Sentiment Analysis on Bengali Comments of YouTube's Bangla Drama to Predict Emotions: A TF-IDF Approach

This Thesis Report
Submitted in partial fulfillment of the requirements for the degree of
Master of Science in Software Engineering

Submitted By

A.B.M. Kibria Sarkar
ID No: 0242220005343006
Department of Software Engineering
Daffodil International University

Supervised by

Dr. Imran Mahmud
Associate Professor and Head
Department of Software Engineering
Daffodil International University

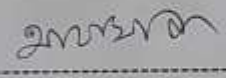
APPROVAL

This thesis entitled on "**Sentiment Analysis on Bengali Comments of YouTube's Bangla Drama to Predict Emotions: A TF-IDF Approach**", submitted by A.B.M. Kibria Sarkar **ID: 0242220005343006** to the department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirement's for the degree of Master of Science in Software Engineering and approval as to its style and contents.

BOARD OF EXAMINERS

Chairman

Dr. Imran Mahmud
Associate Professor and Head
Department of Software Engineering
Daffodil International University

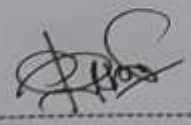


Internal Examiner 1

Afsana Begum
Assistant Professor
Department of Software Engineering
Daffodil International University

Internal Examiner 2

Dr. Md. Fazla Elahe
Assistant Professor and Associate Head
Department of Software Engineering
Daffodil International University



External Examiner

Md. Tanvir Quader
Senior Software Engineer,
Technology Team,
a2i Program

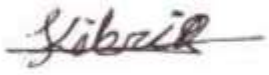
DECLARATION

I hereby declare that, this thesis has been done by us under the supervision of **Dr. Imran Mahmud, Associate Professor and Head, Department of SWE** Daffodil International University. I also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Certified by:

Dr. Imran Mahmud
Associate Professor and Head
Department of Software Engineering
Daffodil International University

Submitted By:



A.B.M. Kibria Sarkar
ID No : 0242220005343006
Department of Software Engineering
Daffodil International University

ACKNOWLEDGEMENT

First, I express my heartiest thanks and thankfulness to almighty God for His divine blessing makes us possible to finish the final year thesis successfully.

I really grateful and wish our profound my indebtedness to **Supervisor** of **Dr. Imran Mahmud** Department of SWE Daffodil International University, Dhaka. Deep Knowledge & keen interest of my supervisor in the field of "*Machine Learning*" to carry out this project. His scholarly guidance, endless patience, constant, continual encouragement and constructive criticism, energetic supervision, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this project.

I would like to express my heartiest gratitude to **Dr. Imran Mahmud, Associate Professor and Head**, Department of SWE, for his kind help to finish my thesis and also to other faculty member and the staff of SWE department of Daffodil International University.

I would like to thank my entire course mate in Daffodil International University, who tookpart in this discuss while finishing the course work.

Finally, I must acknowledge with due respect the continuous support and patients of my parents.

Table of Contents

List of Figures

Figures **Page No**

List of Tables

Tables

List of Equations

Equations

ABSTRACT

Sentiment Analysis (SA) is indeed a rapidly growing field within computer science, particularly in the realm of natural language processing (NLP). It involves the automated analysis of text to determine the underlying sentiment or emotion expressed within it. This capability holds significant importance in various applications such as social media monitoring, customer feedback analysis, market research, and more.

In my specific study, I have focused on sentiment analysis in the context of the Bengali language, which is a valuable contribution considering the relatively limited research in this area. By examining emotions such as happiness and anger, I am addressing fundamental aspects of human expression within the linguistic context of Bengali.

My methodology involves employing different machine learning techniques to train a dataset for sentiment analysis. Let's delve into the techniques I have utilized and the corresponding accuracies:
Logistic Regression (LR): This is a statistical method used for modeling binary outcomes, making it suitable for sentiment analysis tasks where the goal is to classify text into positive or negative sentiments. My LR model achieved an accuracy of 79.80%, indicating its effectiveness in capturing the nuances of sentiment in Bengali text.

Decision Tree (DT): Decision trees are a popular machine learning algorithm for classification tasks. They partition the feature space into smaller regions based on certain criteria, making them interpretable and easy to visualize. My DT model achieved an accuracy of 78.44%, demonstrating its capability in discerning sentiment patterns in Bengali text.

Random Forest (RF): Random Forest is an ensemble learning technique that combines multiple decision trees to improve predictive performance and reduce overfitting. My RF model achieved a similar accuracy to LR, further validating its effectiveness in sentiment analysis tasks.

Bernoulli Naive Bayes (BNB): Naive Bayes classifiers are based on Bayes' theorem and assume independence between features. Bernoulli Naive Bayes specifically works well with binary features, making it suitable for sentiment analysis where the presence or absence of certain words may indicate sentiment. My BNB model achieved an accuracy of 79.73%, demonstrating its competitiveness with other techniques.

K-Nearest Neighbors (KNN): Classifying instances according to the majority class among their K nearest neighbors (KNN) is the basis of the straightforward and user-friendly KNN classification technique. While my KNN model achieved a lower accuracy of 68.02%, it still provides valuable insights into sentiment patterns in Bengali text.
Support Vector Classifier (SVC): SVC is a powerful classification algorithm that works by finding the hyperplane that best separates different classes in the feature space. My SVC model outperformed the other techniques with an accuracy of 81.48%, indicating its effectiveness in capturing complex sentiment patterns in Bengali text.

Keywords: *Bengali Comment, TF-IDF, SVM, Sentiment, KNN*

CHAPTER 1
Introduction

1.1 Introduction

Sentiment analysis (SA), also known as opinion mining (OP), constitutes the examination of human emotions through systematic processes involving the detection, extraction, measurement, and interpretation of affective states and contextual information within text, facilitated by Natural Language Processing (NLP). This method aims to discern the emotions conveyed within a text, thereby enhancing the understanding of attitudes, opinions, and sentiments expressed in online content. In today's dynamic environment, the ability to identify underlying emotions holds paramount importance. By discerning the motives behind human actions, ranging from consumer choices to detecting criminal intent and improving workplace dynamics, various industries stand to benefit significantly. Leveraging the highly automated technique of sentiment analysis enables the determination of an individual's perspective from spoken or written communication, even extending to video content. Human emotions stem from biochemical changes in the neurophysiological system, influencing feelings, thoughts, and behavioral responses (Panksepp, 2004; Damasio, 1998; Ekman et al., 1994).

Psychologists have extensively studied the vast spectrum of human emotions, attempting to categorize and define them through various concepts. Among these efforts, researchers have proposed six fundamental emotions: regret, joy, anxiety, disgust, fury, and surprise. Notably, within computer science circles, the term "Sentiment Analysis" gained prominence as researchers sought to discern these core human emotions from diverse forms of human expression, such as text, speech, and facial expressions.

The origins of sentiment analysis can be traced back to academic research conducted during and after World War II, initially motivated by political interests. However, it was not until the mid-2000s that modern sentiment analysis, which primarily focuses on analyzing product reviews available online, gained widespread recognition (Viking et al., 2016). Subsequently, sentiment analysis applications expanded to encompass areas such as stock market predictions and assessing public sentiment following events like terrorist attacks.

Moreover, sentiment analysis serves as a valuable tool for monitoring public opinion on specific topics, particularly through social media channels. Its effectiveness spans across various applications, prompting organizations worldwide to capitalize on social data for information gathering purposes.

However, upon reviewing numerous research projects on human emotion analysis, it becomes apparent that, possibly due to the linguistic intricacies of the Bangla language, there is a notable scarcity of studies compared to those conducted for English and other languages (Drovo et al., 2019; Hossain et al., 2021).

Bangla, spoken by approximately 250 million native speakers worldwide, ranks as the fourth most widely used language globally. Given its widespread usage, I directed my focus towards leveraging Bangla for emotion extraction, recognizing its prevalence as a language through which around 250 million individuals communicate their emotions. However, the detection of slang or derogatory expressions in Bangla presents a greater challenge compared to English, owing to a scarcity of available information on this linguistic aspect (Mumu et al., 2021; Islam et al., 2019).

Furthermore, Bengali speakers vary in their religious beliefs. Religious violence is rising daily as a result of people disparaging other religious beliefs on social media. Social media platforms like Facebook, Twitter, and others are typically used by people to share their opinions and ideas, and a large amount of information is available reading a many different topics (Mahmudun et al., 2016). In this research, I suggested a method to estimate sentiment from Bangla-written text using the SVM algorithm, inspired by the use of sentiment analysis.

1.2 Problem around the world

People of different languages live in the world. We know that English is an international language. Apart from international languages, people also express their opinions in their own language. People feel more comfortable to comment in his own language on different social media, mass media of his country etc (van et al., 2008). Because of English is an international language, sentiment analysis of this language is easier than other languages. Sentiment analysis in other languages is quite difficult as compared to English language ((Bautin et al., 2008)). For example, when Japanese people comment in Japanese language, sentiment analysis should be done separately for Japanese language, because sentiment analysis program for English language, it will not work for Japanese language. Sentiment is not match one language to another language, so this is very big problem to analyses the sentiment (Varghese & Jayasree, 2013). Therefore, sentiment analysis has to be done separately for different languages.

1.3 Problem around Bangladesh

Bangladesh is moving forward in keeping with the world. The present age is the age of globalization. The people of Bangladesh have also entered the age of the Internet and the use of the Internet is increasing along with it (Mamun et al., 2019). The trend of publishing comments using the Internet is also increasing day by day. People are expressing their comments on different social media and different platforms. Bengali language is different from English language so English language sentiment analysis is not suitable for Bengali language (Varghese & Jayasree, 2013). At this time we need appropriate sentiment analysis process for Bengali language. Need different filtering system, stop word and other process which is separate from English language sentiment analysis.

1.4 Application

In essence, opinion mining (OP) and sentiment analysis (SA) provide an opportunity to delve into the attitudes of audience members and evaluate the state of a product from an adversarial perspective. As a result, sentiment analysis (SA) emerges as a valuable technique for a multitude of purposes (Mejova, 2009; Mehta & Pandya, 2020), including:

- Enhanced product analytics
- Market research
- Reputation management
- Precision targeting
- Marketing analysis
- Public relations (PR)
- Product reviews
- Net promoter scoring
- Product feedback
- Customer Service

CHAPTER 2
Review of Literature

2. Related Works

Viking et al. (2016) conducted a thorough review and found that sentiment analysis, which has its roots in early surveys on public opinion research conducted in the early 20th century and the development of text analysis by computational language research groups in the 1990s, is a rapidly expanding field of study within computer science. The proliferation of text data on the internet, combined with advancements in computer-based sentiment analysis techniques, notably accelerated during significant events such as outbreaks. This progress has had profound implications across various industries by enabling the identification of underlying sentiments in textual or audio messages. As a result, sentiment analysis has been the subject of extensive exploration in numerous studies utilizing diverse methodologies (Khan et al., 2021).

I discovered a number of publications that were pertinent to my field of study while doing my research. Das and Bandyopadhyay (2010) conducted a remarkable work in which they suggested a hybrid method to extract views from literature written in both Bangla and English. This mechanism integrates rule-based and automated systems. Their methodology comprised the application of Support Vector Machines (SVM) in conjunction with Natural Language Processing (NLP) approaches.

Furthermore, Mahmudun et al. (2016) conducted research aimed at augmenting existing solutions by integrating the concepts of Term Frequency (TF) and Inverse Document Frequency (IDF) values. Through the utilization of these values, they effectively extracted different attributes of negative, neutral, and positive terms from Bangla text, thereby enhancing the accuracy of results.

Moreover, Das (2011) introduced an emotion monitoring methodology tailored to specific subjects or events. This approach integrated sense-based affect rating techniques with SentiWordNet for analyzing text in both English and Bangla languages.

Similarly, Hasan and Rahman (2014) used WordNet and SentiWordNet to determine the semantic meanings and polarity of terms in the Bangla language. Their sentiment recognition techniques that were adapted for Bangla text analysis were based on these conclusions.

Additionally, Chowdhury and Chowdhury (2014) used the Maximum Entropy (MaxEnt) technique in conjunction with Support Vector Machine (SVM) in a different study to automatically extract sentiments from messages on Bangla

microblogs (Twitter). Whether the polarity was positive or negative, their methodology made it easier to identify the attitudes present in the text.

In a separate study, Go et al. (2009) employed three machine learning techniques, namely Support Vector Machine (SVM), Maximum Entropy (MaxEnt), and Naive Bayes, to discern emotions in Twitter messages containing emoticons.

Additionally, Pandey and Govilkar (2015) proposed a method for sentiment analysis of Hindi movie reviews, utilizing HindiSentiWordNet (HSWN) and the Synset Replacement Algorithm (SRA) in their approach.

In their work on extracting emotions from Bangla text, Tuhin et al. (2019) proposed two machine learning techniques: the Naive Bayes Classification Algorithm and the Topical Method, applied at both article and sentence levels. Additionally, in a separate study, Tuhin et al. (2019) devised a sentiment analysis model leveraging Latent Dirichlet Allocation (LDA), which they employed to assess the popularity of tweets.

Moreover, Umamaheswari et al. utilized Support Vector Machine (SVM) and Latent Dirichlet Allocation (LDA) to identify viewpoints from the IMDB movie analysis dataset.

Taboada (2016) gives a general review of sentiment analysis, a branch of linguistics and computer science that focuses on automatically identifying text's emotional content. It goes on sentiment analysis's uses, how it interacts with linguistics, and how it fits into social media analysis.

Bhadane et al. (2015) investigate techniques for categorizing text according to the viewpoints represented within it, with a focus on whether the sentiment is good or negative. Additionally, a two-step technique for sentiment analysis is covered.

Mishra et al. (2021) delve into sentiment analysis within code-mixed social media texts for Dravidian languages, with a specific focus on Tamil, Malayalam, and Kannada. Their study explores the utilization of both machine learning and deep learning models to effectively recognize sentiments expressed in code-mixed text.

Using sentiment analysis of spoken conversation transcripts, Ojamaa et al. (2015) estimates speakers' attitudes based on their utterances. With the help of conventional sentiment analysis techniques, it reports encouraging results.

An approach to assessing sentiment is provided by Shaikh et al. (2008), which focuses on identifying sentences' positive or negative valence. Semantic

dependency analysis and rule-based techniques for text sentiment analysis are covered.

For sentiment analysis, Hogenboom et al. (2014) suggest translating the sentiment expressed in natural language text to standardized star ratings. It examines several mapping techniques and how well they work with datasets in both Dutch and English.

Awatramani et al. (2021) use a variety of techniques, such as lexicon-based, rule-based, and machine learning, to analyze sentiment in mixed-case languages like Hinglish (Hindi-English). It assesses how well these methods work in categorizing sentiment.

Verma and Thakur's (2018) discussion of sentiment analysis emphasizes text reviews, methods, lexicons, and machine learning techniques for examining people's attitudes and sentiments.

The efficacy of psychological and linguistic variables for sentiment classification in the Spanish language is assessed by Salas-Zárate et al. (2017). It analyzes sentiment in traveler evaluations using a variety of metrics and classifiers.

From a cognitive standpoint, Panikar et al. (2022) present a thorough taxonomy of sentiment analysis, distinguishing many sentiment components in natural language. It goes over how these classifications are used in many fields.

In order to reveal hidden meaning in unstructured data, Solanki (2019) uses the Natural Language Toolkit and Python Libraries to demonstrate the sentiment analysis procedure.

In their study, Bhattacharya and Banerjee (2017) looked into the use of linguistic cues to identify sentiment in Twitter tweets. To evaluate sentiment in tweets, it uses polarity classification models and n-gram language models.

According to Dhuria (2015), sentiment analysis and opinion mining are two key areas of focus within the field of natural language processing. The research underscores the importance of sentiment analysis, particularly in the realm of analyzing digital data.

Giatsoglou et al.'s (2017) discussion of sentiment analysis and opinion mining focuses in particular on the difficulties of sentiment identification across linguistic boundaries. It suggests a machine learning strategy utilizing different document vectorization techniques.

Bilal (2021) emphasizes the significance of comprehending sentiment in diverse languages by focusing on sentiment analysis in Arabic tweets. It assesses various algorithms for Arabic text sentiment prediction.

Tan et al. (2023) highlight the importance of sentiment analysis in natural language processing and its uses in a variety of industries, including marketing and politics. They also give an overview of recent developments, difficulties, and potential future directions for sentiment analysis research.

Zia et al. (2018) analyze various strategies for sentiment classification, including algorithms and application domains, with a focus on sentiment analysis as a text mining technique.

Verma & Thakur (2018) use lexical and machine learning techniques to examine sentiment analysis. It emphasizes how critical it is to consider user emotions and viewpoints, particularly on social networking platforms.

With a focus on social media data, Jindal & Aron (2021) investigate sentiment analysis as a method for recovering human emotions from unstructured text. It talks about several sentiment analysis methods and algorithms.

The need for sentiment analysis is discussed by Attri & Dutta (2020) in relation to the growing trend of opinion sharing on social media. It gives an examination of various sentiment analysis techniques.

The study by Dorothy & Rajini (2016) examines several techniques and methodologies for sentiment analysis, focusing on its use in social media, consumer feedback, and areas like movie reviews.

In their article, Kaushik & Naithani (2015) cover the methods and tools utilized in sentiment analysis as well as the difficulties with the technique. It underlines how crucial it is to comprehend what the general public thinks of goods and services.

Alasmari et al. (2017) talk about sentiment analysis as a way to extract irrational information from text data, with a particular emphasis on attitudes and opinions regarding specific topics. It highlights how crucial it is for both consumers and producers.

Various sentiment analysis algorithms are examined by Sinha & Kaur (2016, November), who compare them based on their accuracy. It draws attention to the value of sentiment analysis in assessing user opinions.

In their study of sentiment analysis methodologies and approaches, Arya et al. (2020, May) place a special emphasis on the extraction and categorization of sentiments or views from various sources.

In the context of natural language processing, Vohra et al. (2013) explore sentiment analysis strategies and give a comparison examination of various techniques.

Feldman (2013) examines sentiment analysis's key uses and difficulties, highlighting its importance in the field of computer science.

Bhadane et al. (2015) concentrate on techniques for categorizing natural language text in accordance with the opinions conveyed in it, whether those opinions are favorable or unfavorable.

In their article, Shukla & Shukla (2015), they go over numerous sentiment categorization and analysis approaches, emphasizing the value of examining user feedback in digital data.

With a focus on improving accuracy through feature selection and current classification procedures, Vaghela et al. (2016) appear to be primarily concerned with examining and analyzing various sentiment classification strategies.

CHAPTER 3
Methodology

3.1 Methodology

In my research on text classification, I employed Supervised Learning, one of the three primary areas of Machine Learning, as outlined by Das et al. (2021). This approach was well-suited for my investigation, considering that my dataset was multiclass labeled, making it suitable for classification tasks. Supervised classification of text necessitates pre-defined classification categories, relying on a pre-labeled dataset with accurate values.

Machines are incapable of comprehending free image, text, or video data; they can only process binary data represented as 0s and 1s. To render data machine-readable, it must be converted or encoded. To achieve this, I applied TF-IDF (Term Frequency-Inverse Document Frequency) using the "Scikit-learn" Python library. The dataset had to be cleaned up and converted into vector format.

My study's methodology consisted of cleaning the dataset, converting it into a machine-readable format, and then using the modified dataset to train machine learning models. When I compared SVM (Support Vector Machine) classifier to other classifiers like K-Nearest Neighbors, Naive Bayes, Random Forest, and Logistic Regression, it yielded very good results for my study. The purpose of using extra classifiers was to make sure the outcomes were legitimate. Because of its kernel technique, which allows it to handle nonlinear input forms, SVM is well recognized (Datacamp.com, 2021).

There are two key steps in supervised learning-based sentiment analysis document categorization. The trained classifier, as shown in Fig. 1, projects the target class of unknown test data after being trained using the training set. This allows the testing phase to evaluate the accuracy of the model.

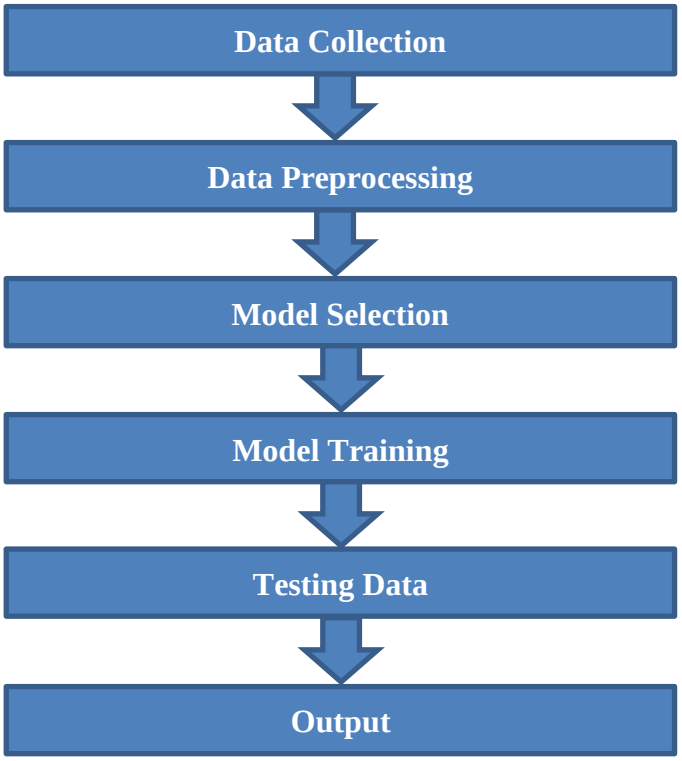


Figure-1: Functioning Mechanism of the Suggested Approach

3.2 Data Collection

For this research, data collection was conducted through GitHub. The gathered data was initially stored in an Excel sheet in its raw format and subsequently exported as a CSV Excel sheet, as described by Mahmudun et al. (2016).

In total, 11705 comments were collected as raw data for this study where 6058 were associated with strongly positive sentiment, 3091 were associated with strongly negative sentiment, 1443 were associated with weakly negative sentiment, 852 were associated with weakly positive sentiment, while the remaining 261 data points were related to neutral sentiment. Upon analyzing the collected data, it will be separated into five categories: strongly positive, strongly negative, weakly negative, weakly positive and neutral sentiments.

Table-1: show the Sample dataset and Figure-2: show the Class Polarity distribution where (-1.0 to -0.61) for strongly negative, (-0.60 to -0.19) for weakly negative, (-1.0 to -0.61) for strongly negative, (-0.20 to 0.20) for neutral, (-1.0 to -0.61) for strongly negative, (0.21 to 0.60) for weakly positive and (0.61 to 1.0) for strongly positive.

	Comment	Polarity_Class	Polarity
0	অসাধারণ নিশো বস্ আর অমি ভাইকেও	stronglyPositive	1.0
1	আমার দেখা বেস্ট নাটক	stronglyPositive	1.0
2	সত্যি অসাধারণ একটি রিলেশন	stronglyPositive	1.0
3	মজা পাইছি ভাষা গুলো কেমন লাগলো	weaklyPositive	0.6
4	খুব সুন্দর হয়েছে	stronglyPositive	1.0
...	...	...	...
11195	সত্যিই শত চেষ্টা করেও ভালবাসা হয়নাএটা জাষ্ট ঘ...	weaklyPositive	0.6
11196	ফালতুনাটক	stronglyNegative	-1.0
11197	জিম করার পর যেই পানীয় টা পান করে সেটার নাম কি	neutral	-0.2
11198	মা মেয়েকে ছেলে খোজার জন্য অনুমতি দেয় এটা কি...	neutral	-0.2
11199	মিশুকে মূল চরিত্রে বেমানান ফালতু অভিনয় আবাল ক...	stronglyNegative	-1.0

11200 rows × 3 columns

Table-1: Sample Data

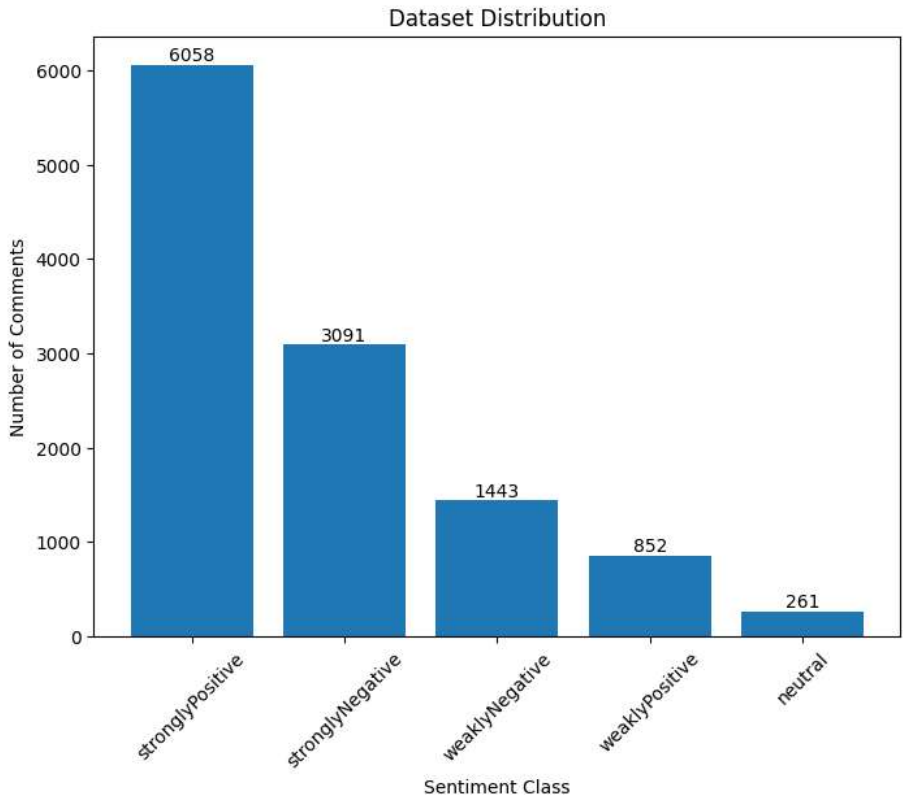


Figure-2: Dataset Distribution

3.3 Data Cleaning

In the process of data preprocessing, the collected raw data consisted of various comments, but the requirement was to retain only pure Bangla text while removing any mixed-language comments to reduce noise and enhance model accuracy. This meant removing irrelevant information and keeping comments that were either entirely written in Bangla or included both Bangla and other characters.

The remaining comments were then subjected to additional cleaning based on certain requirements, such making sure that the Unicode value of every character was between 2432 and 2559 or 32. Bangla fonts are represented by Unicode values between 2432 and 2559, with 32 signifying space. After then, the dataset was split into Happy and Anger, two separate classes.

Furthermore, a set of Bangla stop words was compiled for the research project and removed from the data. These stop words, totaling around 388, included terms like 'অতএব', 'অথচ', 'অথবা', 'অনুযায়ী', 'অনেক', 'অনেকে', 'অনেকেই', 'অন্তত', and 'অন্য'.

To implement this preprocessing, Python code was utilized for removing punctuations, emojis, and non-Bengali words, as depicted in Figure-3. Additionally, Figure-4 illustrates the Python code for eliminating Bangla stop words. Table-2 displays the cleaned data after preprocessing.

```
def remove_punc_imoji_non_bengali_word(text):
    t = regex.findall(r"[\p{Bengali}]+", text)
    text = ' '.join(t)
    text = re.sub('[^\u0980-\u09FF]',' ',str(text))

    return text
```

Figure-3: Removing punctuations, emoji and non-Bengali word

```
BengaliComments['Cleaned_Comment'] = BengaliComments['Comment'].apply(lambda x: ' '.join([word for word in x.split() if word not in (stop_words)]))
BengaliComments.head(10)
```

Figure-4: Removing Bangla stop word

	Comment	Polarity_Class	Polarity	Emotion	EmotionNum	Cleaned_Comment
0	অসাধারণ নিশো বস্ আর আমি ভাইকেও	stronglyPositive	1.0	Happy	1	অসাধারণ নিশো বস্ আমি ভাইকেও
1	আমার দেখা বেস্ট নাটক	stronglyPositive	1.0	Happy	1	বেস্ট নাটক
2	সত্যি অসাধারণ একটি রিলেশন	stronglyPositive	1.0	Happy	1	সত্যি অসাধারণ রিলেশন
3	মজা পাইছি ভাষা গুলো কেমন লাগলো	weaklyPositive	0.6	Happy	1	মজা পাইছি ভাষা গুলো কেমন লাগলো
4	খুব সুন্দর হয়েছে	stronglyPositive	1.0	Happy	1	সুন্দর
...	...	...	...	...	...	...
11195	সত্যিই শত চেষ্টা করেও ভালবাসা হয়না,এটা জাষ্ট ঘ...	weaklyPositive	0.6	Happy	1	সত্যিই শত করেও ভালবাসা হয়না,এটা জাষ্ট ঘটে যায়...
11196	ফালতুনাটক	stronglyNegative	-1.0	Anger	0	ফালতুনাটক
11197	জিম করার পর যেই পানীয় টা পান করে সেটার নাম কি	neutral	-0.2	Anger	0	জিম যেই পানীয় টা পান সেটার নাম
11198	মা মেয়েকে ছেলে খোজার জন্য অনুমতি দেয় এটা কি...	neutral	-0.2	Anger	0	মা মেয়েকে ছেলে খোজার অনুমতি চরিত্র
11199	মিশুকে মুল চরিত্রে বেমানান ফালতু অভিনয় আবাল ক...	stronglyNegative	-1.0	Anger	0	মিশুকে মুল চরিত্রে বেমানান ফালতু অভিনয় আবাল ক...

11200 rows x 6 columns

Table-2: Sample of Cleaned data

3.4 Machine learning algorithms and statistical analysis

In my dataset, there are 6058 records for strongly positive conversations, 3091 records for strongly negative conversations, 1443 records for weakly negative conversations, 852 records for weakly positive conversations, 261 records for neutral conversations. I used the train-test split function to separate the dataset. Techniques for supervised machine learning were used. I used 80% of my data to train our model, and 20% of the data for testing. I used some classifier-based techniques to determine the accuracy on our dataset. These include the Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Bernoulli Naive Bayes (BNB), K-Nnearest Neighbors (KNN) and Support Vector Machine (SVM). Figure-6: Illustrates how I conducted our research in brief terms.

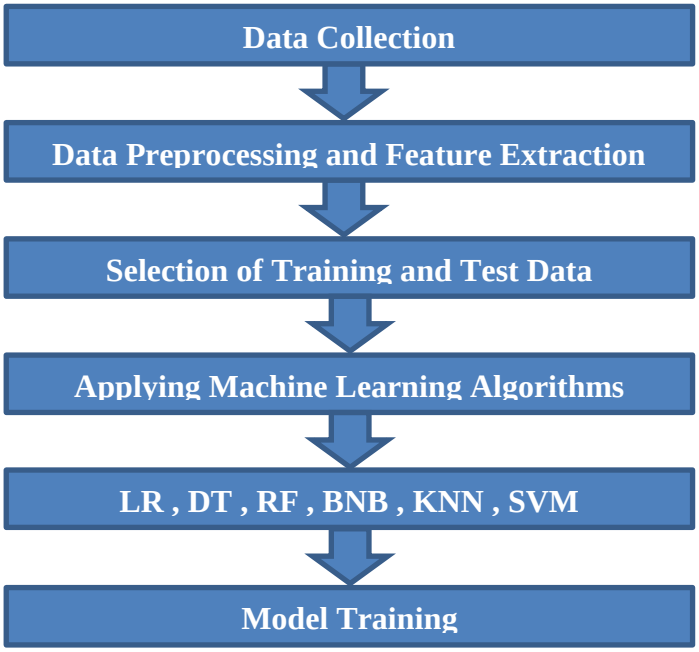


Figure-5: Proposed Model Structure

3.4.1 Feature extraction

In this context of natural language processing tasks, machine learning techniques are applied to accomplish my objectives. My model is trained by a feature extraction process that involves two primary characteristics. I utilize a method known as tokenizer to implement this technique, which involves dividing phrases into individual word parts. These words possess unique properties and are common across the text. Additionally, I make use of TF-IDF a numerical representation that assesses the importance of a word within a text. This approach has been successfully adopted by prominent publications for various languages. Motivated by their achievements, I observed that our learning algorithms also achieved high accuracy.

3.4.2 Classifier algorithms

In the context of classification tasks, several decision trees are often constructed during the training process. Naïve Bayes classification operates under the assumption of independence between different characteristics within a class. This model is known for its simplicity and efficiency, especially when dealing with large datasets. In fact, Naïve Bayes can outperform more complex classification algorithms in certain scenarios. Conversely, logistic regression is capable of generating a probability model for a particular class or event. For instance, it can be employed to classify groups of images containing pictures of various animals into distinct classes using a single model.

CHAPTER 4
Results & Analysis

4.1 Results

Four crucial indicators were used in order to evaluate the suggested classifiers' performance: Accuracy, Recall, Precision, and F1-Score. These metrics provided comprehensive insights into the model's performance, allowing for a clear understanding of its efficacy.

For each metric, five classifiers were evaluated alongside SVM to determine the optimal classifier. By comparing the performance of these classifiers across multiple metrics, the most effective model could be identified, ensuring robust sentiment analysis.

This approach enabled a thorough examination of each classifier's ability to accurately classify sentiment, providing valuable insights into their strengths and weaknesses.

4.1.1 Accuracy

The accuracy, as defined by Mahmudun et al. (2016), represents the ratio of correctly predicted True Negatives (TN) and True Positives (TP) to the total test dataset. Mathematically, accuracy (Acc) can be calculated using the following equation (Eq. 1):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

Where:

- TP represents the number of True Positives,
- TN represents the number of True Negatives,
- FP represents the number of False Positives, and
- FN represents the number of False Negatives.

This formula allows for the quantification of the model's overall correctness in predicting the sentiment of the test dataset. A higher accuracy score indicates better performance, as it reflects a higher proportion of correctly predicted sentiments.

After training the algorithms with a dataset labeled with five classes, the accuracy scores obtained were as follows: Logistic Regression (LR) achieved an accuracy of 79.80%, Decision Tree (DT) attained 78.44%, Random Forest (RF) achieved 79.80%, Bernoulli Naive Bayes (BNB) obtained 79.73%, K-Nearest Neighbors (KNN) achieved 68.02%, and Support Vector Classifier (SVC) yielded an accuracy of 81.48%. These accuracy scores reflect the proportion of correctly predicted sentiments by each classifier, providing insights into their respective performances in sentiment analysis.

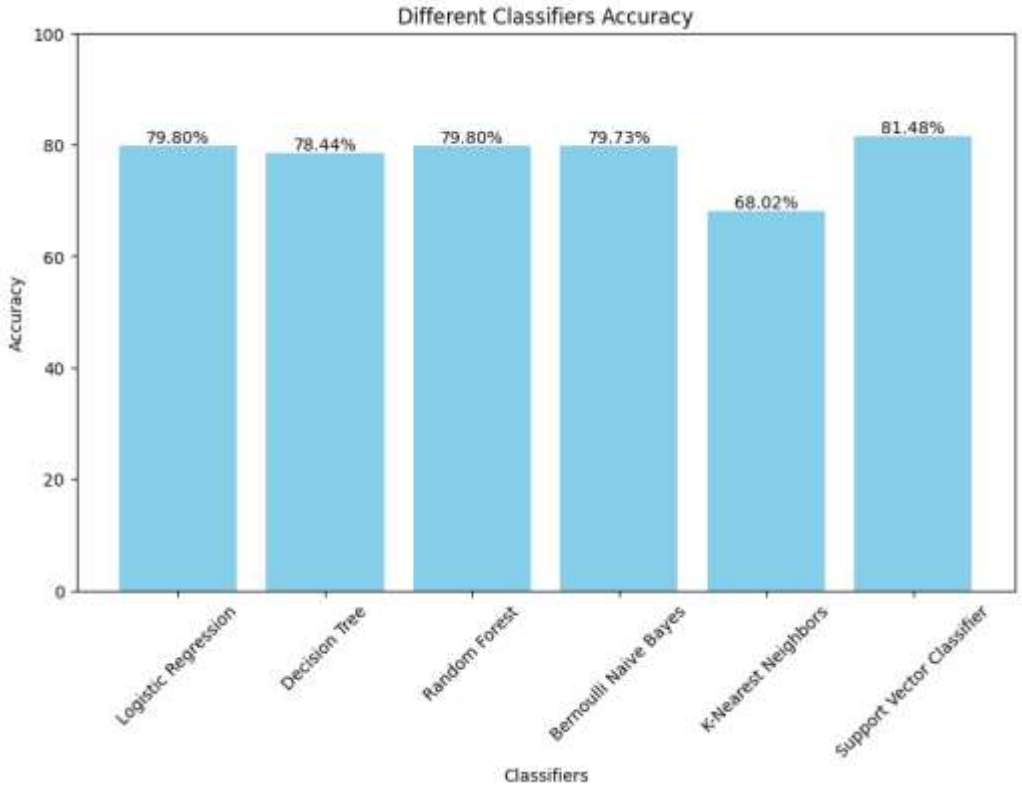


Figure-6: Different Classifier's Accuracy

4.1.2 Precision

The precision metric, as described by Mahmudun et al. (2016), assesses the accuracy of a class prediction. It shows whether or not the positive class's predictions are accurate. Precision measures the proportion of true positive predictions out of all positive predictions made by the classifier. Mathematically, precision (Prec) can be calculated using the following equation (Eq. 2):

$$Precision = \frac{TP}{TP + FP} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

Where:

- TP represents the number of True Positives,
- FP represents the number of False Positives.

This formula quantifies the correctness of positive predictions made by the classifier, providing insights into its ability to accurately identify instances of the positive class. A higher precision score indicates a lower rate of false positives and higher accuracy in predicting the positive class.

After training the algorithms with a dataset labeled with five classes, the precision scores obtained for each classifier were as follows: Logistic Regression (LR) achieved a precision of 82.46%, Decision Tree (DT) attained 80.62%, Random Forest (RF) achieved 81.97%, Bernoulli Naive Bayes (BNB) obtained 82.84%, K-Nearest Neighbors (KNN) achieved 77.05%, and Support Vector Classifier (SVC) yielded a precision of 82.19%.

Precision is defined as the ratio of correctly predicted positive views to all positively predicted observations. A lower false positive rate corresponds with a greater precision score, which denotes a higher prediction accuracy of positive opinions. Consequently, classifiers that score higher on accuracy metrics are better at correctly identifying pleasant emotions.

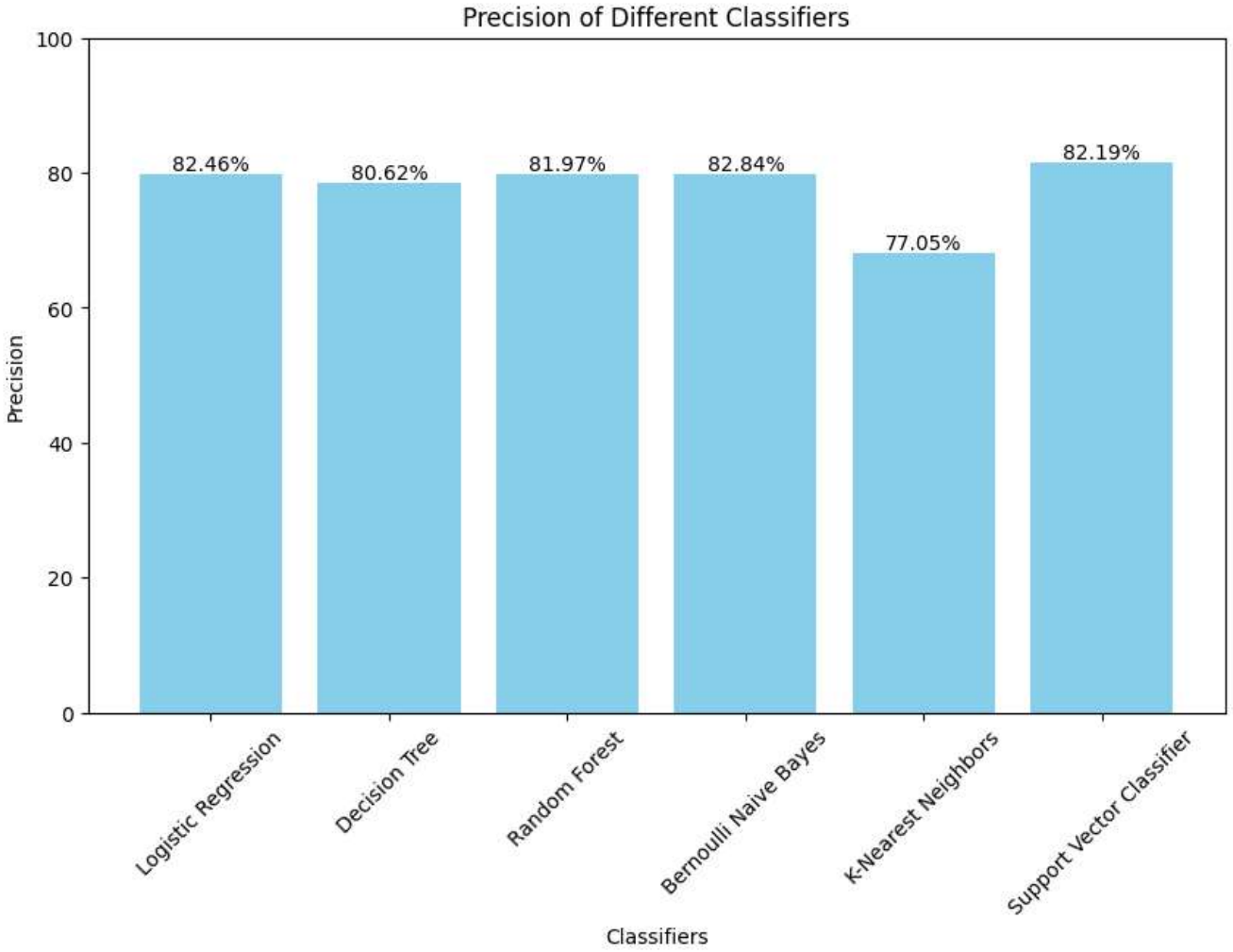


Figure-7: Precision of Different Classifiers

Figure-7 illustrates that Bernoulli Naive Bayes (BNB) exhibits high precision across all five classes. Conversely, the K-Nearest Neighbors classifier demonstrates lower precision values compared to BNB, although it successfully classifies all classes. Precision, while informative, may not provide a comprehensive understanding on its own, as it primarily focuses on the positive class. Therefore, It is frequently used in conjunction with the Recall metric to offer a more comprehensive evaluation of a classifier's effectiveness.

4.1.3 Recall

The recall, as defined by Mahmudun et al. (2016), represents the ratio of correctly predicted positive observations to all observations in the actual positive class. Unlike precision, recall considers false negatives (FN) in its calculation. Mathematically, recall (Rec) can be calculated using the following equation (Eq. 3):

$$Recall = \frac{TP}{TP + FN} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots. (3)$$

Where:

- TP represents the number of True Positives, and
- FN represents the number of False Negatives.

This formula quantifies the ability of the classifier to correctly identify all positive instances within the dataset, providing insights into its sensitivity to the positive class. A higher recall score indicates a lower rate of false negatives and higher coverage of positive instances by the classifier.

After training the algorithms with a dataset labeled with five classes, the recall scores obtained for each classifier were as follows: Logistic Regression (LR) achieved a recall of 79.80%, Decision Tree (DT) attained 78.44%, Random Forest (RF) achieved 79.80%, Bernoulli Naive Bayes (BNB) obtained 79.73%, K-Nearest Neighbors (KNN) achieved 68.02%, and Support Vector Classifier (SVC) yielded a recall of 81.48%.

Recall quantifies the model's ability to identify positive classes; that is, the percentage of correctly detected positive examples among all positive instances that actually occur. A higher recall score signifies a lower false-negative rate, indicating that the classifier effectively captures positive instances within the dataset. Therefore, classifiers with higher recall scores are more adept at identifying positive sentiments.

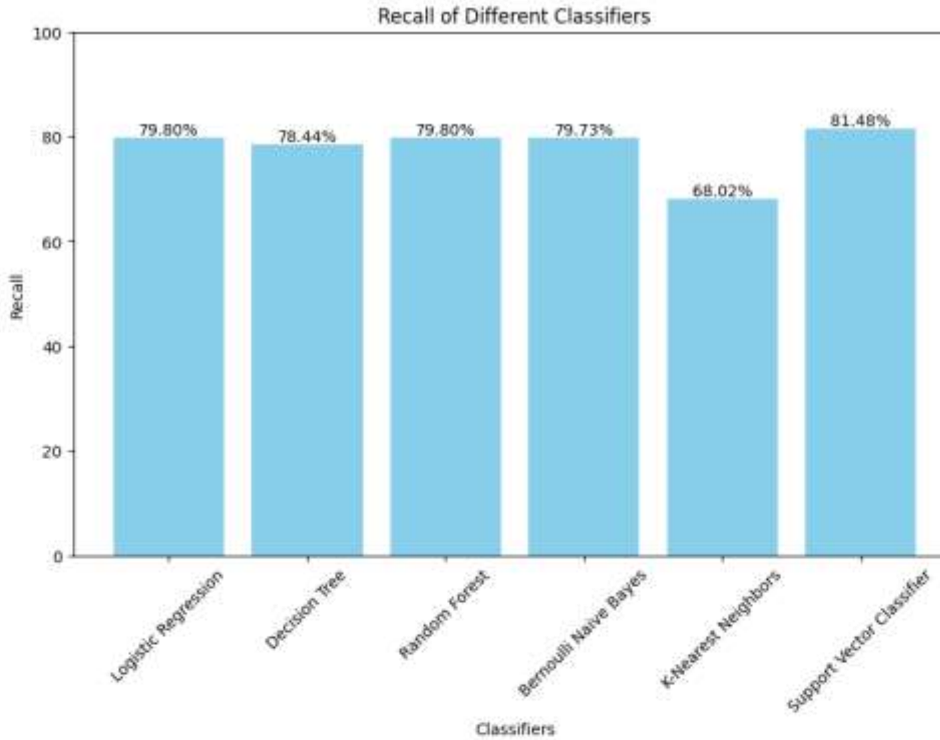


Figure-8: Recall of Different Classifiers

Figure-8 illustrates that the Support Vector Classifier (SVC) exhibits an overall higher recall score compared to other classifiers. This indicates that SVC has a greater ability to recognize positive classes and capture positive instances within the dataset. The higher recall score suggests that SVC effectively minimizes false negatives, contributing to its robust performance in identifying positive sentiments.

4.1.4 F1 Score

The F1 Score, as described by Ullah et al. (2017) and Bhuiyan et al. (2019), represents the weighted average of Recall and Precision, taking into account both FN (false negatives) and FP (false positives). This score provides a balanced measure of a classifier's performance, particularly in scenarios where the distribution of classes is uneven.

Mathematically, the F1 Score (F1) can be calculated using the following equation (Eq. 4):

$$F1\ Score = \frac{2\ X\ (Precision\ X\ Recall)}{Precision + \ Recall} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots. (4)$$

Where:

- Precision is the proportion of true positive predictions out of all positive predictions made by the classifier, and
- Recall is the ratio of correctly predicted positive observations to all observations in the actual positive class.

This formula considers both false positives (FP) and false negatives (FN), providing a comprehensive evaluation of a classifier's effectiveness. The F1 Score is particularly useful when false positives and false negatives do not have equal costs. If false positives and false negatives have equal costs, accuracy may suffice. However, in scenarios where their costs differ, Precision and Recall, and subsequently the F1 Score, offer more insightful metrics for evaluation.

Using a weighted average for calculating the F1 Score is a suitable approach, especially when dealing with imbalanced datasets. This method ensures that predictions are weighted based on the proportion of samples in each class, thus providing a more balanced evaluation of the classifier's performance.

It's important to recognize that while a higher F1-Score generally indicates better model accuracy, it does not automatically deem one classifier superior to another. The F1 Score combines both Precision and Recall, offering a holistic measure of the classifier's effectiveness. Therefore, a high F1-Score signifies high Precision and Recall, suggesting that the classifier achieves a good balance between minimizing false positives and false negatives.

Ultimately, the choice of the best classifier depends on the specific requirements and context of the task at hand. It's essential to consider various factors, such as the distribution of classes, the cost associated with false positives and false negatives, and the overall objectives of the analysis, when determining the most suitable classifier for a particular application.

After training the algorithms with a dataset labeled with five classes, the F1-Scores obtained for each classifier were as follows:

- Logistic Regression (LR): 79.39%
- Decision Tree (DT): 78.07%
- Random Forest (RF): 79.46%
- Bernoulli Naive Bayes (BNB): 79.25%
- K-Nearest Neighbors (KNN): 65.16%
- Support Vector Classifier (SVC): 81.38%

These F1-Scores provide a comprehensive evaluation of each classifier's performance, considering both Precision and Recall. A higher F1-Score suggests better overall model accuracy, as it indicates a balance between Precision and Recall. Therefore, the Support Vector Classifier (SVC) achieved the highest F1-Score among the classifiers evaluated in this study.

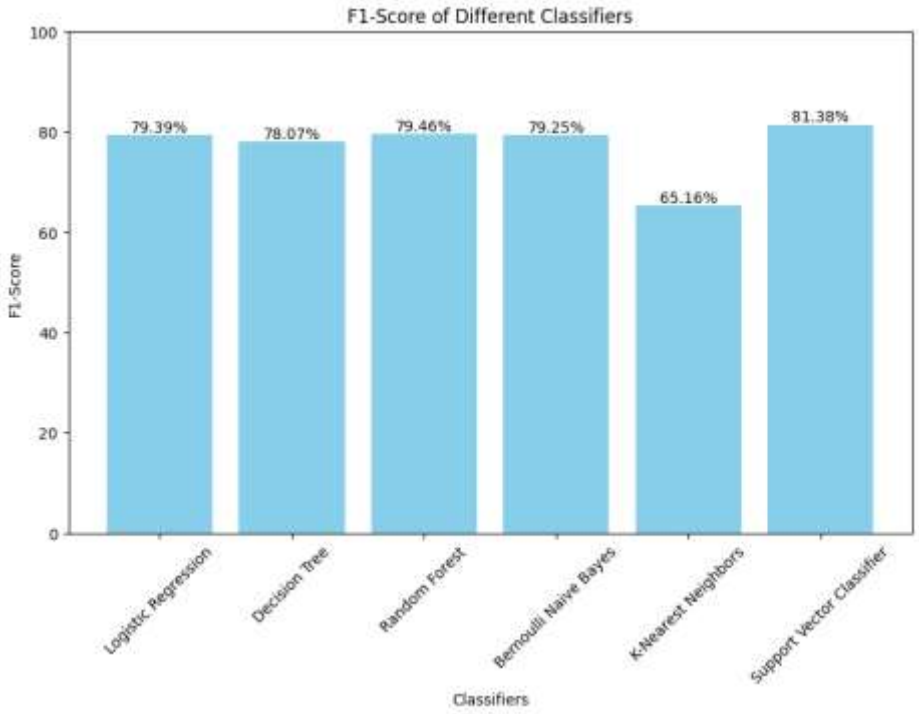


Figure-9: F1-Score of Different Classifiers

Indeed, the Support Vector Classifier (SVC) obtained the highest F1-Score among the classifiers evaluated in this study, achieving a score of 81.38%. This indicates that the SVC correctly predicted 81.38% of the data, while taking into account both false positive (FP) and false negative (FN) values.

The high F1-Score of the SVC suggests that it strikes a good balance between minimizing false positives and false negatives, resulting in strong overall performance in classifying sentiment. As a result, the SVC demonstrates its effectiveness in accurately predicting sentiment across multiple classes in the dataset.

4.2 Performance Table for N-gram Feature

A (n-1) order Markov model is a type of n-gram model used in probabilistic language modeling to predict the next item in a sequence. N-gram models, including (n-1) order Markov models, find applications in various fields such as probability theory, communication theory, computational linguistics (e.g., statistical natural language processing), computational biology (e.g., biological sequence analysis), and data compression.

One of the key advantages of n-gram models is their simplicity and scalability. These models, along with the algorithms that utilize them, are straightforward to implement and can scale efficiently to handle large datasets. Additionally, as the value of n increases in an n-gram model, the model can capture more context from the input data. This allows for a more accurate prediction of the next item in a sequence, striking a balance between the amount of context considered and the computational resources required.

Overall, n-gram models, including (n-1) order Markov models, are versatile and widely used in various applications due to their simplicity, scalability, and ability to capture contextual information from data efficiently.

4.2.1 Performance Table for Unigram Feature

A 1-gram (or unigram) is a one-word sequence.

	Classifier	Accuracy	Precision	Recall	F1-Score
0	Logistic Regression	84.903640	85.207714	84.903640	84.856460
1	Decision Tree	82.155603	82.230198	82.155603	82.135383
2	Random Forest	85.581727	85.702000	85.581727	85.560834
3	Bernoulli Naive Bayes	83.975732	84.368841	83.975732	83.911816
4	K-Nearest Neighbors	75.481799	78.404989	75.481799	74.740041
5	Support Vector Classifier	86.438258	86.725412	86.438258	86.399638

Table-3: Performance table for Unigram feature

From Table 3, I got the result:

Highest Accuracy achieved by Support Vector Classifier = 86.44
Highest Precision achieved by Support Vector Classifier = 86.73
Highest Recall achieved by Support Vector Classifier = 86.44
Highest F1-Score achieved by Support Vector Classifier = 86.40

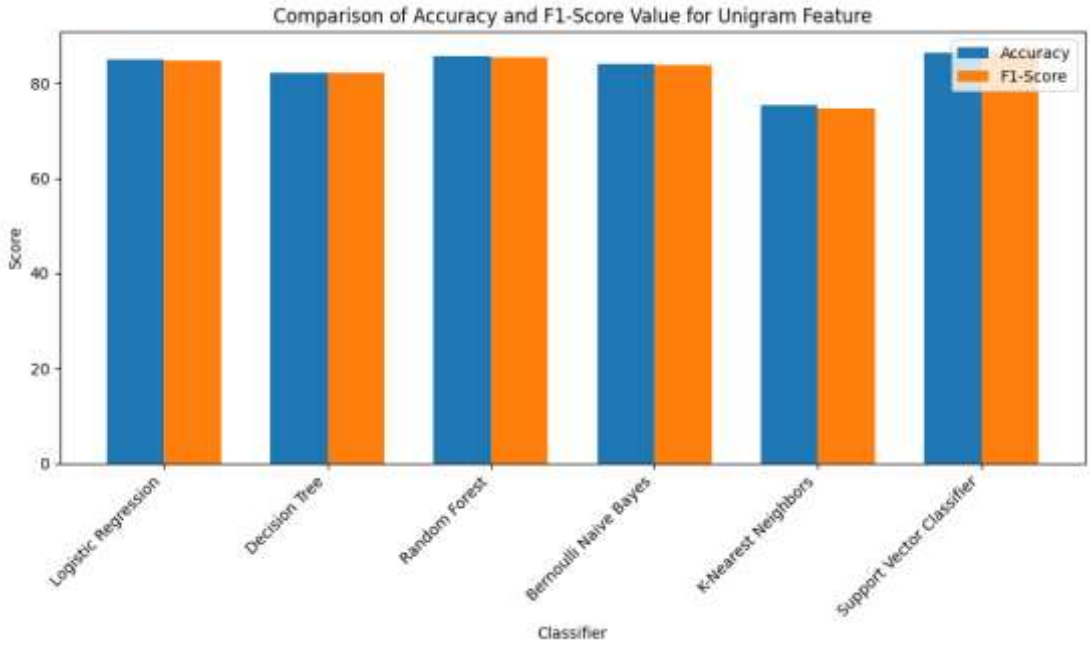


Figure-10: Comparison of Accuracy and F1-Score Value for Unigram Feature

4.2.2 Performance Table for Bigram Feature

A two-word string is known as a 2-gram (or bigram).

	Classifier	Accuracy	Precision	Recall	F1-Score
0	Logistic Regression	79.800143	82.604435	79.800143	79.291846
1	Decision Tree	79.086367	81.140070	79.086367	78.675819
2	Random Forest	80.406852	82.736901	80.406852	79.995414
3	Bernoulli Naive Bayes	81.513205	83.423799	81.513205	81.198203
4	K-Nearest Neighbors	64.061385	64.201025	64.061385	64.026709
5	Support Vector Classifier	82.369736	83.685306	82.369736	82.159146

Table-4: Performance table for Bigram feature

From Table 4, I got the result:

Highest Accuracy achieved by Support Vector Classifier = 82.37
Highest Precision achieved by Support Vector Classifier = 83.69
Highest Recall achieved by Support Vector Classifier = 82.37
Highest F1-Score achieved by Support Vector Classifier = 82.16

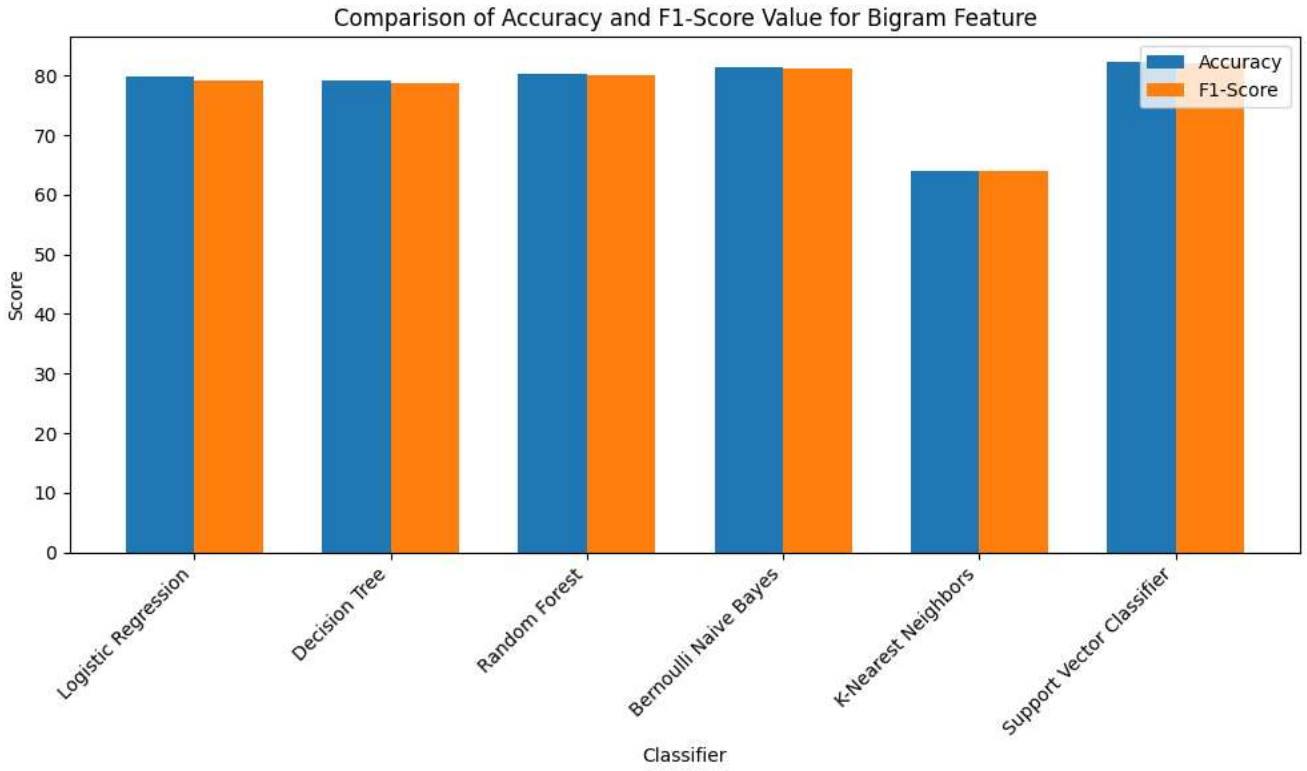


Figure-11: Comparison of Accuracy and F1-Score Value for Bigram Feature

4.2.3 Performance Table for Trigram Feature

A three-word string is known as a 3-gram (or trigram).

	Classifier	Accuracy	Precision	Recall	F1-Score
0	Logistic Regression	69.771592	79.364955	69.771592	66.866005
1	Decision Tree	68.772305	78.272384	68.772305	65.674023
2	Random Forest	69.700214	78.687044	69.700214	66.909885
3	Bernoulli Naive Bayes	71.163455	79.846751	71.163455	68.710640
4	K-Nearest Neighbors	60.349750	76.851754	60.349750	52.688290
5	Support Vector Classifier	73.162027	73.164050	73.162027	73.153488

Table-5: Performance table for Trigram feature

From Table 5, I got the result:

Highest Accuracy achieved by Support Vector Classifier = 73.16
Highest Precision achieved by Logistic Regression = 79.36
Highest Recall achieved by Support Vector Classifier = 73.16
Highest F1-Score achieved by Support Vector Classifier = 73.15

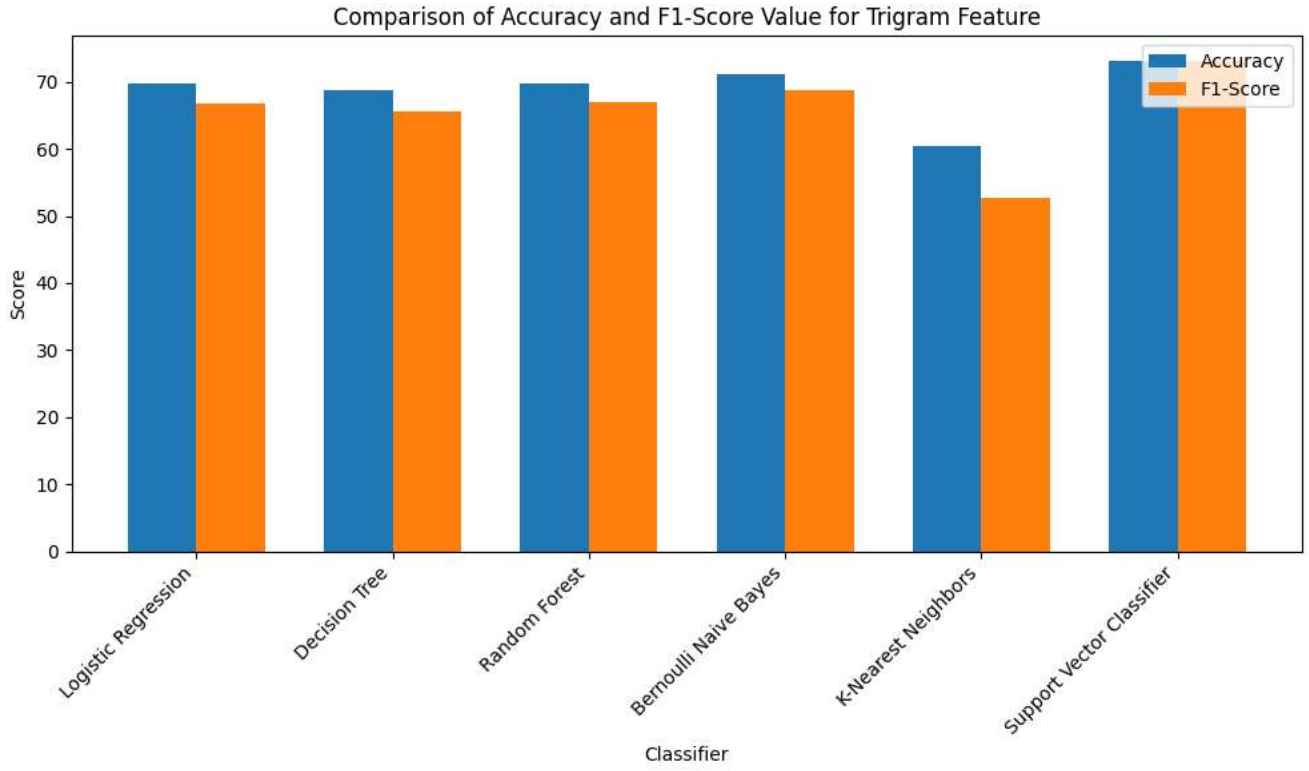


Figure-12: Comparison of Accuracy and F1-Score Value for Trigram Feature

4.3 Analysis

Based on the analysis presented above, it is evident that the Support Vector Classifier (SVC) outperformed other classifiers in terms of accuracy, precision, recall, and F1-Score. Therefore, SVC emerges as the most suitable choice for this research, indicating its potential for predicting human emotions towards various subjects or individuals.

This research aims to extract emotions from comments posted on YouTube videos, where individuals express their opinions and sentiments. By collecting comments from YouTube and utilizing the proposed Support Vector Classifier algorithm, it becomes feasible to predict human expressions or sentiments effectively. However, before deploying the algorithm for prediction, it is essential to train it with sample data to enhance its accuracy and effectiveness.

With further development and training, it is conceivable to create a fully automated tool capable of detecting human sentiment accurately. Such a tool could find applications in various domains, including social media analysis, market research, and customer feedback analysis, enabling organizations to gain valuable insights into public sentiment and opinion.

4.4 Prediction

I have tried to test our model by using a random comment data and I got a result. In Figure- 13 and Figure- 14, I can see happy and anger prediction from comment. That Mean's, I can see that my proposed model can extract sentiment from Bangla comment data.

```
#Enter your Bengali comment to predic emotion Happy or Anger:

comment = 'গল্প টা খুব সুন্দর ছিল'

sentiment = predict_sentiment(comment)

if sentiment == "0":
    print("Your Entered comment emotion is : Anger")
elif sentiment == "1":
    print("Your Entered comment emotion is : Happy")
else:
    print("Your Entered comment ,we can't calculate Sorry...")
```

```
Your Entered comment emotion is : Happy
```

Figure-13: Predicting the happy emotion

```
#Enter your Bengali comment to predic emotion Happy or Anger:

comment = 'সব থেকে বাজে নাটক'

sentiment = predict_sentiment(comment)

if sentiment == "0":
    print("Your Entered comment emotion is : Anger")
elif sentiment == "1":
    print("Your Entered comment emotion is : Happy")
else:
    print("Your Entered comment ,we can't calculate Sorry...")
```

```
Your Entered comment emotion is : Anger
```

Figure-14: Predicting the anger emotion

CHAPTER 5
Conclusion

5.1 Conclusion

Sentiment analysis, also known as opinion mining, is a process of extracting and analyzing human emotions, opinions, and attitudes from various sources such as text, speech, or facial expressions. It has become increasingly important in understanding public sentiment and attitudes towards different topics, products, or services.

Here's a more detailed explanation of its significance and applications:

Social Media Tracking: Social media platforms like Twitter, Facebook, and Instagram have become major channels for people to express their opinions on a wide range of topics. Sentiment analysis helps in tracking and analyzing these conversations to understand public sentiment towards brands, events, or social issues.

Product Analysis: Understanding customer sentiment towards products and services is crucial for businesses to make informed decisions. Sentiment analysis of product reviews, feedback, and comments helps companies gauge customer satisfaction, identify areas for improvement, and make strategic decisions in product development and marketing.

Customer Reviews Analysis: Customer reviews on platforms like Amazon, Yelp, or TripAdvisor contain valuable insights into consumer preferences and experiences. Sentiment analysis helps businesses analyze large volumes of reviews to identify trends, sentiments, and customer preferences, enabling them to enhance their products or services accordingly.

Clinical Service Analysis: In healthcare, patient feedback and satisfaction are vital for improving clinical services and patient experience. Sentiment analysis of patient reviews, surveys, and feedback forms helps healthcare providers identify areas of improvement, address patient concerns, and enhance overall service quality.

Research and Academia: Sentiment analysis is also widely used in research and academia to analyze public opinion on various topics, track sentiment trends over time, and study the impact of events or policies on public sentiment.

Given the widespread use of social media and the increasing popularity of platforms where people express their opinions, sentiment analysis has become a crucial tool for businesses, researchers, and organizations across various domains. In the context of the research article I mentioned, focusing on sentiment analysis for the Bengali language fills a significant gap and contributes to a deeper understanding of public sentiment and opinions in that linguistic context.

5.2 Limitations

Small Dataset: With only around 12,000 data points and a significant class imbalance, traditional deep learning algorithms like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) might not perform optimally. These algorithms typically require larger datasets to generalize well. However, there are techniques I can employ to address this issue:

- **Data Augmentation:** Generate synthetic data by applying transformations like rotation, translation, scaling, or adding noise to existing data points. This can help increase the size and diversity of your dataset.
- **Transfer Learning:** Utilize pre-trained models on larger datasets for feature extraction, then fine-tune them on your smaller dataset. This approach can leverage knowledge from larger datasets and adapt it to your specific task.
- **Ensemble Methods:** Combine predictions from multiple models trained on different subsets of the data or with different architectures. Ensemble methods can often improve overall performance compared to individual models.

Class Imbalance: The unequal distribution of labeled data for different classes (happy and anger) can lead to biased models that perform poorly on minority classes. To address this issue:

- **Resampling Techniques:** Over-sample the minority class or under-sample the majority class to achieve a more balanced distribution. Techniques like SMOTE (Synthetic Minority Over-sampling Technique) can generate synthetic samples for the minority class.
- **Class Weighting:** Assign higher weights to minority class samples during training to penalize misclassifications on these samples more heavily.

Unintentional vs. Intentional Spelling Errors: It seems like my system struggles to distinguish between genuine emotional expressions and errors caused by misspellings. To handle this:

- **Text Preprocessing:** Implement robust text preprocessing techniques to handle spelling errors, such as spell correction algorithms or incorporating context-based language models like BERT or GPT.
- **Error Detection:** Develop or integrate an error detection module into my system to identify and correct misspellings before emotion classification.

By addressing these challenges with appropriate techniques and methodologies, I can improve the accuracy and robustness of my emotion classification system, even with a small and imbalanced dataset. Experimenting with different approaches and combinations thereof can help identify the most effective strategies for your specific task.

5.3 Future Work

This paper suggests expanding the emotional scope of a system, presumably an AI or machine learning model, from just two emotions to include additional ones such as Neutral, Sad, and Amused. This expansion aims to provide a more nuanced understanding of human emotions, allowing the system to better interpret and respond to user input.

The dataset used for training the system is described as small and unbalanced, indicating that it may not adequately represent the full range of emotions and their frequencies in natural language. To address this limitation, the article proposes enriching the dataset by adding more balanced data. This could involve collecting and incorporating a larger and more diverse set of text samples that cover a wider spectrum of emotions.

In addition to expanding the emotional repertoire, the paper suggests improving the system's ability to handle variations in spelling and language usage. This includes detecting intentional or unintentional spelling mistakes as well as recognizing abbreviations or short forms of words. Enhancing the system's language processing capabilities in these areas can lead to more accurate and robust performance, particularly when dealing with user-generated content where spelling errors and informal language are common.

Overall, the proposed enhancements aim to make the system more sophisticated and versatile in understanding and responding to human emotions expressed through text, while also improving its ability to handle variations in language usage.

References

6. References

Alasmari, S. F., & Dahab, M. (2017). Sentiment detection, recognition and aspect identification. *International Journal of Computer Applications*, *975*, 8887.

Arya, A., Shukla, V., Negi, A., & Gupta, K. (2020, May). A review: Sentiment analysis and opinion mining. In *Proceedings of the International Conference on Innovative Computing & Communications (ICICC)*.

Attri, I., & Dutta, M. (2020). Review of various sentiment analysis approaches. In *Proceedings of International Conference on IoT Inclusive Life (ICIIL 2019), NITTTR Chandigarh, India* (pp. 223-234). Springer Singapore.

Awatramani, P., Daware, R., Chouhan, H., Vaswani, A., & Khedkar, S. (2021, September). Sentiment analysis of mixed-case language using natural language processing. In *2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA)* (pp. 651-658). IEEE.

Bautin, M., Vijayarenu, L., & Skiena, S. (2008). International sentiment analysis for news and blogs. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 2, No. 1, pp. 19-26).

Bhadane, C., Dalal, H., & Doshi, H. (2015). Sentiment analysis: Measuring opinions. *Procedia Computer Science*, *45*, 808-814.

Bhattacharya, S., & Banerjee, P. (2017). Towards the exploitation of statistical language models for sentiment analysis of twitter posts. In *Computer Information Systems and Industrial Management: 16th IFIP TC8 International Conference, CISIM 2017, Bialystok, Poland, June 16-18, 2017, Proceedings 16* (pp. 253-263). Springer International Publishing.

Bhadane, C., Dalal, H., & Doshi, H. (2015). Sentiment analysis: Measuring opinions. *Procedia Computer Science*, *45*, 808-814.

Bhuiyan, M., Rahman, A., Ullah, M. and Das, A.K., 2019. iHealthcare: Predictive model analysis concerning big data applications for interactive healthcare systems. Applied Sciences, 9(16), p.3365.

Bilal, S. (2021, March). A Linguistic System for Predicting Sentiment in Arabic Tweets. In *2021 3rd International Conference on Natural Language Processing (ICNLP)* (pp. 134-138). IEEE.

Chowdhury, S. and Chowdhury, W., 2014, May. Performing sentiment analysis in Bangla microblog posts. In 2014 International Conference on Informatics, Electronics & Vision (ICIEV) (pp. 1-6). IEEE.

Das, A.K., Al Asif, A., Paul, A. and Hossain, M.N., 2021. Bangla hate speech detection on social media using attentionbased recurrent neural network. Journal of Intelligent Systems, 30(1), pp.578-591.

Das, A. and Bandyopadhyay, S., 2010. Phrase-level polarity identification for Bangla. Int. J. Comput. Linguist. Appl.(IJCLA), 1(1-2), pp.169-182.

Das, D., 2011, March. Analysis and tracking of emotions in english and bengali texts: a computational approach. In Proceedings of the 20th international conference companion on World wide web (pp. 343-348).

Damasio A.R., 1998. Emotion in the perspective of an integrated nervous system. Brain research reviews, 26(2-3), pp.83-86.

Datacamp.com,2021.[Online].Available:https://www.datacamp.com/community/tutorials /svmclassification-scikit-learn-python.

Dhuria, S. (2015). Sentiment analysis: An approach in natural language processing for data extraction. *International Journal of New Innovations in Engineering and Technology*, *2*(4), 27-31.

Dorothy, M., & Rajini, S. (2016). The Various Approaches for Sentiment Analysis: A Survey. *vol*, *5*, 2014-2016.

Drovo, M.D., Chowdhury, M. Uday, S.I. and Das. A.K., 2019, June. Named entity recognition in bengali text using merged hidden markov model and rule base approach. In 2019 7th International Conference on Smart Computing & Communications (ICSCC) (pp. 1- 5). IEEE.

Ekman. P.E. and Davidson, R.J., 1994. The nature of emotion: Fundamental questions. Oxford University Press.

Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM*, *56*(4), 82-89.

Giatsoglou, M., Vozalis, M. G., Diamantaras, K., Vakali, A., Sarigiannidis, G., & Chatzisavvas, K. C. (2017). Sentiment analysis leveraging emotions and word embeddings. *Expert Systems with Applications*, *69*, 214-224.

Go, A., Bhayani, R. and Huang, L., 2009. Twitter sentiment classification using distant supervision. CS224N project report, Stanford, 1(12), p.2009.

Hasan, K.A. and Rahman, M., 2014, December. Sentiment detection from bangla text using contextual valency analysis. In 2014 17th International Conference on Computer and Information Technology (ICCIT) (pp. 292-295). IEEE.

Hogenboom, A., Bal, M., Frasincar, F., Bal, D., Kaymak, U., & De Jong, F. (2014). Lexicon-based sentiment analysis by mapping conveyed sentiment to intended sentiment. *International Journal of Web Engineering and Technology 7*, *9*(2), 125-147.

Hossain, M.T., Hasan, M.W. and Das, A.K., 2021, January. Bangla Handwritten Word Recognition System Using Convolutional Neural Network. In 2021 15th International Conference on Ubiquitous Information Management and Communication (IMCOM) (pp. 1- 8). IEEE.

Islam, J., Mubassira, M., Islam, M.R. and Das, A.K., 2019, February. A speech recognition system for Bengali language using recurrent neural network. In 2019 IEEE 4th international conference on computer and communication systems (ICCCS) (pp. 73-76). IEEE.

Jindal, K., & Aron, R. (2021). WITHDRAWN: A systematic study of sentiment analysis for social media data.

Kaushik, A., & Naithani, S. (2015). A study on sentiment analysis: methods and tools. *Int. J. Sci. Res.(IJSR)*, *4*(12), 2319-7064.

Khan, M. S. S., Rafa, S. R., & Das, A. K. (2021). Sentiment analysis on bengali facebook comments to predict fan's emotions towards a celebrity. *Journal of Engineering Advancements*, *2*(03), 118-124.

Mahmudun, M., Altaf, M.T. and Ismail, S., 2016. Detecting sentiment from Bangla text using machine learning technique and feature analysis. Int. J. Comput. Appl., 975, p.8887.

Mamun, M. A., Hossain, M. S., Siddique, A. B., Sikder, M. T., Kuss, D. J., & Griffiths, M. D. (2019). Problematic internet use in Bangladeshi students: the role of socio-demographic factors, depression, anxiety, and stress. *Asian Journal of Psychiatry*, *44*, 48-54.

Mishra, A. K., Saumya, S., & Kumar, A. (2021). Sentiment analysis of Dravidian-CodeMix language. In *Working Notes of FIRE 2021-Forum for Information Retrieval Evaluation (Online). CEUR*.

Mejova, Y. (2009). Sentiment analysis: An overview. *University of Iowa, Computer Science Department*.

Mehta, P., & Pandya, S. (2020). A review on sentiment analysis methodologies, practices and applications. *International Journal of Scientific and Technology Research*, *9*(2), 601-609.

Mumu, T.F., Munni, I.J. and Das, A.K., 2021. Depressed people detection from bangla social media status using lstm and cnn approach. Journal of Engineering Advancements, 2(01), pp.41-47.

Ojamaa, B., Jokinen, K., & Muischenk, K. (2015, May). Sentiment analysis on conversational texts. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)* (pp. 233-237).

Panikar, R., Bhavsar, R., & Pawar, B. V. (2022, March). Sentiment analysis: a cognitive perspective. In *2022 8th International Conference on advanced computing and communication systems (ICACCS)* (Vol. 1, pp. 1258-1262). IEEE.

Panksepp J., 2004. Affective neuroscience: The foundations of human and animal emotions. Oxford university press.

Pandey, P. and Govilkar, S., 2015. A framework for sentiment analysis in Hindi using HSWN. International Journal of Computer Applications, 119(19).

Salas-Zárate, M. P., Paredes-Valverde, M. A., Rodríguez-García, M. Á., Valencia-García, R., & Alor-Hernández, G. (2017). Sentiment analysis based on psychological and linguistic features for Spanish language. *Current trends on knowledge-based systems*, 73-92.

Shaikh, M. A. M., Prendinger, H., & Ishizuka, M. (2008). Sentiment assessment of text by analyzing linguistic features and contextual valence assignment. *Applied Artificial Intelligence*, *22*(6), 558-601.

Shukla, A., & Shukla, S. (2015). A Survey on Sentiment Classification and Analysis using Data Mining. *International Journal of Advanced Research in Computer Science*, *6*(7).

Sinha, H., & Kaur, A. (2016, November). A detailed survey and comparative study of sentiment analysis algorithms. In *2016 2nd International Conference on Communication Control and Intelligent Systems (CCIS)* (pp. 94-98). IEEE.

Solanki, M. S. (2019). Sentiment analysis of text using rule based and natural language toolkit. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, *8*(12S).

Taboada, M. (2016). Sentiment analysis: An overview from linguistics. *Annual Review of Linguistics*, *2*, 325-347.

Tan, K. L., Lee, C. P., & Lim, K. M. (2023). A survey of sentiment analysis: Approaches, datasets, and future research. *Applied Sciences*, *13*(7), 4550.

Tembhurnikar, S.D. and Patil, N.N., 2015. Sentiment Analysis using LDA on Product Reviews: A Survey. International Journal of Computer Applications, 975, p.8887.

Tuhin, R.A., Paul, B.K., Nawrine, F., Akter, M. and Das, A.K., 2019, February. An automated system of sentiment analysis from bangla text using supervised learning techniques. In 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS) (pp. 360-364). IEEE.

Umamaheswari, K. and Karthiga, R., Sentiment Classification based on Latent Dirichlet Allocation. International Journal of Computer Applications, 975, p.8887.

Ullah, M.R., Bhuiyan, M.A.R. and Das, A.K., 2017, September. IHEMHA: Interactive healthcare system design with emotion computing and medical history analysis. In 2017 6th International Conference on Informatics, Electronics and Vision & 2017 7th International Symposium in Computational Medical and Health Technology (ICIEV-ISCMHT) (pp. 1-8). IEEE.

Van den Berg, R. M., & Berg, R. M. (2008). *Proclus' Commentary on the Cratylus in context: ancient theories of language and naming* (Vol. 112). Brill.

Vaghela, V. B., Jadav, B. M., & Scholar, M. E. (2016). Analysis of various sentiment classification techniques. *International journal of Computer applications*, *140*(3), 975-8887.

Varghese, R., & Jayasree, M. (2013). A survey on sentiment analysis and opinion mining. *International journal of Research in engineering and technology*, *2*(11), 312-317.

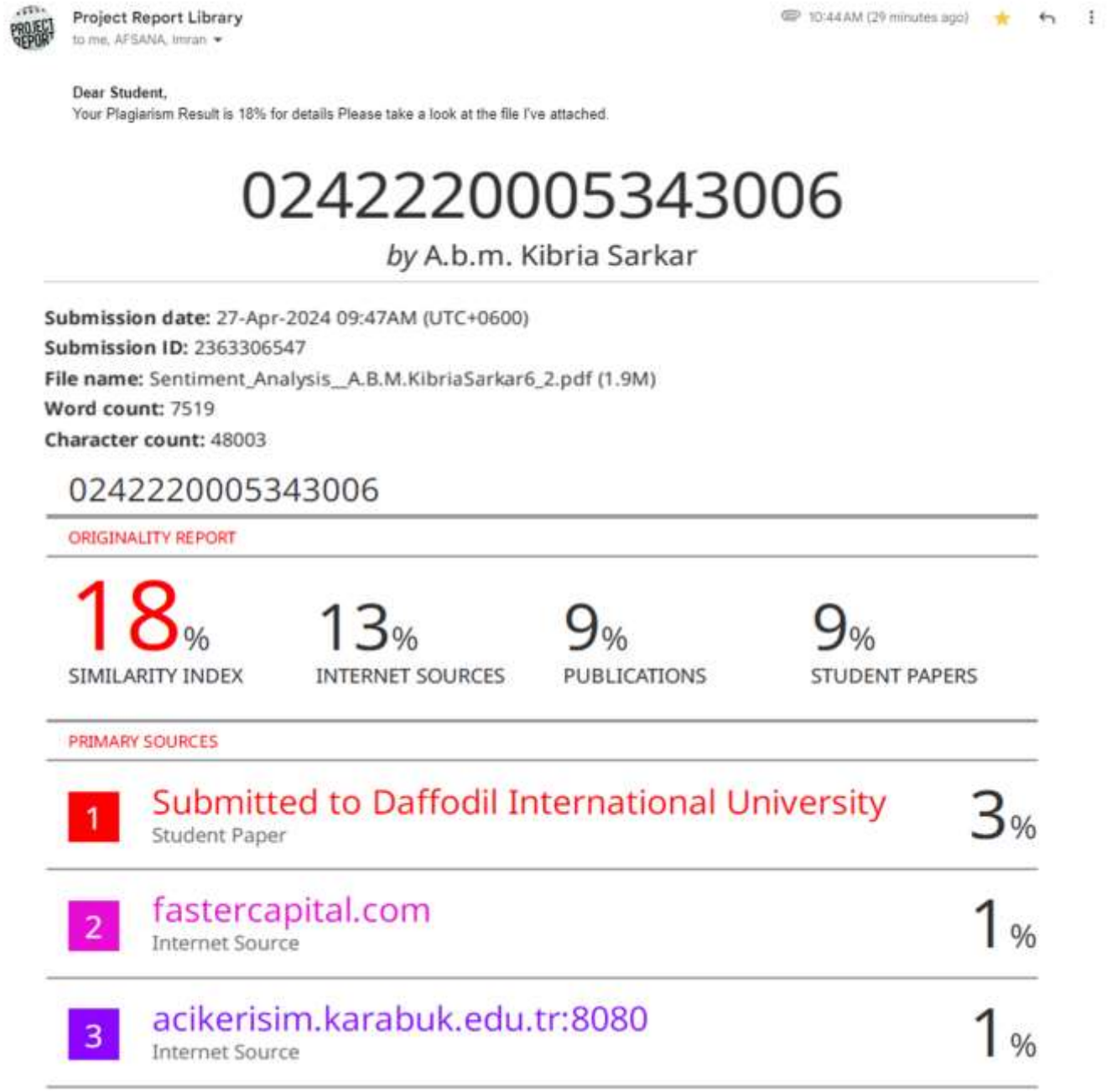
Verma, B., & Thakur, R. S. (2018). Sentiment analysis using lexicon and machine learning-based approaches: A survey. In *Proceedings of International Conference on Recent Advancement on Computer and Communication: ICRAC 2017* (pp. 441-447). Springer Singapore.

Verma, B., & Thakur, R. S. (2018). Sentiment analysis using lexicon and machine learning-based approaches: A survey. In *Proceedings of International Conference on Recent Advancement on Computer and Communication: ICRAC 2017* (pp. 441-447). Springer Singapore.

Viking Mäntylä., M. Graziotin, D. and Kuutila M., 2016. The Evolution of Sentiment Analysis-A Review of Research Topics, Venues, and Top Cited Papers. arXiv e-prints, pp.arXiv-1612.
Vohra, S. M., & Teraiya, J. B. (2013). A comparative study of sentiment analysis techniques. *Journal Jikrce*, *2*(2), 313-317.

Zia, S., Fatima, S., Mala, I., Khan, M. S. A., Naseem, M., & Das, B. (2018). A survey on sentiment analysis, classification and applications. *Int J Pure Appl Math*, *119*(10), 1203-1211.

Plagiarism Result: 18%



Project Report Library
to me, AFSANA, Imran

10:44 AM (29 minutes ago)

Dear Student,
Your Plagiarism Result is 18% for details Please take a look at the file I've attached.

0242220005343006

by A.b.m. Kibria Sarkar

Submission date: 27-Apr-2024 09:47AM (UTC+0600)
Submission ID: 2363306547
File name: Sentiment_Analysis__A.B.M.KibriaSarkar6_2.pdf (1.9M)
Word count: 7519
Character count: 48003

0242220005343006

ORIGINALITY REPORT

18%	13%	9%	9%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	3%
2	fastercapital.com Internet Source	1%
3	acikerisim.karabuk.edu.tr:8080 Internet Source	1%

Account Clearance