবে।... তোফায়েল সাহেব কে? কমার্স মিনিস্টার তোফায়েল সাহেব। নেন- কফির কাপ চুমুক দিতে দিতে হাঁটা শুরু করেন। হাঁটাখাটি 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 ইউনিলয়েন্স টেক্সটাইলসের ভূমিকা। আমি কিছুক্ষণ তবদা মেয়ে থাকলাম। ‘যেই লাইন বললেন, এইটার বাংলা কী?’ এইটাই বাংলা, একটু গুছাইয়া এইটা এঁটা বইল্যা এইটো, আর শেষে বর্তমান সরকারের ইতিবাচক ভূমিকা ক্রিয়াকর দিবেন, এতটুকুই, পারবেন না? স্যার, আপনার পরকটে চিক্রনি আছে? না তো, কেন? আমার চুলটা একটু আচড়াইয়া নিতাম, চুল এনেদেলে থাকলে আপনার ভাবী আবার রাগ করে, আমার ক্যামেরা ফেস ভালো, খালি এলেমেলো 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!</p>	0
<p>2</p> <p>অদভূত বিরোধীদলহীনতায় ভুগছে সরকার। এ এক অন্যরকম শূন্যতা। বাবা-মায়ের সাথে থেকে আজকালকার ছেলেমেয়েরা যেমন একাকীত্বে ভোগে, তেমনি সর্বাঙ্কু ঘারা বেষ্টিত থাকলেও বর্তমান সরকারের পাশে কী যেন নেই। কী যেন নেই! প্রচলিত অনিচ্ছায় জাতীয়পাটি বিরোধীদের চেয়ারে বসলেও, তারা ঠিক যেন বিরোধীদল নয়। অনেকটা বিরোধীদের রেক্রিটার 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। অভিসম্বল 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। অনলাইন অর্ডার করলে কখনোই ভাল মাল পাওয়া যায় না। জাতীয়পাটিকে 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 বাই। সেখানে গিয়ে আমরা প্রচলিত ধর্মমত পরিবেশ দেখতে পাই। যেন পুরো অফিসেরই মন খারাপ। অফিসের এক কোণায় স্পিকারে গ্লো সাউন্ডে গান বাজতে শুনি- ‘তোমরা কেউ কি দিতে পারো বিরোধীদের ভালবাসা?’</p>	0
<p>3</p> <p>রাশিয়া বিশ্বকাপ নকআউট পর্ব ফ্রান্সের সাথে ৪-৩ গোলে পরাজয়ের প্রান্তিতে ফুটবল থেকে অবসরের ঘোষণা দিলেন মেসি। খেলা শেষে মাঠ ছাড়তে ছাড়তে বিভ্রাট করে এমনটাই বলছিলেন আর্জেন্টাইন এই সুপারস্টার। তাহলেই চলছিলো ম্যাচ। শুরুতে ফ্রান্স গোল দিলে মাঠ কাঁপলেও পরবর্তীতে দুটি গোল মুক্ত হয় আর্জেন্টাইনের খুলিতা। বুক চেতিয়ে খেলা শুরু করেন মেসি। হঠাৎই বললে যায় ম্যাচের দৃশ্যপট। একে একে ফ্রান্স আরো তিনটি গোল চুকিয়ে দেয় আর্জেন্টাইনরা জালে। এরই সাথে বিশ্বকাপ থেকে বিদায় ঘন্টা বেজে যায় আর্জেন্টাইনরা 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 লিপি রিডিং করে বুঝতে পারেন, মেসি বলছে, ‘আমি আর ফুটবল খেলবোনা। আমি আর ফুটবল খেলবোনা।’ বাতাসের বেগে কথটি স্টেডিয়ামজুড়ে ছড়িয়ে পরে। পরবর্তীতে মেসির অবসর নেয়ার ব্যাপারে দ্রুত এ খবরটি স্টেডিয়াম প্রান্তনে ভাইরাল হয়ে যায়। এবং অনেকেরই তা বিশ্বাস করে বলেন। যদিও আনুষ্ঠানিকভাবে মেসি এখনো কিছু জানাননি। এখন শুধুই এ কিংবদন্তীর মুখ থেকে অবসর নেয়ার ঘোষণাটি শোনার পালা...</p>	0

Table 4.9 and 4.10 show the results we found before merging headline and content columns-

Table 4.9: Results Obtained from ‘Headline’ Column for Different Classifiers

	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

Table 4.10: Results Obtained from ‘Content’ Column for Different Classifiers

	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 Results with Different Data Split Ratios

There are a couple of standards for choosing the correct ratio for the data test train. For our model, we tried different split ratios to see how the model is behaving according to them. In real life, we see fake news pretty often but logistically real news appears more in our feeds. Here are some of the results with different train test ratios.

With 50:50 Train Test:

Table 4.11: 50:50 Train Test Result

	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

With 60:40 Train Test:

Table 4.12: 60:40 Train Test Result

	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

With 80:20 Train Test:

Table 4.13: 80:20 Train Test Result

	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

With 70:30 Train Test:

Table 4.14: 70:30 Train Test Result

	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

We can see passive-aggressive classifier (PAC) and support vector machine (SVM) performed well in all scenarios. Other classifiers have struggled through different ratios but at 70:30 they showed their best results.

4.8 Gradient Boosting Algorithm

Gradient boosting algorithm is a type of boosting machine learning algorithm. It gives results on the basis of the best upcoming model combined with the previous models. This boosting algorithm reduces

the error from prediction. We applied this boosting algorithm but found that our existing models are performing better.

Here in Table 4.15, we show the classification report found from Gradient Boosting Algorithm -

Table 4.15: Classification Report of Gradient Boosting Algorithm

Gradient Boosting Classifier Accuracy 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 Manual Testing of News

In order to experiment with our model we have built a manual testing program where a user can input the news and different classifiers would predict whether the news is fake or not.

Table 4.16: Manual Test of News

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

এবার রাজধানীতে ঘটলো আরেকটি চাঞ্চল্যকর হত্যাকাণ্ডের ঘটনা। ধরের টাকা পরিশোধ না করায় প্রচণ্ড গরমের মধ্যে সোয়েটার পড়িয়ে মেরে ফেলা হয়েছে খায়রুল আলম নামে এক যুবককে। রাজধানীর মানিকনগর পাওয়ারহাউস এলাকায় এই হত্যাকাণ্ডের ঘটনাটি ঘটেছে।\n\nজানা যায়, দীর্ঘদিন যাবৎ খায়রুলের কাছে টাকা পায় হারুন নামের এক প্রতিবেশী। অনেকবার টাকা দিবো দিচ্ছি বলেও খায়রুল ধরের 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 বখাটে যুবকেরা। ফলে অতিরিক্ত ডিহাইড্রেশনের কারণে মারা যায় খায়রুল।\n\nখায়রুলের গা থেকে সোয়েটার খুলে তার মরদেহটি টাকা মেডিকেল মর্গের নিকট হস্তান্তর করা হয়েছে।\n\nধারণা করা হচ্ছে, চাঞ্চল্যকর এ খুনের দায়ের 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
```

In Table 4.16, we can see that we have to input fake data into the model. Different classifiers have predicted the model and all of the classifiers are able to detect it is fake news apart from Multinomial Naive Bayes.

Chapter 5

Result Analysis

To assess the performance of our model, we will show the accuracy, precision, recall, and F-1 score of all the classifiers. We used 30% of the dataset as our testing dataset and 70% as the training dataset. Here, we describe the performance of each of the classifiers-

I. Passive Aggressive Classifier

Accuracy = 93.8%; Precision = 94.4%; Recall = 95.4%; F1-score = 94.9%.

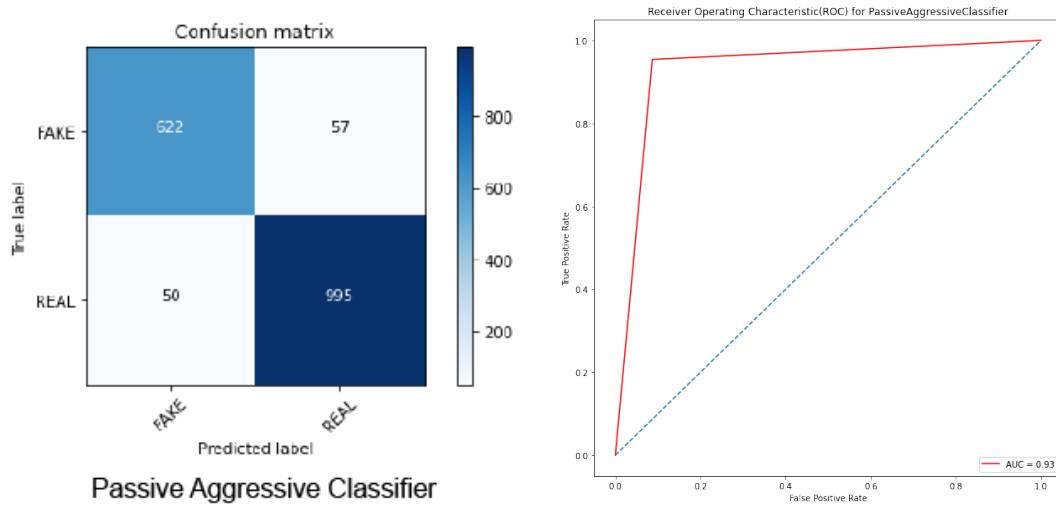


Figure 5.1: Confusion Matrix and ROC Curve for PAC

We see from Figure 5.1 that, accuracy is approximately 93.8%. Precision which is also called positive predictive value is the resultant of the division operation where true positive is considered as the numerator and the sum of true positive and false positive is considered as the denominator, gave a score of 94.4%. We found out that 622 fake news were correctly identified as fake news and 995 true news were correctly identified as authentic news. The Area Under the ROC curve(AUC) came to be 0.93, which shows that predictions were good. Furthermore, The recall and F-1 score of the classifier is 95.4% and 94.9% which shows the validity of the classifier.

II. Multinomial Naive Bayes

Accuracy = 86.9%; Precision = 82.6%; Recall = 99.2%; F1-score = 90.2%.

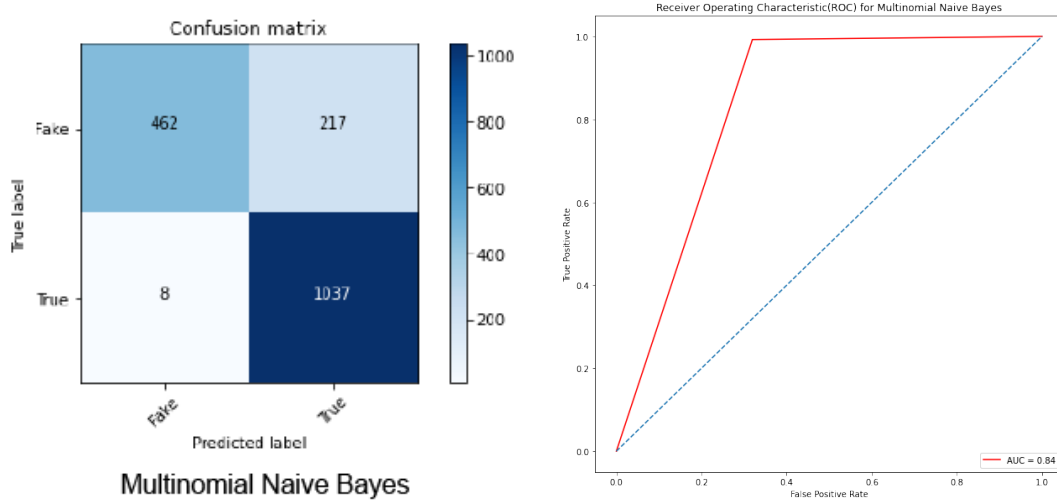


Figure 5.2: Confusion Matrix and ROC Curve for MNB

We see from Figure 5.2 that, accuracy is approximately 86.9%. Precision gave a score of 82.6%. We found out that 462 fake news were correctly identified as fake news and 1037 true news were correctly identified as authentic news. The Area Under the ROC curve(AUC) came to be 0.84, which shows that predictions were good. Furthermore, The recall and F-1 score of the classifier is 99.2% and 90.2% which shows the validity of the classifier.

III. Support Vector Machine

Accuracy = 93.5%; Precision = 93.3%; Recall = 96.2%; F1-score = 94.7%.

We see from Figure 5.3 that, accuracy is approximately 93.5%. Precision gave a score of 93.3%. We found out that 607 fake news were correctly identified as fake news and 1005 true news were correctly identified as authentic news. The Area Under the ROC curve(AUC) came to be 0.93, which shows that predictions were good. Furthermore, The recall and F-1 score of the classifier is 96.2% and 94.7% which shows the validity of the classifier.

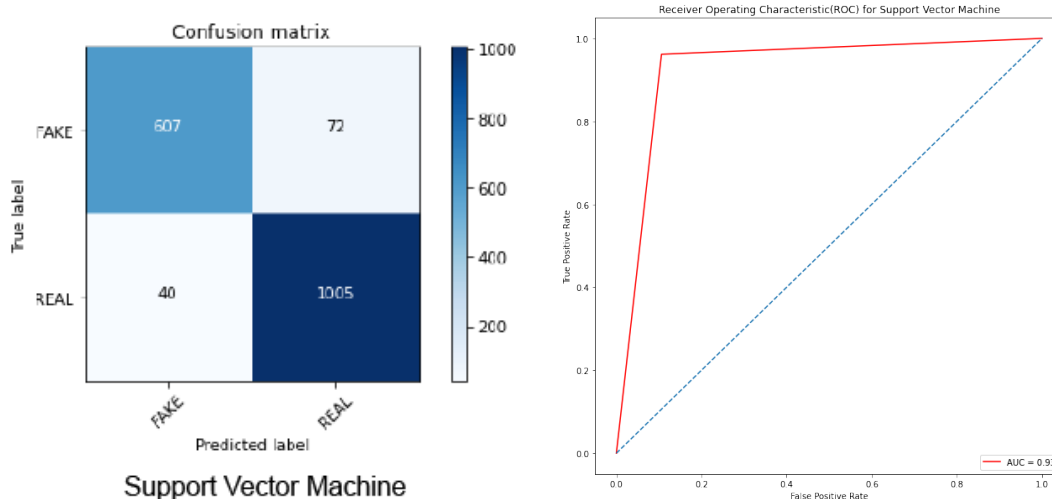


Figure 5.3: Confusion Matrix and ROC Curve for SVM

IV. Logistic Regression

Accuracy = 92.5%; Precision = 91.9%; Recall = 96.2%; F1-score = 94.0%.

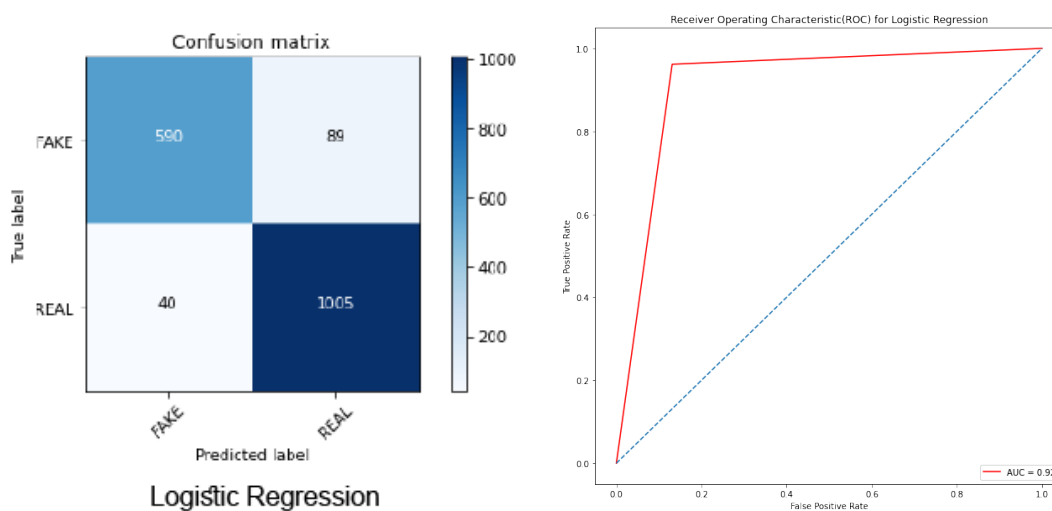


Figure 5.4: Confusion Matrix and ROC Curve for LR

We see from Figure 5.4 that, accuracy is approximately 92.5%. Precision gave a score of 91.9%. We found out that 590 fake news were correctly identified as fake news and 1005 true news were correctly identified as authentic news. The Area Under the ROC curve(AUC) came to be 0.92, which shows that predictions were good. Furthermore, The recall and F-1 score of the classifier is 96.2% and 94.0% which shows the validity of the classifier.

V. Decision Tree Classifier

Accuracy = 86.8%; Precision = 89.5%; Recall = 88.7%; F1-score = 89.1%.

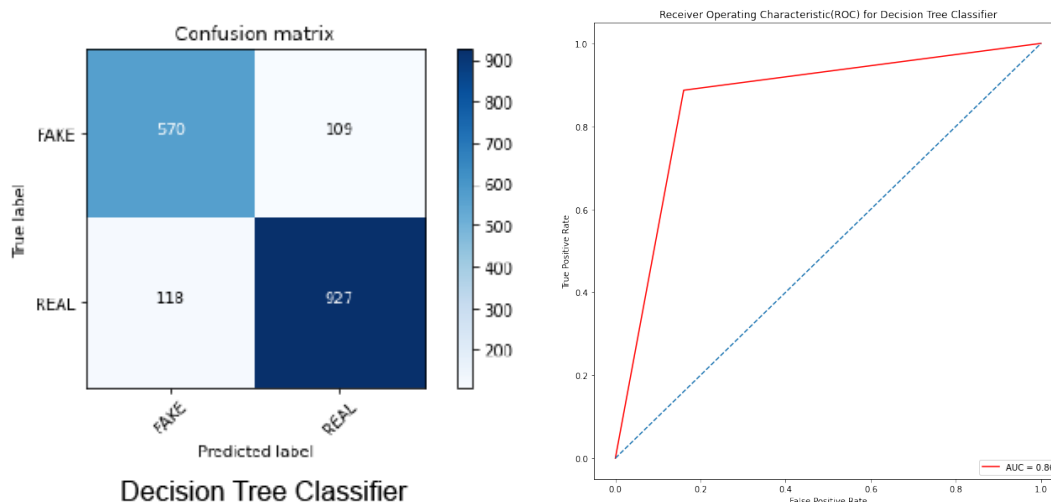


Figure 5.5: Confusion Matrix and ROC Curve for Decision Tree Classifier

We see from Figure 5.5 that, accuracy is approximately 86.8%. Precision gave a score of 89.5%. We found out that 570 fake news were correctly identified as fake news and 927 true news were correctly identified as authentic news. The Area Under the ROC curve(AUC) came to be 0.86, which shows that predictions were good. Furthermore, The recall and F-1 score of the classifier is 88.7% and 89.1% which shows the validity of the classifier.

VI. Random Forest Classifier

Accuracy = 93.2%; Precision = 90.8%; Recall = 98.9%; F1-score = 94.6%.

We see from Figure 5.6 that accuracy is approximately 93.2%. Precision gave a score of 90.8%. We found out that 574 fake news were correctly identified as fake news and 1033 true news were correctly identified as authentic news. The Area Under the ROC curve(AUC) came to be 0.92, which shows that predictions were good. Furthermore, The recall and F-1 score of the classifier is 98.9% and 94.6% which shows the validity of the classifier.

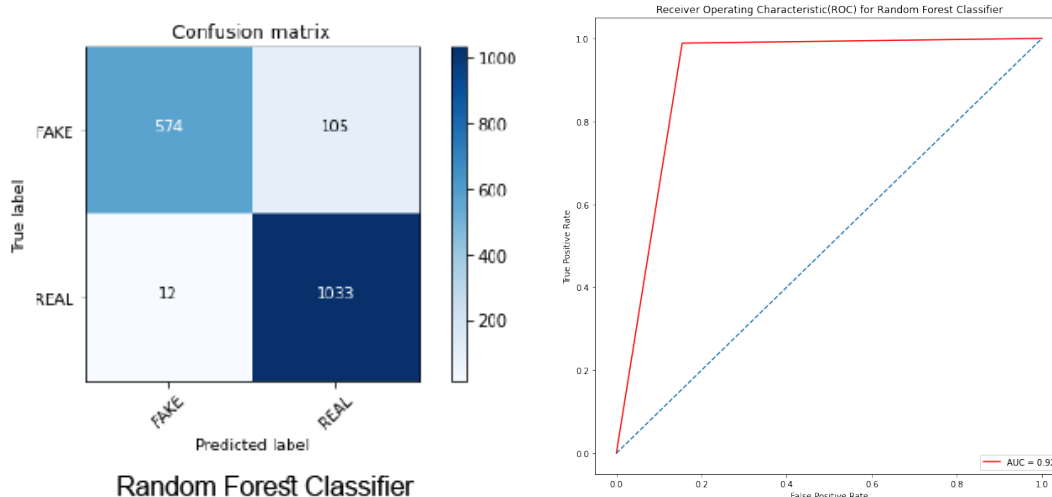


Figure 5.6: Confusion Matrix and ROC Curve for Random Forest Classifier

Most of the classifiers we have used gave an accuracy rate of over 90%. The following table shows the detailed result of our model-

Table 5.1: Result with different classifiers

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

We see from Table 5.1 that, Passive Aggressive Classifier showed the highest accuracy of 93.8% and it also gave the highest precision. Support Vector Machine and Random Forest Classifier also showed outstanding accuracy of 93.5% and 93.2% respectively. We also see that the Decision Tree Classifier showed the lowest accuracy of 86.8%, this occurred because of the nature of the algorithm. The algorithm is deterministic and greedy in nature. Also, they have a tendency to overfit. We further see that Multinomial Naive Bayes(MNB) gave low accuracy compared to other classifiers which is 86.9%. This occurred because MNB presumes that the features are independent of each other.

Result representation in column chart (Figure 5.7) which we achieved through the classifiers -



Figure 5.7: Result Histogram

Chapter 6

Conclusion

Through our research, we have classified Bangla fake news precisely with different machine learning classifiers. After collecting the datasets of fake and authentic news, we used them to train and test the model. Furthermore, we pre-processed the dataset by removing punctuations, numbers, stopwords etc. Next, we merged the headlines and contents of the datasets. After that, we implemented TF-IDF as a part of feature extraction. Finally, we used the classifiers to get our result. Most of the classifiers achieved accuracy of over 90%. Among all the classifiers Passive Aggressive Classifier and Support Vector Machine worked better than the other machine learning classifiers. Moreover, multiple classifiers have been used to get a big picture of how these classifiers work on Bangla fake news. In the future, if researchers want to research in this field they can get a decent idea about which classifier to use for their model. In this model, it is evident that to detect Bangla fake news the classifiers work better with TF-IDF. Other techniques like ‘Word2Vec’ also can extract features although it works better with big datasets. Due to the lack of a fake dataset, we could not explore more feature extraction techniques. Moreover, there are very few tools that work with Bangla natural language processing. We believe that if we can solve this issue, our model will perform better. We wish to increase our fake dataset as much as possible. We also want to experiment with various methods that have shown to be useful in other widely spoken languages and see if we can adapt them to the Bangla language. We will work on a browser extension that would show the result immediately if someone types in bogus or true news. Finally, we wish to adopt BERT (Bidirectional Encoder Representations from Transformers) [23] for Bangla because it has a lot of potential and is now regarded as one of the best transformers. We are confident that these future efforts will significantly improve our research.

Bibliography

- [1] M Martinez. (2018). “Burned to death because of a rumour on whatsapp,” [Online]. Available: <http://www.bbc.com/news/world-latin-america-46145986>.
- [2] M. S. A. Chowdhury, A. Hossain, and M. J. Rime, “News literacy in bangladesh,” *Management and Resources Development Initiative (MRDI)*,
- [3] E. Team. (2020). “Rumor: Thankuni will prevent coronavirus,” [Online]. Available: <https://rumorsscanner.com/fact-check/archives/750>.
- [4] R. Rafe. (2019). “Bangladesh: Fake news on facebook fuels communal violence,” [Online]. Available: <https://www.dw.com/en/bangladesh-fake-news-on-facebook-fuels-communal-violence/a-51083787>.
- [5] P. Bhikkhu. (2014). “Who will be tried for ramu destruction?” [Online]. Available: <https://en.prothomalo.com/opinion/Who-will-be-tried-for-Ramu-destruction>.
- [6] M. Z. Hossain, M. A. Rahman, M. S. Islam, and S. Kar, “Banfakenews: A dataset for detecting fake news in bangla,” *arXiv preprint arXiv:2004.08789*, 2020.
- [7] D. M. Eberhard and F Gary, “Simons, and charles d. fennig (eds.). 2021,” *Ethnologue: Languages of the world*, vol. 24,
- [8] M. G. Hussain, M. R. Hasan, M. Rahman, J. Protim, and S. Al Hasan, “Detection of bangla fake news using mnb and svm classifier,” in *2020 International Conference on Computing, Electronics & Communications Engineering (iC-CECE)*, IEEE, 2020, pp. 81–85.
- [9] S. B. S. Mugdha, S. M. Ferdous, and A. Fahmin, “Evaluating machine learning algorithms for bengali fake news detection,” in *2020 23rd International Conference on Computer and Information Technology (ICCIT)*, IEEE, 2020, pp. 1–6.
- [10] S. Lorent *et al.*, “Master thesis: Fake news detection using machine learning,” Université de Liège, Liège, Belgique, 2019.
- [11] N. Ruchansky, S. Seo, and Y. Liu, “Csi: A hybrid deep model for fake news detection,” in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 2017, pp. 797–806.
- [12] E. Shoemaker, “Using data science to detect fake news,” JMU Scholarly Commons, 2019, p. 714.
- [13] A. Thota, P. Tilak, S. Ahluwalia, and N. Lohia, “Fake news detection: A deep learning approach,” *SMU Data Science Review*, vol. 1, no. 3, p. 10, 2018.

- [14] A. Benamira, B. Devillers, E. Lesot, A. K. Ray, M. Saadi, and F. D. Malliaros, “Semi-supervised learning and graph neural networks for fake news detection,” in *2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, IEEE, 2019, pp. 568–569.
- [15] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer, “Online passive aggressive algorithms,” *Journal of Machine Learning Research*, 551–585, 2006.
- [16] J. D. Rennie, L. Shih, J. Teevan, and D. R. Karger, “Tackling the poor assumptions of naive bayes text classifiers,” in *Proceedings of the 20th international conference on machine learning (ICML-03)*, 2003, pp. 616–623.
- [17] R. Gandhi, “Support vector machine — introduction to machine learning algorithms,” URL: <https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47> (hämtad 14 maj 2020), 2018.
- [18] D. Jurafsky and J. H. Martin, *Speech and language processing (3rd edition draft ed.)* 2020. [Online]. Available: <https://web.stanford.edu/~jurafsky/slp3/5.pdf>.
- [19] *Decision tree classification algorithm*. [Online]. Available: <https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm>.
- [20] T. Yiu, “Understanding random forest,” *Towardsdatascience.com. June*, vol. 12, 2019.
- [21] C. Snijders, U. Matzat, and U.-D. Reips, “Big data: Big gaps of knowledge in the field of internet science,” *International journal of internet science*, vol. 7, no. 1, pp. 1–5, 2012.
- [22] Techopedia, *Data preprocessing*, 2021. [Online]. Available: <https://www.techopedia.com/definition/14650/data-preprocessing>.
- [23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.