THE CLASSIFICATION OF YOUTUBE BANGLA COMMENTS USING SENTIMENT ANALYSIS

By
Md Mofazzul Islam

161-35-1439

This Report Presented in Fulfillment of the Requirements for the Degree of B.Sc. in Software Engineering

DEPARTMENT OF SOFTWARE ENGINEERING

DAFFODIL INTERNATIONAL UNIVERSITY

Approval

This thesis, project, or internship was completed by Md Mofazzul Islam (ID: 161-35-1439) and was titled “THE CLASSIFICATION OF YOUTUBE BANGLA COMMENTS USING SENTIMENT ANALYSIS." It has been approved for partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and for its presentation and content.

BOARD OF EXAMINERS

Chairman

Dr. Imran Mahmud
Associate Professor and Head
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 1

Dr. Md. Mostafijur Rahman
Associate Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 2

Afsana Begum
Lecturer (Senior)
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

External Examiner

Professor Dr. Mohammed Nasir Uddin
Professor and Chairman
Department of Computer Science and Engineering
Jagannath University

THESIS DECLARATION

This statement is declaring that the study document being referred to was created by the person making the statement, under the supervision of Md Shohel Arman, and that it is original work and has not been submitted elsewhere for a Bachelor's degree or any other graduation program.

Supervised by

Md Shohel Arman

Associate Professor

Department of Software Engineering

Daffodil International University

Submitted by

Md Mofazzul Islam

ID: 161-35-1439

Department of Software Engineering

Daffodil International University

ACKNOWLEDGEMENT

It's great that you are expressing gratitude for the support and guidance you received during your studies. Acknowledging the contributions of others is an important aspect of any successful endeavor, and it's clear that you have a lot of people to thank for helping you achieve your goals. Keep up the good work!

TABLE OF CONTENTS

THESIS DECLARATION iii

ACKNOWLEDGEMENT iv

LIST OF FIGURES vi

LIST OF TABLES vii

ABSTRACT vii

Chapter 1 2

Introduction 2

1.1 Background 2

1.2 Problem Statement 3 1.3

Research Questions 4 1.4

Research Objective 4

1.5 Research Scope 4

1.7 Thesis Organization 5

CHAPTER 2 6

LITERATURE REVIEW 6

2.1 Previous literature 6

2.5 Summary 8

CHAPTER 3 9

RESEARCH METHODOLOGY 9

3.1 Collection of bangla comments 10

3.2 Data Preparation 10
3.3 remove stop words 11

3.4 eliminating irrelevant words 12

3.5 feature extractions 12

3.6 calculating tf-idf 13

3.7 classified fitting 15

3.8 MACHINE LEARNING BASED SYSTEM 15

3.8.1 Naïve Bayes (MNB) .... 16 **3.8.2**.

STOCHASTIC GRADIENT DESCENT (SGD) 17

3.8.2 RANDOM FOREST 18 3.9

SUMMARY 18

Chapter 4 19

Results and discussion 19

4.1 collecting dataset 19

4.2 CLEANING RAW DATA 21

4.3 FEATURE SELECTION AND EXTRACTION 22

4.4 BUILDING A MODEL 22

4.5 EVALUATION METRICS 23

$ff1=2\times pppppppppppppppppp \quad \times pppppprrrrrpppppppppppppppppp+pppppprrrrr$
.... 25

4.6 Visualization 25

4.6.1 unigram 25

4.6.2 bi-gram 27

4.6.3 trigram 29

4.7 COMPARE PRECISION 31

4.8 COMPARE RECALL.. 31

4.9 F1 Score 32

4.10 COMPARE ALGORITHM ACCURACY 32

4.11 performance evalution 32

*rrppaaaarrrrppppppppaappaappppppaarraappaappppppppppaappaapp*255114*ppppaar*
*raappaapp*111320 32

4.12 prediction result 32

4.13 SUMMARY 33
CHAPTER 5 34

CONCLUSIONS AND RECOMMENDATIONS 34

5.1 introduction 34

5.2 conclusions 34

5.3 Recommendations.. 34

5.4 Future directions 35

LIST OF FIGURES

Figure 1: methodology of study 9
Figure 2: all stop words.. 12
Figure 3: feature extraction model 16
Figure 4: class of datasets 19
Figure 5: length frequency distribution 20
Figure 6: data statistics 21
Figure 7: cleaning raw data for positive 21
Figure 8: cleaning raw data for negative 22
Figure 9: f1 score for unigram feature 26
Figure 10: roc curve for unigram 27
Figure 11: f1 score for bigram feature.. 28
Figure 12: roc curve for bigram feature 28
Figure 13: f1 score for trigram feature 30
Figure 14:roc curve for trigram feature 31

LIST OF TABLES

Table 1: all positive data 10
Table 2: all negative data 11
Table 3: tf-idf result for positive data 14
Table 4: tf-idf result for negative data 15
Table 5: evaluation metrics 23
Table 6: performance table for unigram 25
Table 7: performance table for bigram 27
Table 8: performance table for trigram 29
Table 9: compare precision 31
Table 10: compare recall 31
Table 11: f1 score 32
Table 12: compare algorithm accuracy 32

ABSTRACT

In this paper, the authors present a machine learning-based approach for sentiment analysis of Bangla language comments on YouTube. They propose an algorithm to classify comments as positive or negative and build models to extract the emotion of the comments. They evaluate the performance of the model using a new dataset of Bangla comments from various YouTube videos. They compare the performance of different algorithms such as Multinomial Naive Bayes (MNB), Stochastic Gradient Descent (SGD), Random Forest (RF), Logistic Regression (LR), Support Vector Machine (SVM), AdaBoost, and XGBoost. The results show that MNB achieves the best accuracy of 70.88%. The paper suggests that there is a need for more research in the field of sentiment analysis of Bangla language.

CHAPTER 1

INTRODUCTION

1.1 BACKGROUND

Sentiment analysis is a key area of research in natural language processing because of its wide range of applications, including opinion mining, emotion extraction, and pattern forecasting in social media. The rise of online social networking sites such as Facebook, Twitter, and Myspace has increased the demand for sentiment analysis research that can determine people's opinions, evaluations, attitudes, and emotions from text. Most of the research on sentiment analysis has been conducted in the English language, but there is limited work in other languages such as Bangla. Sentiment analysis techniques are often developed for English and are difficult to adapt to other languages. Current sentiment analysis systems are typically designed to work in a single language, usually English. However, effectively mining global opinion requires text analysis in a variety of local languages. While sentiment analysis systems specific to each language can be built, these approaches are labor-intensive and complicated by the lack of semantic resources such as WordNet for many languages. The availability of large amounts of online data and the advancement of machine learning techniques have accelerated the development of a wide range of methods for sentiment analysis and emotion extraction from text in various languages including English, French, Arabic, and more.

Number of users in Bangladesh. Sentiment analysis in the Bangla language is an understudied area, despite the large number of speakers and internet users in the country. The use of online platforms and social media is increasing, with a high number of active Facebook users and a growing presence on platforms like YouTube. This presents an opportunity to study sentiment analysis in the Bangla language to better understand the emotions and thoughts expressed by users on these platforms.

This passage is discussing the use of YouTube in Bangladesh and the potential for using machine learning techniques to analyze sentiments expressed in the comments section of YouTube videos in the Bangla language. The author notes that deep learning has shown promising results in sentiment analysis, and proposes using various machine learning techniques, such as Stochastic Gradient Descent, K-Nearest Neighbor, Naive Bayes, logistic regression, Random Forest, Decision tree, Support Vector Machine, and Xgboost, to build a sentiment analyzer for Bangla comments. The main challenge is obtaining enough training data to build the model. The author also mentions that previous work has been done with some of these techniques, and the goal is to compare different algorithms to determine the best performing sentiment analysis result.

Motivation of the Research

It is clear that a large number of people use the internet, with many of them accessing and viewing videos on YouTube. Many of these users also leave comments on the videos, and the analysis of these comments can provide insight into the reactions and opinions of Bengali people on YouTube

1.2 PROBLEM STATEMENT

It sounds like you have conducted a literature review and found that there is a lack of research on sentiment analysis of Bangla comments, and that previous studies have mostly used Twitter data and English comments. You have also noted that the most common machine learning algorithms used in previous studies are SVM and RF, while you have chosen to use MNB for sentiment detection. Additionally, you have chosen to use Tf-Idf with n-gram tokens to create a list of features for each sentence.

1.3 RESEARCH QUESTIONS

The research question was

- RQ1: To study how to classify or analyzing Bangla YouTube comments using some classifier algorithm?
- RQ2: Compare different Classifier algorithm to know better classification result.

1.4 RESEARCH OBJECTIVE

The objectives of this research are

- To find out the best performing Neural Network method.
- To know better classification result.

1.5 RESEARCH SCOPE

1. The study "Detecting Multilevel Sentiment and Emotions from Bangla YouTube Comments" by Tripto, Nafis & Ali, Mohammed Eunus (2018) used English, Bangla and Romanized comments and applied support vector machine and naïve Bayes algorithms to detect sentiment and emotions. They found that the final outcome was better when comparing all algorithms.

1.7 THESIS ORGANIZATION

It sounds like the structure of your document includes an introduction, a literature review, a methodology section, results, and a discussion/conclusion section. In the next chapter, you plan to delve deeper into the existing research on the topic, identify any gaps in the literature, and propose your own research methodology. In chapter three, you will describe the methodology you have chosen in more detail. In chapter four, you will present and analyze the results of your research. Finally, in chapter five, you will discuss the implications of your findings, including any assumptions, limitations, and suggestions for future research.

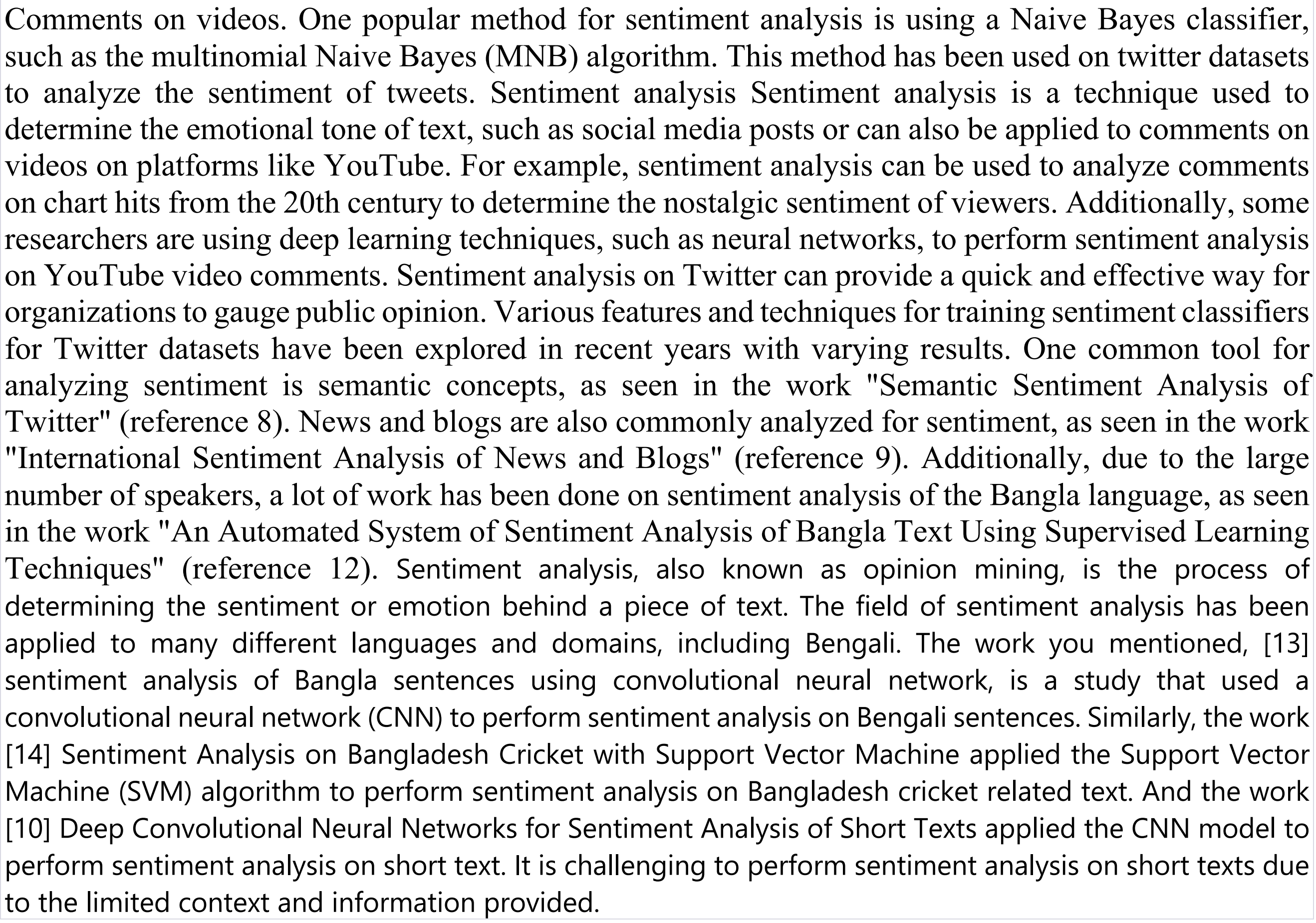
.

CHAPTER 2

LITERATURE REVIEW

2.1 PREVIOUS LITERATURE

Sentiment analysis is a natural language processing technique used to determine the sentiment or emotion expressed in text, such as whether it is positive, negative, or neutral. It is often used by organizations to analyze customer feedback and social media data to understand customer sentiment towards their brand or products. Sentiment analysis models can focus on polarity (positive, negative, neutral) or emotions (happy, sad, angry, etc.). Different machine learning algorithms, such as Naive Bayes, logistic regression, CNN, LSTM, SVM, Random Forest, and SGD, can be used for sentiment analysis. While a lot of sentiment analysis work has been done with English language text, some work has been done with other languages, such as Bangla. Some research has focused on detecting multilevel sentiment and emotion from Bangla YouTube comments, and using NLP to analyze sentiment in comments on videos in different languages, such as English and Italian.



Comments on videos. One popular method for sentiment analysis is using a Naive Bayes classifier, such as the multinomial Naive Bayes (MNB) algorithm. This method has been used on twitter datasets to analyze the sentiment of tweets. Sentiment analysis Sentiment analysis is a technique used to determine the emotional tone of text, such as social media posts or can also be applied to comments on videos on platforms like YouTube. For example, sentiment analysis can be used to analyze comments on chart hits from the 20th century to determine the nostalgic sentiment of viewers. Additionally, some researchers are using deep learning techniques, such as neural networks, to perform sentiment analysis on YouTube video comments. Sentiment analysis on Twitter can provide a quick and effective way for organizations to gauge public opinion. Various features and techniques for training sentiment classifiers for Twitter datasets have been explored in recent years with varying results. One common tool for analyzing sentiment is semantic concepts, as seen in the work "Semantic Sentiment Analysis of Twitter" (reference 8). News and blogs are also commonly analyzed for sentiment, as seen in the work "International Sentiment Analysis of News and Blogs" (reference 9). Additionally, due to the large number of speakers, a lot of work has been done on sentiment analysis of the Bangla language, as seen in the work "An Automated System of Sentiment Analysis of Bangla Text Using Supervised Learning Techniques" (reference 12). Sentiment analysis, also known as opinion mining, is the process of determining the sentiment or emotion behind a piece of text. The field of sentiment analysis has been applied to many different languages and domains, including Bengali. The work you mentioned, [13] sentiment analysis of Bangla sentences using convolutional neural network, is a study that used a convolutional neural network (CNN) to perform sentiment analysis on Bengali sentences. Similarly, the work [14] Sentiment Analysis on Bangladesh Cricket with Support Vector Machine applied the Support Vector Machine (SVM) algorithm to perform sentiment analysis on Bangladesh cricket related text. And the work [10] Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts applied the CNN model to perform sentiment analysis on short text. It is challenging to perform sentiment analysis on short texts due to the limited context and information provided.

2.5 SUMMARY

It sounds like your study aims to compare the performance of various machine learning algorithms, specifically Naive Bayes, Random Forest, and SGD, on a dataset of YouTube comments in Bengali language. The goal is to classify the comments into two categories, and evaluate which algorithm performs best in this task.

CHAPTER 3

RESEARCH METHODOLOGY

The methodology for this study involves using the Multinomial Naive Bayes (MNB) algorithm to analyze YouTube Bangla comments. The comments are collected from various Bangla videos and the dataset is preprocessed and feature extraction is performed. The MNB algorithm is compared with other algorithms to determine which one works best for this task.

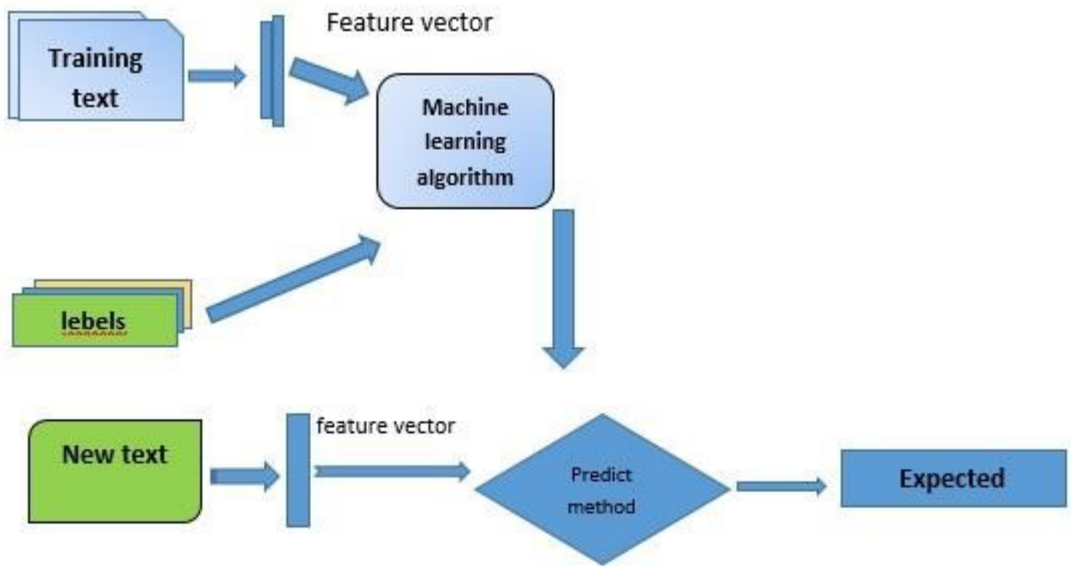


FIGURE 1: METHODOLOGY OF STUDY

3.1 COLLECTION OF BANGLA COMMENTS

It sounds like you have developed a scraper to collect comments from Bengali videos on YouTube, and have stored the data in a CSV file. Is there a specific question or task you would like me to assist with regarding this data collection process?

3.2 DATA PREPARATION

Information Collection is a colossal and troublesome cycle. To collect the total amount of data and measure it. Into a correct structure is a major undertaking. And our data processing has many steps as describe in below. Class name: positive Number

of document: 2000

Number of words: 14556

Number of unique words: 5518

Most frequent words

ভাই	494
ভাে লা	228
ি িভডও	186
গান	117
টা	117
কথা	92
লাে গ	80
হেয়	77
সুর	72
ি েেরা	69

TABLE 1: ALL POSITIVE DATA

Class name: negative

Number of document: 1999

Number of words:14659

number of unique words: 6063

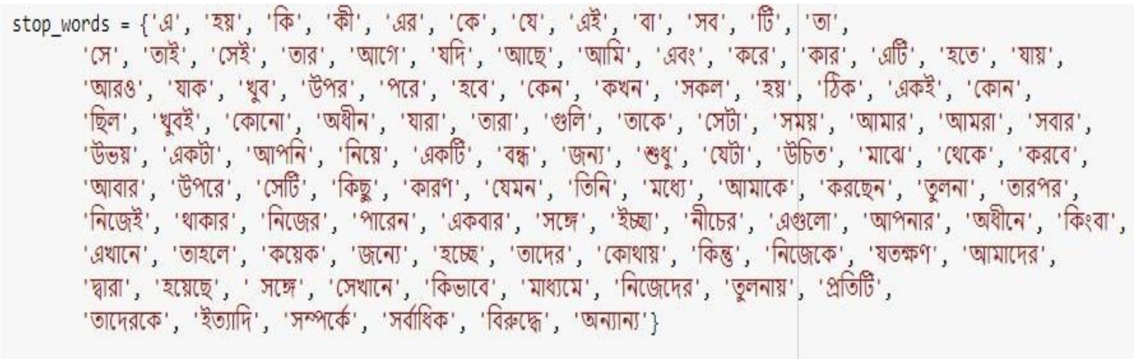
Most frequent words

ভাই	302
ি িভডও	93
গান	91
হেয়	82
টা	82
ভাে লা	75
ি েেয়	65
ি েেরা	64
সাে থ	58
কথা	52

TABLE 2: ALL NEGATIVE DATA

3.3 REMOVE STOP WORDS

After collecting Bangla comments our first undertaking is to eliminate stop words and accentuation mark from all comments. And meaningless symbol like (. , ” „ | ‟ { } ( ) !, ; ?) have been remove to clean our dataset. And make our dataset noise free.



```
stop_words = {'এ', 'হয়', 'কি', 'কী', 'এর', 'কে', 'যে', 'এই', 'বা', 'সব', 'টি', 'তা',
        'সে', 'তাই', 'সেই', 'তার', 'আগে', 'যদি', 'আছে', 'আমি', 'এবং', 'করে', 'কার', 'এটি', 'হতে', 'যায়',
        'আরও', 'যাক', 'খুব', 'উপর', 'পরে', 'হবে', 'কেন', 'কখন', 'সকল', 'হয়', 'ঠিক', 'একই', 'কোন',
        'ছিল', 'খুবই', 'কোনো', 'অধীন', 'যারা', 'তারা', 'গুলি', 'তাকে', 'সেটা', 'সময়', 'আমার', 'আমরা', 'সবার',
        'উভয়', 'একটা', 'আপনি', 'নিয়ে', 'একটি', 'বন্ধ', 'জন্য', 'শুধু', 'যেটা', 'উচিত', 'মাঝে', 'থেকে', 'করবে',
        'আবার', 'উপরে', 'সেটি', 'কিছু', 'কারণ', 'যেমন', 'তিনি', 'মধ্যে', 'আমাকে', 'করছেন', 'তুলনা', 'তারপর',
        'নিজেই', 'থাকার', 'নিজের', 'পারেন', 'একবার', 'সঙ্গে', 'ইচ্ছা', 'নীচের', 'এগুলো', 'আপনার', 'অধীনে', 'কিংবা',
        'এখানে', 'তাহলে', 'কয়েক', 'জন্যে', 'হচ্ছে', 'তাদের', 'কোথায়', 'কিন্তু', 'নিজেকে', 'যতক্ষণ', 'আমাদের',
        'দ্বারা', 'হয়েছে', ' সঙ্গে', 'সেখানে', 'কিভাবে', 'মাধ্যমে', 'নিজেদের', 'তুলনায়', 'প্রতিটি',
        'তাদেরকে', 'ইত্যাদি', 'সম্পর্কে', 'সর্বাধিক', 'বিরুদ্ধে', 'অন্যান্য'}
```

FIGURE 2: ALL STOP WORDS

3.4 ELIMINATING IRRELEVANT WORDS

It sounds like you have created a list of "stop words" that will be removed from text before it is analyzed or classified. Stop words are common words that do not provide much information and are often removed to improve text processing. Examples of stop words include pronouns, conjunctions, and prepositions. It is also common to remove numbers, single letter words and punctuations as these are considered insignificant in the context of text classification.

3.5 FEATURE EXTRACTIONS

In summary, an n-gram is a sequence of n items, typically words that are extracted from a corpus of text. Unigrams are n-grams of size 1, bigrams are n-grams of size 2, and trigrams are n-grams of size 3. The purpose of using n-grams is to convert text into a sequence of words, where the term frequency-inverse document frequency (tf-df) can be calculated easily.

3.6 CALCULATING TF-IDF

Yes, that is correct. TF-IDF is a method used to measure the importance of a word in a document or corpus. It combines the term frequency (TF), which represents the number of times a word appears in a document, with the inverse document frequency (IDF), which represents the rarity of a word across the entire corpus. By combining these two factors, TF-IDF can determine the weight of a word in a document, and it is often used in search engines, digital libraries, and content-based recommender systems. Additionally, it is useful for text mining, natural language processing and information retrieval.

$$TTTT(aa) = \frac{ppaannnnpppp\ ppffaappnnpppp\ aappppnn\ aa\ rrpppppppprrpppp\ pppp\ ddppppaannppppaa}{aappaarrrr\ ppaannnnpppp\ ppff\ aappppnnpp\ pppp\ aahpp\ ddppppaannppppaa}$$

In simpler terms, the inverse document frequency (IDF) is a measure of how informative a word is within a collection of documents. It is calculated by taking the logarithm of the ratio of the total number of documents to the number of documents containing the word. The more rare a word is across the collection of documents, the higher its IDF value will be

$$IIIIT(aa) = \log pp \frac{aappaarrrr\ ppaannpppp\ ppff\ ddppppaannppppaapp}{ppaannnnpppp\ ppff\ ddppppaannppppaapp\ wwppaah\ aappppnn\ aa\ pppp\ ppaa}$$

TF-IDF is a commonly used method for determining the importance of words in a document or dataset. It stands for "term frequency-inverse document frequency" and is a measure of how often a word appears in a document, but also takes into account the number of times the word appears in the entire dataset. This allows it to identify words that are important or significant within a specific document, but not necessarily throughout the entire dataset. In your case, you are using TF-IDF to find significant words for every report in your dataset, which consists of 40000 lines

আা◌ে ি◌িম আমা◌ে ◌ের সবা◌েইক হকরানাভাইরাস হ ◌েথক হ ◌েফাজি কর-িআমন

	Tf idf

হকরানাভাইরাস	0.584863
হ েফাজি	0.537729
সবােইক	0.475421
আাে	0.377835

TABLE 3: TF-IDF RESULT FOR POSITIVE DATA

েআসল ভাই িিে ও আমার গাি ল হ েওয়ার অভাস ি ু ব অ িবুও গাি ল হ েওয়ার হই হে।

	Tf idf
হ েওয়ার	0.523979
গাি ল	0.498762
িবুও	322300
অভাস	0.296325
ুি ব	0.287964
অ	0.261990

হই	0.252158
আোসল	0.244377
ভাই	0.100654

TABLE 4: TF-IDF RESULT FOR NEGATIVE DATA

3.7 CLASSIFIED FITTING

It sounds like you are using various machine learning algorithms to analyze comments or text data. Naive Bayes, logistic regression, stochastic gradient descent, random forest, and k-nearest neighbors are all common algorithms used in natural language processing and text classification tasks. By "importing" and "fitting" the algorithm, you are likely using a library or package in a programming language like Python to apply the algorithm to your data.

3.8 MACHINE LEARNING BASED SYSTEM

Machine learning can be used to classify content by analyzing past data and identifying patterns between the content and its associated output (such as labels). This is done by converting the content into a mathematical representation called a vector, which is often done using a technique called bag of words. The vector represents the frequency of words in a predefined vocabulary. For example, if the vocabulary contains the words {this, is, the, not, stunning, terrible, basketball}, the content "This is wonderful" would be represented by the vector (1, 1, 0, 0, 1, 0, 0). The machine learning algorithm is then trained on this data to create a classification model.

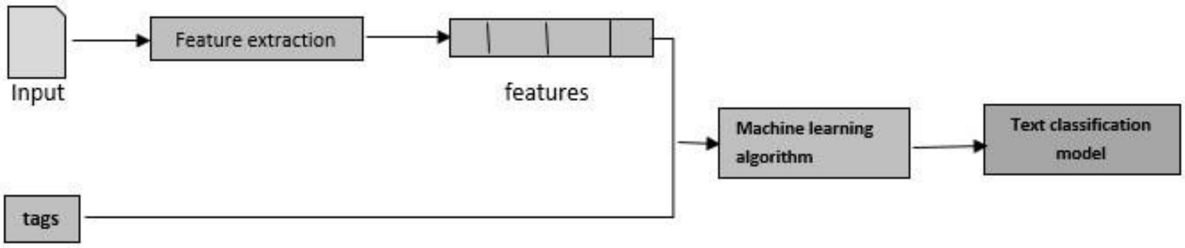


FIGURE 3: FEATURE EXTRACTION MODEL

3.8.1 Naïve Bayes (MNB)

In simple terms, Naive Bayes is a family of algorithms used for classification tasks, and one of its members is Multinomial Naive Bayes (MNB). MNB is particularly useful when working with limited data and computational resources. It relies on Bayes' Theorem, which helps calculate the probability of an event occurring based on the probability of individual events. To classify a piece of text, MNB looks at the probabilities of certain words appearing in a given class and uses this information to determine the likelihood of the text belonging to that class. Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c). Look at the equation below:

$$PP(ppcxx) = \frac{PP(xxcpp)PP(pp)}{pp(xx)}$$

Above

- *P*(*c*|*x*) is the posterior probability of *class* (c, *target*) given *predictor* (x, *attributes*).
- *P*(*c*) is the prior probability of *class*.
- *P*(*x*|*c*) is the likelihood which is the probability of *predictor* given *class*.
- *P*(*x*) is the prior probability of *predictor*.

3.8.2. STOCHASTIC GRADIENT DESCENT (SGD)

Stochastic gradient descent is a variation of gradient descent that is used for large datasets. Instead of updating the coefficients after processing all the training examples, the coefficients are updated after each training example. The order of the training examples is randomized to avoid getting stuck in local minima. The update formula for the coefficients is the same as in standard gradient descent, but the cost function is calculated for one training example at a time.

3.8.2 RANDOM FOREST

Yes, that is correct. Random Forest is a type of ensemble learning algorithm that builds multiple decision trees and combines their outputs to improve the prediction accuracy of the model. It is commonly used for both regression and classification problems, with the difference being that in regression, the target variable is continuous, and in classification, the target variable is categorical. Additionally, Random Forest algorithm is known for its simplicity and high accuracy.

3.9 SUMMARY

In this part we have discuss our research methodology and approach about data collection, information cleaning and how we need to implement classification models. Likewise, how we can distinguish best model for YouTube Bangla comments.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 COLLECTING DATASET

The chapter is discussing the process of collecting and analyzing data for a research project. The data is being gathered from Bangla YouTube videos in various categories such as drama, news, comedy, songs, movie and review videos. A total of 4000 comments were collected, with 2000 being labeled as positive and 2000 as negative. The data was stored in a CSV file and labeled according to its class.

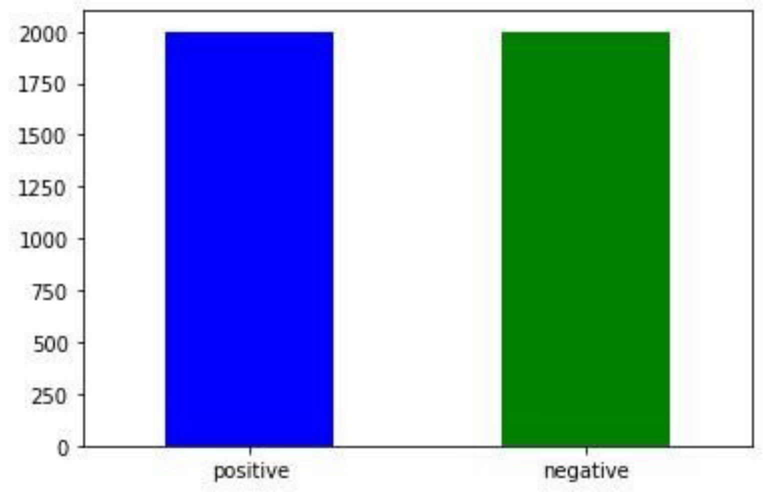
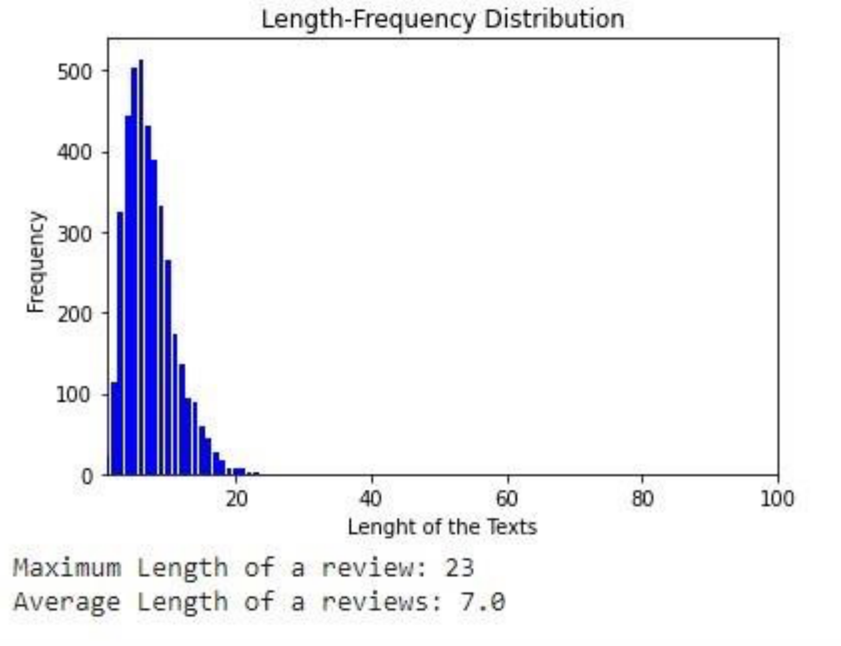


FIGURE 4: CLASS OF DATASETS



```
Maximum Length of a review: 23
Average Length of a reviews: 7.0
```

FIGURE 5: LENGTH FREQUENCY DISTRIBUTION

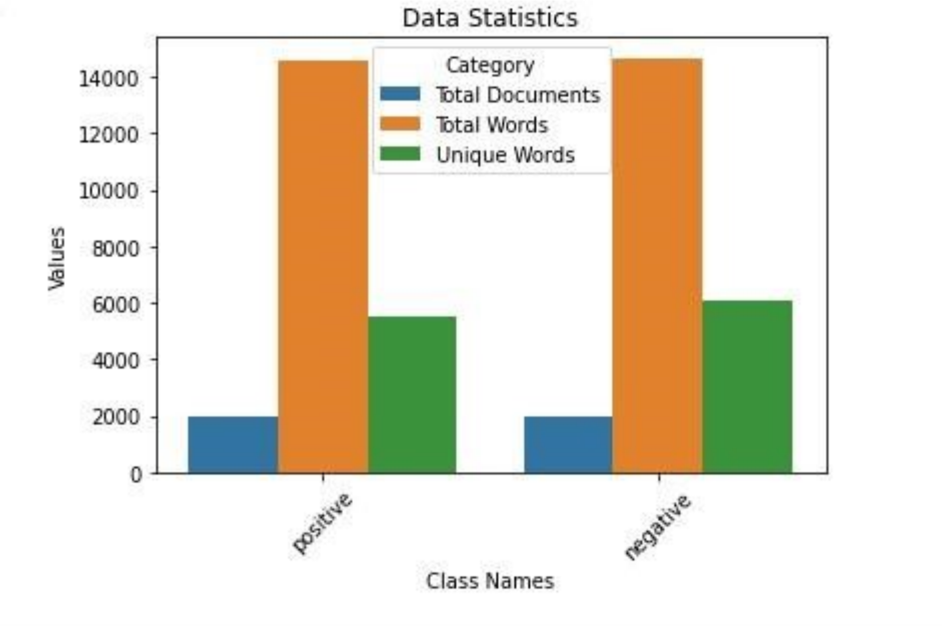


FIGURE 6: DATA STATISTICS

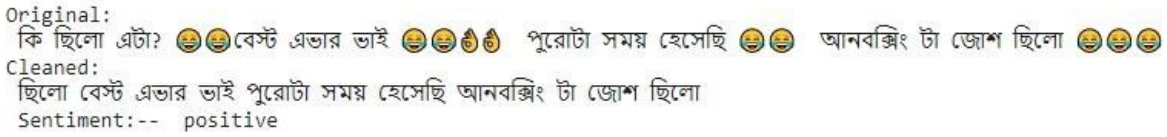
4.2 CLEANING RAW DATA

We use a content record to be useful of our data pre-preparing task.

- Remove all html tags.

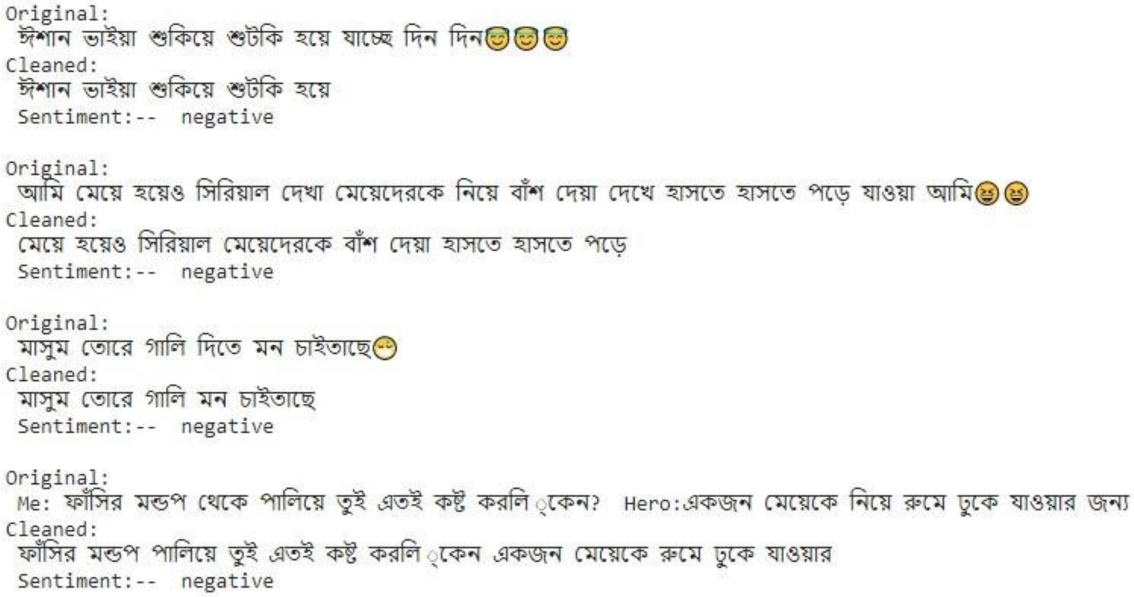
- Remove all types of emoji's

- Remove unnecessary spaces from text



```
Original:
 কি ছিলো এটা? 😂😂বেস্ট এভার ভাই 😂😂👌👌  পুরোটা সময় হেসেছি 😂😂  আনবক্সিং টা জোশ ছিলো 😂😂😂
Cleaned:
 ছিলো বেস্ট এভার ভাই পুরোটা সময় হেসেছি আনবক্সিং টা জোশ ছিলো
 Sentiment:--  positive
```

FIGURE 7: CLEANING RAW DATA FOR POSITIVE



```
Original:
 ঈশান ভাইয়া শুকিয়ে শুটকি হয়ে যাচ্ছে দিন দিন😊😊😊
Cleaned:
 ঈশান ভাইয়া শুকিয়ে শুটকি হয়ে
 Sentiment:--  negative

Original:
 আমি মেয়ে হয়েও সিরিয়াল দেখা মেয়েদেরকে নিয়ে বাঁশ দেয়া দেখে হাসতে হাসতে পড়ে যাওয়া আমি😆😆
Cleaned:
 মেয়ে হয়েও সিরিয়াল মেয়েদেরকে বাঁশ দেয়া হাসতে হাসতে পড়ে
 Sentiment:--  negative

Original:
 মাসুম তোরে গালি দিতে মন চাইতাছে😁
Cleaned:
 মাসুম তোরে গালি মন চাইতাছে
 Sentiment:--  negative

Original:
 Me: ফাঁসির মন্ডপ থেকে পালিয়ে তুই এতই কষ্ট করলি ্কেন?  Hero:একজন মেয়েকে নিয়ে রুমে ঢুকে যাওয়ার জন্য
Cleaned:
 ফাঁসির মন্ডপ পালিয়ে তুই এতই কষ্ট করলি ্কেন একজন মেয়েকে রুমে ঢুকে যাওয়ার
 Sentiment:--  negative
```

FIGURE 8: CLEANING RAW DATA FOR NEGATIVE

4.3 FEATURE SELECTION AND EXTRACTION

The stage you are describing is the feature selection and extraction stage in a classifying approach. It involves selecting the most relevant features and extracting them for use in the classification process. Word count is often used as a feature in this stage, and the order in which features are chosen can affect the overall classification performance.

4.4 BUILDING A MODEL

Yes, that is correct. In machine learning, it is common practice to divide the dataset into three parts: training, validation, and test. The training data is used to train the model, the validation data is used to tune the hyper parameters, and the test data is used to evaluate the performance of the model. The scikit-learn library, or sklearn, is a popular library for machine learning in Python and it provides a variety of classifiers that can be used for different types of problems. The classifiers are used to classify the data into different classes or categories.

4.5 EVALUATION METRICS

Yes, that is correct. The accuracy of a model is typically measured by comparing the predicted values to the actual values. This can be done using metrics such as mean absolute error (MAE), mean squared error (MSE), or coefficient of determination (R-squared). Additionally, assessment indicators, such as precision, recall, and F1-score, are used to evaluate the performance of a model and identify areas that may need improvement.

	Predicted class		
Actual class		Yes	No
	Yes	True Positive	False Negative
	No	False Positive	True Negative

TABLE 5: EVALUATION METRICS

Here,

- TP= True Positive
- TN= True Negative
- FP= False Positive

Accuracy

Accuracy is the most natural indicator of performance. It is simply a ratio of correctly predicted observation to the total observations. For our Naïve Bayes we got 0.70 that means our model predict 70% accurately.

$$AAppppaapprrppAA = TTPP + TTTT \div TTPP + TTPP + TTTT + TTTT$$

Precision

In pattern recognition data recovery and classification, precision is the portion of relevant occasions among the recovered occurrences, precision is used as an estimation of the relevance.

Precision= (true positive) / (true positive + false positive)

Recall

The recall is the proportion of our model effectively identifying True Positives. recall is used as an estimation of the relevance.

$$pppppprrrrrr = \frac{aappaapp\ ppppppppaappaapp(TTPP)}{aappaapp\ ppppppppaappaapp(TTPP) + ffrrrrpppp\ ppppaarraappaapp(TTTT)}$$

F1 Score

In factual analysis of binary classification, the F1 score (likewise F-score or F-measure) is a proportion of a test's accuracy. It considers both the accuracy p and the review r of the test to process the score.

$$ff1=2\times\frac{pppppppppppppppppp\times ppppppprrrrrr}{pppppppppppppppppp+ppppppprrrrrr}$$

4.6 VISUALIZATION

Yes, that is correct. N-grams are a way to represent a sequence of words or characters in a text by grouping them into chunks of a specific size (determined by the value of "n"). These chunks can then be used to analyze language patterns and predict the likelihood of certain words or phrases appearing in a given context. They are commonly used in tasks such as language modeling, text classification, and information retrieval.

4.6.1 UNIGRAM

	Accuracy	Precision	Recall	F1 score	Model name
0	67.25	75.00	59.45	66.32	LR
1	62.88	67.79	60.14	61.74	DT
2	64.00	73.40	52.76	61.39	RF
3	71.13	74.58	70.97	70.69	MNB
4	63.62	65.78	68.66	63.19	KNN
5	60.88	83.43	34.79	49.11	Linear SVM
6	62.62	78.24	43.09	55.57	RBF SVM
7	69.50	75.13	65.44	68.95	SGD
8	58.88	72.15	39.40	50.97	Adab
9	63.50	73.36	51.38	60.43	Xgb

TABLE 6: PERFORMANCE TABLE FOR UNIGRAM

The unigram model is otherwise called the pack of words model. Assessing the overall probability of various expressions is helpful in numerous common language handling applications, particularly those that produce text as an output. Using unigram our model accuracy is multinomial Naïve Bayes 71.13%, stochastic Gradient Descent(SGD) accuracy 69.50%. Naïve Bayes are best for our model because its show most accuracy.

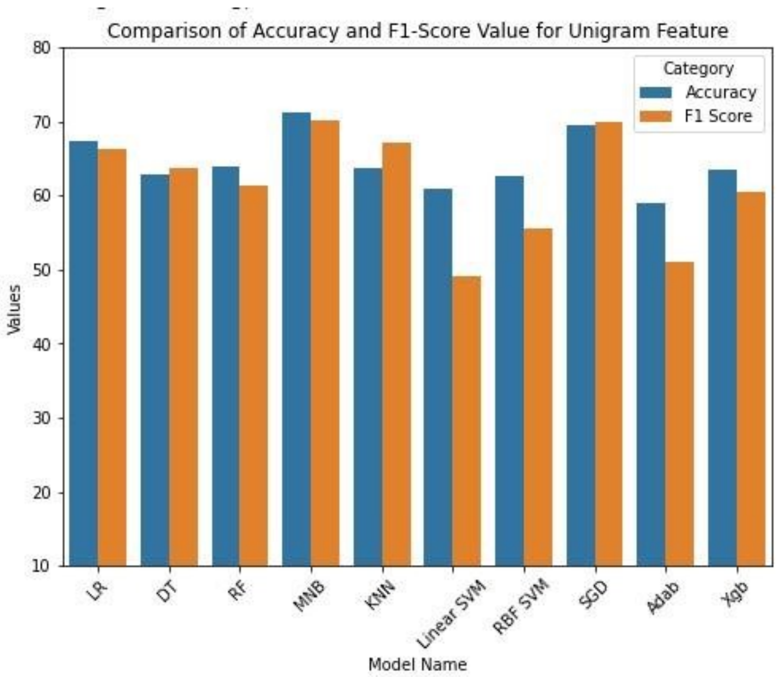


FIGURE 9: F1 SCORE FOR UNIGRAM FEATURE This model shows the best accuracy.

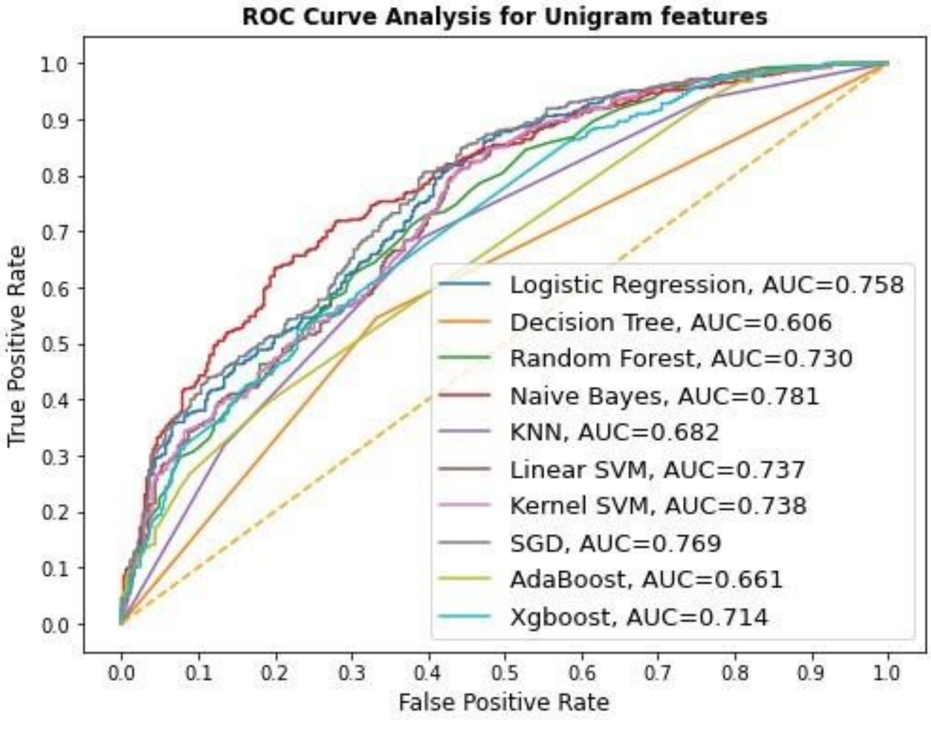


FIGURE 10: ROC CURVE FOR UNIGRAM

4.6.2 BI-GRAM

	Accuracy	Precision	Recall	F1 score	Model name
0	64.38	73.35	53.92	62.15	LR
1	61.50	68.00	54.84	60.71	DT
2	62.75	74.46	47.70	58.15	RF
3	70.75	73.36	72.35	70.10	MNB
4	64.25	66.89	67.51	63.20	KNN
5	52.88	88.00	15.21	25.29	Linear SVM
6	58.88	85.23	29.26	43.57	RBF SVM
7	67.00	73.10	61.98	66.08	SGD
8	58.12	70.54	39.17	50.37	Adab
9	62.88	72.46	50.92	59.81	Xgb

TABLE 7: PERFORMANCE TABLE FOR BIGRAM

Bigram show 70.75% accuracy for Multinomial Naïve Bayes(MNB).

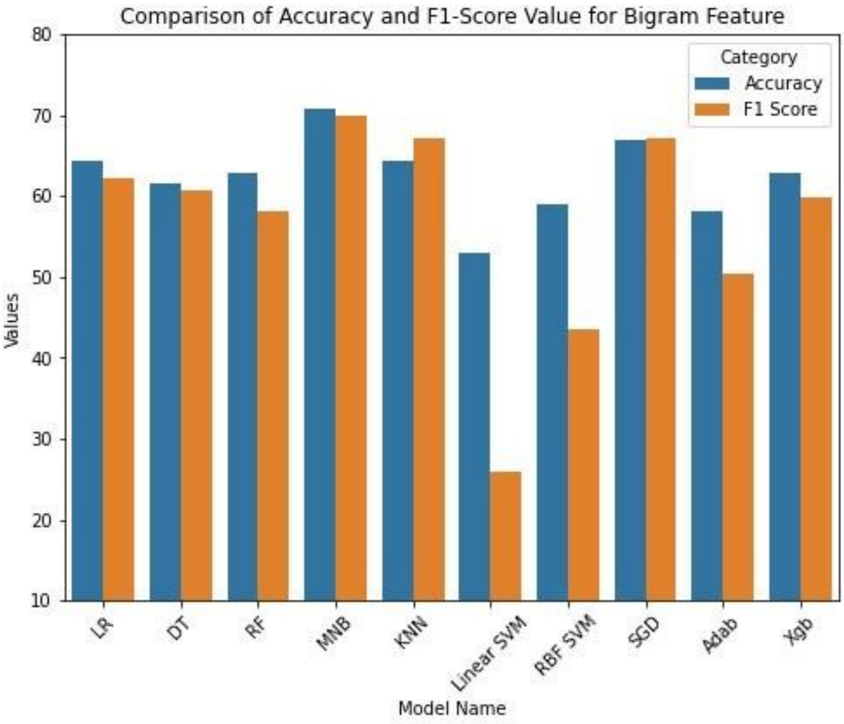


FIGURE 11: F1 SCORE FOR BIGRAM FEATURE

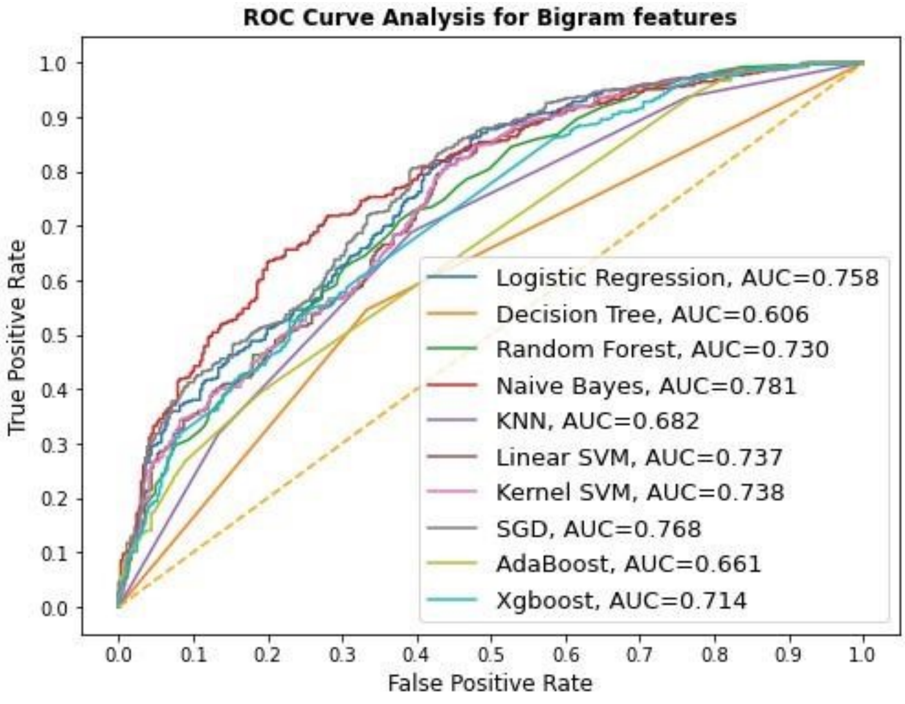


FIGURE 12: ROC CURVE FOR BIGRAM FEATURE

4.6.3 TRIGRAM

	Accuracy	Precision	Recall	F1 score	Model name
0	64.12	73.79	52.53	61.37	LR
1	60.12	66.02	54.61	59.77	DT
2	62.12	74.17	46.31	57.02	RF
3	70.75	73.15	72.81	70.12	MNB
4	64.88	67.43	68.20	62.81	KNN
5	48.25	88.46	5.30	10.00	Linear SVM
6	54.75	85.29	20.05	32.46	RBF SVM
7	66.12	72.09	61.29	46.25	SGD
8	58.50	70.73	40.09	51.18	Adab
9	62.75	72.37	50.69	59.62	Xgb

TABLE 8: PERFORMANCE TABLE FOR TRIGRAM

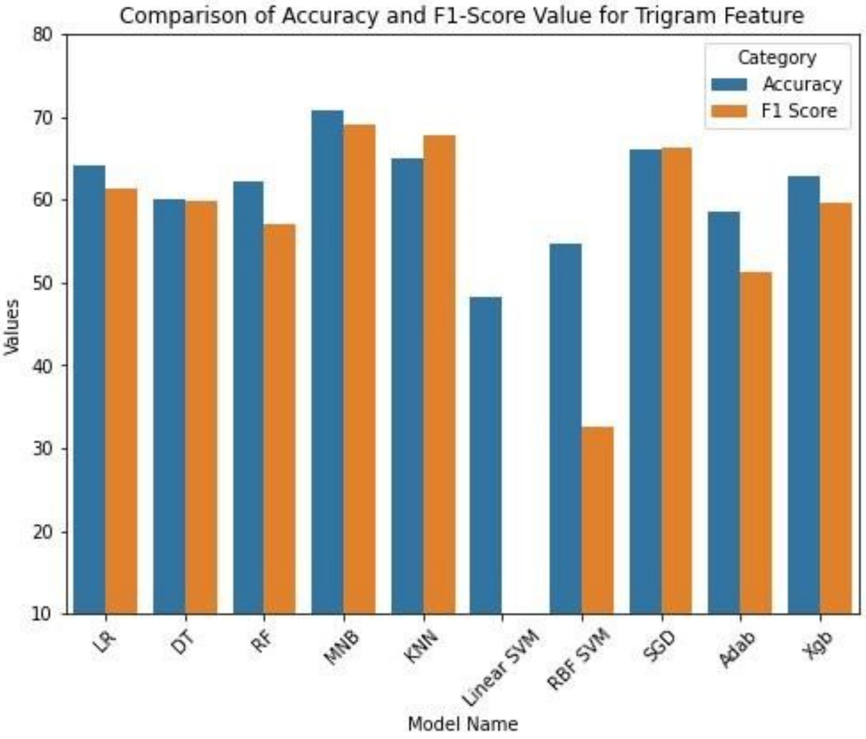


FIGURE 13: F1 SCORE FOR TRIGRAM FEATURE

This model also shows 70.75% accuracy of Naïve Bayes.

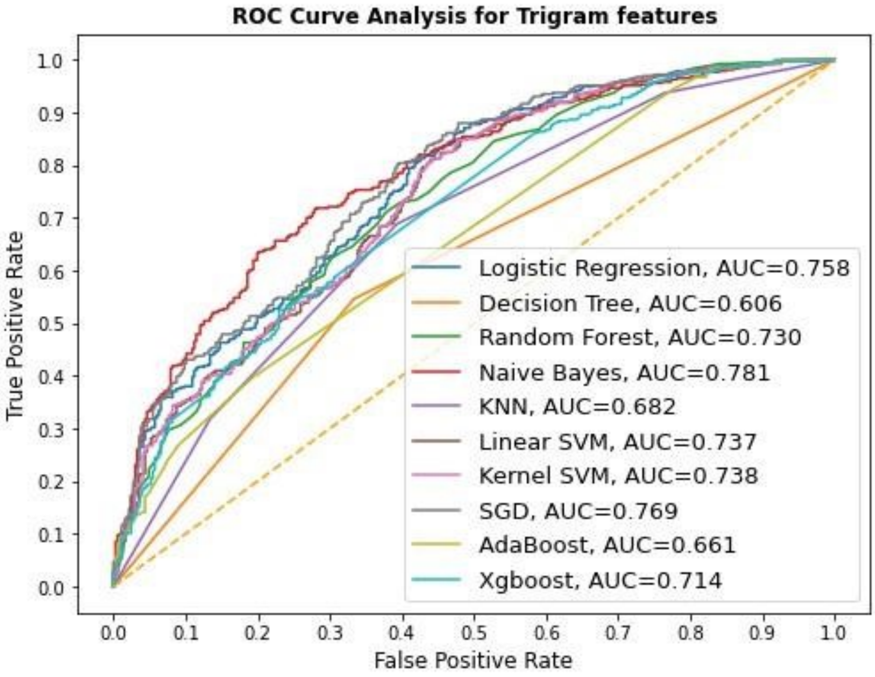


FIGURE 14:ROC CURVE FOR TRIGRAM FEATURE

4.7 COMPARE PRECISION

ALGORITHM	ACCURACY
NAÏVE BAYES	73.69%
STOCHASTIC GRADIENT DESCENT	73.44%
RANDOM FOREST	65.01%

TABLE 9: COMPARE PRECISION

4.8 COMPARE RECALL

ALGORITHM	ACCURACY
NAÏVE BAYES	72.04%
STOCHASTIC GRADIENT DESCENT	62.90%
RANDOM FOREST	48.92%

TABLE 10: COMPARE RECALL

4.9 F1 SCORE

ALGORITHM	ACCURACY
NAÏVE BAYES	70.30%
STOCHASTIC GRADIENT DESCENT	61.01%
RANDOM FOREST	58.85%

TABLE 11: F1 SCORE

4.10 COMPARE ALGORITHM ACCURACY

ALGORITHM	ACCURACY
NAÏVE BAYES	70.88%
STOCHASTIC GRADIENT DESCENT	67.54%
RANDOM FOREST	62.09%

TABLE 12: COMPARE ALGORITHM ACCURACY

4.11 PERFORMANCE EVALUTION

The confusion matrix of our model.

$$\begin{array}{lll} & rrppaaaarrrrpppppppppaappaapp & ppppaarraappaapp \\ ppppppppaappaapp255 & 114 & \\ ppppaarraappaapp111 & 320 & \end{array}$$

4.12 PREDICTION RESULT

From the above comparison tables, we see that, on account of Precision, Recall, f1-score and precision Naive Bayes classifier is the awesome. The estimations of accuracy, review and f1-score are separately 0.73, 0.72, and 0.70 and the accuracy of this model is 70.8% that is highest value in comparison with all classifier.

4.13 SUMMARY

To improve the accuracy of a Naive Bayes classifier, it is important to properly prepare the dataset. This includes ensuring that all categorical comments are equally numbered, and performing data cleaning to remove any irrelevant or inaccurate information. The more data is preprocessed, the more accurate the predictions of the classifier will be

CHAPTER 5

CONCLUSIONS AND RECOMMENDATIONS

5.1 INTRODUCTION

The study outlined in this passage discusses the abundance of research on Natural Language Processing (NLP) with a focus on the English language, and how it has led to advancements in computing. The author expresses that there is a lack of research on Bangla language, but that there is hope that more experts from different countries will begin to conduct research in this field..

5.2 CONCLUSIONS

It sounds like you have gained valuable experience and knowledge from your project, even though the classifier algorithm's accuracy was not as high as desired. You have learned how to work with Bangla text and preprocess raw data, which will be valuable for future research in this area. It's important to remember that even though a project may not have the desired outcome, it can still provide valuable learning opportunities and insights.

5.3 RECOMMENDATIONS

A couple of remarkable suggestions for this are as per the following

- To create a better dataset can produce a better output for this research work.
- Applying some deep learning method could be improving the model.

5.4 FUTURE DIRECTIONS

- Adding more classes in this model, can make this more effective.
- using more classifiers on this dataset, can improve understanding on which classifier can be the best for this work.

REFERENCES

1 Tripto, Nafis & Ali, Mohammed Eunus. (2018). Detecting Multilabel Sentiment and Emotions from Bangla YouTube Comments. 1-6. 10.1109/ICBSLP.2018.8554875.

0. Olga Uryupina , Barbara Plank SenTube: A Corpus for Sentiment Analysis on YouTube Social Media, Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)

1. Rumman Rashid Chowdhury , Mohammad Shahadat Hossain Analyzing Sentiment of Movie Reviews in Bangla by Applying Machine Learning Techniques, International

Conference on Bangla Speech and Language Processing(ICBSLP), 27-28 September, 2019

2. A. N. Muhammad, S. Bukhori and P. Pandunata, "Sentiment Analysis of Positive and Negative of YouTube Comments Using Naïve Bayes – Support Vector Machine (NBSVM) Classifier," *2019 International Conference on Computer Science, Information Technology, and Electrical Engineering (ICOMITEE)*, Jember, Indonesia, 2019, pp. 199-205, doi: 10.1109/ICOMITEE.2019.8920923.

3. H. Parveen and S. Pandey, "Sentiment analysis on Twitter Data-set using Naive Bayes algorithm," 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), Bangalore, 2016, pp. 416-419, doi: 10.1109/ICATCCT.2016.7912034.

4. @inproceedings{Timoney2018NostalgicSA, title={Nostalgic Sentiment Analysis of YouTube Comments for Chart Hits of the 20th Century}, author={J. Timoney and B. Davis and A. Raj}, booktitle={AICS}, year={2018}}

5. Cunha, Alexandre & Costa, Melissa & Pacheco, Marco. (2019). Sentiment Analysis of YouTube Video Comments Using Deep Neural Networks. 10.1007/978-3-030209124_51.

6. Saif, Hassan; He, Yulan and Alani, Harith (2012). Semantic sentiment analysis of twitter. In: The 11th International Semantic Web Conference (ISWC 2012), 11-15 Nov 2012, Boston, MA, USA.

7. @inproceedings{Bautin2008InternationalSA, title={International Sentiment Analysis for News and Blogs},

8. Dos Santos, Cicero & Gatti de Bayser, Maira. (2014). Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts.

9. S. Chowdhury and W. Chowdhury, "Performing sentiment analysis in Bangla microblog posts," 2014 International Conference on Informatics, Electronics & Vision (ICIEV), Dhaka, 2014, pp. 1-6, doi: 10.1109/ICIEV.2014.6850712.

10. R. A. Tuhin, B. K. Paul, F. Nawrine, M. Akter and A. K. Das, "An Automated System of Sentiment Analysis from Bangla Text using Supervised Learning Techniques," 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS), Singapore, 2019, pp. 360-364, doi: 10.1109/CCOMS.2019.8821658.

11. M. H. Alam, M. Rahoman and M. A. K. Azad, "Sentiment analysis for Bangla sentences using convolutional neural network," 2017 20th International Conference of Computer and Information Technology (ICCIT), Dhaka, 2017, pp. 1-6, doi: 10.1109/ICCITECHN.2017.8281840.

12. S. Arafin Mahtab, N. Islam and M. Mahfuzur Rahaman, "Sentiment Analysis on Bangladesh Cricket with Support Vector Machine," 2018 International Conference on Bangla Speech and Language Processing (ICBSLP), Sylhet, 2018, pp. 1-4, doi: 10.1109/ICBSLP.2018.8554585.