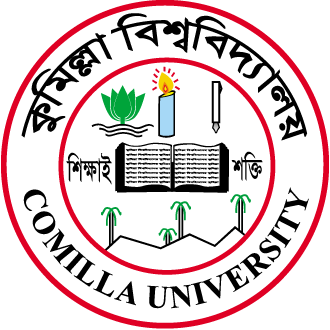
**Detecting Bangla Fake News With Machine Learning: Multi-Algorithm Approach**



**A PROJECT REPORT**

***Submitted by***

|  |  |  |
| --- | --- | --- |
| ***MD. Samrat Hossen*** | ***Maymuna Akter*** | ***Jannatul Ferduse*** |
| *ID: 11808050* | *ID: 11808026* | *ID: 11808002* |

***Supervised By***

***Mahmuda Khatun***

*Assistant Professor*

*Department of Computer Science and Engineering*

***In partial fulfillment for the course Research Project (Part-II)***

***of***

**BACHELOR OF SCIENCE(ENGG.)**

**IN**

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**COMILLA UNIVERSITY :: CUMILLA – 3506**

**September 2023**

21

**ACKNOWLEDGEMENTS**

We consider ourselves extremely lucky to have the help of a number of persons during this project. First and foremost, we want to express our gratitude to the Almighty God for his neverending greatness to complete this challenging task.

We'd want to express our heartfelt gratitude to our honorable supervisor, **MAHMUDA KHATUN**, for providing us with the excellent chance to work on this fantastic project on **" Detecting Bangla Fake News With Machine Learning: Multi-algorithm Approach "**. She was always encouraging, friendly, and motivating. Her willingness to freely give of her time has been greatly appreciated. Thank you for your constant encouragement, help, knowledge, and experiences over the last year.

We learned a lot of new information. After that, we'd like to thank our friends for their assistance in completing this project in such a short period time.

Finally, we wish to thank our family for their support and encouragement throughout our study.

**ABSTRACT**

Today, fake news is a significant problem, causing confusion worldwide. Therefore, creating the most accurate algorithm feasible would be a discovery with significant implications for social issues that are prevalent as well as the current political environment. People utilize social networks and web-based news items as their main source of information and news because they are quick, simple, and inexpensive to access—all with just one click. The fact that there is no means to confirm the origin, dependability, or legitimacy of the opinions being approved is one of its many shortcomings. The validity of the arguments being made Six machine learning techniques (Logistic regression, Decision tree classification, Gradient Boosting Classifier, Random Forest Classifier, KNN and SVM) for comparison to see which algorithm is more effective and efficient and better for predicting the result. In our project, we use these algorithms to predict. But there may casue one confusion like some algorithms may detect one news as Real when it is fake, vice-versa. Then we may take decisions like when most of the methods detect as Fake news, it is Fake news and vice-versa.

**Keywords:** Real News, Fake News, Machine Learning, Classifications, Confusion matrix.

**Contents**

|  |  |
| --- | --- |
| **Contents** | **Page** |
| **Acknowledgement** | **iii** |
| **Abstract** | **Iv** |
| **Table of Contents** | **v** |
| **List of Tables** | **vi** |
| **List of Figures** | **viii** |
| **Chapter 1: Introduction** | **1-3** |
| 1.1 Introduction | 2 |
| 1.2 Motivation | 2 |
| 1.3 Objectives | 3 |
| 1.4 Contribution | 3 |
| 1.5 Paper Structure | 3 |
|  |  |
| **Chapter 2: Literature Review** | **4-6** |
| **Chapter 3: Overview of Study** | **9-15** |
| 3.1 Research Subject and Instrument | 10 |
| 3.2 Algorithm used | 10 |
| 3.2.1 Linear Regression | 10 |
| 3.2.2 Decision Tree Classifier | 10 |
| 3.2.3 Gradient Boosting Algorithm | 11 |
| 3.2.4 Confusion Matrix | 12 |
| 3.2.5 Random Forest Algorithm | 13 |
| 3.2.6 KNN | 14 |
| 3.2.7 SVM | 15 |
|  |  |
| **Chapter 4: Methodology** | **16-19** |
| 4.1 Dataset Creation | 17 |
| 4.2 Data Preprocessing | 17 |
| 4.3 Traing the data | 18 |
| 4.4 Analyze the performance | 19 |
|  |  |
| **Chapter 5: Experimental Result and Discussion** | **20-26** |
|  |  |
| **Chapter 6: Conclusion** | **27-28** |
| 6.1 Conclusion | 28 |
| 6.2 Limitations | 28 |
| 6.3 Future Works | 28 |
|  |  |
| **REFERENCES** | **29** |

## **LIST OF TABLES**

|  |  |
| --- | --- |
| **List Of Tables** | **Page** |
| Table 3.1: Confusion Matrix | 12 |
| Table 4.1: Sample Dataset For Real News | 17 |
| Table 4.2: Sample Dataset For Real News | 17 |
| Table 5.1 Accuracy of Algorithms | 21 |
| Table 5.2 Classification report for LR | 21 |
| Table 5.3 Classification report for Decision Tree classifier | 22 |
| Table 5.4 Classification report for Gradient Boosting classifier | 23 |
| Table 5.5 Classification report for Random Forest Classifier | 23 |
| Table 5.6 Classification report for KNN | 24 |
| Table 5.7 Classification report for SVM | 24 |

## **LIST OF FIGURES**

|  |  |
| --- | --- |
| **List of Figures** | **Page** |
| 3.1 Logistic Regression Curve | 10 |
| 3.2 Decision Tree Classifier | 11 |
| 3.3 Random Forest Clssifier | 13 |
| 3.4 KNN Classification Block Diagram | 14 |
| 3.5 Hyperplane in SVM | 15 |
| 4.1 Workflow diagram | 17 |
| 5.1 Confusion Matrix for Logistic Regression | 22 |
| 5.2 Confusion Matrix for Decision Tree | 22 |
| 5. 3 Confusion Matrix for Gradient Boosting Classifier | 23 |
| 5.4 Confusion Matrix for Random Forest Classifier | 24 |

###### **Chapter 1**

###### **Introduction**

**1.1 Introduction**

Fake news is information that is inaccurate and erroneous but is intended to lead readers to believe it is factual and accurate. False information is spread to incite racism, feed people's anxieties, and fuel bullying and violent acts against the defenseless. Even bogus news has significant democratizing effects. When Donald Trump used the phrase to deny any accusations against him throughout the 2016 U.S. presidential campaign, "fake news" became widely known. The reason of increasing fake news are media distortion and misinformation, governmental and social power, incitement and social conflict, and economical benefit. Fake news manipulates people’s minds to think a certain way and makes people support a particular opinion.

On the other side, individuals and groups with potentially harmful intentions are known to propagate false information with the intention of influencing events and policies around the globe. Due to rumors, there have been numerous fatal events in Bangladesh during the past few years. Because of widespread reports about impending human sacrifice during the construction of the Padma Bridge, ten people were hurt and five people died after being battered by mobs in July 2019. In this electronic era, it is one of the biggest challenges to control the spreading of false or misleading news due to the free flow of information through social networking sites such as Facebook, Twitter, YouTube, micro-blogging, and others. For instance, a new report says, nearly one in three citizens across the United States, Spain, Germany, United Kingdom, Argentina, and South Korea claim they have seen false or misleading information on social media related to COVID-19.

In Bangladesh, the 2012 Ramu incident is an exemplary event where almost 25 thousand people participated in destroying the Buddhist temples based on a Facebook post from a fake account (Manik, 2012). About 12 Buddhist temples and monasteries and 50 houses were destroyed by the angry mob. Fake news that contains blasphemy can easily repeat these types of incidents where people are very sentimental to their religions.

So, the identification of fake news is becoming more and more essential. Many Researchers have recently been very interested in identifying bogus news. When the question is to identify fake manipulated news, various methods have been used so far.

In our study, we provide a revolutionary classifier-based approach for detecting bogus news in Bangla. The substance of recent social situations is a basis for the works that are already in existence. To assess whether a news item is false or legitimate, we employed a larger collection of news-related attributes and a multi-source Bangla news dataset with work on 48 thousand data. In our research, we develop a model that uses machine learning algorithms to determine whether some news is real or false, with an improved probability of preventing negativity and confusion across the nation. We achieved very high accuracy using six machine learning algorithms, including support vector machines (SVM), gradient boosting classifiers, decision trees, random forests, and logistic regression.

**1.2 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. 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.

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. Another 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.

So, we see that if anyone receives any 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. To resist this spread of fake news, we are proposing a model using different machine learning classifiers which will detect Bangla fake news more accurately.

**1.3 Objectives**

The following are the research's key goals:

1. Preserving information integrity
2. Enhancing Media Literacy
3. Maintaining Trust in Media
4. Mitigating Social Polarization
5. Contributing to technological Advancements

**1.4 Contribution**

We discussed papers about fake news where various solutions were given. These solutions are not very efficient due to they only discussed within . Our contribution is regarding as follows.

* Creating two Dataset for real and fake news from various online portals and social media
* Enhancing Performance
* Multi-model integration
* Real time detection system
* Specilizing for Bangla news
* Ethical and Fair Solutions

**1.5 Report structure**

The organization of this research report is as follows: Chapter 2 presents the literature review regarding our research topic. We try to give an overview of all algorithms in Chapter 3. In Chapter 4 we described the methodology stepwise. Chapter 5 discusses the experimental results and we conclude the report mentioning the limitations and recommending of some future works in Chapter 6.

###### **Chapter 2**

###### **Literature Review**

They proposed a novel framework based on a deep neural network architecture that exploits the information from the news articles and the social contexts to detect fake news. They identified and addressed two unique challenges related to fake news: a. early fake news detection and b. label shortage. Here the proposed model is based on a transformer architecture, which has two parts: the encoder part to learn useful representations from the fake news data and the decoder part that predicts the future behavior based on past observations. They included a broader set of news-related features and social context features compared to the previous works.[1]

They can only represent a limited amount of fake news detection analysis over a given period.[1]

Another limitation in this study is that they were only use a small portion of users profiles from the currently available dataset. Another limitation of this study is the availability of limited resources like GPUs, memory, data storage, due to which we could not perform many experiments on other large-scale data sources. [1]

In this research, A hybrid model combining recurrent neural networks (RNN) and support vector machines (SVM) was developed to detect real and fake news.[2]

The radial basis function kernel of an SVM was fed the encoded features. The suggested framework outperforms state-of-the-art techniques, according to experiments on real-world datasets, which could have an impact on businesses and governments. A hybrid algorithm that outperforms previous models in terms of accuracy and F1 score metrics has been presented to identify rumours in the PolitiFact and GossipCop datasets.[2]

One limitation is that the researchers use SVM, and the size of the feature vector affects the performance of the SVM; larger feature vectors improve performance while smaller ones degrade it. The output of a feature vector must be at least 512 units in size. [2]

In this work, detection of fake news is proposed using two different modalities available in an efficient manner using Support Vector Machine (SVM) and Multinomial Naive Bayes (MNB) and two classifiers SVM and MNB have been used to classify fake news from Bangla news articles.

Finally, they show SVM with the linear kernel with an accuracy of 96.64% outperform MNB with an accuracy of 93.32%.[3]

In this study classifiers are limited to the Bangla language and generate poor macro F1 scores and can be used only for Bangla language. Need to require additional text capabilities and fusion features. [3]

This paper focuses on identifying fake news in web-based informal organizations using various information mining perspectives. Four categories—Application Oriented, Data Oriented, Model Oriented, and Features Oriented—are used to categorise the research. This paper presents a fake news recognizable proof model using man-made intelligence methods. Eight machine learning methods were explored, with the Direct Help Vector Classifier being the best with a 96% precision.[4]

This paper achieved 95.7% accuracy in testing and 99.3% accuracy in training using the SVC model.

For this system to become more effective at detecting things, it needs additional data. One limitation is that they need to detect fake information more accurately than they do now. [4]

In this research, the researcher focuses on identifying fake news on social media platforms by combining Subjective Opinions with two fresh approaches such as a Subjective Opinion based model and a Probability based model and experiments on two popular real-world datasets, BuzzFeed and PolitiFact.[6]

On the BuzzFeed dataset, SO fnd had the highest Accuracy (87.1%), F1 score (88.0%), and second-best precision (81.6%); on the PolitiFact dataset, SO fnd has the highest Accuracy (95.3%), F1 score (95.3%), and third-best precision (95.3%). Accuracy was 94.1%, while recall was second-best at 96.7%.[6]

One thing lacking is that they didn’t improve news content mining to predict news veracity and introduce user profile features like age, account verification, registration time, and follower count.[6]

They aimed to provide Deep Learning-based detection models for differentiating fake news from real news. They built topic identification models to retrieve topics appearing in fake news and in real news regarding COVID-19. They identified feature differences between fake news and real news items about the pandemic. The goal of this paper is to identify fake news on social media to help stop the spread. They discovered feature differences between fake news and real news items and added them into the sentence embeddings, they found that they affected the model performance.[7]

They found that half of the identified topics are common in fake news and real news and shows that the existence of fake news could make readers confused and make them mistake fake news for real news. [7]

They were not able to work on achieving better transferability for fake news data from different domains, and not building detection models that can identify cross-domain fake news. [7]

In this Research, they try to build the "Bangla fake news" dataset that reduces redundant data by combining newly acquired data with secondary datasets. They used the most effective models, such as CNN, CNN-LSTM, and BiLSTM, in experiments involving diverse machine learning, deep neural networks, and transformer models. The datasets' application increased accuracy by 1.4% to 3.4%. They applied the BNLP toolkit for text preprocessing purposes and used Count Vectorizer and TF-IDF Vectorizer to extract numeric features from unstructured text data.[8]

In this research, ML, DL, and Transformer models achieved over 90% accuracy in fake news classification, with LSTM model showing the highest recall value at 95.9%.One limitation is that they only use previous benchmark datasets by collecting new fake news data and combining previous Bangla fake news datasets available online. They didn’t work with English news datasets.They need to further enhance the Bangla fake news dataset and add limited fake news data from different sources. [8]

In this research they provide a comprehensive analysis of several machine and deep learning methods, as well as how the systems perform to detect fake Urdu news, which are submitted to the shared tasks UrduFake@FIRE2020 and UrduFake@FIRE2021.[9]

In this study two datasets prepared for the shared tasks and the datasets contained five kinds of news: (v) Technology, (ii) Showbiz (entertainment), (iii) Business, (iv) Health, and (v) Sports.

In this paper the shared tasks were organized to entice other academics to encourage and develop to automated systems for detection and classification of Urdu fake news instances in digital media.The objective of the shared tasks was to classify a given Urdu news instance as either real or fake, and the teams proposed numerous algorithms to achieve this goal.

In this study, they investigated different methods of machine learning, and these approaches range from different text re-presentation techniques, such as pre-trained embedding, contextual representation, and end-to-end neural network-based methods. They only work on fake Urdu news and show that limited amount of fake news detection analysis over a given period of time. They were not able to work on huge dataset. [9]

This study compares many proposed approaches and offers solutions for the imbalance in fake news identification in Bangla. It also provides a method for enhancing performance even with an unbalanced dataset and present stacked generalisation as a substitute for manipulation-free fake news detection. By overcoming the dataset restrictions of BanFakeNews, this project hopes to considerably advance the Bangla false news detection system. They created the baseline model for performance evaluation, and three strategies were used to handle unbalanced data. Six classifiers were selected as baseline models, including logistic regression, support vector machine, multinomial Nave Bayes, Bernoulli Nave Bayes, random forest, and decision tree classifier.In comparison to baseline models, the results reveal a 93.1% F1-score when employing data manipulation techniques like SMOTE and a 79.1% F1-score when not.

They only concentrated on resolving the data imbalance issue. They didn't work with other dataset features. Additionally, they didn't experiment with different models, including deep learning models such as RNN, CNN, LSTM, and GRU. [10]

In this paper, they proposed a hybrid fake news detection system that makes use of both linguistic-based such as number of words, lexical diversity, title, reading ease, and knowledge-based approaches. This study includes three different types of data: (i) the credibility of the news outlet’s Website; (ii) coverage, or the amount of information used to report the news; and (iii) facts verification.

The detection of users based on shared interests and behaviors can also help in identifying fake news and a real-time fake news detection system will be utilized to benefit a range of new activities, such as preventing the dissemination of fake information during elections or even pandemics. They used an ensemble technique yields superior results in comparison with a straightforward classifier by using DL and ML methods to build an ensemble mode.

Using multimodal data to identify fake news is still a difficult and unexplored topic that requires more research.

In this research they can’t include more perspectives and not develop a more effective model that runs faster whilst still offering promising results. [13]

###### **Chapter 3**

###### **Overview of Study**

**3.1 Research Subject and Instrument**

Our analysis focuses on detecting fake and real news. We're doing it with a variety of machine learning algorithms. “A comparative analysis on detecting Bangla fake news on social media using machine learning algorithms" is our research subject. By browsing news portals, we gathered all of the data from Facebook and Twitter. We used laptops and mobile devices, as well as our own social media accounts, as instruments for our study. Following the data collection, several preprocessing techniques were used to generalize the data. After that, we prepared the preprocessed data for our research. The prepared data was then fed into machine learning (ML) algorithms in order to predict our result.

**3.2 Overview of Algorithms**

The algorithms mentioned below were used and described briefly:

**3.2.1 Logistic Regression**

Logistic regression is a method of supervised machine learning for categorization assignments that uses a sigmoid function to estimate the likelihood of an instance belonging to a given class. In contrast to linear regression, it predicts the likelihood of an instance belonging to a certain class, whereas linear regression produces a continuous number. It is a method for predicting a dependent variable with a categorical category using a set of independent variables. It delivers probability values ranging from 0 to 1, allowing for a variety of outcomes. A S-shaped logistic function is fitted in logistic regression, predicting two maximum values.

In figure 3.1, is showing the logistic function curve:



Fig 3.1: Logistic Regression Curve

**3.2.2 Decision Tree Classifier**

Decision Tree is a supervised learning method that can be applied to classification and regression problems, but it works best when dealing with classification-related issues.

There are two nodes in Decision Tree:

1. Decision Node: helps to take decisions
2. Leaf Node: ensures that doesn’t need to make branches

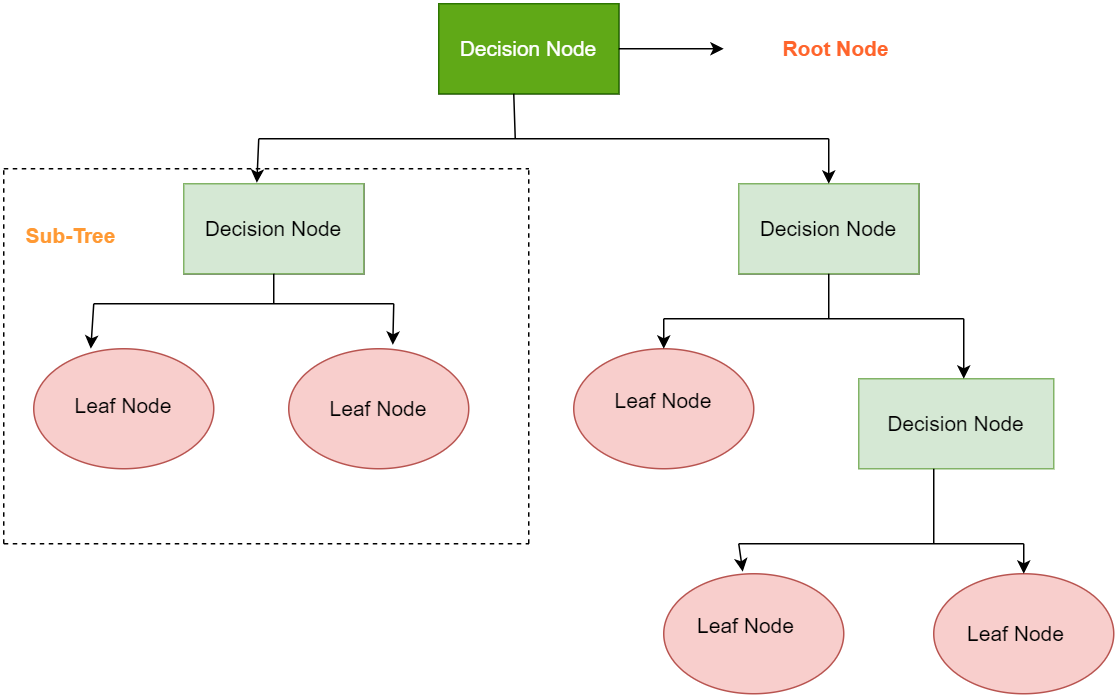


Fig 3.2: Decision Tress

**3.2.3 Gradient Boosting Classifier**

Gradient boosting classifier is a machine learning technique that combines several weak learners to produce a strong learner. Decision trees are frequently used for disadvantaged learners, however, alternative models can also be utilized. The approach operates by adding new trees to the ensemble repeatedly, each of which is taught to repair the errors of the prior trees. The gradient boosting classifier works by minimizing a loss function. The log loss function, which evaluates the difference between expected probabilities and actual labels, is the most often used loss function for classification.

Steps of Gradient Boosting Classifier:

1. Initialize the model parameters.
2. Fit a weak learner to the data.
3. Calculate the residual errors.
4. Fit a new tree to the residual errors.
5. Update the model parameters.
6. Repeat steps 2-5 until a stopping criterion is met.

Advantages of Gradient Boosting Classifier:

1. It is a very accurate algorithm for classification tasks.
2. It can handle complex data sets.
3. It is robust to overfitting.
4. It is relatively easy to interpret.

Disadvantages of Gradient Boosting Classifier:

1. Training can be computationally expensive.
2. It is susceptible to hyperparameter selection.
3. Tuning the algorithm for best performance can be tough.

**3.2.4 Confusion Matrix**

A confusion matrix is a crucial tool in machine learning and statistics that aids in evaluating classification methods. It gives a straightforward and comprehensive way to assess the correctness of a model's predictions by categorizing expected and actual class labels.

Let's delve into the components and terminology associated with a confusion matrix:

1. **True Positives (TP):** These are the instances where the model correctly predicted a positive class when the actual class was indeed positive.
2. **True Negatives (TN):** These are the instances where the model correctly predicted a negative class when the actual class was indeed negative.
3. **False Positives (FP)**: These are the instances where the model incorrectly predicted a positive class when the actual class was negative. This type of error is also known as a "Type I error."
4. **False Negatives (FN):** These are the instances where the model incorrectly predicted a negative class when the actual class was positive. This type of error is also known as a "Type II error."

Based on these components, a confusion matrix is often represented in a tabular form like this:

Table 3.1: Confusion Matrix

|  |  |  |
| --- | --- | --- |
|  | **Predictive Positive(P)** | **Predictive Negative(N)** |
| Actual Positive(P) | True Positives(TP) | False Negatives(FN) |
| Actual Negative(N) | False Positives(FP) | True Negatives(TN) |

Using the values in this matrix, you can calculate various performance metrics:

* **Accuracy**: It is calculated as (TP + TN) / (TP + TN + FP + FN).
* **Precision**: It is calculated as TP / (TP + FP). High precision indicates low false positives.
* **Recall (Sensitivity or True Positive Rate**): It is calculated as TP / (TP + FN). High recall indicates low false negatives.
* **Specificity (True Negative Rate):** It is calculated as TN / (TN + FP).
* **F1-Score**: It is calculated as 2 \* (Precision \* Recall) / (Precision + Recall).
* **False Positive Rate (FPR**): It is calculated as FP / (FP + TN).
* **False Negative Rate (FNR):** It is calculated as FN / (FN + TP).

**3.2.4 Random Forest Classifier**

A Random Forest classifier is a common and powerful ensemble learning technique used in machine learning for tasks such as regression and classification. It is an algorithm with a decision tree extension that combines the predictions of numerous independent decision trees to enhance their performance, resilience, and generalization.

Random Forest operates in two stages: the first is to generate the random forest by mixing N decision trees, and the second is to make predictions for each tree generated in the first phase.

The Fig 3.2.4.1 shows that how work Random Forest Classifier:

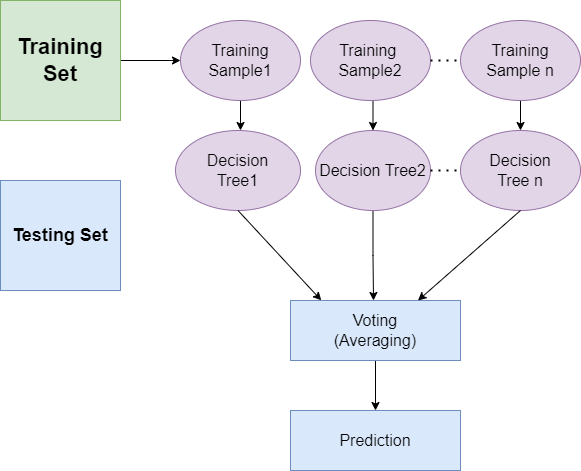


Fig 3.3: Random Forest Classifier

Steps of Random Forest Classifier:

1. Data Pre-processing
2. Fitting the Random forest algorithm to the Training set
3. Predicting the test result
4. Test accuracy of the result (Creation of Confusion matrix)
5. Visualizing the test set result.

Advantages of Random Forest Classifier:

1. Random Forest can handle both classification and regression tasks.
2. It can handle huge datasets with high dimensionality.
3. It improves the accuracy of models and avoids the overfitting problem.

Disadvantages of Random Forest Classifier:

1. Although random forest may be used for tasks such as classification and regression, it is not ideal for regression.

**3.2.5 K-Nearesr Neighbor(K-Nn):**

K-NN means K-Nearest Neighbor, is a supervised machine algorithm technique. It is a non-parametric algorithm and it’s also called lazy learner algorithm. It is used in regression and classification. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. [The Below figure shows how to classifiy KNN:](https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning)



Fig 3.4: KNN Block Diagram[14]

Steps of K-NN algorithm:

1. Choose the Kth neighbor's number.
2. Determine the Euclidean distance between K neighbors.
3. Pick the K closest neighbors based on the Euclidean distance estimate.
4. Determine how many data items fall into each category among these k neighbors.
5. Assign the additional data points to the category where the neighbor count is at its highest.
6. Our model is complete.

Advantages of K-NN algorithm:

1. It is easy to implement.
2. It is robust to the noisy training data.
3. It can be more effective if the training data is large.

Disvantages of K-NN algorithm:

1. Always needs to determine the value of K which may be complex some time.
2. Because the distance between each data point for each training sample must be calculated, the computation cost is considerable.

**3.3.6 Support Vector Machine**

A Support Vector Machine (SVM) is a supervised machine learning algorithm used for classification and regression tasks. It works by finding a hyperplane that best separates the data into different classes while maximizing the margin between the classes. The main objective of the SVM algorithm is to find the optimal hyperplane in an N-dimensional space that can separate the data points in different classes in the feature space.

To create the hyperplane, SVM selects the extreme points and vectors which are called Support vectors. Look at the diagram below, where two distinct categories are separated using a decision boundary or hyperplane:



Fig 3.5: HyperPlane in SVM[15]

The hyperplane tries to ensure that the margin between the closest points of different classes should be as maximum as possible. The dimension of the hyperplane depends upon the number of features.

###### **Chapter 4**

###### **Methodology**

* 1. **Dataset Creation:** Preparing the dataset is the first task for this research. We collect dataset from Kaggel website. We need to organize the dataset as needs of related features. The collected data are then used in preprocessing phase.  
     We show below some sample data collected from kaggle catagorical dataset. There are two types dataset: 1. Real News at Table 4.1. 2. Fake news at Table 4.2.

**Dataset Kaggle Link :** [Fake News](https://www.kaggle.com/c/fake-news/data?select=train.csv)

Table 4.1: Sample Dataset For Real News

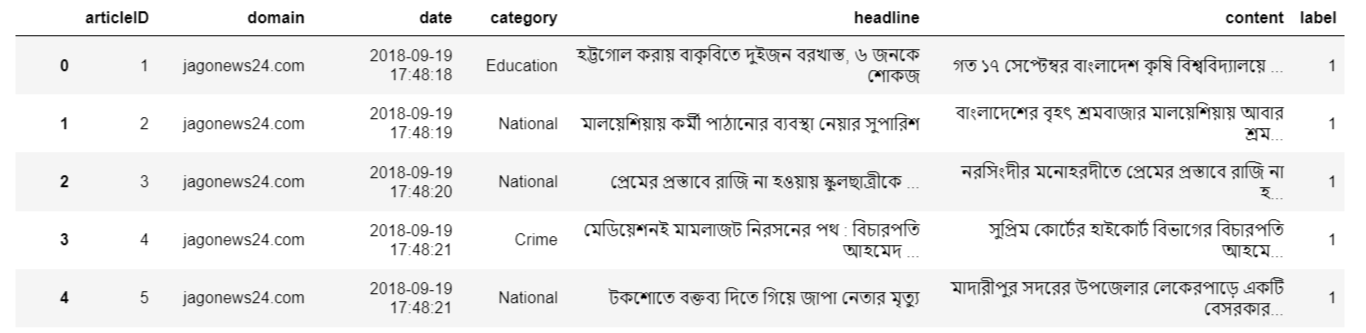
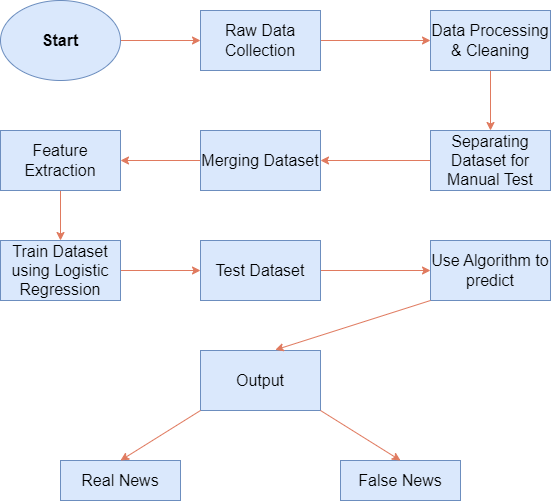


Table 4.2: Sample Dataset For Fake News



**4.2 Data Preprocessing:** After collecting raw data from kaggle as input, we need to preprocess these data and make these clean. Illustrating the whole procedure in the Figure 4.2.1:



**Fig 4.1: Workflow diagram**

**4.3 Training the Model:** Once all of the preprocessing processes have been completed, the dataset is ready to be fed into the model for training and testing. The trained model employs a number of techniques such as Linear Regression, Gradient Boosting Classifier, Decision Tree, Random Forest Classifier, KNN and SVM. See the followig steps to prepare the dataset to tain and test:

**Step 1**: We imported the datasets into jupyter notebook . The datasets named Real datasets and fake datasets respectably we have created a manual testing dataset for testing manually consisting last 10 rows from each dataset that shortens the datasets by 10 rows.

**Step 2**: We dropped the unnecessary column from the merged dataset.



**Step 3:** We constructed a function after randomly shuffling the data frame to convert the text to lowercase, remove unnecessary space, special characters, URLs, and links. We divided the dataset into training and testing datasets after defining dependent and independent variables as x and y.

**Step 4**: For text processing we have used tfidfvectorizer. All described broadly below:

**Tfidfvectorizer:**

The TF-IDF is a subtask of information retrieval and extraction that seeks to convey the value of a word to a text that is part of a corpus (a collection of documents). Few search engines use it to get improved results that are more important to a particular enquiry. In this section, we'll go over what TF-IDF is and how it works and explain the math behind this one, and then show how to use the SK-Learn library to implement it in Python.

**Term Frequency (TF):** The Term Frequency of a word is the amount of times it comes in a sentence. If a word appears more often than others, a higher value indicates that the text is a good fit when the term is part of the search words.

**IDF (Inverse Document Frequency**): Words that appear frequently in one document but not in others may be irrelevant. The IDF is a metric for determining how important a word is across the entire corpus.

The TfidfVectorizer transforms a pile of raw documents into a TF-IDF function matrix. This is done by counting the amount of times a word happens to be in a document as well as the number of times the identical word occurs in other documents in the corpus.

The reasoning for this as follows:

•A term that pops up often in a news has greater significance for that news, implying that the document is more likely to be directing towards that particular word.

•A word is that appears regularly in more news will make it difficult to find the right news, the word is appropriate for every news or no news at all. It won't help us strain out a single or a tiny portion of news from the entire collection in either case.

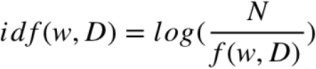
•As a result, in our dataset TF-IDF is indeed a result that is applied to each and all words in each text. And the TF-IDF value for each word increases with each occurrence in a news, but progressively decreases with each arrival in other papers or documents. The math for it is in the following section:

Then let's have a look at the TF-IDF numerical measure's basic formula. Let us start by defining few terms:

N, (the number of documents we have in our dataset) d, (a given document from our dataset)

D, (the collection of all documents) w, (given word in a document)

now the term frequency equation is:

The frequency of the word w in document d is f (w,d). Now the inverse frequency:

Now the final step tfidf computing:

**Step 5**: After that apply the algorithms mentioned in chapter 3 for training dataset and making prediction.

**4.4 Analyzing Performance:** We see that every algorithms works differently, and gives different type of accuracy and outcomes. But most of the case, it is very obvious to detect fake or false news.

###### **Chapter 5**

###### **Experimental Results and Discussion**

**5.1 Experiment Results:**

* + 1. **Accuracy of Algorithms:**

We list out accuracy for the given dataset of six machine learning algorithms which are used to train and predict. The table 5.1 shows the accuracy of all algorithms:

Table 5.1: Accuracy of Algorithms

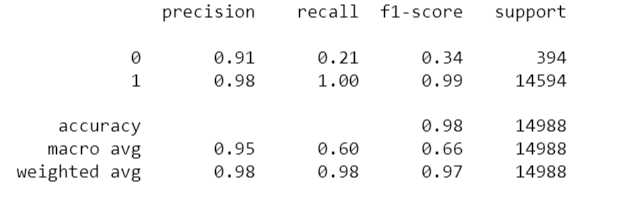
|  |  |
| --- | --- |
| **Algorithm** | **Accuracy** |
| Logistic Regression | 0.981051507872965 |
| Decision Tree | 0.9776487856952228 |
| Gradient Boosting Classifier | 0.9835201494528957 |
| Random Forest Classifier | 0.9784494262076328 |
| KNN | 0.9766479850547104 |
| SVM | 0.982586068855084 |

**5.1.2 Confusion Matrix**

**Linear Regression:**

The below table shows Total classification report by linear regression:

Table 5.2: Classification Report for LR



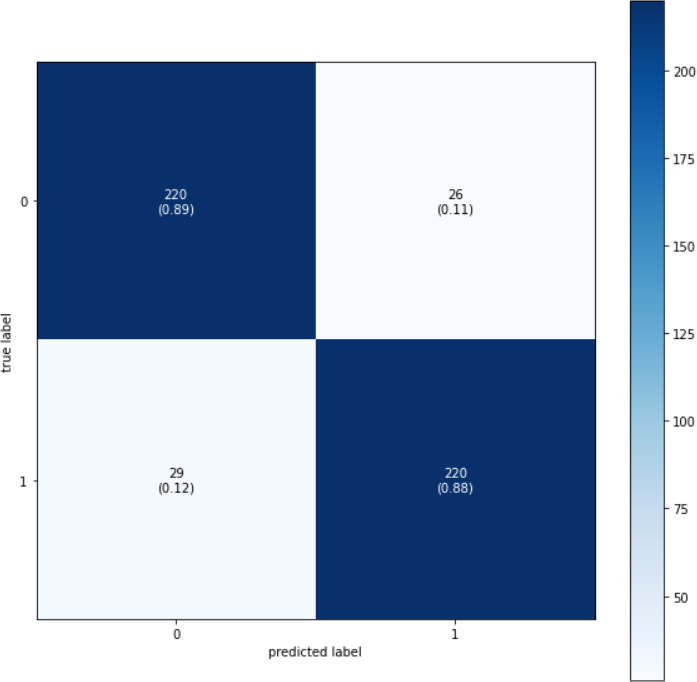
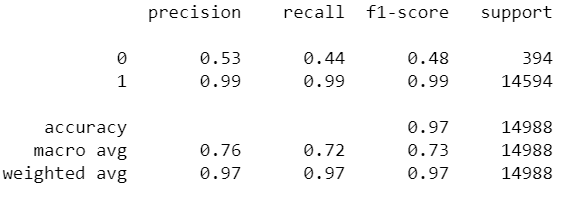
Confusion Matrix for Linear Regression is shown in below figure:

Fig 5.1: Confusion Matrix For Linear Regression

**Decision Tree:**

The table 5. shows total classification report by Decision Tree:

Table 5.3: Classification Report by Decision Tree



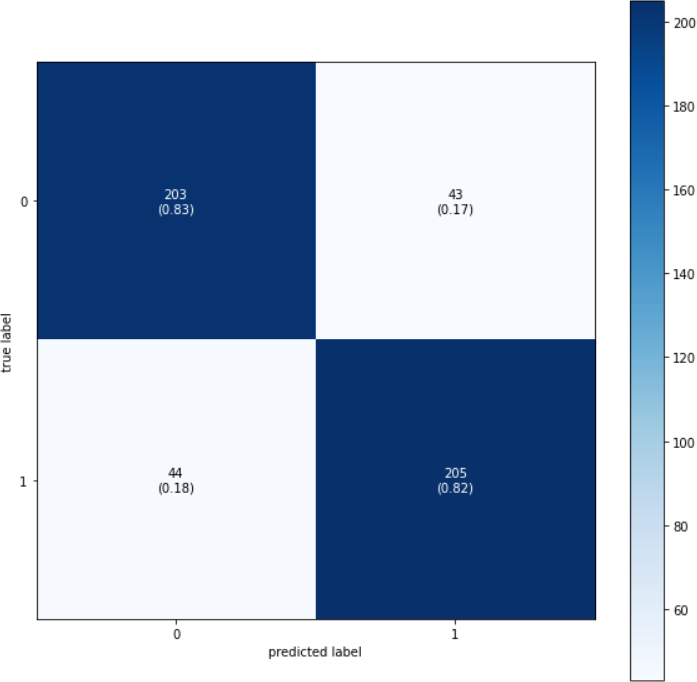
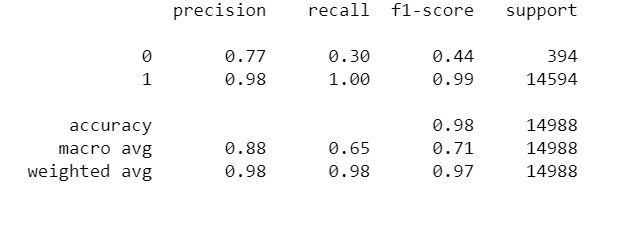
Confusion Matrix for Decision Tree is shown in below figure:

Fig 5.2: Confusion Matrix for Decision Tress

**Gradient Boosting Classifier:**

The below table shows Total classification report by Gradient boosting classifier:

**Table 5.4: Classification report by Gradient Boosting Classifier**



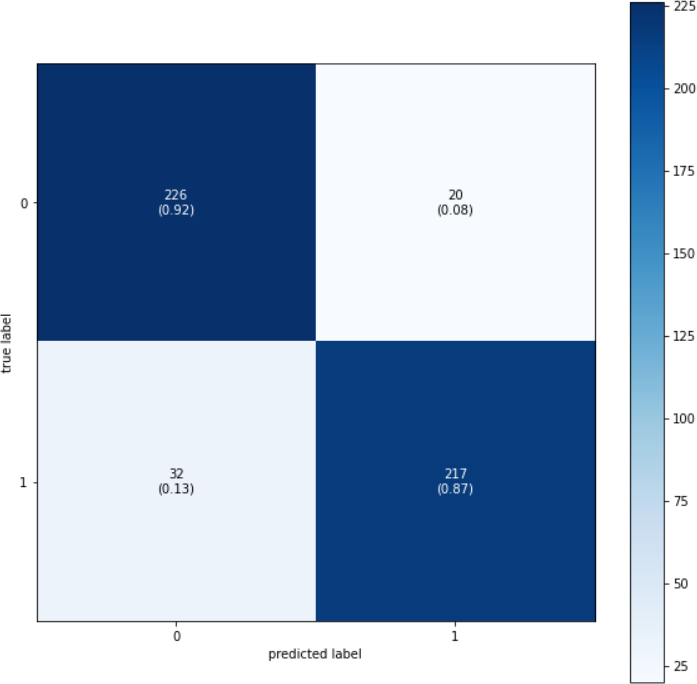
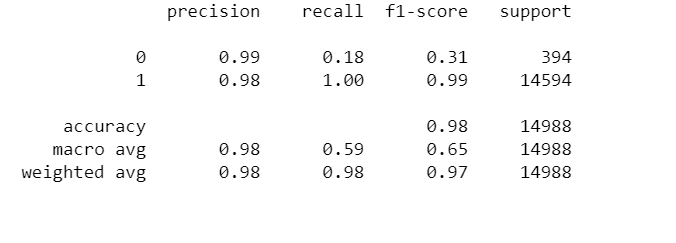
 Confusion Matrix for Gradient Boosting is shown in below figure:

Fig 5.3: Confusion Matrix for Gradient Boosting

**Random Forest Classifier:**

The below table shows Total classification report by Random Forest Classifier:

Table 5.5: Classification Report by Random Foresr Classifier



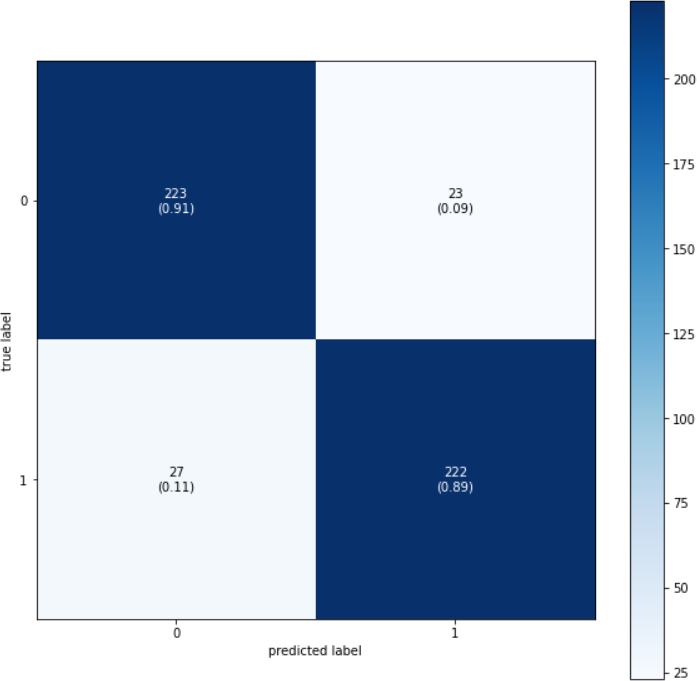
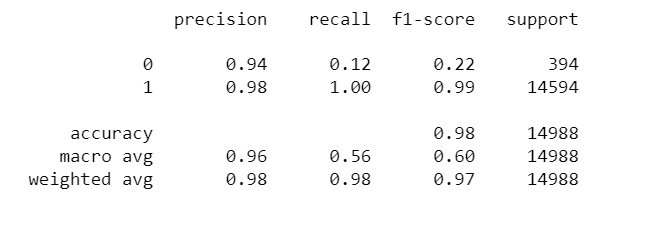
Confusion Matrix Random Foresrt Classifier is shown in below figure:

Fig 5.4: Confusion Matrix for Random foresrt Classifier

**K-Nearest Neighbor(KNN):**

The below table shows Total classification report by KNN:

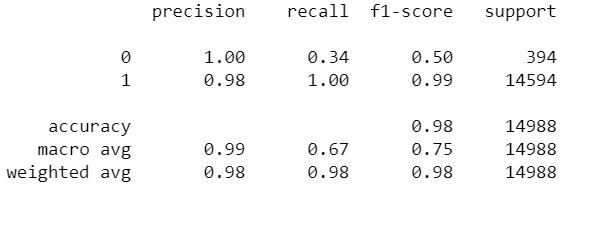
Table 5.6: Classification Report by KNN

****

**Support Vector Machine(SVM):**

The below table shows Total classification report by SVM:

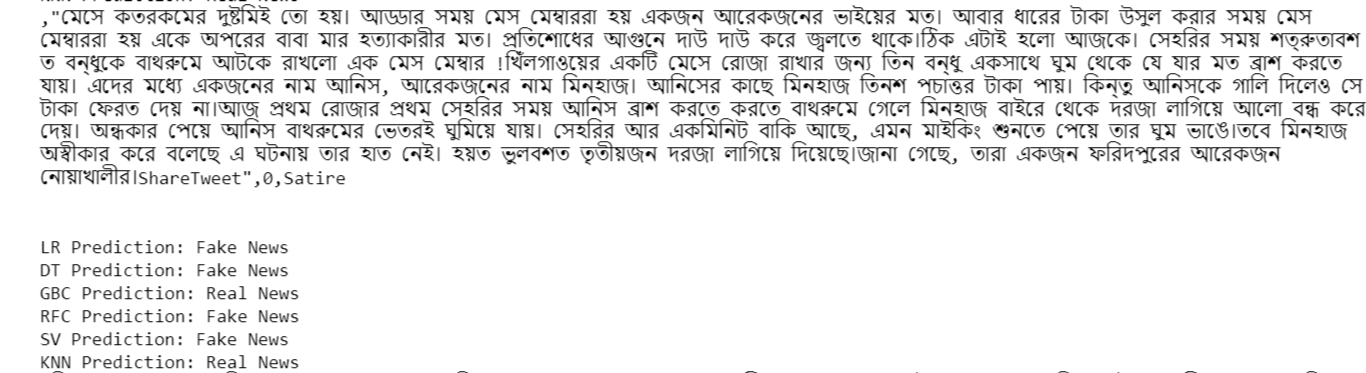
Table 5.7: Classification Report by SVM



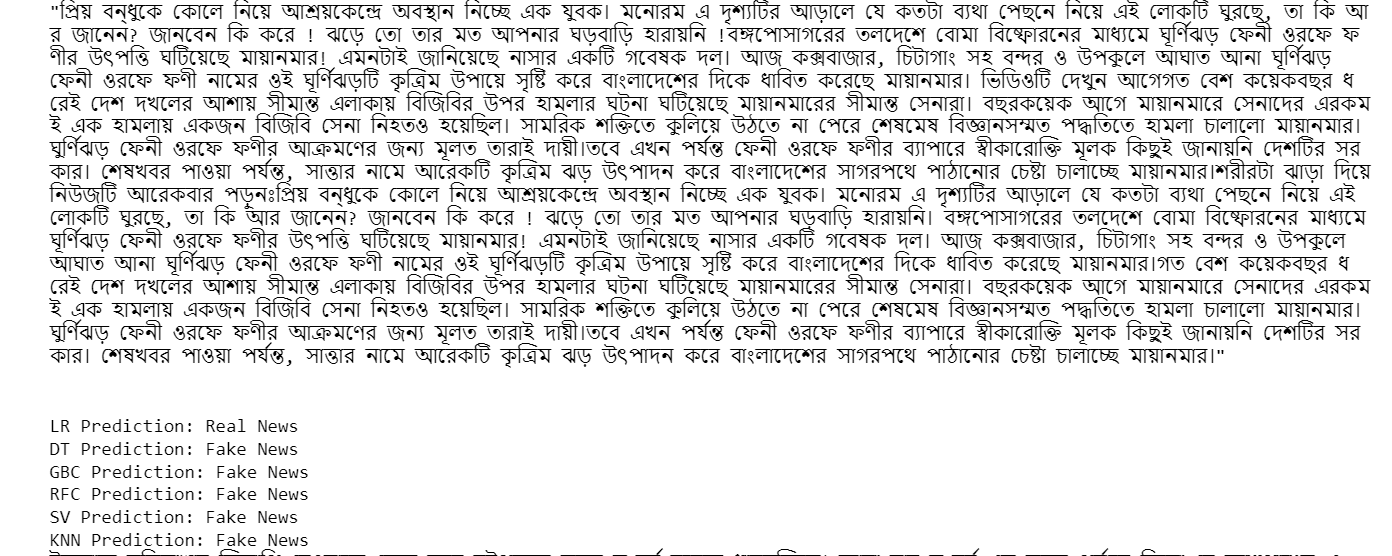
**5.1.2: Sample Input/Output**

We show below some experiment results where we read news input as Bangla text, and predict using mentioned machine learning algorithms the taken input news as real or fake.

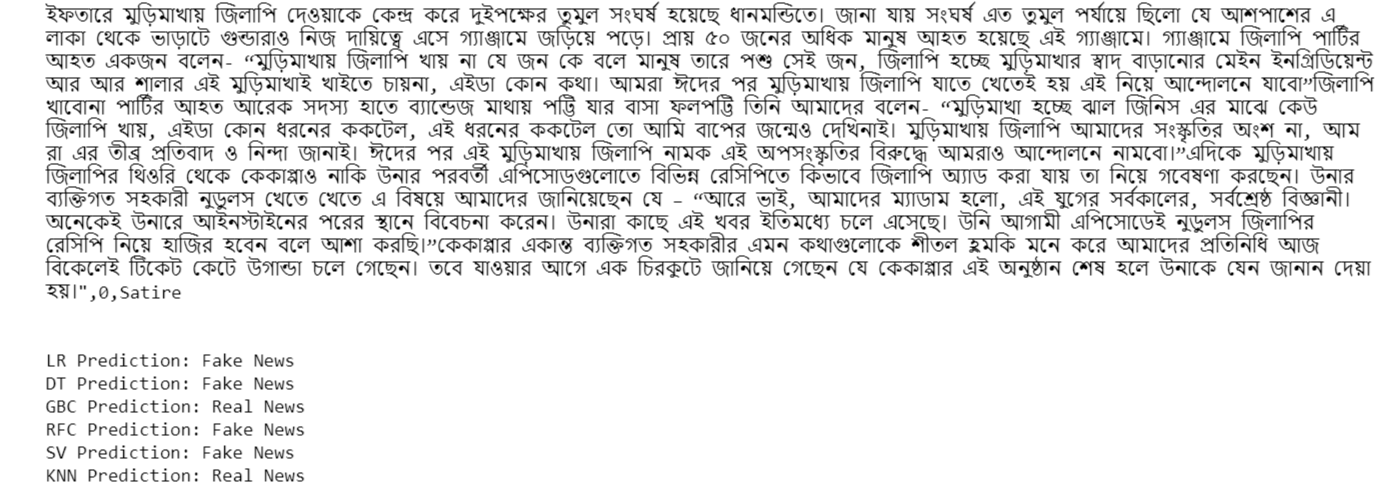
**Test 1:**



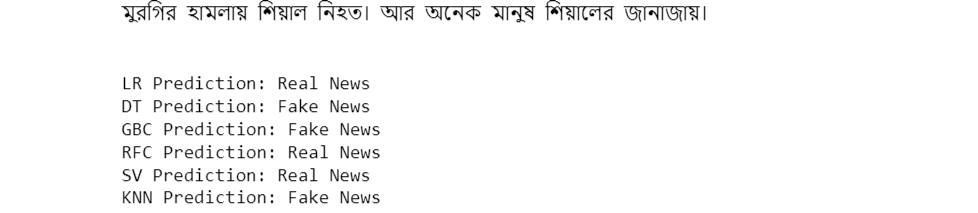
**Test 2:**



**Test 3:**



**Test 4:**



**Discussion**

Since, there are used multiple machine learning algorithms and some of algorithms predict as Real News and some algorithm predict as Fake News. So we may say that, where most of the algorithm say as Real, then we can say that as Real News, otherwise Fake News. But there is a problem to use only one algorithm, which may predict as Real when it is fake, vice-versa. That’s why, we shouldn’t use only one algorithms considering it’s accuracy.

###### **Chapter 6**

###### **Conclusion**

**6.1 Conclusion**

In conclusion, the utilization of multiple machine learning algorithms for detecting fake news represents a significant advancement in our ongoing battle against misinformation and disinformation in the digital age. This multifaceted approach has shown great promise in improving the accuracy and robustness of fake news detection systems. The battle against fake news requires ongoing vigilance, adaptability, and a commitment to ethical considerations, such as privacy and bias mitigation.

The synergy between these algorithms can enhance the overall reliability of fake news detection, as they can cross-validate each other's findings and reduce the risk of false positives or false negatives. This collaborative effort helps strike a balance between precision and recall, ensuring that fewer instances of fake news slip through the cracks while minimizing the risk of flagging legitimate information as false.

**6.2 Limitations**

We work with six algorithms,yet there is some limitations:

* We simply used a dataset of false news in Bengali from various sources. However, no English news data was developed.
* We can't get 100% accuracy while utilising models.
* Sometimes there are shown inappropiate results.

**6.3 Future works**

Our research intended to open new avenues for fake news detection in Bangla. We expect that researchers will be able to apply any of our suggested approaches with confidence in future studies using this dataset for the purpose of detecting false news.

We will try to work with English newspaper in future and will try to improve our project by using more data sets. Predicting false news patterns will be a future priority in order to stop it from spreading online. Through the collection of more data, this will assist individuals in evaluating the veracity of news and enhance accuracy. We also investigated the sensitivity of our model to the parameters, the impact of training data size on model performance, and the winning and losing situations of our models. In addition, we want to test a variety of models, including deep learning models like RNN, LSTM, and GRU.

**REFERENCES**

[1] Raza, S. and Ding, C., 2022. Fake news detection based on news content and social contexts: a transformer-based approach. International Journal of Data Science and Analytics, 13(4), pp.335-362.

[2]Albahar, M., 2021. A hybrid model for fake news detection: Leveraging news content and user comments in fake news. IET Information Security, 15(2), pp.169-177.

[3] Gulzar Hussain, M., Rashidul Hasan, M., Rahman, M., Protim, J. and Al Hasan, S., 2020. Detection of Bangla Fake News using MNB and SVM Classifier. arXiv e-prints, pp.arXiv-2005.

[4] Monti, F., Frasca, F., Eynard, D., Mannion, D. and Bronstein, M.M., 2019. Fake news detection on social media using geometric deep learning. arXiv preprint arXiv:1902.06673.

[6] Zhang, D. and Zadorozhny, V.I., 2020. Fake news detection based on subjective opinions. In Advances in Databases and Information Systems: 24th European Conference, ADBIS 2020, Lyon, France, August 25–27, 2020, Proceedings 24 (pp. 108-121). Springer International Publishing.

[7] Li, C.Y., Kollapally, N.M., Chun, S.A. and Geller, J., 2023. Fake News Detection and Behavioral Analysis: Case of COVID-19 (preprint).

[8] Rasel, R.I., Zihad, A.H., Sultana, N. and Hoque, M.M., 2022, December. Bangla Fake News Detection using Machine Learning, Deep Learning and Transformer Models. In 2022 25th International Conference on Computer and Information Technology (ICCIT) (pp. 959-964). IEEE.

[9] Amjad, M., Vitman, O., Sidorov, G., Zhila, A. and Gelbukh, A., 2023, June. Analysis of Fake News Detection Methods. In Recent Developments and the New Directions of Research, Foundations, and Applications: Selected Papers of the 8th World Conference on Soft Computing, February 03–05, 2022, Baku, Azerbaijan, Vol. II (pp. 131-144). Cham: Springer Nature Switzerland.

[10] Hossain, M.M., Awosaf, Z., Prottoy, M.S.H., Alvy, A.S.M. and Morol, M.K., 2022, October. Approaches for Improving the Performance of Fake News Detection in Bangla: Imbalance Handling and Model Stacking. In Proceedings of International Conference on Fourth Industrial Revolution and Beyond 2021 (pp. 723-734). Singapore: Springer Nature Singapore.

[11] Ahmed Foysal, M.F., Tangim Pasha, S., Abujar, S. and Akhter Hossain, S., 2021. Bengali News Classification Using Long Short-Term Memory. In Emerging Technologies in Data Mining and Information Security: Proceedings of IEMIS 2020, Volume 2 (pp. 329-338). Springer Singapore.

[12] Gulzar Hussain, M., Rashidul Hasan, M., Rahman, M., Protim, J. and Al Hasan, S., 2020. Detection of Bangla Fake News using MNB and SVM Classifier. arXiv e-prints, pp.arXiv-2005.

[13] Shu, K., Mahudeswaran, D., Wang, S., Lee, D. and Liu, H., 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. Big data, 8(3), pp.171-188.

[14] [“KNN Block Diagram”](https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning)

[15] [“Hyperplane in SVM”](https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm)

[16] “[Logistic Regression Curve](https://www.javatpoint.com/logistic-regression-in-machine-learning)”