“ BANGLA TEXT CLASSIFICATION THROUGH MACHINE LEARNING ALGORITHM”

A Dissertation Submitted

in Fulfilment of the Requirements for the Degree of

Bachelor of Science (B.Sc.)

in

Computer Science and Engineering (CSE)

by

ASHRAF UDDIN TAREK ,

SATIRTHA CHOWDHURY, MOSHARAF HOSSAIN

(Metric ID: C191011, C191020, C183032)

to

Faculty of Science and Engineering International Islamic University Chittagong

DECLARATION

Being it is hereby announced that:

1. We have completed the thesis' original work as part of our degree program at International Islamic University Chittagong.

2. Unless properly cited in the references, the thesis does not incorporate any previously published or written by a third party content.

3. No information from the thesis has ever been accepted or submitted to a university or other institution for consideration for another degree or diploma.

4. All key contributors have been given credit.

Student's Full Name and Metric ID:

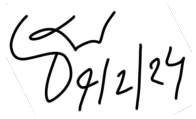
1. ASHRAF UDDIN TAREK **[C191011]**

2. SATIRTHA CHOWDHURY **[C191020]**

3. MOSHARAF HOSSAIN **[C183032]**

SUPERVISOR'S DECLARATION

I hereby certify that I have carefully read this thesis and that it meets the require- ments for the Bachelor of Science in Computer Science and Engineering.



Dr. Shahidul Islam Khan

Associate Professor

Computer Science and Engineering, IIUC

ACKNOWLEDGEMENT

Glory to the Great Allah first and foremost, with whose assistance we were able to write and conclude our thesis without too many difficulties. Second, we would like to express our gratitude to **Dr. Shahidul Islam Khan**, our research supervisor, for his considerate leadership and ongoing assistance. When we initially required help, he made accommodations for us. inspired us during the challenging times. And last but not least, without the prayers and well wishes of our parents, it would not have been possible.

ETHICS STATEMENT

To be successfully completed, an undergraduate thesis must conform rigorously to all university policies and procedures as well as the ethical standards for doing research. Original data served as evidence for our thesis. We double-checked all of our citations and references. Collectively, the article's three writers admit any instances in which the thesis rule was broken. We used a range of conference papers and journal papers to get around numerous findings. As a reaction, we would want to thank everyone who has supported us during the voyage. No unethical techniques were employed in the completion of our thesis.

.

ABSTRACT

The increasing number of social media users and e-commerce platforms have a great impact on people's daily life. People share their emotions and opinion through the social media platform. These emotions and comments now take the important place of analysis because based on these emotions, opinion, business farm make them plan what will produce, consumer decide what will he/she buy etc. Lots of work have been carried out to analysis the sentiment or emotion in English language. Due to the complexity of the Bangla language, few work has been done on it but in recent years several researchers have been carrying out various research based on the Bangla language. In this paper we conduct sentiment analysis (SA) on Bangla language by using the machine leaning model (ML). Most of the work basically divide the sentiment into three categories in this paper we divide the sentiment into four categories namely strong positive, positive, negative, and strongly negative. In this paper we use logistic regression (LR), decision tree (DT), Random Forest (RF), linear support vector machine (LSVM), confusion matrix, and kernel support vector machine (KSVM) algorithm of machine learning (ML). From support vector machine (SVM) we mainly used gaussian kernel radial basis function (RBF). The sentiments are converted to NumPy array to use the sentiment in machine learning. Since the NumPy array is numeric we train our model by these data to get the proper prediction about any given sentiment whether that sentiment positive or negative. The data used are all raw data or primary data collected from the different microblog websites and social media platforms. The total number of raw data 10851 most of them collected from Facebook and YouTube due to their popularity in our country some of the data collected from twitter as well. All the ML model applied for the single word of the sentences first, known as unigram feature analysis best on the single word the logistic regression model and RBF SVM provide the highest accuracy 71.26% and 71.91% respectively. By using two words each model works almost like unigram models, in bigram models LR again shows the highest accuracy, but the accuracy level little bit dropped for RBF SVM. The accuracy for LR and RBF SVM 72.04% and 68.79% respectively. Later we used models for three words defined as trigram feature analysis in that time get highest 70.87% accuracy for LR. Most of the papers basically use SentiWordNet to assess the polarity of the sentiments but in this paper, we use word by word analysis which hardly

seen to any paper. This paper will help the business farm as well as the consumers to make their decision and will work as guideline for the new researcher in this topic.

Keywords: Facebook, YouTube, Twitter, Unigram, Bigram, Trigram, Machine learning Model, Primary data, Feature.

DEDICATION

This thesis paper is dedicated to our family member's and our supportive friends. Through their continuous support they become part of the thesis paper. Our respectful supervisor was the part and parcel to our work. He helps us to continue the work and shows the way to overcome the barrier in our work. All these supports help us to make our way easier.

Table of Contents

List of Figures

List of Tables

CHAPTER I

INTRODUCTION

1.1 OVERVIEW

Sentiment analysis (SA) is also called opinion mining. Understanding emotions is one of the most important aspects of personal development and growth and as such, it is a key tile for the emulation of human intelligence [1]. Apart from these emotions play an important role in successful and effective human-human relationships. In our paper we try to find the polarity of the emotion of people's comment made in social media. We mainly classify the sentiment in two broad categories, positive and negative to analyze the human behavior in different subjects.

1.2 PROBLEM STATEMENT

As the growing number of internet user sentiment analysis become relevant topic to natural language processing (NLP) in machine learning and deep learning. Due to the vast Bangla spoken population and large scale of opinionated data on internet that's why researchers are increasing their interest in this field day by day. In modern time information is much more important than others thing in the world. To carry out any kinds of business, new theory, new vaccine, new product, or others decision on national or international scale the reaction of people is important. Another important thing is that this is the world of democracy where the people's polarity is playing a significant role in choosing their ruler. The Internet and information technology bring us closer to one another. one can easily share one's opinion with the world through the social media. In economic world consumers play the vital role to make thrives by any kinds of business farm. If the people review about any product of a farm are all good, this farm will thrive in their business. Shortly, sentiments or opinions delivered by people are crucial in the technical world. For this reason, we try to carry out sentiment analysis on Bangla language to help the people from every sector to make a proper decision. In this

paper we carry out word by word analysis to make proper prediction about the sentiment's polarity.

1.3 MOTIVATION AND SCOPE OF RESEARCH

At present, there are about 2,500 e-commerce companies in the country and at least 50,000 business pages on Facebook [2] and Bangladesh is the third largest country of Facebook user along with 66.94M active internet user. The summary of social media user in Bangladesh as follows.

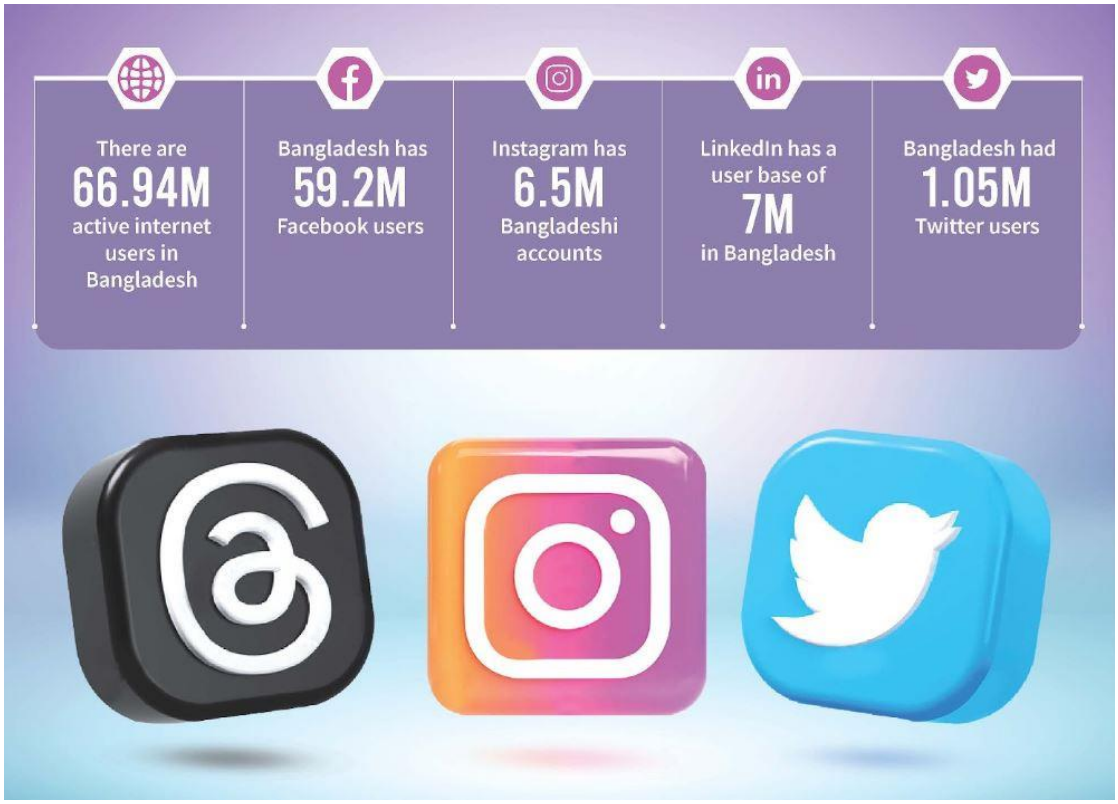


Figure 1 Summary of social media user in Bangladesh

Bangla speakers number about more than 230 million today, making Bangla the seventh language after Chinese, English, Hindi-Urdu, Spanish, Arabic and Portuguese. [3]

For this vast population of Bangla language, it is important to understand the emotions

of this population. These emotions play an important role in business, politics, e-commerce, and preference.

Here in this paper, we try to understand the emotions of Bangla language. How does it polarize? and where it can be used? It will help the business farm to make decisions, improve their product quality, produce new products and the political movement, people's preference etc. The ML and Deep learning model were also tested, and this paper will help the new researcher to this field.

1.4 OBJECTIVE OF RESEARCH

The subject we try to reveal through this paper are:

1. To get the more accurecy in predictiong the polarization of sentiment on Bangla language throurg different approches.

2. Utilize the maching learnig and Deep learning tools for better performance

3. Utilize the real-life data for proper appliction

4. Predict the human behaviour or preferecnces

5. Pre-train dataset of one word , two word and three word to prove its usefulness.

1.5 THESIS OUTLINE

An overview of the total work is given in the first chapter. In the second chapter we discuss the related works of sentiment analysis on Bangla language. In the third chapter we discuss the methodology and different models applied in our paper. The result of the applied model discussed in the fourth chapter. Finaly the conclusion and limitations of our paper are mentioned in the last chapter.

CHAPTER II

LITERATURE REVIEW

Several works have been carried out regarding sentiment analysis on Bangla Language. Now a days sentiment analysis takes significant place to research. In this regard, like other language sentiment analysis has been carried out to our bangle language as well. Several methods have been introduced to carry out proper sentiment analysis, like author [4] used two Machine Learning (ML) algorithm, called Naive Bayes Classification algorithm and Topical approach on six individual emotions, but they get highest accuracy as they use a smaller number of emotions classification. They get above 90% accuracy for topical approach. Afif Hasan, Mohammad Rashedul Amin [5] applied Deep Recurrent Model Specially Long Short-Term Memory (LSTM) on bangle and Randomizer Bangla text.

S Chowdhury and W Chowdhury [6] Used Support vector machine (SVM) and Maximum Entropy (MaxEnt) to classify Bangla Microblog post, into two categories, collected from twitter, get the accuracy 93%.

Purba, M.R Akter [7], used a hybrid convolutional long short-term memory (CNN-LSTM) based natural language Processing (NLP) model.

Sentiment analysis introduced on different sector like Ecommerce Business platform [8]

Covid-19 vaccination [9] and YouTube comment [10].

Here Some other works on Bangla sentiment analysis

NR Bhowmik M Arifuzzaman use Bangla text sentiment Score (BTSC) to score the whole sentence after analysing each word to detect the emotion of whole sentence with the help of Lexicon Data Dictionary (LDD) [11]

Authors of [12] use the valency of word to detect emotion from sentence, for this work they used SentiWordNet which is basically designed for English language.

MA Safin M Hasan [13] made a product review analysis on Bangla Language with the help of support vector machine (SVM) , Random Forest and Logistic Regression . They basically classify the comment as positive and negative. They show the accuracy level of different models among them SVM, shows the highest accuracy almost 90%.

S Sharmin and D Chakma [14] for the first time use attention mechanism to analyze the Bangla sentiment to minimize the false detection rate and enhance the prediction accuracy.

Authors of [15] applied Adaptive Neuro_Fuzzy Inference System (ANFIS) to predict the polarity of Bangla text collected from twitter tweets. They didn't handle the spam tweets which cause the low accuracy of their model.

NI Tripto ME Ali [16] uses five levels of sentiment including strongly positive and strongly negative level but for 3 class they get 65.97% accuracy where 54.24% for 5 class. They use Bangla, English, and Randomized text from you tube comments.

Eshita Khatun and T Rabeya get 98.39% accurecy for Random Forest model (RF), among all others ML model like AdaBoost, Decision tree, SVM and LightGBM for book reviews in Bangla language. This is the highest accuracy I have ever seen.
While authors of [17] get maximum accuracy, 76.48% for Linear Support Vector machine (SVM) They also implemented TF-IDF algorithm on data crawled from google play store. They also showed that naive bayes model works well even in heavy context situation while simple linear kernel works well and first.

Author [18] used Deep Neural Network (DNN) and Natural Language Processing (NPL) for product review classification in E-Commerce platform for bangle text.

S.m.s salehin and R miah conduct a comparative sentiment analysis on Bangli face book post, they use Deep learning and machine learning approach together among them ML approach gives the high accuracy,86.7% for SVM on the other hand Recurrent

Neural Network namely LSTM gives 72% accuracy, but the deep learning models are potential in Bangli sentiment analysis also [19]

Serial no.	Author(s)	Technique(s)	Findings	Strength or Limitations
1	Hasan K.M.A et al.	SentiWordNet, WordNet	Contextual valency of texts or paragraph	Major limitations is use of SentiWordNet
2	Tuhin R.A et al.	Naïve Bayes, Tropical Approach	Six emotions detection	Lack of others ML models
3	Hasan A et al.	RNN, LSTM	Extraction emotions from Romanized text	Use of the Deep learning model
4	Safin M.A et al.	NLP, SVM, KNN, Decision Tree, RF, and LR	Product reviews sentiments classification and prediction	ML and Deep learning models application
5	Tripto N.I et al.	CNN, LSTM, NB, and SVM	Detection of multilabel sentiment	Excluding of multiple aspects and topic information in sentiment

Table 1 Comparative Analysis Referring Papers

CHAPTER III

METHODOLOGY

3.1 WORKING TOOLS

The following tools were used in this thesis:

1) Python 3.10

2) Google Colaboratory

3) Google Sheet

4) Microsoft Excel

5) Microsoft Visual Studio

6) MS office

3.2 WORKFLOW

In this article, we proposed different supervised machine learning (ML) models. We first depict about our data set, the source of raw data, the raw data is not directly useable so we go through some data processing methods like cleaning the unnecessary punctuation from the data and removing the low length data. After processing the data, we make the summary report about the dataset and also visualize the distribution of the response feature. In tag feature there are four classes for each of these classes we figure out the summary and the list of words which most frequently used in a particular class. The data later converted into NumPy array to make it applicable to the ML models. Finaly, we applied our proposed ML models on unigram, bigram, and trigram feature of sentiments. We present details of our models results in the following chapter.

3.3 DATA COLLECTION

In this research paper we use the raw data from primary sources. This all data is collected one by one from fakebook, YouTube and twitter also. We collect total 10851 comments from Facebook, you tube and twitter, among them 5051 comments from Facebook, 4400 from YouTube and rest of them from twitter. Since most people in our country are popular with Facebook and YouTube, we take these platforms with high priority to collect data. People in our country hardly use twitter so we just take few data from that platform. Various kinds of news published through this social media platform and people willingly make their opinions on that news, these opinions we use as raw data of our research after collecting them randomly and later we classified those as positive, negative, very positive or very negative. To collect data no scientific procedure or probabilistic measure has been used, we collect data based on our personal judgement.

3.4 DATASET ATTRIBUTES

In our dataset there are only two attributes namely comment, and tag. Each of them contains 10851 rows. Comment attributes contains the comment or opinion collected from Facebook, YouTube, and twitter, where the tag attributes hold the label of sentiment with respect to the data of comment attributes of dataset.

3.5 DATASET PROPERTIES

In the row dataset, the total number of reviews are 10851 from which there are 2280 number of very positive reviews, 1445 number of positive reviews,3928 number of very negative reviews, and 3198 number of negative reviews.

In short,

Total Reviews: 10851

Total Very Positive Reviews: 2280

Total Positive Reviews: 1445

Total Very Negative Reviews: 3928

Total Negative Reviews: 3198

3.6 REPRESNTATION OF SCANNED RAW DATASET

Raw dataset sample of our study presented at below,

Comment	Tag
লিখার সময় পারলে সত্য লিখার অভ্যাস শিখুন।	Negative
এটা কেন হচ্ছে? সংশ্লিষ্ট সকলের ডিপ্রেশনের ফলে? নাকি সরকার মনোনিত পরিচালনা পর্ষদের ব্যর্থতার কারণে?	Negative
আমাদের দেশের স্বাভাবিক অর্থনৈতিক গতিপ্রবাহকে বাধাগ্রস্ত করা হয়েছে। তাই এসব ঘটছে।	Very Negative
চুরি নয় লুটপাট।	Very Negative
একটা ভাল প্রতিষ্ঠান কিভাবে লুট হয়ে যাচেছ!	Very Negative
সরকার যাদের এই ব্যাংকে নিয়গ দিয়েছে তারা ব্যাংকিং এর 'ব' ও জানে না. এইটা ধবংস হবার পর এদের কলিজা ঠাং	Very Negative
পুরোপুরি আওয়ামীকরন হলেই আর এর অসতিতই থাকবেক না!	Negative
এককথায় ব্যাংকটিতে ডাকাতি হয়েছে এবং হতেই থাকবে।	Negative
ইসলামি ব্যাংক প্রারম্ভ থেকেই গ্রাহকদের পছন্দের ব্যাংক ছিল। যার সাফল্য ইর্ষনীয়।এটা কার দ্বারা পরিচালিত সেটা	Negative
এরা যেখানেই যাবে সেখানেই চুরি হবে।	Very Negative
এটা ভুলে গেলে চলবে না যে ব্যাংকটিতে নুতন পরিচালনা পর্ষদ এসেছে তাদের ব্যাংকটির সব কিছু বুঝে উঠতে একটু	Positive
শেয়ার প্রতি আয় কমেছে। শূন্য হয় নি এখনও। সরকারী [অর্থ মন্ত্রণায়ের নির্ধারিত] মনোনীত পর্ষদ যেদিন দায়িত্বভা	Negative
পুরোপুরি জামাতমুক্ত করা গেলেই লাভের মুখ দেখবে ইসলামি ব্যাংক	Very Positive
বিগত কয়েক বছরের অভিজ্ঞতা বলে ব্যাংকসহ শ্যেনদৃষ্টি যেখানে সর্বনাশ সেখানে।	Negative
ইদানিং যে গ্রুপ ইসলামি ব্যাংকের শেয়ার কিনছে, তাদের কাছে কি অালাদিনের চেরাগ অাছে? এত ব্যাংক তারা বি	Negative
প্রবাসী বাংলাদেশিরা চলতি অর্থবছরের (২০১৮-১৯) প্রথম ৮ মাসে ১০,৪১০ দশমিক ২৯ মিলিয়ন মার্কিন ডলার রেমি	Negative
গরীবদের বেশী টানটান হতে নাই।	Negative
পাকিস্তান ভিক্ষা করে পেয়েছে। ভারত পেয়েছে তার অর্থনৈতিক শক্তির জন্য। আমরা কি ভিক্ষা করবো, নাকি নিজে	Positive

Figure 2 Sample data part 1

আমরা যারা বিদেশে মধ্যবিত্ত তারা মাসের আয়ে মাসের সংসার খরছ চললেই স্বচ্ছল ভাবি। খরছ বাদে হাতে উদ্বৃত্ত থ	Positive
২০০৫ সালে নকশা চূড়ান্ত করা গভীর সমুদ্র বন্দর নির্মাণ কাজ এখনো শুরু করা যায়নি কার কারনে ?	Negative
এত কিছু করা লাগেনা, ব্যবসায়িদের একটু সুবিধা দিন দেখবেন সব ঠিক হয়ে গেছে, এক সময় ঢাকার আই সি টি পা	Positive
ভারতে যে সেবার খরচ ৪ রুপি বা বাংলাদেশী টাকায় ৪টাকা ৬০ পয়াসা (আজকের ভারতীয় মুদ্রার বাংলাদেশী টাকা দ	Negative
এমএনপি করতে খরচ হবে ১৫৮ টাকা, একটা নতুন সিম কিনতে কত লাগে এটা সবাই জানে। শুধু চলতি নাম্বারের ে	Very Positive
আপনার নাম্বার যদি ০১৭ দিয়ে শুরু হয়, তবে গ্রামীন থেকে রবি বা বাংলালিঙ্ক এ সুইচ করলেও আপনার নাম্বার ০১৭	Very Negative
পুরাটা ই গ্রাহকের জন্য যন্ত্রনাদায়ক এবং ফলাফল শূন্য ব্যবস্হা।	Positive
আমাদের দেশ ভারতের থেকে অনেক ছোট আয়তনে, তাই বিনিয়োগ অনেক কম করতে হয় টাওয়ার বসানোর জন্য	Positive
আন্তর্জাতিক কলরেট দেড় সেন্ট থেকে দুই সেন্ট করলে দেশে মোবাইল কল রেট ও বাড়ান দরকার। মোবাইল কল রা	Negative
অবইধ ভিওয়াইপি ব্যাবসা বন্ধের জন্য আন্তর্জাতিক কল রেট কমানো উচিত। কল রেত বাড়ালে ভীওয়াইপি ব্যাবসা ব	Positive
আমাদের দেশে এটা কল্পনাই করা যায় না। আমাদের এখানে জিজ্ঞেস ই করা হয়, অমুক কোম্পানির মালিক কে? অ	Negative
রাবিশ ক্ষমতা লোভীর মতই কথা।	Negative
দেশের অর্থনীতির যা অবস্থা আপনাকে খুব বেশী চালাক লোক বলে মনে হয় না।	Very Positive
একতরফা নির্বাচন কইরা নিজেরাই।	Very Positive
তা মি: আকুব : আপনি নিজেকে প্রশ্ন করুন কেন তারা আসলো না ? এ সরকারের অধীনে নির্বাচনে যাওয়া সুইসাইডা	Very Negative
বেকুব না হলে কি নির্বাচন ঠেকানোর নামে দেশের মাঝে সন্ত্রাসী কর্মকান্ড করে সাধারণ মানুষ গুলিকে পেট্রোল বোম	Very Positive
ধরেন ২০১৯ সালের নির্বাচনে আওয়ামীলীগ নির্বাচন বয়কট করলো, ভোটাররা কি বিএনপিকে জেতানোর জন্য আর	Positive
ক্ষমতা লোভী দের জন্যেই দেশে একটি নির্বচন হয়েছিলসেখানে লোভী লোকগুলি দেশের ক্ষমতা দখল করেছিল।	Very Negative

Figure 3 Sample data part 2

3.7 DATA PRE-PROCESSING

The crucial part of the research is the pre-processing of the data where we will apply our machine learning model and other methods to reach our goal. The various methods applied to the raw data to make it fit for the machine learning model. These methods include cleaning the unnecessary data, cleaning the unnecessary punctuations and encoding the string data by NumPy array.

3.7.1 HANDLING UNNECESSARY PUNCTUATIONS

Though punctuation is a vital part of every language is always neglected by native speakers or not properly used. Since ML models are factionalized, they can't read unnecessary punctuation, to make the sentence readable my ML model we must remove this unnecessary punctuation.

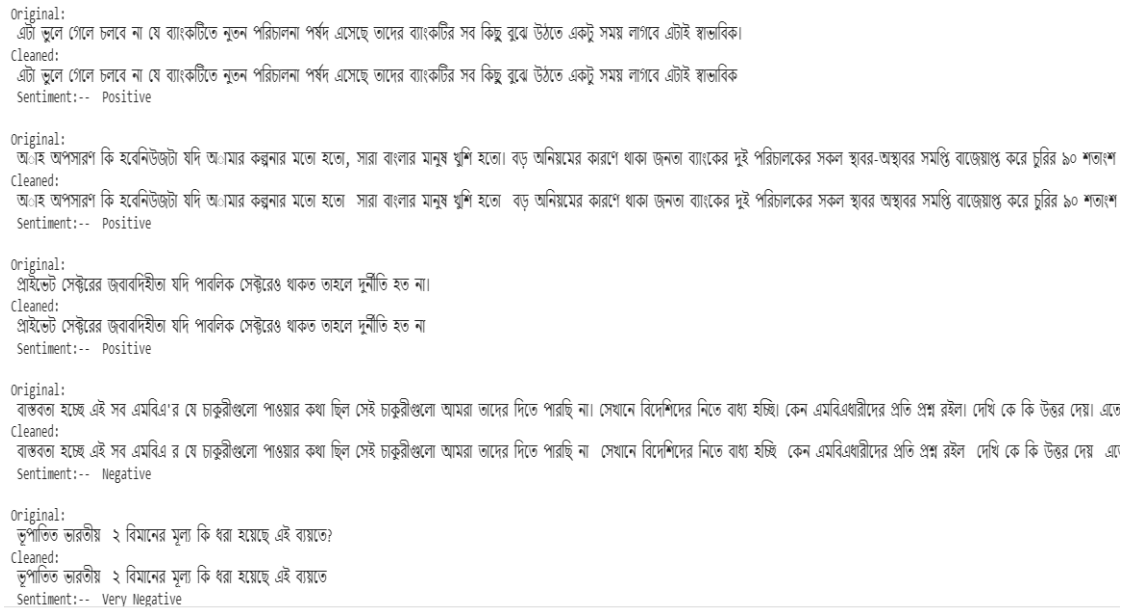


Figure 4 Data after removing unnecessary punctuation.

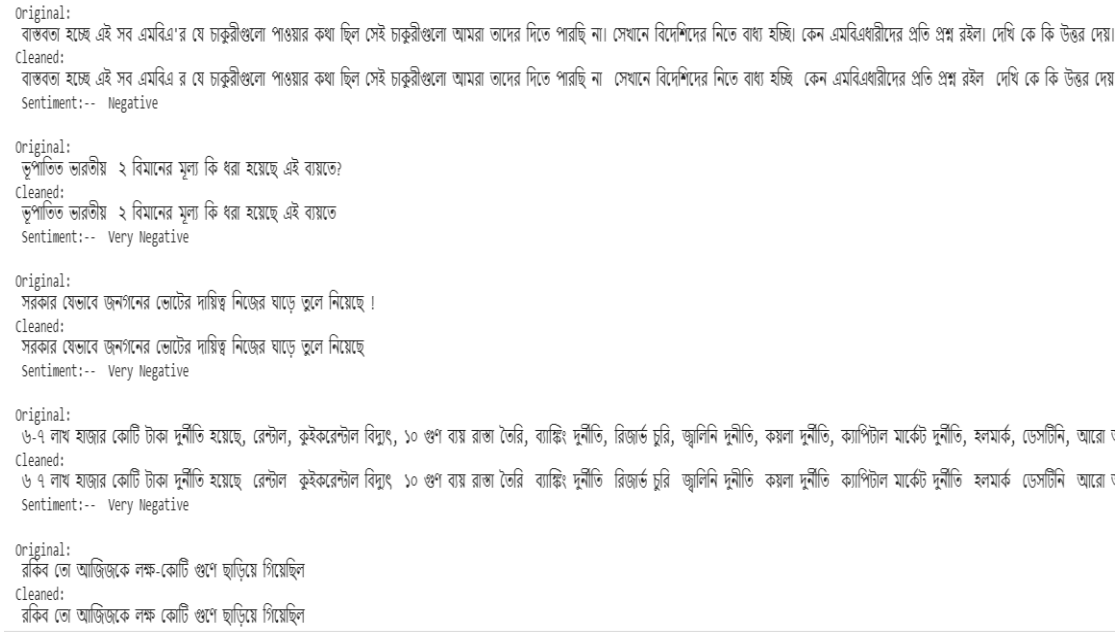


Figure 5 Data after removing unnecessary punctuation part 2

3.7.2 HANDLING THE LOW LENGTH DATA

It is usual sometimes the native speakers to express their opinions by using 2/3 letters without completing a word, which is only understandable by another native speaker. Here we remove such low length data from our data set, which have length 2 or less. The total number of small reviews is 72 out of 10851 reviews. After cleaning these small reviews we get 10779 net reviews.

After Cleaning:

Removed 72 Small Reviews

Total Reviews: 10779

3.7.3 LABLE ENCODING

Since the computer and machine learning model can't read the string data so we need to convert this string data or comment into numerical one to use the comment in ML model fruitfully. In this section we convert the whole comment into NumPy array to use it as numerical number.

3.8 APPLIED ALGORITHMS

The ML models, we applied in this paper are:

a) Logistic Regression

b) Decision Tree

c) Random Forest

d) Naïve Bayes

e) SVM K Neighbors Classifier (KNN)

f) Linear Support Vector Machine (Linear SVM)

g) Kernel Support Vector Machine (Kernel

3.8.1 LOGISTIC REGRESSION

Logistic regression model or simply Logit model is a statistical method used as machine learning model where the response variable is categorical or binary. This model calculates the probability of an event by taking having the log-odds of the other independent attributes. The independent attributes may or may not be the categorial variable and one or more than one but the number of response variable always one and it must be a categorical variable. The model coded the outcome of response variable by 0 and 1. This model also shows the probability as well. Since most of the models in statistics are based on numerical response variable so logit model takes a significant place due to its characteristics of handling the categorical response variable.

The logistic function written as $p(x) = 1/(1 + e^{-\frac{(x-\mu)}{s}})$,where μ is a parameter of location measurement this function also known as sigmoid function

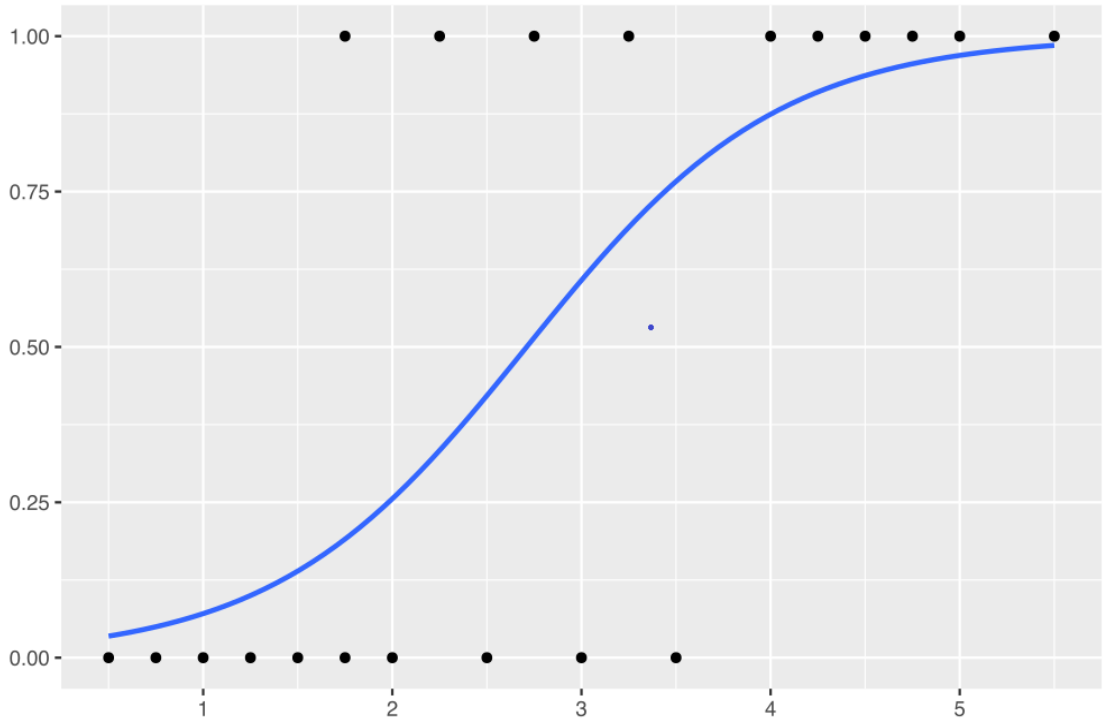


Figure 6 Logistic regression

Since in this paper the response variable is categorical variable logistic regression play an important role to this paper.

3.8.2 DECISION TREE

Decision tree is a powerful supervised machine learning algorithm, also known as regression tree. This algorithm sequentially divided the data into two subsets to make predictions. There is root node in the top which contains all the data, this root node later subdivided into internal nodes. Internal nodes are the features in the datasets. To work with the model, we first select a threshold value, based on this value each internal node subdivided into another terminal node or internal node and this process continues until certain condition satisfied. Each terminal node of the tree does not split later; they present the result or predictive value of the algorithm on problem. Decision trees are

easily understandable by visualizing and it easy to use. Decision tree algorithm used for the quantitative attributes or categorical attributes, is the most useful statistical tool for analyzing data. Here the graphical shape of the decision tree

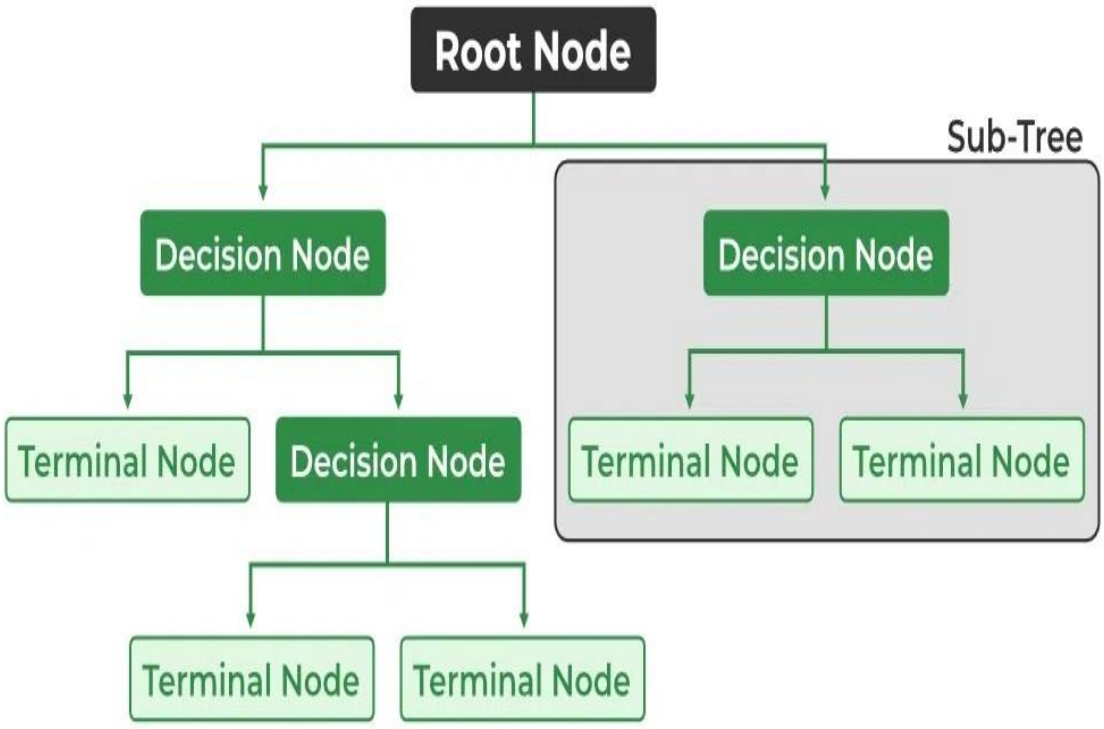


Figure 7 Decision tree

3.8.3 RANDOM FOREST

Random forest model is like the decision tree algorithm, they both work almost similarly but Random Forest is upgraded from the decision tree model, overcoming some of the drawbacks of decision tree. Basically, random forest algorithm is a collection of decision treesit takes name as random because here we apply to random selection process namely bootstrap sampling and selection of features. In the random forest algorithm first, we select data from the original dataset with random replacement process, this process is called bootstrapped sampling, better to say that the size of the datasets should be equal to the original datasets. After sampling we allocate the features

randomly to each of the bootstrapped sampling data, the allocation size of the feature should be equal to log or square root of the total number of features. For each of the allocated data and features we will train a decision tree model, Now the set of observation will pass through each decision tree. The voting or averaging of each decision tree will be the outcome or prediction value of the model. This method reduces the variation, and it is also applicable for the features which have small amount of effect to the response variable. The change in the original datasets hardly affected the whole model. Graphical representation of the random forest algorithm as follows.

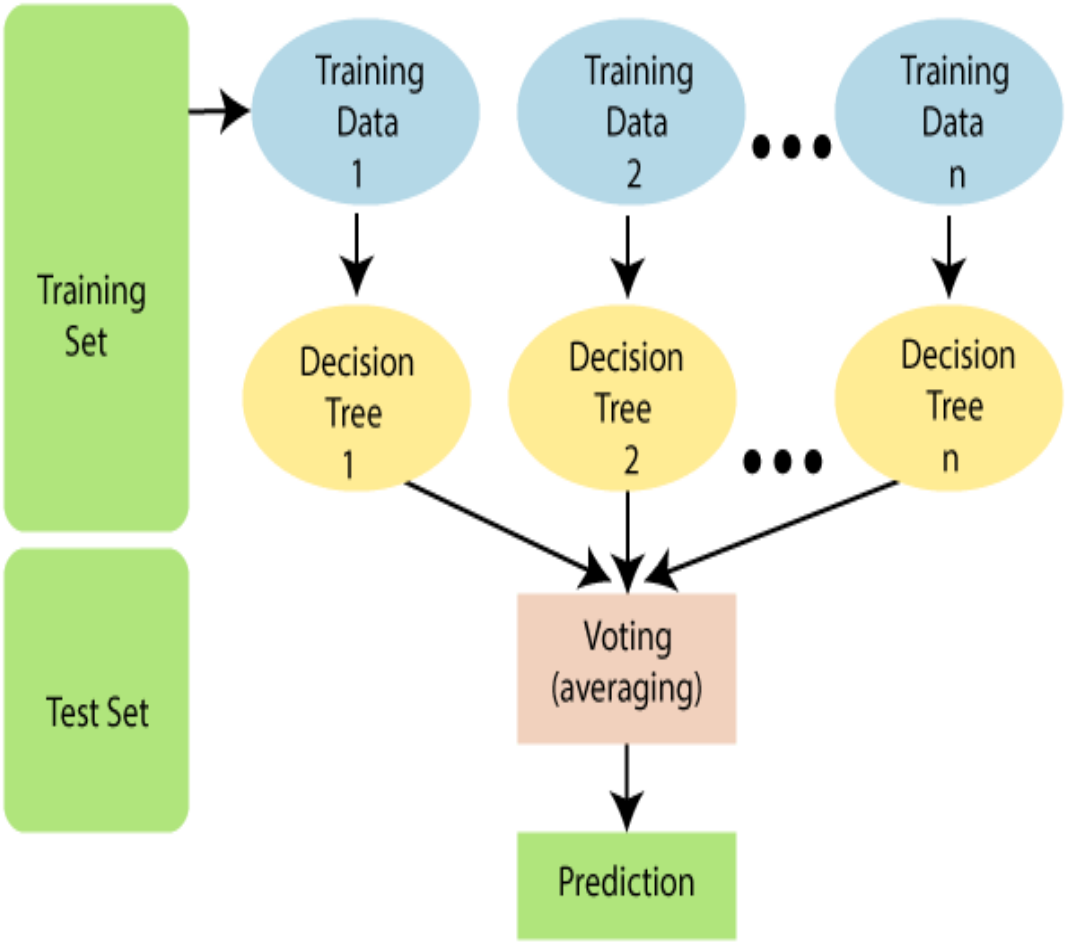


Figure 8 Random forest

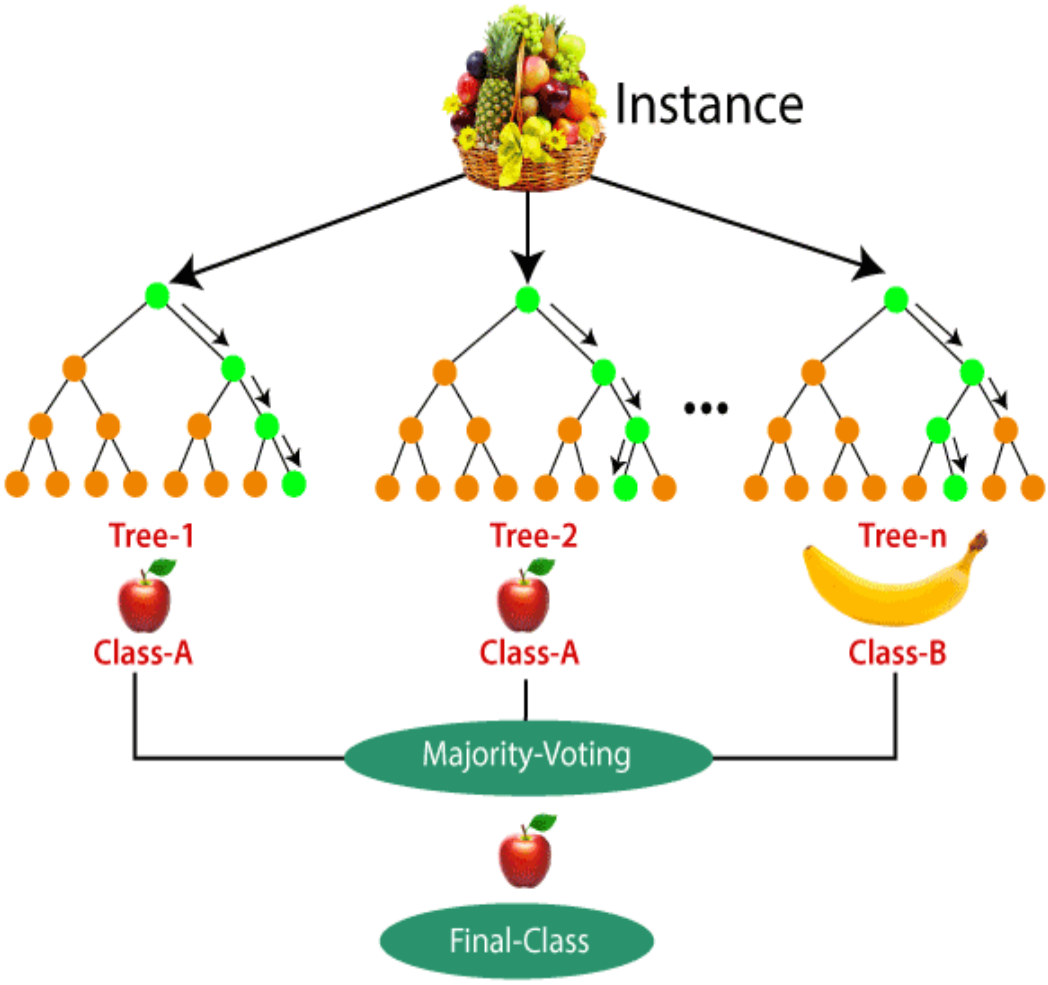


Figure 9 Random forest part 2

3.8.4 KNEIGHBORS CLASSIFIER(KNN)

K-Nearest Neighbors (KNN) is a simple and important machine learning algorithm which frequently applied for both classification and regression tasks. The algorithm is based on the principle that similar data points tend to fall into similar categories or have similar values. The key considerations for this model first to select the value of K, it should be the odd number. Once we select the number of K our second task to calculate minimum distance the of the given observation from the original datasets. The minimum K number distance selected among the category which selected maximum number, the given data will be classified according to this category.

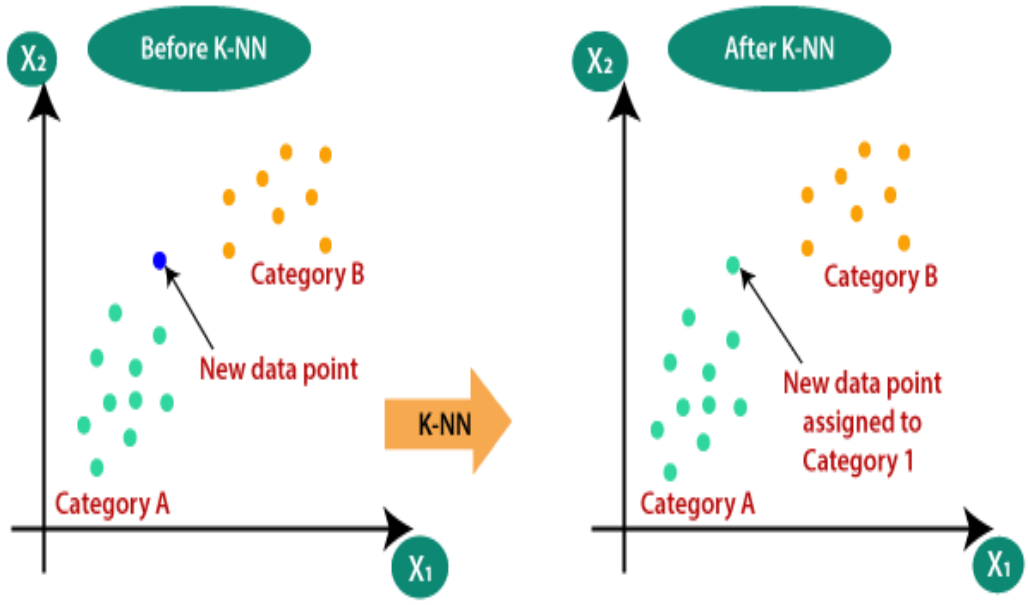


Figure 10 K-Nearest Neighbors

3.8.5 LINEAR SUPPORT VECTOR MACHINE(LSVM)

The linear support vector machine is a widely used machine learning algorithm. SVMs are used in various sorts of tasks like data classification on N-dimension where N is number features, image classification, outlier detection, text classification, spam detection, face detection, handwriting identification etc. LSVM is usually used when a dataset has 2 or 3 features only where a single line divides the whole dataset into two or more classes. The number of singles can be more than one. This line is drawn in such a way that it maximizes the marginal distance from each class. The distance between the support vector and hyperplane is called marginal distance. support vectors are the

nearest data of the hyperplane. Once the line is drawn, we use regression and classification to predict and classify new data respectively. The final line which is set for decision making is known as hyperplane.

In figure 11 there are two features, one identified as blue color and another with green color. in next figure we separated the feature by a single line, there are several lines can be drawn but we must choose the best one. in figure three the optimal line maximizes the marginal distance so this line is taken as threshold thus a new data will pass through the train model it will classify the data either as blue or as green

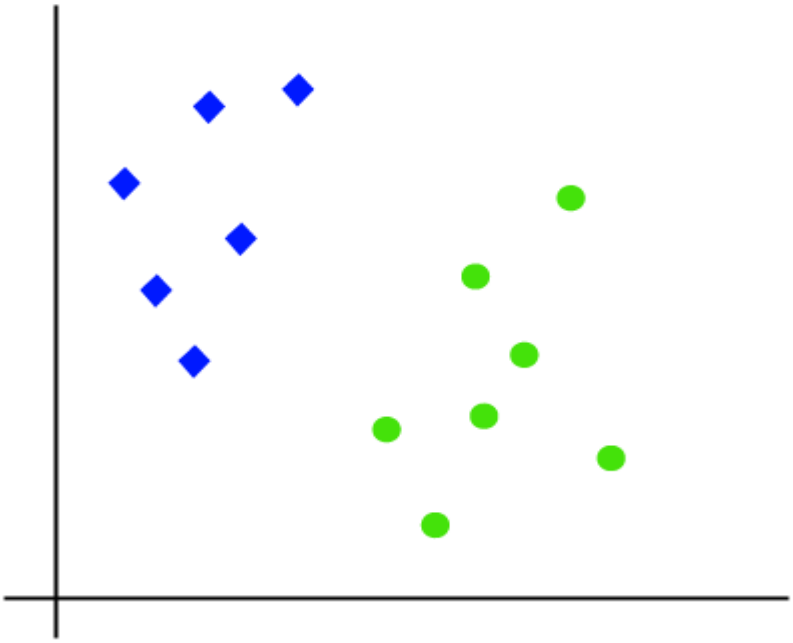


Figure 11 linear support vector machine part 1

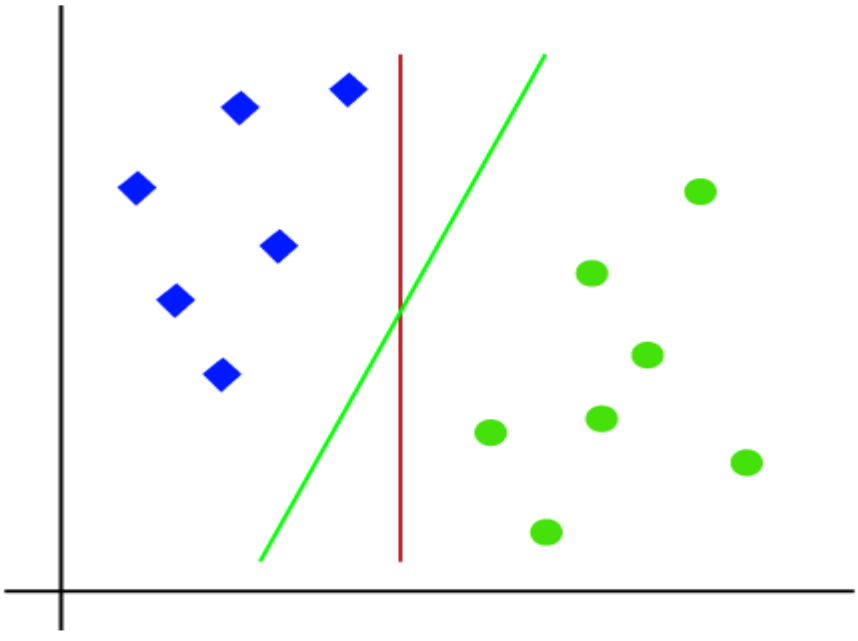


Figure 12 linear support vector machine part 2

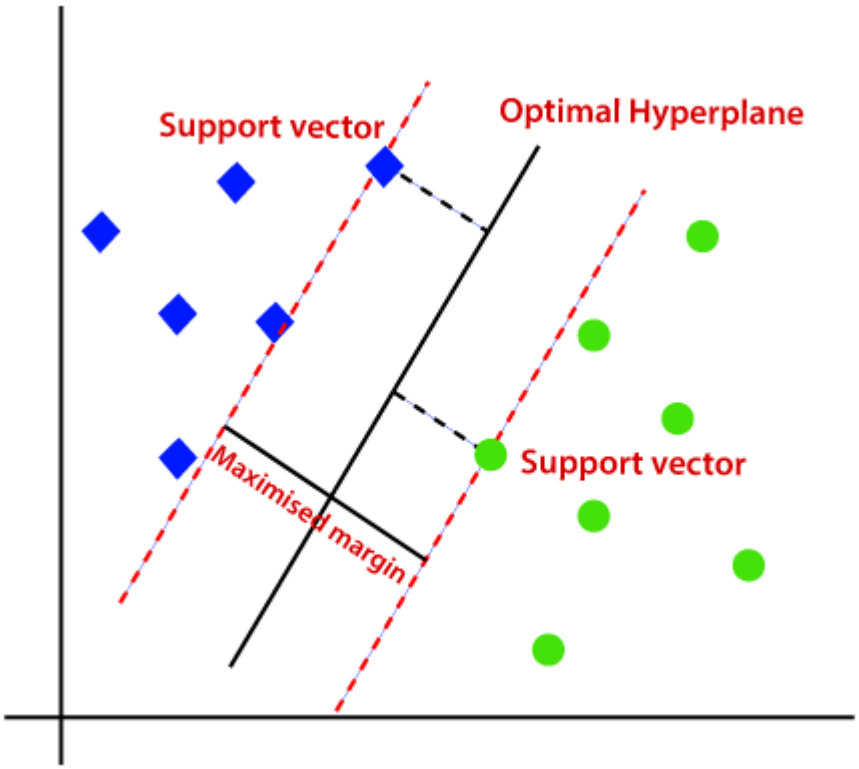


Figure 13 linear support vector machine part 3

3.8.6 KERNEL SUPPORT VECTOR MACHINE(KSVM)

The kernel support vector machine is a crucial machine learning algorithm for classification and regression tasks. Kernel SVM method used in that time when the linear separation is not possible to the datasets, which means dataset consists of more than 3 features. Kernel is defined as a set of mathematical functions that are used to transform the training datasets into linear equations in a higher number of dimension spaces where one can't use linear SVM directly. Kernel function can be written as

$$K(\bar{x}) = \begin{cases} 0 & if\ \|\bar{x}\| \leq 0 \\ 1 & otherwise \end{cases}$$

There are several kernel functions, among them RBF or gaussian kernel radial basis function used widely than others. The RBF defined as

$$K(x, y) = e^{-(\gamma\|x-y\|^2)}$$

Where "γ" is a parameter that controls the shape of the decision boundary.

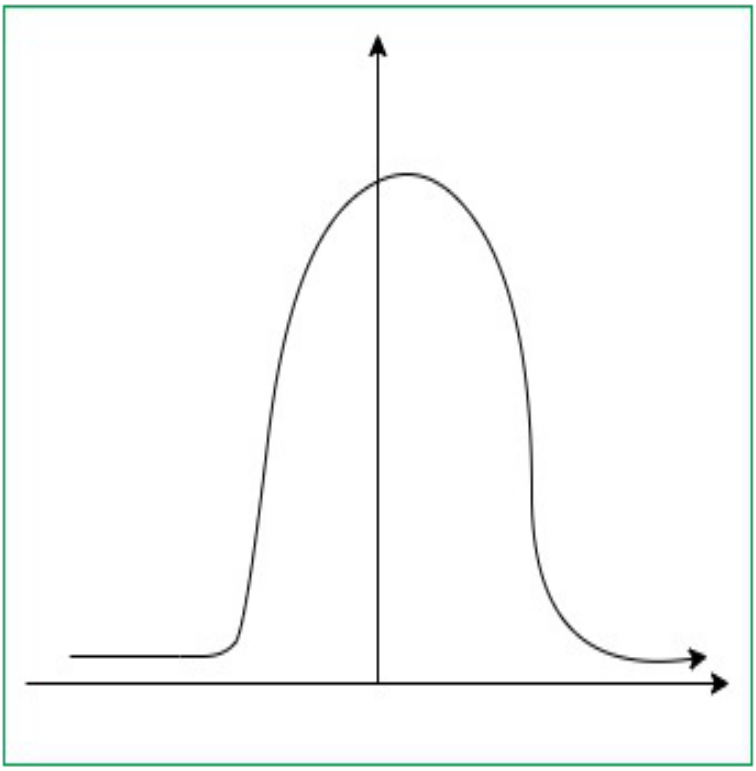


Figure 14 Gaussian Kernel Graph

3.9 SUMMARY

To make the analysis easier at first, we should visualize the dataset distribution. Since our data are categorical, we set a bar diagram for the response feature. The height of the bars based on the frequency of 'tag' feature in the dataset.

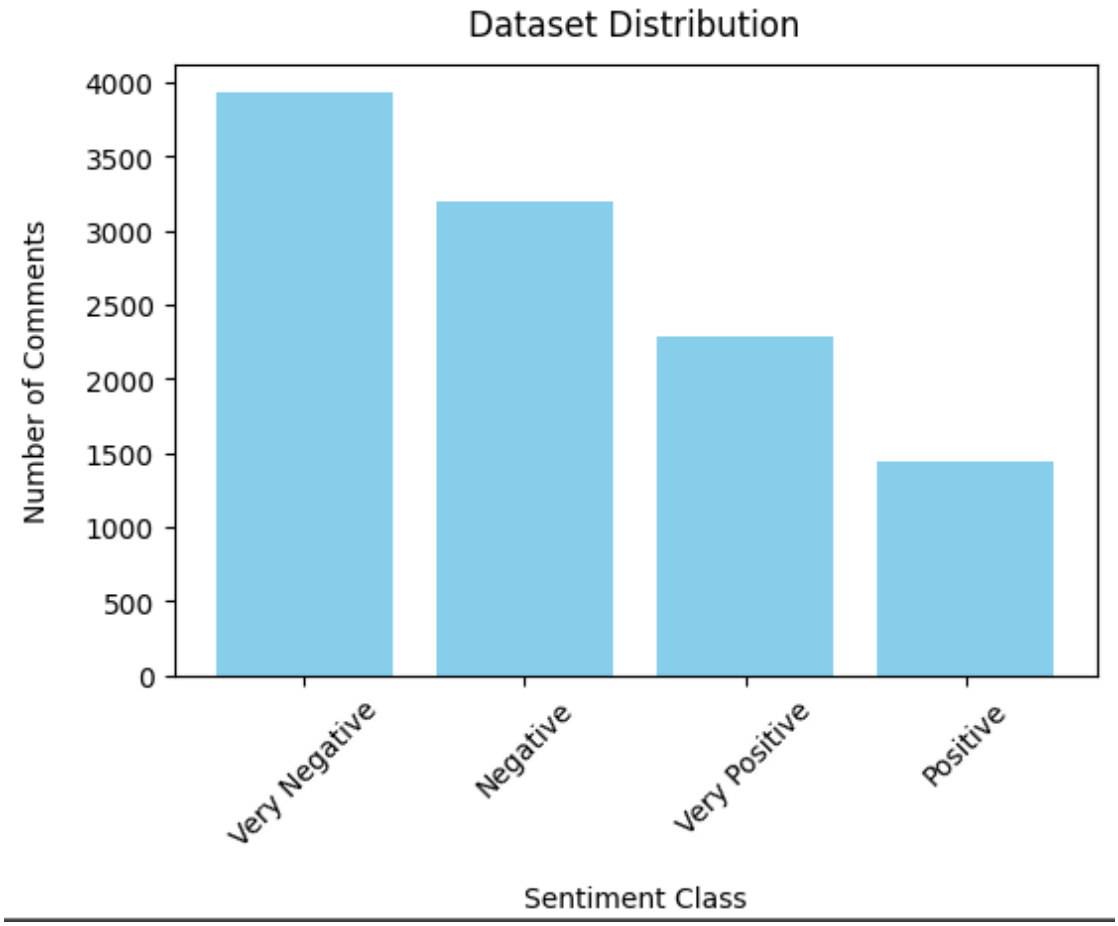


Figure 15 Dataset Distribution

Here from the distribution, we can see that there are many the number of very negative and negative comments in our dataset.

It is better to see the class wise dataset summary of the tag attribute. There are four class in the tag attributes namely very positive, positive, very negative, negative. Now we show here the first 12 words which have been frequently used in the comment for each of these classes.

Class Name: Negative

Number of Documents: 3185

Number of Words: 59697

Number of Unique Words: 13375

Most Frequent Words	
না	1125
কি	657
করে	612
আর	586
এই	452
হবে	382
হয়	330
তো	314
যে	295
টাকা	289

Table 2 Most Frequent Negative Words

Class Name: Very Negative

Number of Documents: 3908

Number of Words: 80743

Number of Unique Words: 16247

Most Frequent Words	
না	1538
আর	868
করে	851
কি	720
এই	644
হবে	529
হয়	455
যে	420
টাকা	399
তো	394

Table 3 Most Frequent Very Negative Words

Class Name: Positive

Number of Documents: 1442

Number of Words: 31129

Number of Unique Words: 8858

Most Frequent Words	
না	561
করে	307
আর	259
হবে	254
এই	211
করা	183
জন্য	178
কি	176
ও	161
হয়	154

Table 4 Most Frequent Positive Words

Class Name: Very Positive

Number of Documents: 2244

Number of Words: 44151

Number of Unique Words: 10904

Most Frequent Words	
না	654
হবে	429
এই	416
করে	372
জন্য	346
আর	296
করা	286
ও	232
করতে	228
আমাদের	227

Table 5 Most Frequent Very Positive Words

Total Number of Unique Words: 29379

The visualization of the summary of the classes' words

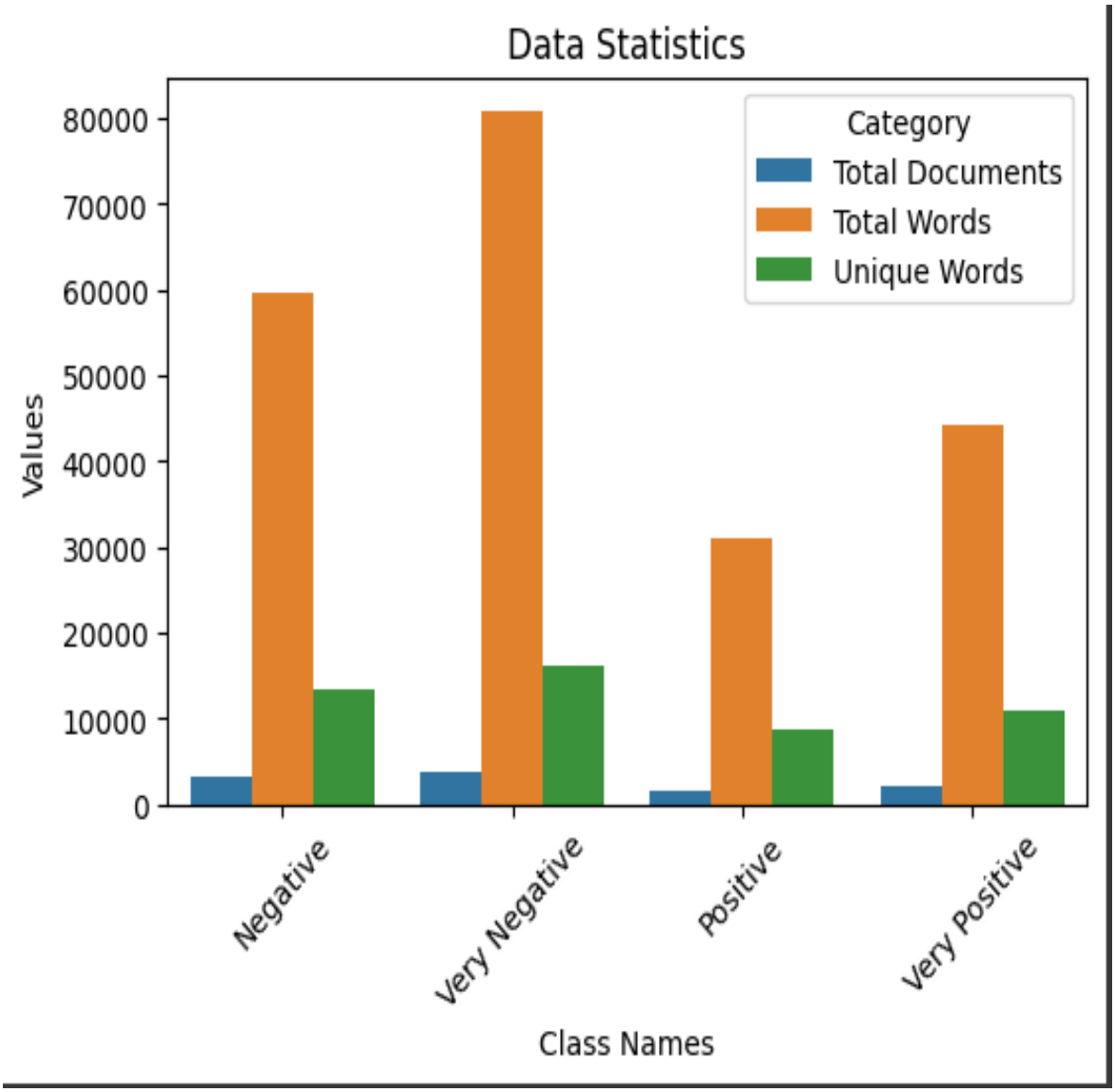


Figure 16 Summary of classes words

The length of comment is another important characteristic to measure. Here we figure out the length distribution of the comments. The height of the bar depends on the number of comments for length of the comments.

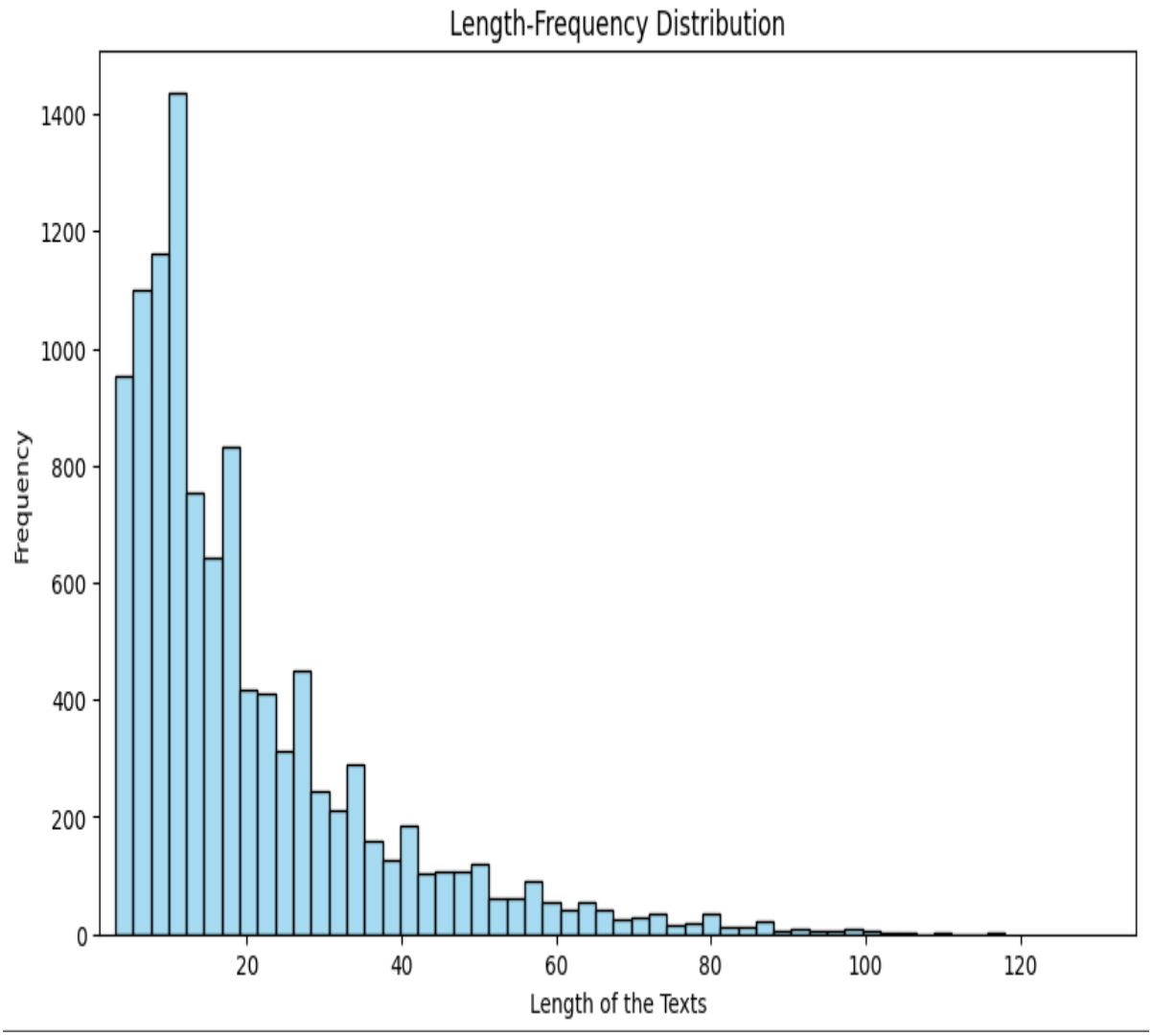


Figure 17 Length of Comments

CHAPTER IV

RESULTS AND DISCUSSION

4.1 OVERVIEW

In this section we will discuss the outcome of the applied models in this research paper. We applied our machine learning models on unigram, bigram, trigram feature of the sentiments in comment variable. For unigram feature logistic regression and RBF kernel support vector machine provides above 70% accuracy. In bigram feature logistic regression models shows 72.04% accuracy and for trigram feature logistic regression model gives 70.87% accuracy.

4.2 EVALUATION METRICS

There are many yardsticks to evaluate the applied model among them confusion matrix, accuracy, precision, and F-1 score are more familiar to us. These metrics used to compare different type of applied model, because any one can fit any kinds of model to make prediction about on the basis of given observation now several question have raised after fitting the model such as, is the predicted result reliable?, Will one should accept it or reject it?, How much one can depends on the result?, Is the outcome effective?. To answer this type of question these metrics have been introduced to know which model efficiently predicts the response feature, and which one is reliable.

Confusion matrix

Confusion matrix is a useful tool in machine learning and statistics to assess the performance of the models. it is a table consists of four separate part which are true positive (TP), true negative (TN), false positive (FP), false negative (FN) respectively.

Accuracy

It is the percentage of predictions that are accurate. ratio of all correctly anticipated cases, whether positive or negative, and all cases in the data is measured, and it explains how frequently the model predicts the right outputs.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

Precision

It can be defined as the number of correct outputs provided by the model or out of all positive classes that have predicted correctly by the model, how many of them were actually true. It can be calculated using the below formula:

$$\text{Precision} = \frac{TP}{TP+FP}$$

Recall

It is defined as the out of total positive classes, how our model predicted correctly. The recall must be as high as possible.

$$\text{Recall} = \frac{TP}{TP+FN}$$

F-measure

If two models have low precision and high recall or vice versa, it is difficult to compare these models. So, for this purpose, we can use F-score. This score helps us to evaluate the recall and precision at the same time. The F-score is maximum if the recall is equal to the precision. It can be calculated using the below formula:

$$\text{F-measure} = \frac{2*Recall*Precision}{Recall+Precision}$$

4.3 RESULTS ANALYSIS

In our dataset, we train based on unigram, bigram and trigram and apply 6 machine learning algorithms. And we split our dataset, 10% for test and 90% for train. Below table shows the accuracy, precision, F1-score and recall values.

Performance Table for Unigram feature					
	Accuracy	Precision	Recall	F1 Score	Model Name
0	71.26	73.93	63.93	68.56	LR
1	59.82	58.85	59.95	59.40	DT
2	70.48	76.22	57.82	65.76	RF
3	71.52	73.80	64.99	69.11	MNB
4	64.76	64.40	62.86	63.62	KNN
5	70.61	83.26	50.13	62.58	Linear SVM
6	71.91	81.32	55.44	65.93	RBF SVM

Table 6 Performance Table for Unigram feature

In case of Unigram feature:

Highest Accuracy achieved by RBF SVM at = 71.91

Highest F1-Score achieved by MNB at = 69.11

Highest Precision Score achieved by Linear SVM at = 83.26

Highest Recall Score achieved by MNB at = 64.99000000000001

Performance Table for Bigram feature					
	Accuracy	Precision	Recall	F1 Score	Model Name
0	72.04	77.18	61.01	68.15	LR
1	62.55	62.13	60.48	61.29	DT
2	69.31	79.75	50.13	61.56	RF
3	71.26	73.08	65.52	69.09	MNB
4	65.28	65.03	63.13	64.06	KNN
5	61.51	85.22	25.99	39.84	Linear SVM
6	68.79	86.24	43.24	57.60	RBF SVM

Table 7 Performance Table for Bigram feature

In case of Bigram feature:

Highest Accuracy achieved by LR at = 72.04

Highest F1-Score achieved by MNB at = 69.08999999999999

Highest Precision Score achieved by RBF SVM at = 86.24000000000001

Highest Recall Score achieved by MNB at = 65.52

Performance Table for Trigram feature					
	Accuracy	Precision	Recall	F1 Score	Model Name
0	70.87	77.42	57.29	65.85	LR
1	59.43	58.90	57.03	57.95	DT
2	69.70	83.96	47.21	60.44	RF
3	71.39	72.62	66.84	69.61	MNB
4	64.24	63.49	63.66	63.58	KNN
5	55.40	92.50	9.81	17.75	Linear SVM
6	62.68	86.29	28.38	42.71	RBF SVM

Table 8 Performance Table for Trigram feature

In case of Trigram feature:

Highest Accuracy achieved by MNB at = 71.39

Highest F1-Score achieved by MNB at = 69.61

Highest Precision Score achieved by Linear SVM at = 92.5

Highest Recall Score achieved by MNB at = 66.84

4.3.1 CONFUSION MATRIX

For Unigram:

Confusion Matrix for Logistic Regression (Unigram feature):

	Predicted 0	Predicted 1
Actual 0	307	85
Actual 1	136	241

Table 9 Logistic Regression (Unigram feature)

Confusion Matrix for Decision Tree (Unigram feature):

	Predicted 0	Predicted 1
Actual 0	234	158
Actual 1	151	226

Table 10 Decision Tree (Unigram feature)

Confusion Matrix for Random Forest (Unigram feature):

	Predicted 0	Predicted 1
Actual 0	324	68
Actual 1	159	218

Table 11 Random Forest (Unigram feature)

Confusion Matrix for Naive Bayes (Unigram feature):

	Predicted 0	Predicted 1
Actual 0	305	87
Actual 1	132	245

Table 12 Naive Bayes (Unigram feature)

Confusion Matrix for KNN (Unigram feature):

	Predicted 0	Predicted 1
Actual 0	261	131
Actual 1	140	237

Table 13 KNN (Unigram feature)

Confusion Matrix for Linear SVM (Unigram feature):

	Predicted 0	Predicted 1
Actual 0	354	38
Actual 1	188	189

Table 14 SVM (Unigram feature)

Confusion Matrix for Kernel SVM (Unigram feature):

	Predicted 0	Predicted 1
Actual 0	344	48
Actual 1	168	209

Table 15 K SVM (Unigram feature)

For Bigram:

Confusion Matrix for Logistic Regression (Bigram feature):

	Predicted 0	Predicted 1
Actual 0	324	68
Actual 1	147	230

Table 16 Logistic Regression (Bigram feature)

Confusion Matrix for Decision Tree (Bigram feature):

	Predicted 0	Predicted 1
Actual 0	253	139
Actual 1	149	228

Table 17 Decision Tree (Bigram feature)

Confusion Matrix for Random Forest (Bigram feature):

	Predicted 0	Predicted 1
Actual 0	344	48
Actual 1	188	189

Table 18 Random Forest (Bigram feature)

Confusion Matrix for Naive Bayes (Bigram feature):

	Predicted 0	Predicted 1
Actual 0	301	91
Actual 1	130	247

Table 19 Naive Bayes (Bigram feature)

Confusion Matrix for KNN (Bigram feature):

	Predicted 0	Predicted 1
Actual 0	264	128
Actual 1	139	238

Table 20 KNN (Bigram feature)

Confusion Matrix for Linear SVM (Bigram feature):

	Predicted 0	Predicted 1
Actual 0	375	17
Actual 1	279	98

Table 21 SVM (Bigram feature)

Confusion Matrix for Kernel SVM (Bigram feature):

	Predicted 0	Predicted 1
Actual 0	366	26
Actual 1	214	163

Table 22 K SVM (Bigram feature)

For Trigram:

Confusion Matrix for Logistic Regression (Trigram feature):

	Predicted 0	Predicted 1
Actual 0	329	63
Actual 1	161	216

Table 23 Logistic Regression (Trigram feature)

Confusion Matrix for Decision Tree (Trigram feature):

	Predicted 0	Predicted 1
Actual 0	242	150
Actual 1	162	215

Table 24 Decision Tree (Trigram feature)

Confusion Matrix for Random Forest (Trigram feature):

	Predicted 0	Predicted 1
Actual 0	358	34
Actual 1	199	178

Table 25 Random Forest (Trigram feature)

Confusion Matrix for Naive Bayes (Trigram feature):

	Predicted 0	Predicted 1
Actual 0	297	95
Actual 1	125	252

Table 26 Naive Bayes (Trigram feature)

Confusion Matrix for KNN (Trigram feature):

	Predicted 0	Predicted 1
Actual 0	254	138
Actual 1	137	240

Table 27 KNN (Trigram feature)

Confusion Matrix for Linear SVM (Trigram feature):

	Predicted 0	Predicted 1
Actual 0	389	3
Actual 1	340	37

Table 28 SVM (Trigram feature)

Confusion Matrix for Kernel SVM (Trigram feature):

	Predicted 0	Predicted 1
Actual 0	375	17
Actual 1	270	107

Table 29 K SVM (Trigram feature)

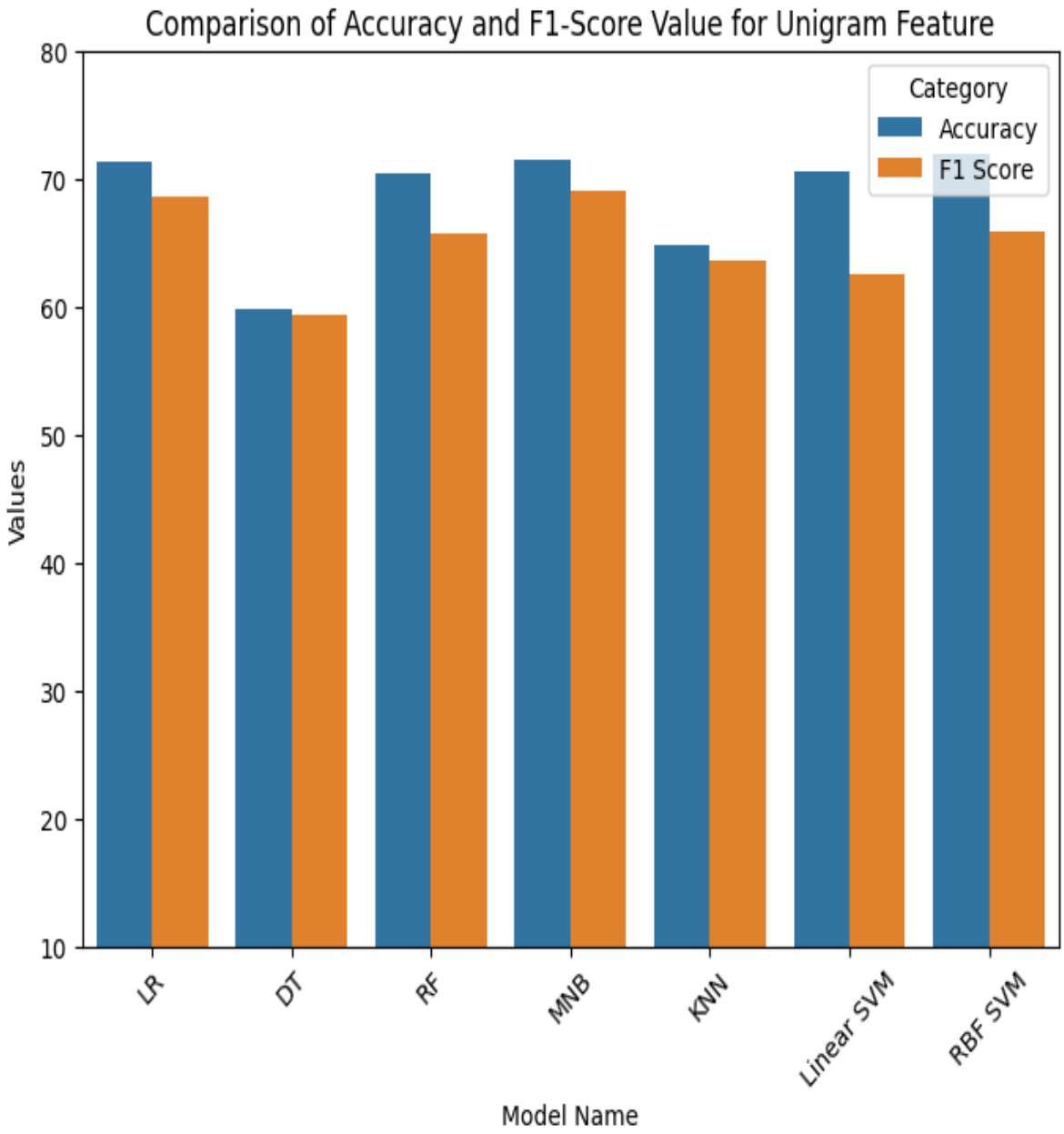


Figure 18 Comparison of Accuracy and F1-Score value for unigram feature

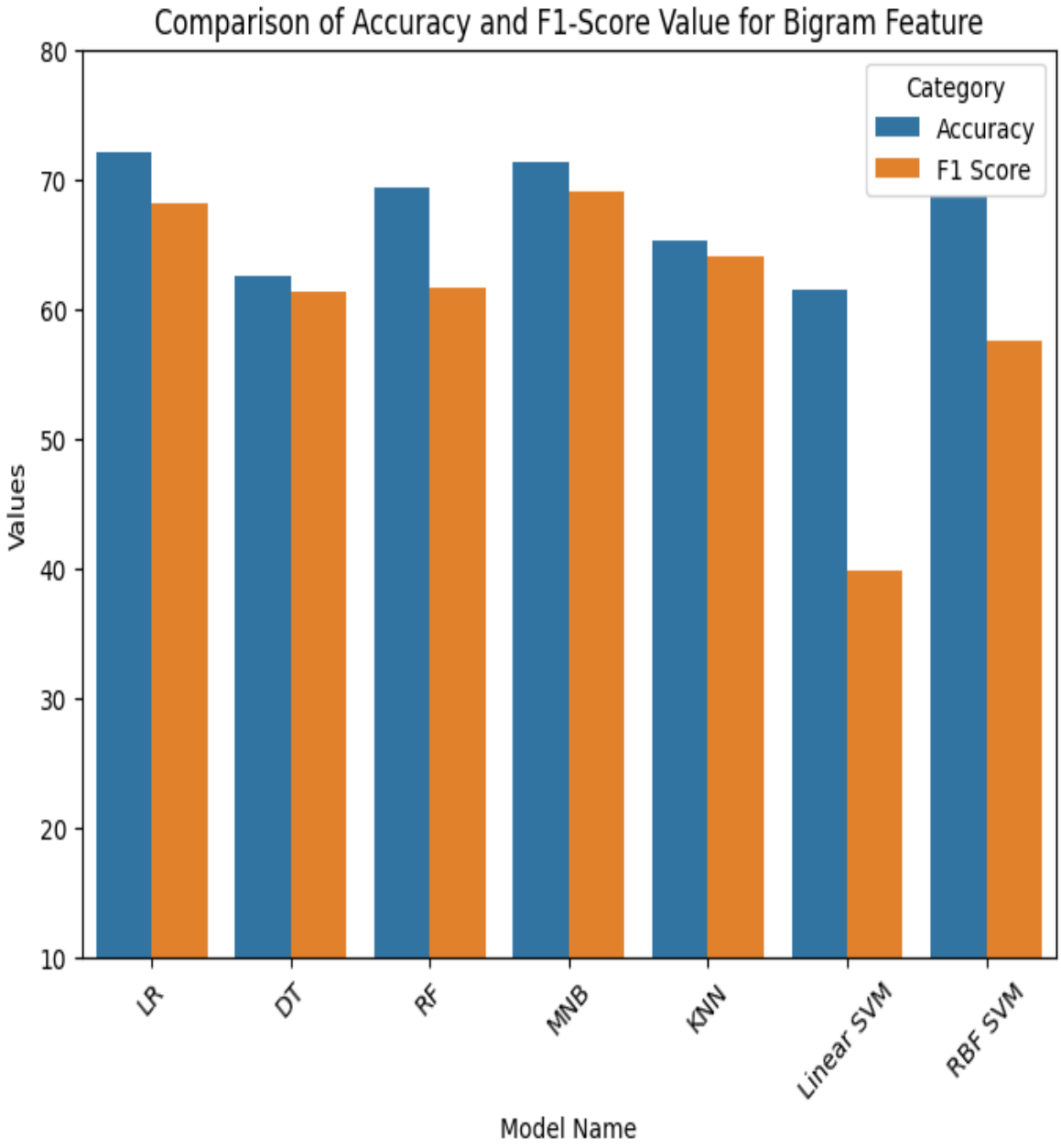


Figure 19 Comparison of Accuracy and F1-Score value for Bigram feature

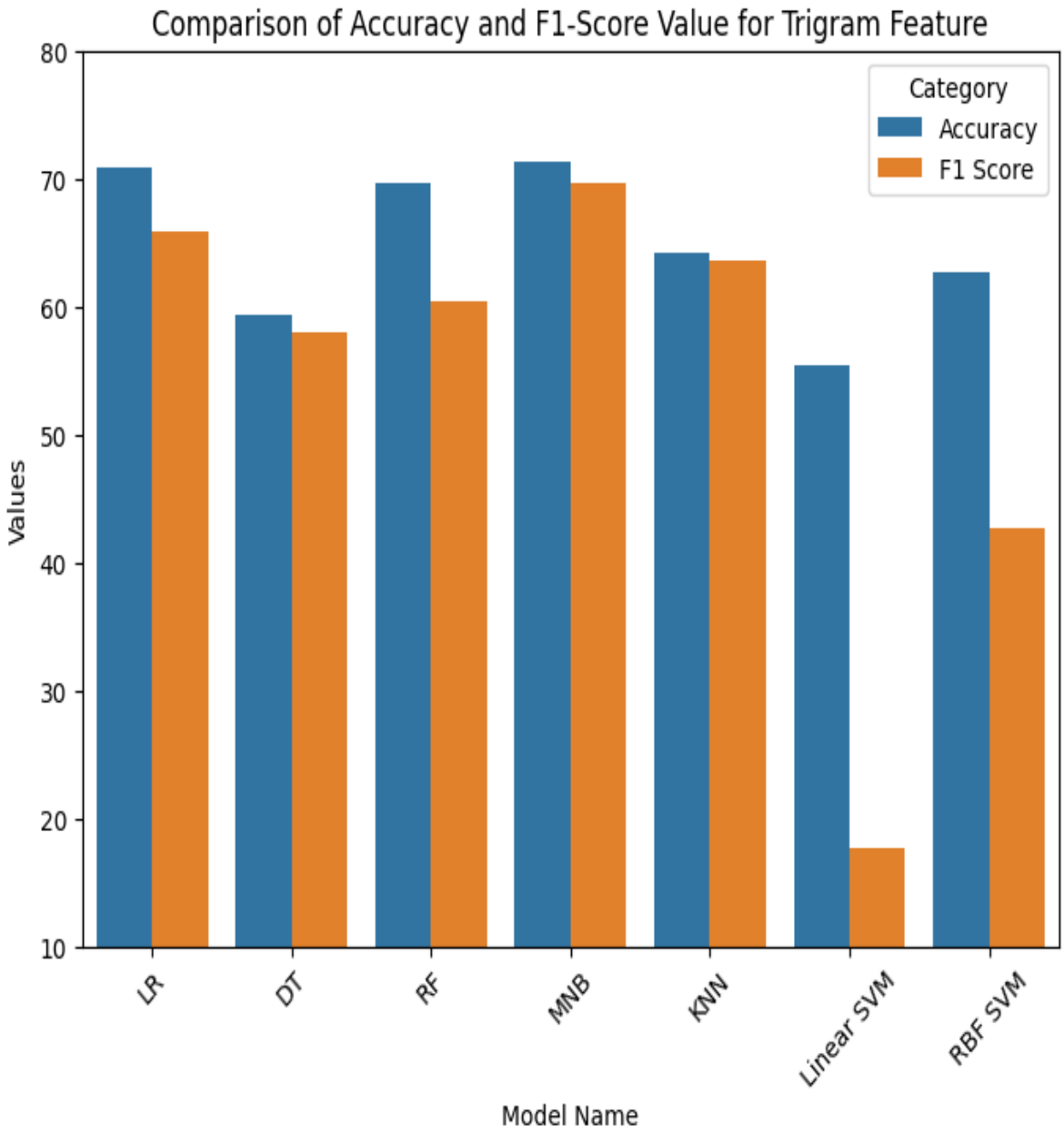


Figure 20 Comparison of Accuracy and F1-Score value for Trigram feature

CHAPTER V

CONCLUSION AND FUTURE WORKS

5.1 CONCLUSION

Since the sentiment analysis (SA) is growing its significance, in our work we try to classify such sentiments as positive, very positive, negative, and very negative. About 10851 raw data have been collected to conduct this research, all of it was raw data collected from Facebook, you tube, and twitter. We mainly try to find the polarization of the sentiments into two broad classes, either positive or negative sentiments. We conduct unigram, bigram, and trigram feature to the sentiments based on one word, two word, and three words respectively. For all the unigram, bigram, and trigram features we set out machine learning models namely LR, RF, SVM. Among them LR provides a high level of accuracy for all three category and it in above the 70%.

5.2 CONTRIBUTION OF THIS THESIS

a) In this thesis paper we successfully processed the raw data by removing unnecessary punctions, low length data, and encoding the sentiments as NumPy array. Since Bangla is a complex language in our work it was our main challenge to process the raw data into useful form. here in this paper, we shown that how to handle the natural language.

b) we applied different kinds of machine learning algorithms to make proper prediction about the polarity of public sentiments. Among the model's logistic regression and BRF support vector machine provide above 70% accuracy. In our work we carry out unigram, bigram, and trigram feature for one word, two words, and three words respectively.

5.3 LIMITATIONS AND FUTURE WORKS

Limitations:

1) Low amount of raw data.

2) Neutral sentiments are not classified.

3) Deep learning models have not been used in our work.

4) Low level of accuracy of our model.

Future works:

a) The use of deep learning models will provide a better result. Since the deep learning model provides superior tools to analysis natural language.

b) The high amount of raw data is also valuable to get high accuracy in ML models.

c) The neutral sentiments can also be classified.

References

[1] E. D. D. B. S. &. F. A. Cambria, "Affective computing and sentiment analysis.," *A practical guide to sentiment analysis (2017),* pp. 1-10, 2017.

[2] U. D. o. C. International Trade Administration, "eCommerce," International Trade Administration, Washington, DC 20230, 2022.

[3] BanglApedia, "Bangla langauge," p. 1, 17 June 2021.

[4] R. A. P. B. K. N. F. A. M. &. D. A. K. Tuhin, "An automated system of sentiment analysis from Bangla text using supervised learning techniques," *IEEE 4th International Conference on Computer and Communication Systems (ICCCS),* pp. 360-364, 2019.

[5] A. A. M. A. A. A. a. M. N. Hassan, "Sentiment analysis on bangla and romanized bangla text using deep recurrent models," *International Workshop on Computational Intelligence (IWCI),* pp. 51-56, 2016.

[6] S. a. C. W. Chowdhury, "Performing sentiment analysis in Bangla microblog posts," *International Conference on Informatics, Electronics & Vision (ICIEV),* pp. 1-6, 2014.

[7] M. A. M. F. R. a. A. F. Purba, "A hybrid convolutional long short-term memory (CNN-LSTM) based natural language processing (NLP) model for sentiment analysis of customer product reviews in Bangla," *Journal of Discrete Mathematical Sciences and Cryptography,* vol. 25(7), pp. 2111-2120, 2022.

[8] M. R. M. a. B. A. Munna, " Sentiment analysis and product review classification in e-commerce platform.," *23rd International Conference on Computer and Information Technology (ICCIT),* pp. 1-6, 2020.

[9] M. R. S. I. M. K. A. a. H. M. Khan, "Sentiment Analysis of COVID-19 Vaccination in Bangla Language with Code-Mixed Text from Social Media.," *12th International Conference on Electrical and Computer Engineering (ICECE),* pp. 76-79, 2022.

[10] N. a. A. M. Tripto, "Detecting multilabel sentiment and emotions from bangla youtube comments.," *International Conference on Bangla Speech and Language Processing (ICBSLP),* pp. 1-6, 2018.

[11] N. A. M. M. M. a. I. M. Bhowmik, "Bangla text sentiment analysis using supervised machine learning with extended lexicon dictionary.," *Natural Language Processing Research,* Vols. 1(3-4), pp. 34-45, 2021.

[12] K. a. R. M. Hasan, " Sentiment detection from bangla text using contextual valency analysis.," *17th international conference on computer and information technology (ICCIT),* pp. 292-295, 2014.

[13] M. H. M. A. M. M. M. N. A. a. F. M. Shafin, "Product review sentiment analysis by using nlp and machine learning in bangla language," *23rd International Conference on Computer and Information Technology (ICCIT),* pp. 1-5, 2020.

[14] S. a. C. D. Sharmin, "Attention-based convolutional neural network for Bangla sentiment analysis," *Ai & Society,* vol. 36, pp. .381-396., 2021.

[15] "]Asimuzzaman, M., Nath, P.D., Hossain, F., Hossain, A. and Rahman, R.M., 2017, July. Sentiment analysis of bangla microblogs using adaptive neuro fuzzy system.," *13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD),* pp. 1631-1638, 2017.

[16] N. a. A. M. Tripto, " Detecting multilabel sentiment and emotions from bangla youtube comments.," *International Conference on Bangla Speech and Language Processing (ICBSLP),* pp. 1-6, 2018.

[17] M. F. S. E. F. S. T. a. A. M. Soumik, "Employing machine learning techniques on sentiment analysis of google play store bangla reviews.," *22nd International Conference on Computer and Information Technology (ICCIT),* pp. 1-5, 2019 .

[18] M. R. M. a. B. A. Munna, " Sentiment analysis and product review classification in e-commerce platform.," *23rd International Conference on Computer and Information Technology (ICCIT),* pp. 1-6, 2020 .

[19] S. M. R. a. I. M. Salehin, "A comparative sentiment analysis on Bengali Facebook posts.," *In Proceedings of the international conference on computing advancements* , pp. 1-8, 2020.

[20] E. Cambria, "Affective Computing and Sentiment Analysis," *IEEE Intelligent Systems,* vol. 31(2), p. 102–107, 2016.

[21] M. a. R. T. Khatun, "A Machine Learning Approach for Sentiment Analysis of Book Reviews in Bangla Language," *6th International Conference on Trends in Electronics and Informatics (ICOEI),* pp. 1178-1182, 2022.

[22] M. H. F. U. U. T. A. K. A. a. F. A. Junaid, "Bangla food review sentimental analysis using machine learning," *12th Annual Computing and Communication Workshop and Conference (CCWC),* pp. 0347-0353, 2022.

[23] M. R. S. I. M. K. A. a. H. M. Khan, "Sentiment Analysis of COVID-19 Vaccination in Bangla Language with Code-Mixed Text from Social Media," *12th International Conference on Electrical and Computer Engineering (ICECE),* pp. 76-79, 2022.

[24] M. M. Z. B. Z. R. A. H. A. a. A. F. Ahmed, "Cyberbullying detection using deep neural network from social media comments in bangla language," *arXiv preprint arXiv,* p. 2106.04506, 2021.

[25] https://en.wikipedia.org/wiki/Logistic_regression#/media/File:Exam_pass_logistic_curve.svg.

[26] D. tree, "https://www.geeksforgeeks.org/decision-tree/".

[27] knn, "https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning".

[28] lsvm, "https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm#:~:text=Linear%20SVM%3A%20Linear%20SVM%20is,called%20as%20Linear%20SVM%20classifier".

[29] rbf, "https://www.geeksforgeeks.org/major-kernel-functions-in-support-vector-machine-svm/".

ORIGINALITY REPORT

16%	12%	8%	8%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.webology.org Internet Source	3%
2	www.slideshare.net Internet Source	2%
3	Submitted to University of Hertfordshire Student Paper	1%
4	dspace.daffodilvarsity.edu.bd:8080 Internet Source	1%
5	rd.springer.com Internet Source	1%
6	dspace.bracu.ac.bd:8080 Internet Source	1%
7	www.coursehero.com Internet Source	<1%
8	Ton Duc Thang University Publication	<1%
9	www.researchgate.net Internet Source	<1%