BANGLA FAKE NEWS DETECTION USING MACHINE LEARNING AND DEEP LEARNING METHODS

BY

Shakib Ahamed Tusher
ID: 181-15-10861

Md.Shameem Alam Shawan
ID: 181-15-11346
And

Mst.Farhana Akter
ID: 181-15-10928

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Narayan Ranjan Chakraborty
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised By

Aniruddha Rakshit
Senior Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

DECEMBER 2022

APPROVAL

This Project/internship titled **"Bangla Fake News Detection Using Machine Learning And Deep Learning Methods"**, submitted by Shakib Ahmed Tusher, ID No: 181-15-10861, Md.Shameem Alam Shawan, ID No:181-15-11346,Mst.Farhana Akter, ID No:181-15-10928 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 4th January 2022.

BOARD OF EXAMINERS

Chairman

Dr. Touhid Bhuiyan
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Dr. Fizar Ahmed
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Nusrat Jahan
Senior Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

External Examiner

Dr. Mohammad Shorif Uddin
Professor
Department of Computer Science and Engineering
Jahangirnagar University

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Narayan Ranjan Chakraborty**, **Assistant Professor Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Narayan Ranjan Chakraborty
Assistant Professor
Department of CSE
Daffodil International University

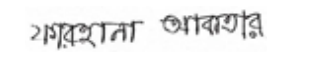
Co-Supervised by:

Aniruddha Rakshit
Senior Lecturer
Department of CSE
Daffodil International University

Submitted by:

Shakib Ahamed Tusher
ID: -181-15-10861
Department of CSE
Daffodil International University

Md.Shameem Alam Shawan
ID: -181-15-11346
Department of CSE
Daffodil International University



Mst.Farhana Akter
ID: -181-15-10928
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project/internship successfully.

We really grateful and wish our profound our indebtedness to **Narayan Ranjan Chakraborty**, **Assistant Professor**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of Natural Language Processing to carry out this project. His endless patience ,scholarly guidance ,continual encouragement , constant and energetic supervision, constructive criticism , valuable advice ,reading many inferior draft and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to Professor Dr. Touhid Bhuiyan, Department Head of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

Fake news detection is as important as cleaning crime from the society. In this digital era spreading violence and manipulating people deception towards an issue become very easy than ever. More and more authentic news are mutilated into fake news for some peoples own mean .In our day to day life we use many online platform for both entertainment and other uses .In those places we are continuously bombarded with fake news. This toxicity needs to be stop for our own good. And to stop the spreading of fake news, detecting it is the most inescapable part. Our research is based on that inescapable part. We tried to create a model for identifying fake news with a very big dataset of Bangla news data. All the data we have collected are from various online sources .Also we have used both machine and deep learning methods for our study. In machine learning method LR, SVM and RF shows 95% of highest accuracy .On the other hand in deep learning method LSTM and BERT give the best accuracy of 93%. Although we have used many methods but few methods we couldn't use for our low performance device. However an application can be develop with this research to identify the news either fake or authentic.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1 Introduction

Fake news can be referred to as a false news or deceiving information presented as a news. Misleading information can damage the reputation of an individual or entity. As the use of social media is growing numerously the amount of false news is also increasing day by day. These days' people use social media for many purposes such as advertisement, politics, Entertainment, News etc. Any type of false news, intentionally or unintentionally can infect the social media platform. Nowadays people follow trends and viral information, they tend to believe what they see but they never inspect or overlook what they are following. This type of scenario can lead us in great danger. By doing this they might harm themselves and others. One false propaganda can destroy a person's career. And false rumors travel faster than anything with the help of social media. Not only social media platforms are getting infected by the fake news but also it is poisoning our society and culture. Nowadays people are so involved with social media that they take their action based on the news presented on any online news portal without judging if the news is fake or true. Because of ongoing pandemic people are more on to online shopping than ever. Product sellers make clickbait intentionally to attract more people to their shop as a result most of the people who are getting fake products are no longer buying anything from online. As a result the reputation of the ecommerce sector is getting hampered. Sometimes politicians spread fake news about their opponent which can create a clash between two parties and again the real victims are the innocent people who have nothing to do with that.

We may not stop the fake news forever but what we can do is stop the spreading of false news as false news tends to spread faster than anything on social media. Humans can detect the fake news if they have already heard the news, sometimes by using common sense we can detect satire news. Moreover if we deep down further investigation we might detect fake news. In the current world people have more to do than time to do, they don't spend their valuable time investigating news if it is true or not. As a result people are getting

affected by the false news. To cope with the situation, what we can do is detect fake news and stop it from spreading. So, our goal is to detect fake news which are presented in online news portals and social media.

A lot of work has been done in this field. With the help of Machine learning and deep learning methods we can create a model which can detect fake news. There are also some websites which can detect fake news. They update the fake news manually by doing logical explanations which is time consuming. Our goal is to detect the fake news automatically and reduce the spreading. Though there are over 250 million Bangla language speaking people in the world, little work has been done in this field using Bangla fake news. According to The Daily Star over two crore and 20 lakh Facebook users remain active in Dhaka, the capital city of Bangladesh this year. Dhaka has been ranked second in terms of having the most active Facebook users in the world according to Global Digital Stats hot of Q2 report of 2017. Most of the fake news is spread through Facebook. So a lot of work needs to be done to stop spreading the fake news. With the help of NLP techniques we have created a model which can detect fake news with the accuracy of 95% in machine learning and 93% in deep learning.

1.2 Motivation

In a busy schedule we never try to seek the evidence of an incident if the incident we heard is true or not. We latch on to anything we hear and reinforce our existing belief. We don't give enough time to challenge the news if it is true or not. Most of the fake news is spreading through online news portals and social media. As the device and internet is getting cheaper day by day, the users of social media are also increasing day by day. People share news without authenticating the source or identifying the news. It is creating a chaotic environment in both online platforms and the physical world. According to News laundry, a recent incident happened in Cumilla, Bangladesh during Durga Puja a hindu festival, a Quran the Islamic holy book was found near the hanuman idol. The news was spreaded all over Bangladesh overnight. Local Muslim demand the puja to be stopped but the Hindu society refused the demand which create a worst communal violence in the history of

Bangladesh. Seven people were killed and more than 450 people were arrested. In this period the media was covering the violence instead of investigating the real cause of the violence. Meanwhile the social media was flooded with photos and videos of the incident. Moreover a blue tick verified Twitter account name “Bangladesh Hindu Unity Council” started posting photos and videos of the attacks. They posted a video and claimed Musilm mob had set fire to a temple in Rangpur. Later on it was found that the video was a fake video of another incident that occurred in Mara Cherra Bazar in Tripura and the twitter account is also a fake account. The whole situation was hit up without verifying the news which caused a violence. There are a lot of incidents like this, occurring day by day because of the spread of fake news. We can use Deep learning and Machine learning techniques and create models which can detect fake news automatically.

Though a lot of work has been done on English news but there is countable work has been done on Bangla fake news detection. According to Wikipedia there are around 230 million native Bengali speakers and 30 million second language speakers which makes Bangla fifth most native speaking language and sixth most spoken language by total number of speakers in the world. But the work done in fake news detection is not enough. We want to contribute more in this field.

1.3 Rational of the study

Though the number of Bengali language speakers in the world is huge, our resource for research using NLP is not rich. Also a little work has been done on fake news detection. Spreading fake news is a common issue in Bangladesh and creating chaotic incidents every day. It is destroying our reputation towards other countries. As people share misleading information unintentionally without verifying. We need to develop a model which can detect fake news automatically. Though a lot of models have been constructed in other languages but for Bangla language it is insufficient. To work on Bangla fake news detection the main obstacle was insufficient dataset which we have overcome by collecting and labeling fake news data in our dataset.

1.4 Research Question

- What type of fake news are there?
- How does Bangla fake news detection work?
- How to process Bangla text data in NLP?
- What are the recent works on Bangla fake news detection?
- What are our future works?

1.5 Expected Output

As we discussed earlier in this chapter, most of the time people can't distinguish between true and fake news without further investigation. Also they intend to believe what they see. It is challenging to stop the spreading of fake news. What we can do is we can make a Model using NLP technique which will detect the fake news automatically. By this we can limit the fake news. There is some previous work which covered this topic. We have increased the amount of data in our dataset and we expect that we will get a better output.

1.6 Report Layout

This report contains 5 chapters. The first chapter starts with the introduction and finishes with the report layout. The next chapter contains the talked about the background knowledge of the thesis. It starts with the introduction and ends with challenges we faced during our exploration. We have talked about the main research materials, all those methods and classifiers for our model on the third chapter. It also starts with introduction and end with the Implementation Requirements. Furthermore all the experimental result and the discussion around the result was done 4th chapter. Moreover all the impact on society, environment and sustainability was mentioned in chapter five. The final chapter contains our whole thesis summary and our research conclusions with the implication for the further study.

CHAPTER 2

BACKGROUND

2.1 Introduction

This era of digital communication spreading of fake news are more organized than ever .This days many standard news Medias on online stretch deception according to their means. Because of our lake of morality and downfall of ethics we don't actually care about the consequence, a fake news can possess to our society. That's the reason that when a fake news site is ban another site which also outreach fake news took the spot. Beside this one news from certain destination can become vague from sharing and re-sharing induvial posts again and again. Moreover sometime it become difficult for us to trace down the authenticity of a news if it shared by our close companion. Also we play blind eye when we consume news from online media which is very easy and facile work for those sites to provide spurious news instead of veritable news .And this has a very serious outcome regarding toward issues like political, occasions and breaking news.

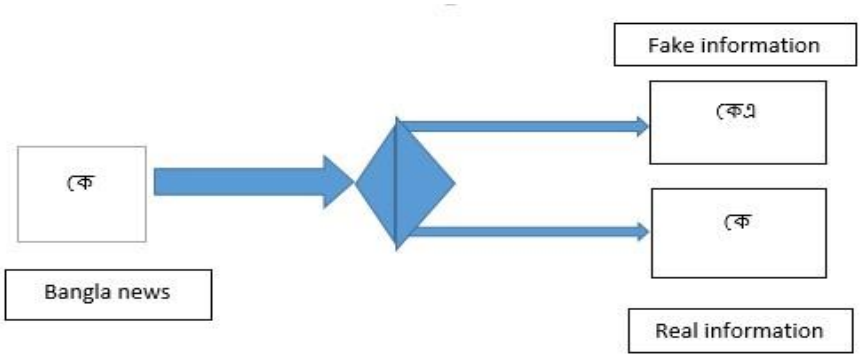


Figure 2.1: Bangla fake news detection

2.2 Related Work

Fake news can lead us in a great danger. Though we are blessed with our modern asset it can mislead us with many ways. As people are getting more involve with the online activities it is becoming easier to reach more people in a second. People shares lot of thing in the social media nowadays without any authentication. Which can affect the both online and physical world. Sometimes people take action according to what they see on online. So the fake news spreading should be stopped. To do that what we can do is, detect those fake news automatically can't stop it from spreading. With the help of AI we can build a model which can detect fake news automatically. A lot of work has been done in this field but there are few work has been done with the Bangla language. Our work is also inspired by some of previous work like [1], used linguistic features and neural network based methods to detect fake news. 8.5k data was manually labelled and the best result was achieve with the SVM. It scored 91 F1-score. Hussain et al. [2], used two supervised machine learning algorithms, MNB and SVC classifiers to detect Bangla fake news with CountVectorizer and Term Frequency - Inverse Document Frequency Vectorizer as feature extraction. In this work 2.5k news were taken as a dataset. Among these classification 96% accuracy gained for Linier kernel SVM. [3], proposed a model to detect fake news in Bangla. In this work used Passive Aggressive Classifier, Multinomial Naive Bayes, Support Vector Machine, Logistic Regression, Decision Tree Classifier, and Random Forest Classifier. The dataset was made with 3.5k authentic data and 2.3k fake data moreover among these classifier Passive Aggressive Classifier and Support Vector Machine achieves 93.8% and 93.5% accuracy respectively. Sharif et al. [4], proposed a Machine Learning (ML)-based classification model to classify Bengali text into non-suspicious and suspicious categories based on its original contents. The dataset was made with 7000 Bengali text documents and the highest accuracy was gained for the SGD classifier 'tf-idf 'with the combination of unigram and bigram features which is 84.57%. George et al. [5], used LSTM and CNN methods to develop deep learning models which can detect fake news from Bangla sentences. In this work compiled a data set from websites which contains about 50k of news. This model gained highest 78% accuracy. .Balo et al.

[6], built a model based on Bag of word, tfidf matrix and Random Forest Classifier for test data whatever this news data real or fake. In this work 500 news were taken as a dataset. This model achieves 86% accuracy. Adib et al. [7], used deep hybrid learning models combining CNN and Machine Learning classifiers. In this model gained highest 99% accuracy. Mahabub et al. [8], proposed a Machine Learning (ML)-based classification model among these classifier select only best three classifier for detect fake news. . In this work 6.5k news were taken as a dataset and best accuracy gained 94%.

2.3 Research Summary

In our research, we used both deep learning and machine learning methods for Bangla fake news detection. We used various deep learning models. For using the model we have collected a vast amount of dataset manually. Several news and social media sites are also in the dataset. Beginning of the data collection we took Bangla news articles form different media sites and make a summary of every induvial information. Therefore there are 8 distinct section in the dataset like domain, date, article id, source, relation, headline, content and label. More than fifty thousand data we collected in total. We preprocess the whole Bangla dataset before any kind of deep learning method we used.in the preprocessing section we try to identify the noises and remove them. After that clearance of unwanted thing we used count Vectorizer which is used to change the text in each word of the whole content into vector based frequency .after that we trained our model several time which conclude with a very desirable outcome.

2.4 Scope of the problem

Although Detection of fake news in NLP is widely common but for Bengali NLP is in development stage. Our research works on utilizing machine and deep learning techniques for identification. There was a huge number of noise was in our dataset. That was a very big issues we had to solve because we weren’t getting desired result for that.in the other hand there was a big crisis of fake news in our data. We tried to full fill it by grabbing more fake data but that wasn’t enough.to counter this problem we mixed few authentic data. For

this research we used deep learning and machine learning methods and get a very delightful result.

2.5 Challenges

To start with this topic is the biggest challenges we faced so far .it's not with the procedure it's the resources is main back draw .there aren't very much work done with Bangla language so there aren't much dataset to work with. We had to collect manually data from a vast amount of sources .to keep the validity and authentic of the data is another kind of challenge we faced. Beside this all the data aren't in a proper manner and structure. Then it comes the labeling part where we had to extra conscious. In the preprocessing we had made a fresh coding step to cop up the model .it wasn't an easy part for running the dataset .there were several breakdown in the device. Along with it Bangla dataset as a string was a problem. There are a lot of work in English language and the dataset is also well mannered in the contrary Bangla language is in primary stage .so that's a problem for us because there aren't much supervised materials for help so we had to do our own. Finally collecting fake news was another major issue .we aimed for as much as fake data we can collect because it favor result to become more precise.

CHAPTER 3
RESEARCH METHODOLOGY

3.1 Introduction

To accomplish our goal we have to go with different procedures. In this chapter, we will explore those procedures. To achieve the best result we have tried different approaches of both Machine learning and deep learning algorithms. We have trained our model with different classifiers such as Multinomial Naive Bayes, K-Nearest Neighbor, Support Vector Machines, and Recurrent Neural Network, LSTM, BERT etc. Before that, we have preprocessed our data because most of the data in the dataset was noisy. Moreover, we are working with NLP and a good dataset is the most important thing in research. To fulfil that need we have collected a dataset online but it wasn't sufficient. So we have collected more data manually to make a decent dataset. In this chapter, we have described all the classifiers we used. Our whole work can be easily visualized by this figure given below.

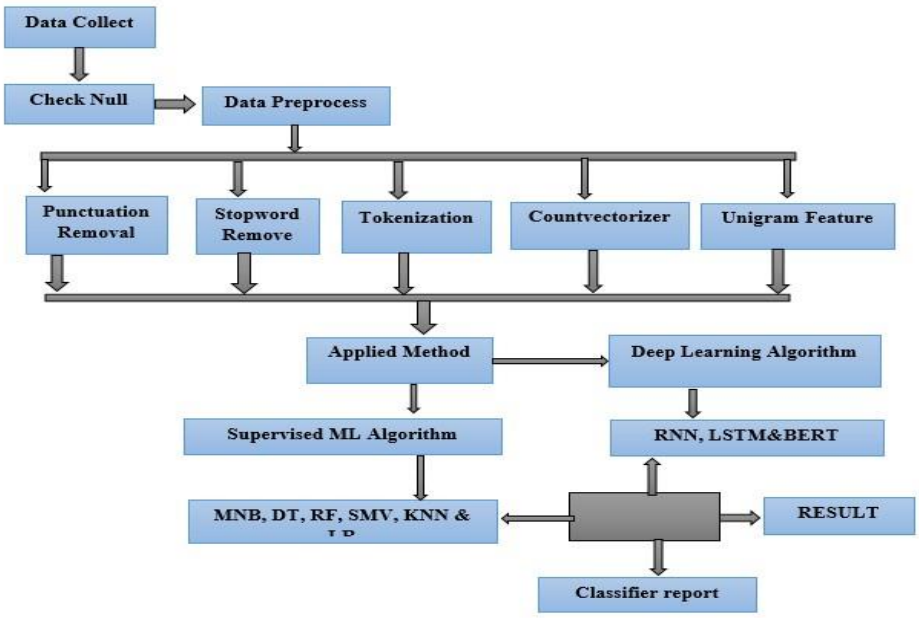


Figure 3.1: Workflow of our model

3.2 Research Subject and Instrumentation

The work wasn't short, since we have used both Machine learning and deep learning classifiers. Sometimes we have to run our program rapidly to achieve the best result which can be time-consuming with poor hardware. To speed up the work, we have used the Hardware and software given below.

Hardware and Software:

- Intel core i5 processor.
- 8GB RAM
- 256GB SSD
- Google Colab

More Tools:

- Windows 10
- Python 3.7
- Pandas
- NumPy

3.3 Data collection & preprocessing

Data is the most important asset of a research work. Data collection is the fundamental and most important step for research. And the main objective of data collection is making a sure that the information we are collecting is reliable. A rich dataset can help to gain more accurate result. As the Dataset is main part of a research we have make sure that we have a decent dataset before we start our work. We are working on Bangla fake news detection with NLP. Scarcity of Bangla dataset made us collect data manually. Though we got a dataset from online but that was not enough. We have a dataset which contains both Bangla

authentic news and Bangla fake news as data. The dataset contains 60K data where 55894 authentic data and the number of fake data 4106. The existing dataset had less data, so we decided to collect some data manually from online. Most of the data was collected from online news media and social media.

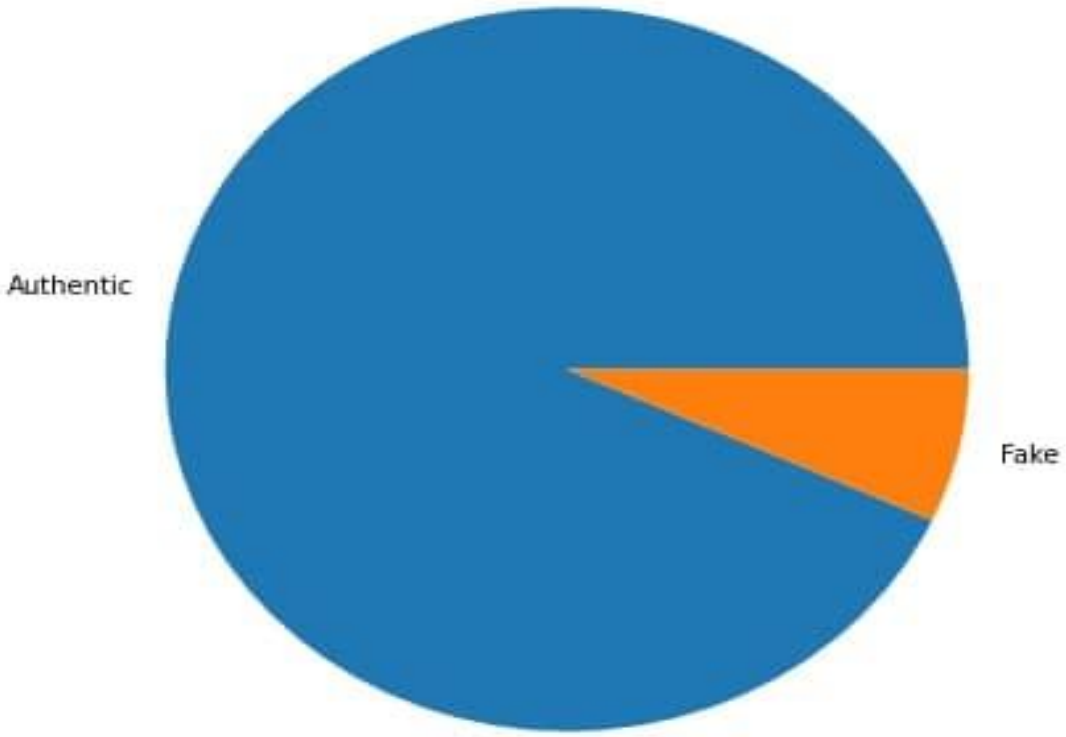


Figure 3.2: Ratio of fake and authentic data

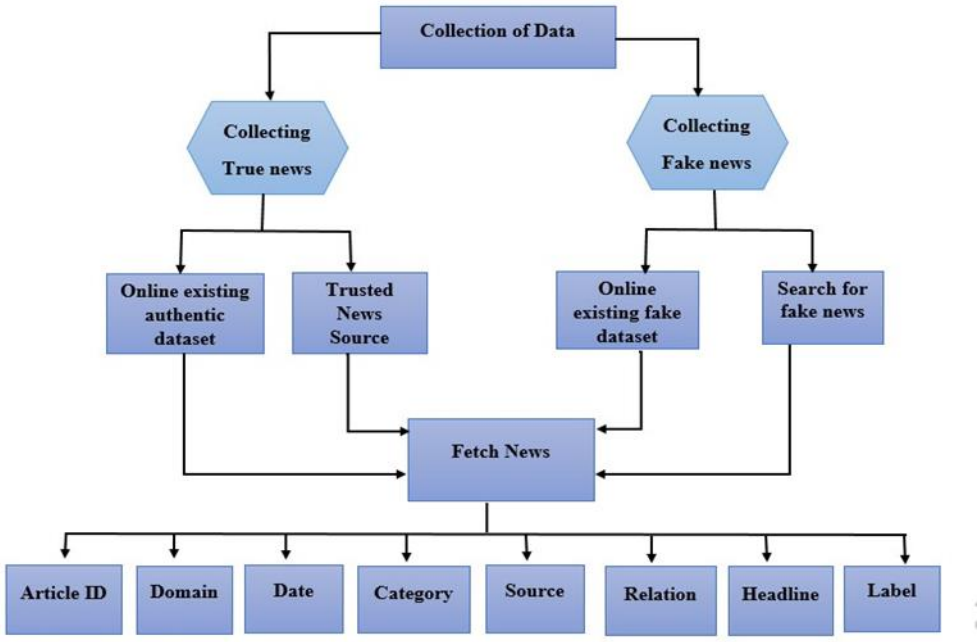


Figure 3.3: Data collection workflow

After creating a proper dataset we have started our work. At the beginning we have preprocessed our dataset. Preprocessing is manipulation of data which enhance the performance of a model. As most of the dataset has noisy data, it is important to preprocess a dataset to create a model easily. We have preprocess our dataset with these steps given below.

Punctuation removal: Like English language Bengali language also uses many punctuation character in text. But this punctuation character is useless for our dataset .so, we removed this character from the data by using libraries.

Stopword removal: In any natural language stop words are the most general words .For the aim of analyzing text data and building NLP models, these stop words might not add much value to the meaning of the document. Stop words are basically articles, prepositions, conjunction and certain pronouns for example ‘on’, ’a’, ’the’, ’an’, ’the’, ’but’ etc. Like

that, Bengali Language has some stop words and we need to discard from the dataset technically for decrease the processing time.

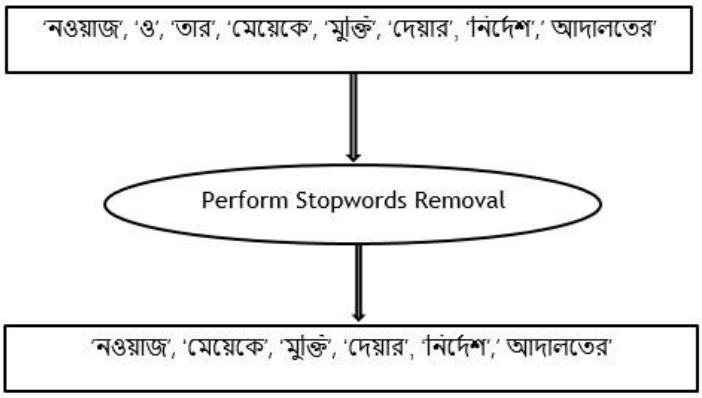


Figure 3.4: Example of stopwords removal

Tokenization: Tokenization of data means that we have to split sentence, phrase, paragraph or an entire text document into words or tokens. Tokenization is mandatory in preprocessing, For Tokenization we need to filter essential words.

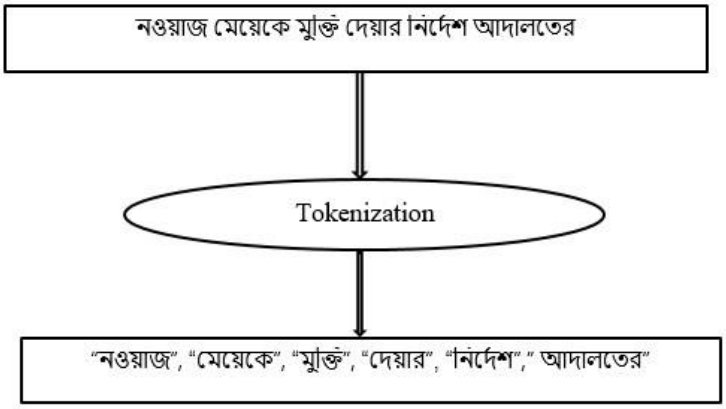


Figure 3.5: Example of tokenization

CountVectorizer: CountVectorization is used to transform a given text into a vector on the basis of the frequency of each word that occurs in the entire text. Machines cannot

understand characters and words. So when dealing with text data we need to represent it in numbers. Countvectorizer is a method to convert text to numerical data.

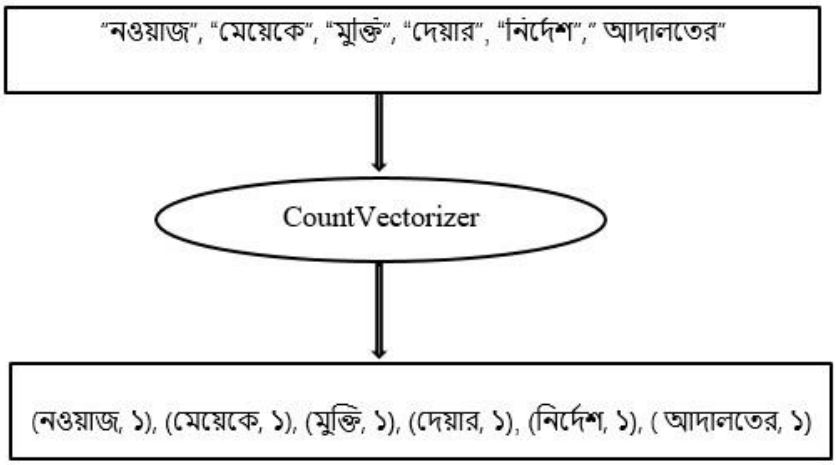


Figure 3.6: Example of countvectorizer

Unigram feature: For feature extraction, we need to apply the character N-gram tokenizer which tokenizes the input sentence into n-grams such as word unigram, word bigram, word Trigram etc. when n=1, 2, 3…….it's called unigram, bigram, trigram... Respectively. In our work, we apply the word unigram to tokenize the given input sentence.

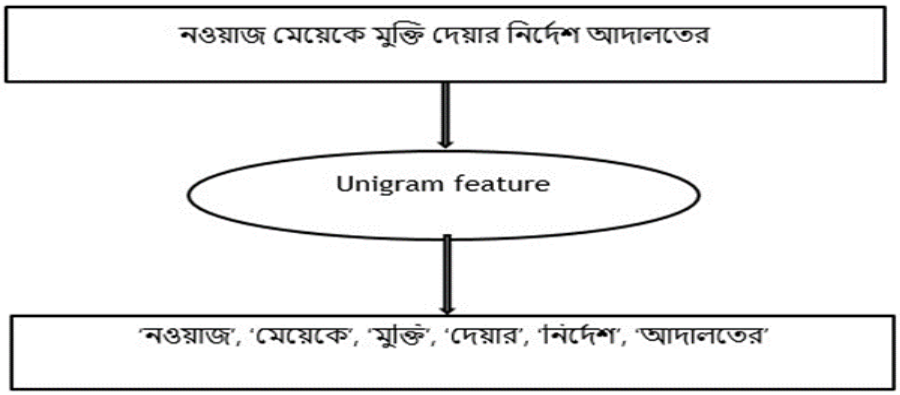


Figure 3.7: Example of unigram feature

3.4 Statistical Analysis

- In our dataset in total there are 60k complete text data. From a number of online media and sites. A very small dataset model of our dataset is given below.

Table 3.1: Sample of dataset

Article ID	Domain	Date	Category	Source	Relation	Headline	Label
1	https://motikontho.wordpress.com	7/25/2016	politics	Reporter	Unrelated	সরকার বিএনপি শাখাকে যৌন হয়রানী করছে: ফখা,	Fake
2	https://www.earki.com	জুন ২৯, ২০২১	Sports	Reporter	Unrelated	ব্যারিস্টার সুমনের কাছে পেনাল্টি শিখতে চান এমবাপ্পে,	Fake
3	https://www.earki.com	১৭:২৬, মে ০৫, ২০২১	miscellaneous	Reporter	Unrelated	সুদমুক্ত 'তরমুজ লোন' দিচ্ছে দেশের কয়েকটি শীর্ষস্থানীয় ব্যাংক,	Fake
4	jagonews24.com	9/19/2018 5:48:18 PM	Education	Reporter	Related	হট্টগোল করায় বাকৃবিতে দুইজন বরখাস্ত, ৬ জনকে শোকজ	Authentic
5	jagonews24.com	9/20/2018 1:00:36 PM	International	Reporter	Related	ইমরানের জন্য খোলা হলো কাবা শরীফের দরজা	Authentic

- The total number of columns is 8
- Authentic data number around 55k.and more than 4k fake data in total.
- Dataset was created in excel format which augmentation is .xlsx.

3.5 Implementation Requirements

We have used split method in using classification and regression models in the research. For our research we used several algorithms in both machine leaning and deep learning. Those algorithms are given below.

Split method

This method is trend to separate the dataset in subset for train set. We split our dataset in 70/30where one subset contains 70% of data and the other one contains 30% .this process takes data randomly. Moreover we only used 2 columns (headline& label) for this research. Label column was divided in two category (fake and authentic).

Machine learning algorithms

Multinomial Naïve Bayes: to classify a new document multinomial naïve bayes is used. It is a supervised learning method where it works with a set of predefine classes. When a new text add it try to predict the new text class with classes previously assigned. This method is widely used in Natural Language Processing (NLP).in mathematically it's highly efficient and also very facile in implementation. Thomas Bayes was the one who formulated the Bayes theorem, where it ascertain the probability of a situation regarding from previous knowledge of the related condition to the situation. The following formula is that it based on.

$$A\ (A|B) = P\ (A|B)/P\ (B)\ldots\ldots\ldots\ (i)$$

Here we are calculating the class probability where predictor B is already provided.

$$P\ (B) = \text{prior probability of B}$$

P (A) = prior probability of A

P (B|A) = Situation of predictor B given class A probability. This formula helps to estimate the probability of the tags in the document. But if the given word was never appear in any classes in the training data then it will be estimate zero by the frequency-based probability. Because the estimate by probability directly proportional to the number appears of a feature's value.it will also collision with the other values because it will multiplied with those and fetch zero in the answer .to avoid this problem there shouldn't be any probability of exact zero in the training data.

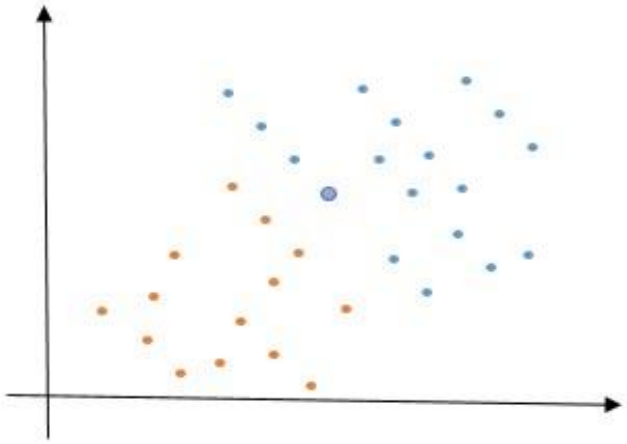


Figure 3.8: Multinomial naïve bayes

Logistic Regression: Logistic regression is a classification algorithm which became an integral part of any data analysis. Because of the uses where it can be appoint in observation for a discrete set of classes. Outcome for the logistic regression is also discrete where the values can be binary, yes or no etc. Instead of 0 and 1 exact values, it gives near result. To accomplish that sigmoid function is use by it to map the predicted value into probabilities.

Threshold value concept is use where threshold value stay below and above from 0 and 1 respectively. Moreover it has an upper hand from liner regression which is classifying given data using continuous and discrete dataset.

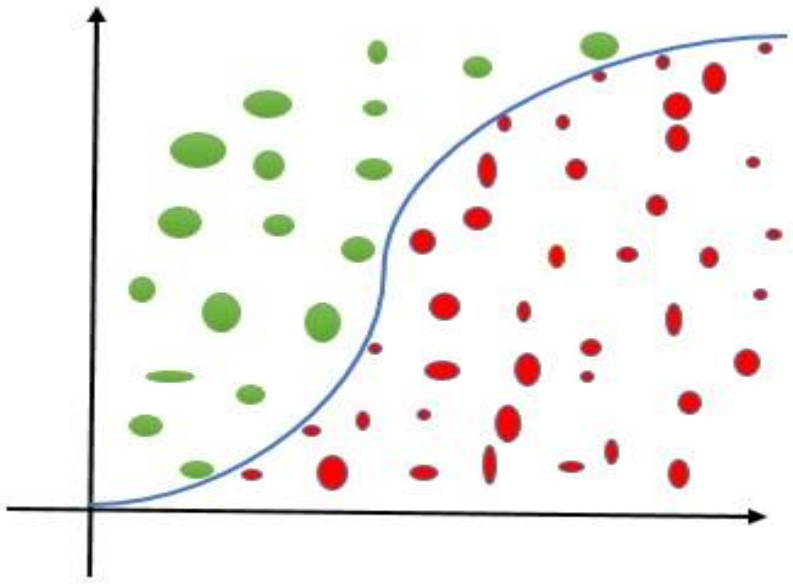


Figure 3.9: Logistic regression

K-Nearest Neighbors: K-Nearest Neighbor (KNN) is a supervised learning algorithm where based on the KNN category, the result of new illustrate query is classified. KNN algorithm infer the possible resemblances among the train data and the data will be given. And facily can put the given data in a class where the given data resemblance the most of train data class. Value of K is the most important part to implement this algorithm. For that reason at the beginning we have to choose the value K.to do that cross-validation can be implement. After that distance between data point will be count. When the distance will be measure will able to see the nearest neighbors of the given data.

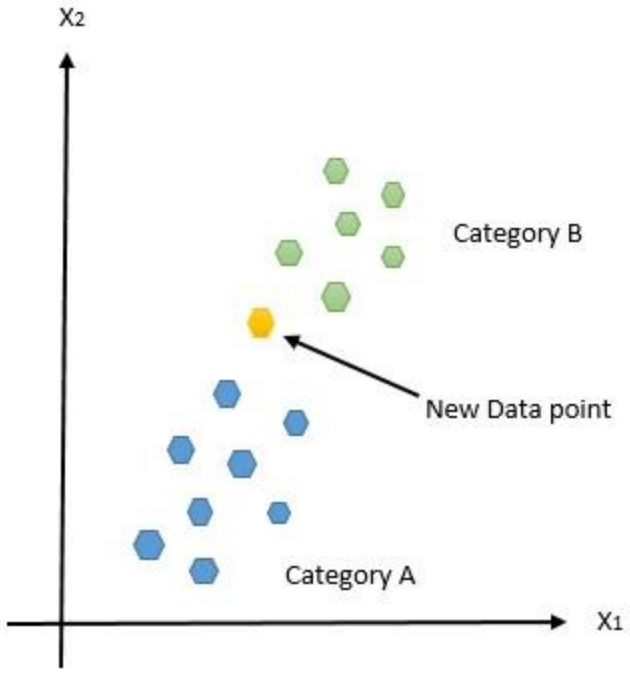


Figure 3.10: K-nearest neighbor

Decision Tree Classifier: Decision Tree is a supervised learning technique. It can be used to solve both classification and regression problems. It works like a flow chart. It is a tree-structured classifier where it has a decision node and leaf node. Decision nodes are used to make the decision and leaf nodes are the output of those decisions. The decision tree mimics human thinking where it asks a question and depending on the yes-no answer it continues its path.

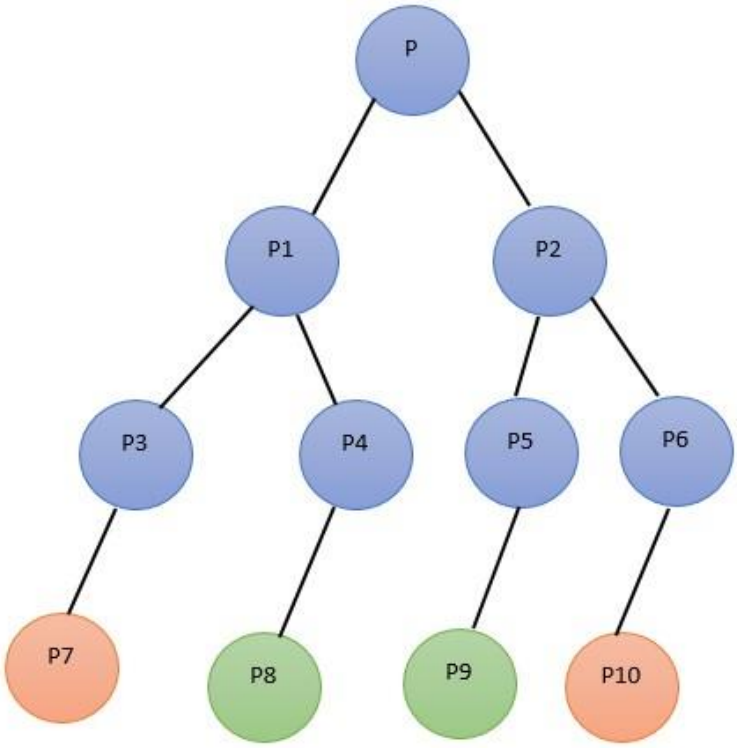


Figure 3.11: Decision tree

Random Forest Classifier: Random forest contains a large number of decision trees. The decision tree is the building block of the random forest classifier. The result of a random Forest is the collective result of all decision trees. It cheeks the prediction of all individual trees and finds the most voted prediction which becomes the prediction of the whole model. The decision tree predicts the result which depends on a set of rules whereas random forest predicts the result based on features chosen randomly. Randomized feature selection of Random forest brings a much accurate result than decision tree classifier.

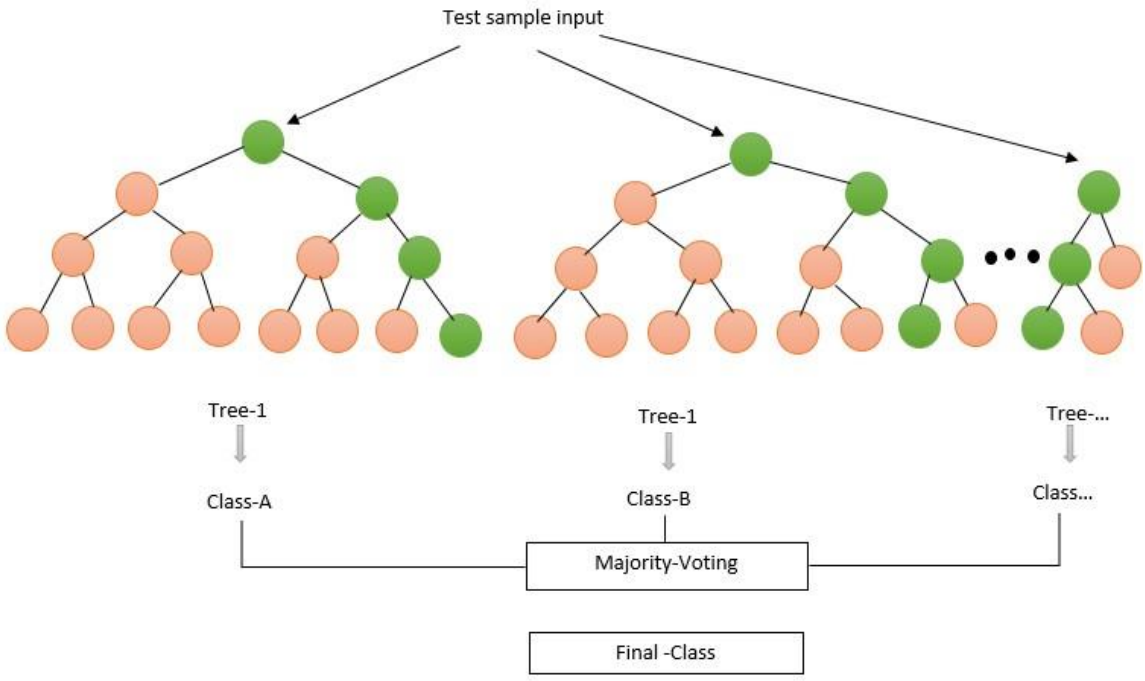


Figure 3.12: Random forest classifier

Support Vector Machine Classifier: Support Vector Machine is a supervised learning technique. It is used for classification, regression and outlier's detection. In terms of text classification Support vector machines considered the best classification technique. It can perform both linear and non-linear classification. The classifier finds the best hyperlink between two classes of data through a line in the middle called the decision line. When a cluster of the same category of data is near the line then it will classify the category as a result. When both categories are so close to one another then it is not possible to classify linearly. At this point, non-linear classification will be the best approach to find the result.

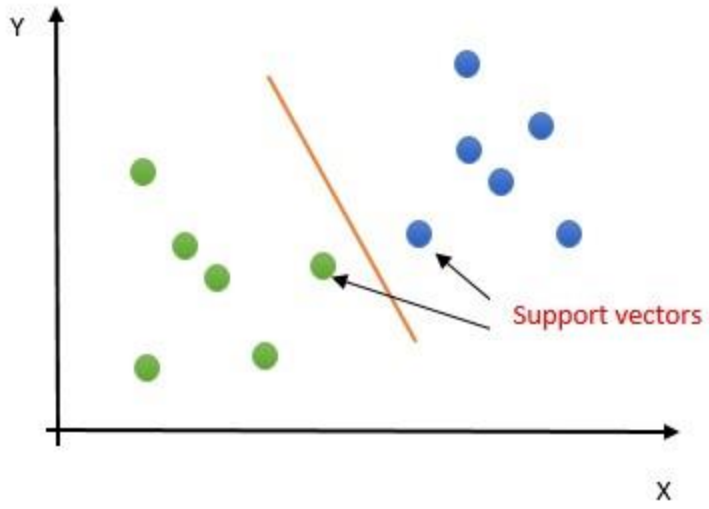


Figure 3.13: Support vector machine

Deep learning algorithms

RNN: Most of the data in the current world are in a sequential manner. A recurrent neural network uses sequential data and predicts the next data. This deep learning algorithm uses previous memory and new input to calculate the output. It has an input layer, a hidden layer and an output layer. Unlike other neural networks, the inputs and outputs are not independent [9].

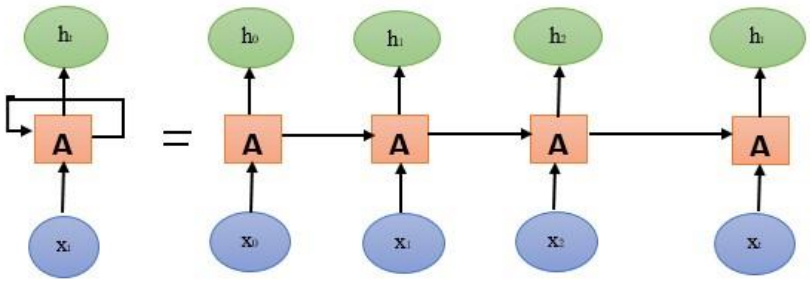


Figure 3.14: Recurrent neural network

BERT: Bidirectional Encoder Representation from Transformers is a transformer-based machine learning technique for NLP. BERT pertained on two task language modelling and next sentence prediction. BERT learns contextual embedding's for words in the training process. Pre training can be a resource-hungry task. In this case, BERT can work on fewer resources on a smaller dataset and optimize the performance on a specific task.

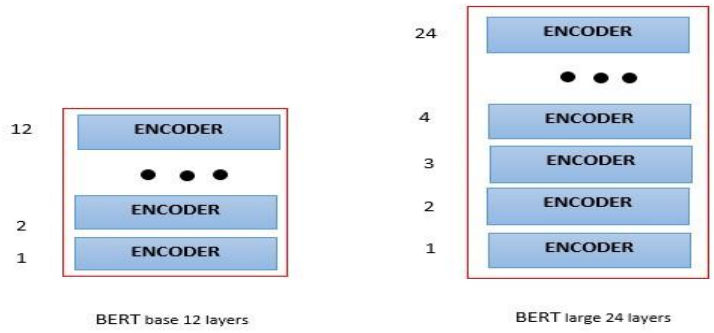


Figure 3.15: Bidirectional encoder representation transformers

Long-short Term Memory: long –short term memory has a great reputation in turning sequential information in efficient structure. LSTM which was introduced by Hochreiter and Schmidhuber in 1997 become extensively used model in the field of text classification and problem generation. LSTM have proven excellent performance by extracting information from both directions text. Beside this attention mechanism of this method can be seen as a very high performable pooling technique for classification task.

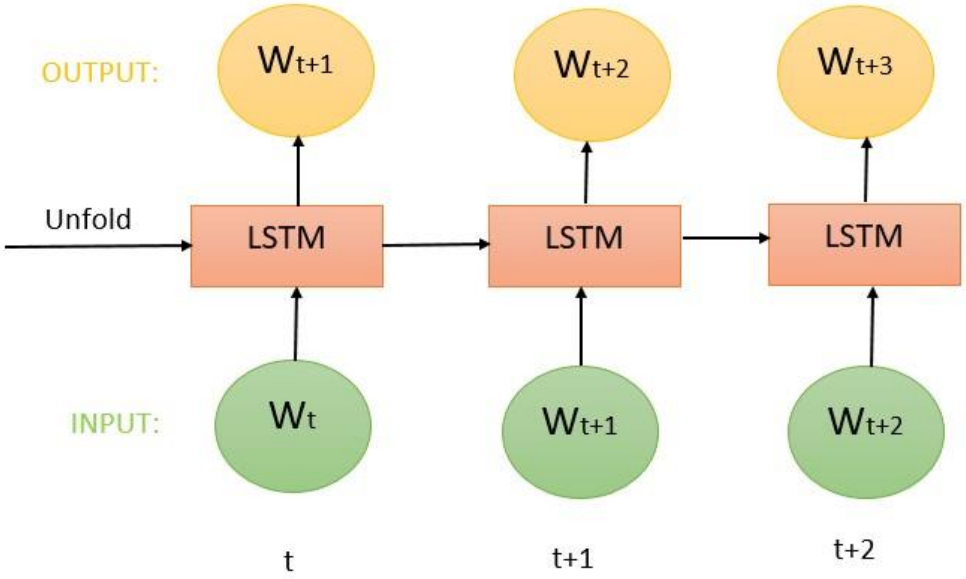


Figure 3.16: long –short term memory

CHAPTER 4
EXPERIMENTAL RESULT AND DISCUSSION

4.1 Experimental Setup

For getting the best result we all know an all-around model is essential piece. Highly capable computer needs to make ready the desire model for fast result.in this research we worked with a very big dataset .it was little complicated for us .But it was a great scope for us to work in Google Colab. Colab is a very popular and widely usable research kit. For dataset training the time is very quick. Although it has so avail but we face some noticeable problem. Because of our vast quantities of data it worked quite slowly. With the spilt method we have train our model. We executed our model with 40 epoch and 42 batch size where we split our dataset in 70/30.

4.2 Experimental Results & Analysis

Mistake is a common phenomenon for a human mind. On the contrary accurate and genuine result is the basic distinction for a machine from human. We get very moderate outcome from our prepared model. But for few the case aren't same. However, we have used different machine learning algorithms and deep learning algorithms for the best accuracy. From machine learning algorithms we get an excellent accuracy from Logistic Regression, Support Vector Machine and Random Forest. On the other hand from the deep learning algorithm Long-short Term Memory and BERT shows great result. Our research utilization of various models and the accuracy we gained from using those are given below.

Table 4.1: Accuracy for machine learning models

Model	Accuracy
Multinomial Naïve Bayes	94%
Logistic Regression	95%
K-Nearest Neighbors	93%
Decision Tree	93%
Support Vector Machine(SVM)	95%
Random Forest	95%

There are three models from where we get the highest accuracy **95%** .Those are respectively **Logistic Regression, Support Vector Machine and Random Forest.**

Table 4.2: Accuracy for deep learning models

Model	Accuracy
Long-short Term Memory	93%
Recurrent Neural Network	77%
BERT	93%

From both **Long-short Term Memory** and **BERT** we get our highest accuracy which is **93%.**

4.3 Discussion

With the spreading of digital media all-around of glob fake news became a major concern for the society to deal with. It's very hard for us to left way the fake news from our daily life .Detection of fake news is very difficult. It's not much difficult for a machine to identify the fake news .But in the same time it's very hard for us to train the machine with proper

material. such as if there are many noise and error in the data we are taking from various online sites then there will be a huge intervening distance in accuracy prediction .to avoid such catastrophe first we preprocessed our data very consciously by removing all error and missing value from it .then several time in both noisy and noise free environment we execute our model and found very hopeful outcome. With the result we feel so confident about our work very enduring and also very promising for future development.

CHAPTER 5

IMPACT ON SOCITY, ENVIRONMENT AND SUSTAINABILITY

5.1 Impact on Society

Fake news is already causing a bad effect in the society. It is polluting the online platform and also creating violence in the physical world. With our work we cannot stop the fake news spread but we can limit the spread. Which will help people aware of a fake news instantly. As the social media is becoming the part and parcel of our life people are taking action according to what they see in social media. A fake news can easily mislead a person which can lead a person in a great danger. On the big picture it is affecting our society directly. By automated fake news detection, we can reduce the problem.

5.2 Impact on Environment

Spread of fake news may affect our Environment indirectly. With our work we can make good impact on environment.

5.3 Ethical Aspect:

Though our work is inspired from some previous work we have tried to fulfil others work deficit. We also have collected more data manually and added to the existing dataset.

5.4 Sustainability Plan:

We have used both Machine learning and deep learning classifier and much bigger dataset to find the best result which make sure sustainability of our work. We have reviewed some previous work of same kind and tried to find the deficit and we also fulfil those.

CHAPTER 6

SUMMARY, CONCLUSION, RECOMENDATION AND IMPLICATION FOR FUTURE RESEARCH

6.1 Summary of the Study

Our main objective based on Bengali NLP can be seen in our whole research work. We have worked with both Machine learning and deep learning algorithm for identifying Bangla fake news. It's a great breakthrough in detecting fake news .nearly two month time we took to finish our research. Our entire research workflow are shown below step by step.

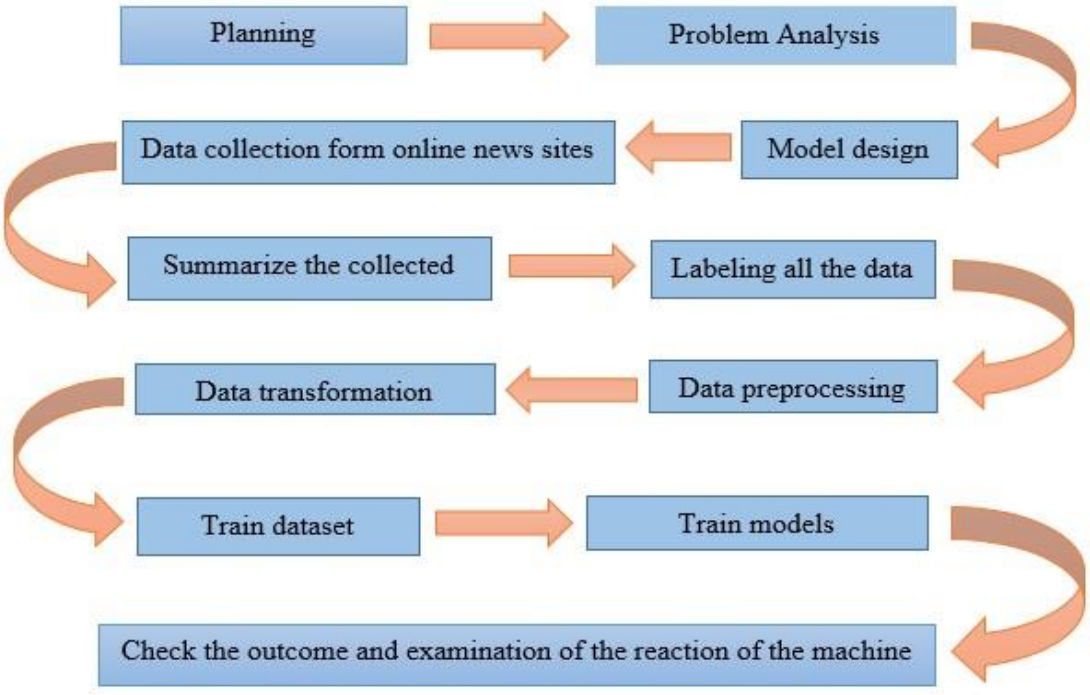


Figure 6.1: Workflow diagram

We can assume that our research will help in Bengali NLP exploration .Moreover in Bengali fake news detection our work can be a helpful material in creating programmable detector.

6.2 Conclusions

In our research in Bengali NLP which is about fake news detection using both machine learning and deep learning .algorithm shows great result. We have collected more than 60k data from various online media and news sites. Our goal was to detect a news whether it's fake or authentic. We are able to dig out very impressive outcome from the dataset. Machine learning algorithm such as Logistic Regression, Support Vector Machine and Random Forest shows accuracy around 95%.Where Long-short Term Memory and BERT for deep learning algorithm shows an impressive mark of 93%. Although our final result are very promising but despite of our painstaking work there are always some flaws. However, after all odds and faults we achieved our desire outcome that our model can detect Bengali fake news with a great ease.

6.3 Implication for Further Study

Fake news are not only harmful for individual but also for our social life too. It's like a cancer for us where it totally deform our judgment. To get rid of such an issue first we need to identify the fake news. As few numbers of work have been done regarding Bangla fake news detection it's an appealing sector where more works need to be done. There was some issues around dataset quantity .we want to increase our dataset more. Beside this we want to use more deep learning algorithm like CNN, Bi-LSTM. Also we want to showcase the difference of fake news detection between English and Bangla language. Furthermore an application like web and portable application can be develop with more time. And this can give us a perfect result of an input which is either fake or authentic.

REFERENCE

[1] 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, 19 Apr 2020.

[2] M. G. Hussain, M. R. Hasan, M. Rahman, J. Protim and S. A. Hasan, "Detection of Bangla Fake News using MNB and SVM Classifier," 2020 International Conference on Computing, Electronics & Communications Engineering (iCCECE), pp. 81-85, 2020.

[3] S. T. Uddin, M. R. Shahriar, F. Rizon, R. A. Polock and S. . I. Shameem, "FakeDetect: Bangla fake news detection model based on different machine learning classifiers," Doctoral dissertation, Brac University, 2021.

[4] O. Sharif, M. . M. Hoque, A. S. M. Kayes, R. Nowrozy and I. H. Sarker, "Detecting Suspicious Texts Using Machine Learning Techniques," Applied Sciences, vol. 10 , no. 18, p. 6527, 2020.

[5] M. Z. H. George, N. Hossain, M. R. Bhuiyan, A. K. M. Masum and S. Abujar, " Bangla Fake News Detection Based On Multichannel Combined CNN-LSTM," 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), pp. 1-5, July 2021.

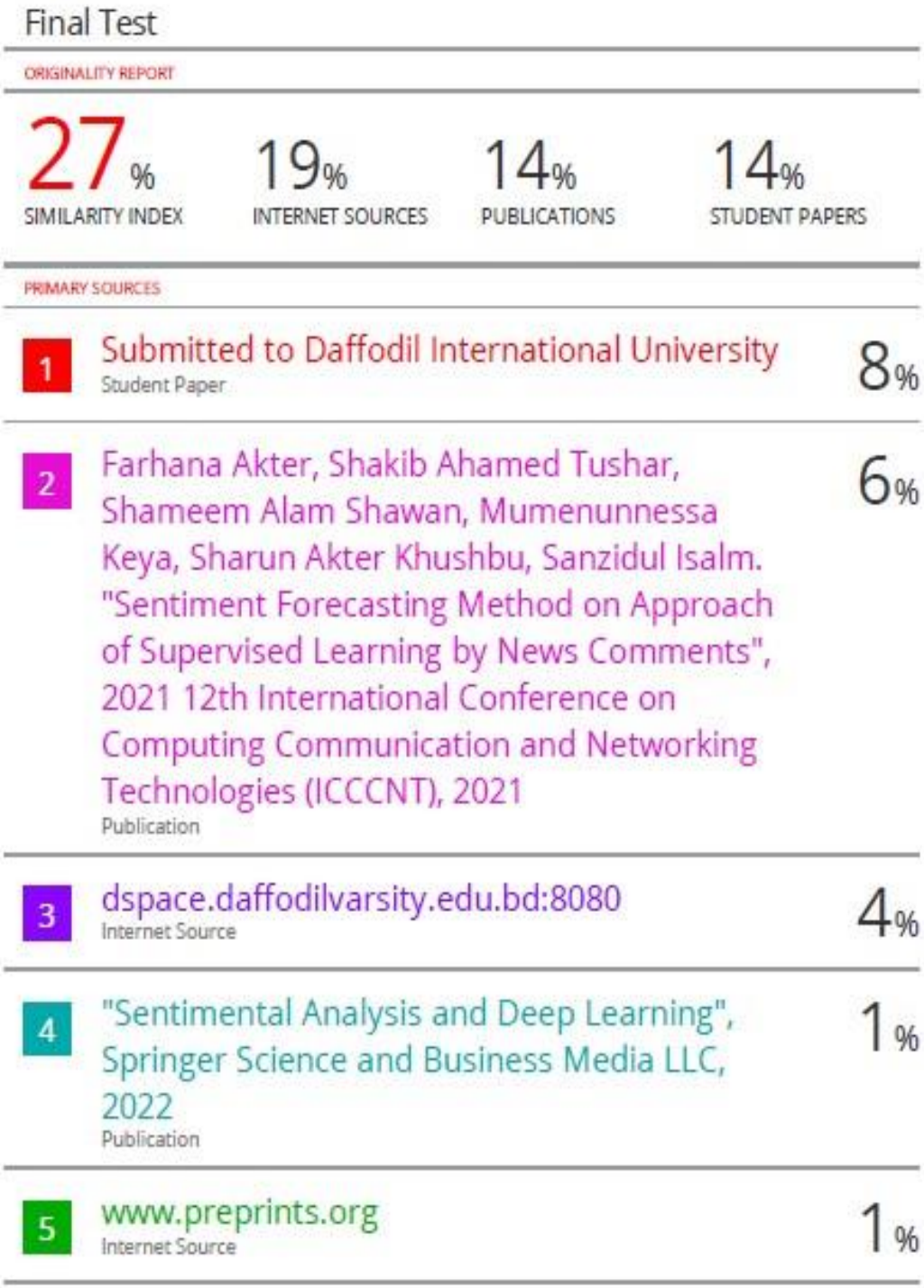
[6] A. Balo, J. Islam and A. . A. Baki , "Bengali Fake News Detection Using Machine Learning," 2019.

[7] Q. . A. R. Adib, M. H. K. Mehedi, M. S. Sakib, K. . K. Patwary, M. S. Hossain and A. . A. Rasel, "A Deep Hybrid Learning Approach to Detect Bangla Fake News," 2021 5th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), pp. 442-447, 2021.

[8] A. Mahabub, "A robust technique of fake news detection using Ensemble Voting Classifier and comparison with other classifiers," SN Applied Sciences volume 2, pp. 1-9, 2020.

[9] F. Akter, S. . A. Tushar, S. A. Shawan, M. Keya, S. A. Khushbu and S. Isalm, "Sentiment Forecasting Method on Approach of Supervised Learning by News Comments," 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), pp. 1-7, July 2021

PLAGIARISM REPORT



Final Test

ORIGINALITY REPORT

27%	19%	14%	14%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	8%
2	Farhana Akter, Shakib Ahamed Tushar, Shameem Alam Shawan, Mumenunnessa Keya, Sharun Akter Khushbu, Sanzidul Isalm. "Sentiment Forecasting Method on Approach of Supervised Learning by News Comments", 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2021 Publication	6%
3	dspace.daffodilvarsity.edu.bd:8080 Internet Source	4%
4	"Sentimental Analysis and Deep Learning", Springer Science and Business Media LLC, 2022 Publication	1%
5	www.preprints.org Internet Source	1%