A Corpus of Recognizing Emotions from Bengali Social Media Comments

BY

ASHIK MAHMOOD
ID: 191-15-12060

AND

FAIROOZ RASHED
ID: 191-15-12405

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

Supervised By

Zerin Nasrin Tumpa
Senior Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

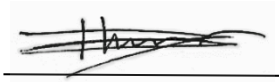
DHAKA, BANGLADESH

JANUARY 2023

APPROVAL

This Project titled **"A Corpus of Recognizing Emotions from Bengali Social Media Comments"**, submitted by ASHIK MAHMOOD and FAIROOZ RASHED to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 02-02-2023

BOARD OF EXAMINERS



Dr. Touhid Bhuiyan — **Chairman**
Professor and Head
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Dr. Sheak Rashed Haider Noorie — **Internal Examiner**
Professor and Associate Head
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Md. Sazzadur Ahmed — **Internal Examiner**
Assistant Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Dr. Md. Sazzadur Rahman — **External Examiner**
Associate Professor
Institute of Information Technology
Jahangirnagar University

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Zerin Nasrin Tumpa, Senior Lecturer, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Sùpervised by:

Zerin Nasrin Tumpa
Senior Lecturer
Department of CSE
Daffodil International University

Sùbmitted by:

Ashik Mahmood
ID: -191-15-12060
Department of CSE
Daffodil International University

Fairooz Rashed
ID: -191-15-12405
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

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

We are really grateful and wish our profound indebtedness to **Zerin Nasrin Tumpa**, **Senior Lecturer**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "*Machine Learning*" to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stages have made it possible to complete this project.

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

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

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

ABSTRACT

Sentiment analysis refers to the evaluation of any text, word, or other. Sentiment analysis, which is also defined as opinion mining, is a method of natural language processing (NLP) that determines the emotional subtext of a body of text. Numerous organizations use this method to identify and categorize customer opinions about a given good, service, or concept. Text is mined for sentiment and qualitative data using machine learning (ML), artificial intelligence (AI), and data mining methods. Opinion mining can extract the content, opinion holder, and polarity (or the quantity of positivity and negativity) from text in addition to identifying sentiment. Additionally, other perspectives, including article, paragraph, sentence, and sub-sentence divisions, can be utilized for sentiment analysis. The expansion of social media platforms is creating a ton of texts and grabbing people's attention. From such data, Sentiment Analysis (SA) derives informative information. The majority of study on SA has been conducted in English, but Bengali and other important languages are also required. Because Bengali is the fifth most often spoken language among native speakers and is extensively implemented on social media, it is crucial to concentrate on Bengali social posts and comments. Despite the large number, Bengali Sentimental Analysis has seen some minor progress. But creating an automated method for classifying emotions in low-resource languages like Bengali is a crucial task. The task is made more difficult by a lack of resources and benchmark corpora. Consequently, creating a benchmark for the development of an emotion classifier for Bengali texts requires a corpus. A total of 5000 texts are labeled into 11 basic emotion categories such as anger, fear, surprise, sad, happy, disgust, funny, abusive, advice, neutral and praise. Our study's primary goal is to more specifically identify the emotions expressed in sentences and comments that are written in Bengali. We selected Eleven classifications for this reason in order to more precisely and accurately define the emotions.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

1.1 Introduction

Social media are interactive media platforms that facilitate the ability to share content such as thoughts, opinions, information and other kinds of expression with others via online communities and networks.

Sentiment Analysis is growing in popularity as a field of study in text mining,especially in the Bengali dialect where little research has been done. Notwithstanding being a significant presence among languages worldwide, Bengali sentiment analysis has received little attention. Enormous datasets are known as big data, and their volume makes them difficult to manage and analyze using standard database tools. Common instances of Big Data are consumer reviews on e-commerce websites or posts on social networks. Users of social media platforms express their opinions on a range of topics, including politics, finance, religion and so forth. Massive amounts of user-generated data are produced as a result, and they might be valuable. Sentiment analysis is developing into a major area of study in the Big Data era. One subfield of sentiment analysis is text mining. Even though Sentimental Analysis has received a great deal of attention in the English language, Bengali social media lacks a substantial amount of resources for various types of speech detection. However, as there is a lack of dataset and other resources for Bengali text categorization, creating and implementing machine learning models to address real-world issues is highly challenging for low resource languages like Bengali. Therefore, there is a need for research on the causes and management of social media. We select Facebook and YouTube as our data platforms in this instance.

This article analyzes how we endeavored to resolve the problem. 5000 Bengali comments from the comment areas of both YouTube and Facebook are included in our dataset. We choose comments from eleven different categories, including newspapers, viral people, sports, entertainment, politics, crime etc. On our dataset, we ran a total of seven deep learning models.

A considerably greater amount of labeled corpus is required to train the machine

learning models in order to gain reasonable performances due to the supervised aspects of the classification approaches. Quality labeled data, however, falls short in a number of areas. A difficult research topic for the resource-constrained language, especially in Bengali, is the processing of these enormous amounts of data to identify underlying sentiment or emotions. When handling languages with scarce resources, like Bengali, NLP technologies have failed to deliver acceptable performance. The complexity results from a number of constraints, including a lack of tools, a dearthof benchmark corpora, and learning strategies. Emotion-based classification based on the eleven emotion classes has not yet been carried out because Bengali is a language with limited resources. The first necessity and ensuing prerequisite for creating an automatic emotion classifier for Bengali literature is an emotion corpus. This study provides an overview of the process of creating a corpus that may be used to categorize emotions in Bengali texts in order to respond to this topic. To more precisely define the emotions represented in Bengali texts and comments, we take intoconsideration eleven different types of textual emotions, including anger, fear, surprise,sad, happy, disgust, funny, abusive, advice, neutral, and praise.

1.2 Motivation

With the explosion of Web 2.0 platforms like social media, blogs, forums etcetera, users choose to share their emotions like happy moments, working environment where they work, celebration of birthdays, sad moments, opinion and so on. As we know that for example Facebook can detect bad comments or harassment comments, post easily because, first of all, the English language comments can detect easily because there are a number of works done of sentimental analysis on English language. Nowadays people comment on Facebook, YouTube and other media in Bengali words. It is hard topredict the sentiment of those comments. There are some reasons behind this. There isnot that much work behind this language and also lacking of a big dataset. Most of thepaper, people just classify only three classes, positive, negative and neutral. Some papers cover a maximum of five to six sentiment classes. We feel that Bengali language sentiment needs to be more specific and emotions can be recognized correctly. So, we decided to make eleven sentiment classes so that it can be more accurate.

1.3 Rationale of the study

Bengali language is the fourth most spoken language in the world. In the research field, especially in comparison to the English language, quite few studies on sentiment analysis have been conducted in Bengali. Most of the paper addressed only three sentiments on Bengali language and they are positive, negative and neutral. Certain studies addressed additional types of sentiment. Therefore, it can be claimed that a Bengali statement cannot accurately express its true meaning. The main aim of our work is to more accurately determine the emotions expressed in Bengali sentences and comments. For this reason, and in order to more precisely and accurately identify the emotions, we choose eleven classifications including anger, fear, surprise, sad, happy, disgust, funny, abusive, advice, neutral, and praise.

1.4 Research Questions

The result of a related study, along with theoretical analysis and data exploration, is a research question. Here, the question of how the task is carried out is answered, along with a brief description of the system.

- How to investigate the several text expressions in the Bengali language to identify the distinguishing characteristics of emotion classes?
- How to develop an emotion corpus for the purpose of emotion classification in Bengali text?
- Which several factors are considered to identify the distinct characteristics of each emotion in Bengali texts?
- How much raw data is collected from social media?
- What kind of algorithms are suited to carry out our tasks more precisely?
- How much information do we need to successfully carry out our work?
- How can this model be trained using the right data set?

1.5 Expected Outcome

As we mentioned that to obtain more precise sentiment, we'll work on several different sentiment classes. We have initially two goals from this work.

1. The main aim of our work is to more precisely define the emotions expressed in Bengali sentences and comments.
2. In our study, we construct a larger corpus that can be utilized by a machine learning model to categorize each text with one of eleven emotion classes anger, fear, surprise, sad, happy, disgust, funny, abusive, advice, neutral, and praise.in order to more precisely and accurately define the emotions

1.6 Project Management and Finance

To do the research, we needed to go outside to printed dataset to get the approval signature from the respected teachers. However, we collected data from Facebook and YouTube comment sections. We used some online platforms like google colab to run code, an extension named “Easy web data scraper” to collect data from Facebook and YouTube and also we also collected some data manually. After that we separated unique values and removed unnecessary colon, semicolon, English words and others. For these purposes, we needed the internet and electricity. The total cost is almost 3,240 taka.

1.7 Report Layout

This project is based on a thesis and has a total of six segments. The work presents a variety of perspectives, each of which is represented by a chapter. Distinct subheadings are divided up into each chapter and presented in an understandable manner. A list of everything in this report is provided.

This report consists of 6 chapters. These are:

Chapter 1: Introduction

1.1 Introduction

1.2 Motivation

1.3 Rationale of the Study

1.4 Research Questions

1.5 Expected Output

1.6 Project Management and Finance

1.7 Report Layout

Chapter 2: Background

2.1 Preliminaries/Terminologies

2.2 Related Works

2.3 Comparative Analysis and Summary

Research

6.1 Summary of the Study

6.2 Conclusions

6.3Implication for Further Study

Chapter 1

Introduces the effort and discusses its goals, drivers, research questions, and expected outcomes. The topics we have discussed: 1.1 Introduction, 1.2 Motivation, 1.3 Rationale of the Study, 1.4 Research Questions, 1.5 Expected outcome 1.6 Project Management and Finance and 1.7 Report Layout. Introduction includes the overview of the study purposes and details. This study provides an overview of the process of creating a corpus that may be used to categorize emotions in Bengali texts in order to respond to this topic. To more precisely define the emotions represented in Bengali texts and comments, we take into consideration eleven different types of textual emotions, including anger, fear, surprise, sad, happy, disgust, funny, abusive, advice, neutral, and praise. The primary driving force for our thesis study was covered in the section on motivation. The study's major goal was covered in the section on the Rationale of the study part. The Research question chapter explains the major questions that pertain to our investigation. The expected outcome section includes a description of the result we've been working toward. The Project Management and Finance document details the financial costs and the correct project management. TheReport layout structure breaks out our whole project into chapters.

Chapter 2

It provides a summary of earlier work that has been done in this situation. Studying earlier work makes it easier to fully comprehend the work that has to be done for our research. We witness the effects of the authors' choice to draw a line through this field of study later in Chapter 2. The following subjects have been covered: Preliminaries/Terminologies, Related Works, Comparative Analysis and Summary, Scope of the Problem, and Challenges are all included in this report.

Chapter 3

It is relevant to this study's theoretical debate. Extending the already used statistical techniques to explore the theoretical element of the research was one aspect of this endeavor. The workings of deep learning are also shown in this chapter.

There are:

3.1 Research Subject and Instrumentation ,3.2 Proposed Methodology

3.2.1 Properties Of Building Bengali Emotion Corpus, 3.3 Bengali Emotions of Corpus Development Schematic Process, 3.4 Data Collection Procedure

3.5 Statistical Analysis, 3.6 Implementation Requirements, 3.6.1. NumPy

3.6.2.Pandas,

3.6.3.train_test_split,

3.6.4. Tokenizer ,

3.6.5. Dataset: sentiment_analysis_bengali_dataset.csv UTF-8,

3.6.6. From the sklearn import allalgorithms.

Chapter 4

Presents the experiment results, analyzes the data, and discusses the conclusions. This section presents some experimental photographs that were taken whilst the project was being developed. It has been covered in 4.1 Experimental Setup, 4.2 Experimental Results & Analysis, and 4.3 Discussion

Chapter 5

Specifies what is trustworthy to appear in the complete project report and proposal. The chapter comes to a close with a discussion of the limits of our study, which may be used as a springboard for other people's future research. We have talked about the following subjects: Impact on Society, Environmental Impact, are all listed in section

Chapter 6

Provides an overview of the research and upcoming projects. We have talked about the following subjects: 6.1 The Study's Executive Summary, 6.2 The Findings, and 6.3 Implication for Future Research The study's executive summary provides an overview of our whole body of work. Our whole study, including the findings is displayed in the conclusion. The subsequent development of our work has been detailed in the section titled Implications for Further Study 5.1.

CHAPTER 2
Background

2.1 Related Works

This present era, people do comments, post a number of things on social media and other platforms. There is a shortage of Bengali resources as not much work is done related to this language although it is spoken widely. Researchers have published various publications on A Corpus of Recognizing Emotions from Bengali Social Media comments. Some of them merely attended three classes on sentiment. Forclassification, several of them attended five or six classes. The paper contains supervised deep learning and supervised machine learning to process each corpus in a publication. Compare and contrast the categorization algorithms Naive Bayes, Support Vector Machine, and Logistic Regression. With a performance rate of around 86.7%, SVM outperforms LSTM for classification when utilizing a deep learning technique..

When applied to a pre-processed corpus, deep learning performs worse than machine learning, providing an accuracy of 72.86%, but it performs better when applied to an unprocessed corpus[1]. Similarly this paper covers six basic emotions named happy, sad, anger, fear, surprise, and disgust. Individuals develop a corpus with 1200 data. SVM classifiers give an output of 73% accuracy when Naive based approach of 60% accuracy[2]. Moreover, there is a paper named "Toxicity Detection on Bengali Social Media Comments using Supervised Models", used five supervised learning models (NaiveBayes, Support Vector Machines, Logistic Regression, Convolu-tional Neural Network, and Long Short Term Memory. They have achieved a highest accuracy of 95.3% from Convolutional Neural Network[3]. Furthermore, in a research work, the two classes named depressive and non-depressive data used for GRU model.

GRU sizes were 64, 128, 256, 518, and 1024. For these sizes, they acquired accuracy, with the accuracy being 59.1%, 70.0%, 67.3%, 74.5%, and 69.1%. Epoches and sizes were employed.

It is demonstrated that the batch size strongly influences the number of epochs needed. High accuracy requires a set of balanced batch sizes with a sizable number of epochs. For a 5 layered GRU with a size of 512, a batch size of 5, and a learning rate of 0.0001 over 3 epochs, the best accuracy we could achieve was 75.7%[4]. A article applies various models named Multiple machine learning and deep learning-based algorithms, including Linear Support Vector Classifier (LinearSVC), Logistic Regression (Logit), Multinomial Naive Bayes (MNB), Random Forest (RF), Artificial Neural Network (ANN), and Recurrent Neural Network (RNN) with a Long Short Term Memory (LSTM) cell, have been evaluated in this paper to identify different sorts of abusive Bengali text. They used seven emotion classes named Slang Religious_hatred, Personal_attack, Politically violated, Antifeminism , Positive and Neutral. RNN, an algorithm built on deep learning, outperforms competing algorithms by achieving the best accuracy of 82.20%[5].

Besides, a self developed corpus named BEmoC (Bengali Emotion Corpus) which covers six different emotions classes. The work contains several metrics including Cohen's Kappa, Zipf's law, coding reliability, most frequent emotion words and density of emotions. The annotator's opinion and the expert's opinions are compared and have a final result. The result is h 0.969 which is called Cohen's κ score[6].

2.2 Comparative Analysis and Summary

As from the beginning, we wanted to make a unique, supportive and more accurate system. So we have proposed a total of seven approaches named LR, DT, RF, MNB, KNN, Linear SVM and RBF SVM. We have a dataset of 5000 data which is taken from Facebook and YouTube comments. We leveled the dataset with eleven different classes. Our main objective is to identify the Bengali comment more accurately and more specific.

Literature Review

Paper Name	Model/Algorithm	Findings	Accuracy
Depression Analysis of Bangla Social Media Data using Gated Recurrent Neural Network	GRU (Gated Recurrent Neural Network)	Dataset: Collected using google form Focus: Their main focus was to find out the reason of depression Limitation: Dataset is small.	75.7%
Hate Speech detection in the Bengali language: A dataset and its baseline evaluation	SVM, LSTM, Bi-LSTM	Dataset: Extract from Facebook and YouTubecomments section using open-source software. Focus: For better labeled dataset to find out the hate speech in Bengali	87.5%
Toxicity Detection on Bengali Social Media Comments using Supervised Models	Naive Bayes, Support Vector Machines, Logistic Regression, Convolutional Neural Network, and Long Short Term Memory	Dataset: Collected manually form social media Focus: Main focus was to find out anti-social behavior and the toxicity from social media which hamper social atmosphere.	Naive Bayes 81.80%, SVM:84.73 %, LR:85.22%, LSTM :94.13%, CNN:95.3%
A Comparative Sentiment Analysis On Bengali Facebook posts	supervised machine learning Approach, supervised deep learning Approach, SVM, LR, NB, RNN	Dataset: Fetch data using Graph API Emotion Classes: Positive, Strong Positive,Negative, Strong Negative and Neutral	SVM:29.68 % LR: 30.25 % NB: 26.64 % Processed_ test :- 60.32 % Unprocessed _test : 72.86 %

BEmoC: A Corpus for Identifying Emotion in Bengali Texts	Cohen's Kappa, Zipf's law, coding reliability, most frequent emotion words and density of emotions.	Dataset: Data collected manually Emotion Classes: Anger, Fear, Surprise,Sadness, Joy, Disgust Focus: Label data by annotators and experts and then compare them.	Score: 0.969 Cohen's k
A Corpus of Recognizing Emotions from Bengali Social Media Comments	LR, DT, RF, MNB, KNN, SVM, RBF SVM	Dataset: Data collected manually from Facebook and YouTube Emotion Classes: anger, fear, surprise, sad, happy, disgust, funny, abusive, advice, neutral, and praise.	MNB UNI: 81.39% MNB BG: 80.17% MNB TRI-GRAM: 78.93%

Table 2.2: Comparative Analysis of Literature reviews

2.3 Scope of the problem

There could be two scopes. We have collected data from Facebook and YouTube comment sections. There are a number of people who comment in Bengali and furthermore they use some emoji which cannot be deleted automatically. So, it could be a problem. Another one is, if any special character input without knowing, the dataset will not work perfectly. So, we have to be careful of these criteria.

2.4 Challenges

There might be some difficulties. Data collection must come first. It is challenging to collect data. After data gathering, the dataset must be processed, and no English words, numeric, or special characters are permitted. It has to be resolved. Second, we need to identify the model with the highest accuracy that can handle this dataset. Thirdly, in order to gain more accuracy, we must keep a large amount of data for training and a lower proportion for testing.

CHAPTER 3

Research Methodology

3.1 Research Subject and Instrumentation

Our research topic name is “A Corpus of Recognizing Emotions from Bengali Social Media comments”. Our main focus is to get more specific results and recognize the Bengali language sentiment more meaningfully. For this purpose, we selected Eleven classifications for this reason in order to more precisely and accurately define the emotions. Therefore, understanding sentiment from the Bengali language will undoubtedly be a remarkable achievement for all.

3.2 Proposed Methodology

Emotion is a complex phenomenon that includes experience, cognition, feelings, behavior, physiology, and conceptualization. Numerous techniques are used to determine human emotions, including body language, facial expressions, blood pressure, heart rate, and text data. This research study focuses on the identification of emotion in Bengali texts. The creation of a Bengali emotion corpus is the main goal of our study, which is discussed in the following subsections. It can also be employed for classification and emotion analysis purposes

Demonstrates the proposed approach of emotion recognition in a schematic manner. The five major parts of this strategy are corpus building, training, classifying, testing, and recognition.

3.2.1 Properties Of Building Bengali Emotion Corpus

The majority of the previous corpora were created with the intention of categorizing Bengali text into three polarities of sentiment: positive, negative, and neutral. Additionally, these datasets are too tiny to be used in an accurate emotion classification model based on machine learning or deep learning. In our study, we present a wider corpus that can be used by a machine learning model to categorize emotion in Bengali language.

These eleven main emotion classes are: anger, fear, surprise, sad, happy, disgust, funny, abusive, advice, neutral, and praise. We tag each text into one of these categories.

Bengali Emotion Expression

Emotion class	Keywords
Anger	রাগ, মেজাজ, শালা
Fear	ভয়, নির্জন, আঁতকে
Surprise	অবাক, আশ্চর্য, হঠাৎ
Sad	কষ্ট, কান্না, খারাপ
Happy	আনন্দ, সুন্দর, ভালোবাসা
Disgust	বিরক্ত,বাজে, ফালতু
Funny	হাসকর, মজার, কৌতুকপূর্ণ
Abusive	গালিগালাজপূর্ণ, অবমাননাকর, অভদ্র
Advice	উপদেশ, পরামর্শ
Neutral	নিরপেক্ষ, উদাসীন
Praise	প্রশংসা, মুগ্ধতা, সাধুবাদ

Table 3.2.1: Commonly used keywords in Bengali emotion expression

We investigated through a number of Bengali text expressions to find the characteristics that differentiate various emotion classes apart. The classes that are taken into consideration in this work are: anger, fear, surprise, sad, happy, disgust, funny, abusive, advice, neutral, and praise. Several factors, including keywords, the intensity of emotion words, semantic of sentence, emotion engagement, think like the person are taken into consideration.

- **Emotion Keywords:** We identified words commonly used in the context of a particular emotion. For example, the words, "joy", "enjoy", "pleased" are considered as seed words for the happy category. Thus, some specific seed words are stored for a specific emotion in Bengali. For example, "রাগান্বিত" (Angry) or "ক্রোধ" (Anger) usually used for expressing "Anger" emotion. Likewise, "খুশি" (joy) or "মন ভালো" (Good mood) are usually used for expression "happy" emotion. Table 3.2.1 shows some commonly used emotion keywords in the context of a particular emotion [6].
- **Intensity of Emotion Word:** In Bengali, different seed words express different emotions in a particular context. In such cases, seed words are compared in terms of intensity and choose the highest intensity seed word, including its emotion class, which is assigned for the emotion of that context [6].

  Consider the following example: আলেকজান্ডারের মৃত্যুর সংবাদ এথেন্সে পৌঁছল। তখন একজন অবাক হয়ে প্রশ্ন করলো, " আলেকজান্ডার মৃত! অসম্ভব। তিনি মারা গেলে প্রতিটা কোণা থেকে মৃত লাশের গন্ধ ভেসে আসত।

  (English translation: when the news of Alexander's death reached at Athens, someone was surprised and asked, "Alexander is dead! Impossible! If he's dead, the smell of his dead body would waft from every corner of the earth.")

  In these texts, several keywords are existed such as, "মৃত্যুর সংবাদ" (death news), "অবাক" (surprised) and "মৃত! অসম্ভব" (dead! Impossible!) Here the words, "অবাক" (surprised) and "মৃত! অসম্ভব" (dead! Impossible!) have more weight than "মৃত্যুর সংবাদ" (death news). Thus, this type of text can be considered as "surprise" because the intensity of this emotion is higher than the intensity of sadness emotion [6].

  • **Semantic of Sentence:** Observing the semantic meaning of the texts is one of the prominent characteristics of ascertaining emotion class. In the previous example, though the sentence started with the death news of Alexander, this sentence turns into astonishment of a regular person in Athens. So, sentence semantics make an essential parameter in designating emotion expression [6].

• **Emotion Engagement**: It is imperative to involve the annotation actively while reading the text for understanding the semantic and context of the emotion expression explicitly. For example, "সেন্টমার্টিনে কাটানো প্রতিটা মুহূর্ত অসাধারন ছিলো। অসংখ্য সেই মুহূর্ত থেকে ক্যামেরাবন্দী কিছু মুহূর্ত।" English translation: moment spent in St. Martin was awesome. Here are some from those countlessmoments captured on camera) [6].

In this particular expression, annotators can feel some happiness as it describes an original moment of someone's experience. This feeling causes annotators engaged with happiness, and the expression designated as "happy" [6].

- **Syntactic Structure:** Sometimes, a syntactic structure plays a vital role during annotation. Let us consider two examples, "কে বলেছে তোমাকে এই কাজ করতে? অনেক কাবিল হয়ে গিয়েছ তাই না?" (English translation: Who told you to do this? How could you do that!) By investigating these sentences, it is found that both the sentences consist of similar words, but their syntactic structures are different. The first example is like someone encountered with rage but the second one with astonishment. Thus, annotators label the first sentenceas "Anger" and the second as "Surprise" [6].
- **Think Like The Person (TLTP):** Usually, an emotion expression is a type of expression of someone's emotion in a particular context. By TLTP, an annotator imagines him/her in the same context where the emotion expression is displayed. By repeatedly uttering, an annotator tried to imagine the situation and annotated the emotion class [6].

3.3 Bengali Emotions of Corpus Development Schematic Process

Here, Figure 1 shows the overview of the development process of Bengali emotion corpus which consists of four major phases: data crawling, prepossessing, data labeling, and label verification, respectively.

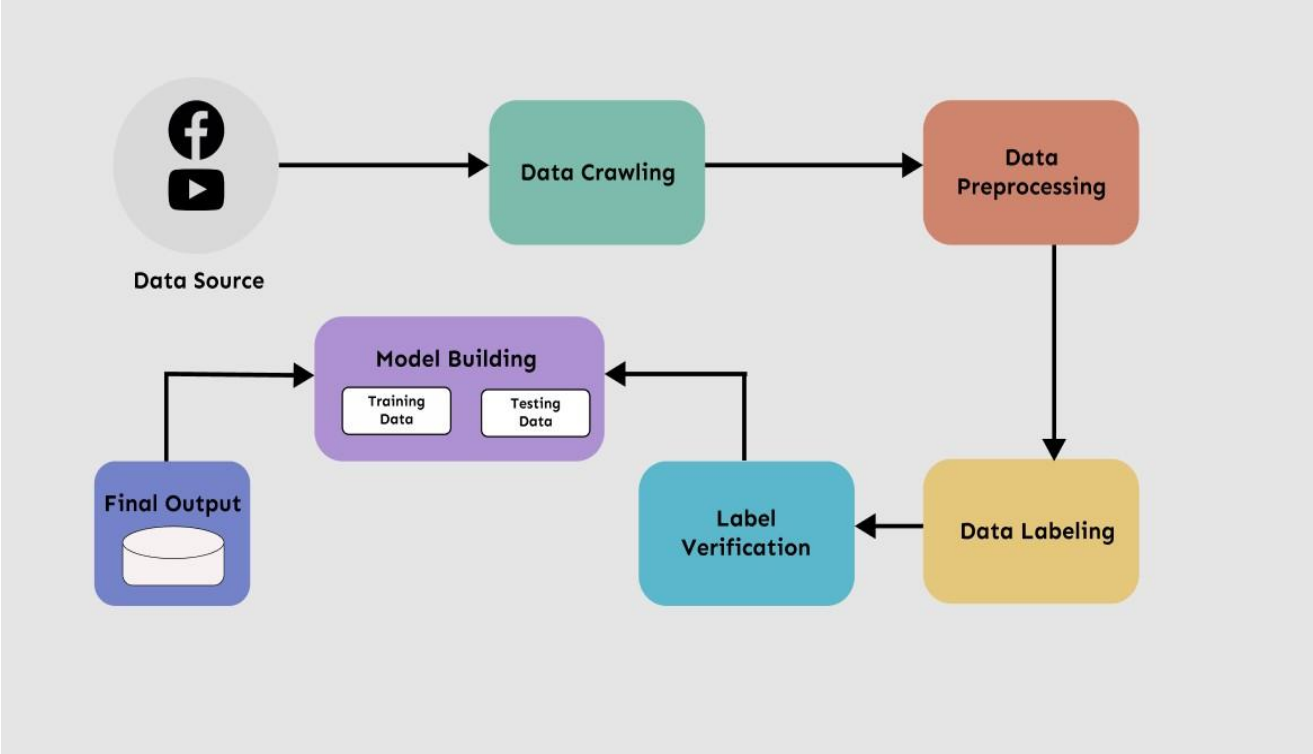


Fig 3.3.1: Process flow diagram for creating the Bengali emotion corpus

A. Data Crawling: Text in Bengali was gathered from a variety of sources, including Facebook comments and posts, YouTube comments, online blog posts, newspapers, and more, on a variety of subjects, including drama, movies, crime news, celebrities, and more. Two participants were assigned to accumulate data. Over the period of two months, we manually collected 5000 text expressions (September 2,2022– November 5, 2022)

B. Data Preprocessing: An essential phase in the data mining process is data preprocessing. It describes the processes of preparing data for analysis by cleansing, converting, and integrating it. The purpose of data preprocessing is to enhance the data's quality and suitability for the particular data mining operation. Data preprocessing is required to fill in missing values, smooth noisy data, and addressinconsistencies.

- **Data cleaning:** in this process, missing, inconsistent, or irrelevant data are found and removed. This can involve eliminating redundant data, adding values when they are missing, and dealing with outliers.

```
Original:
 হাত দেখে মনে হচ্ছে এই আগুন সিগারেটের আগুন
Cleaned:
 হাত দেখে মনে হচ্ছে এই আগুন সিগারেটের আগুন
 Sentiment:--  neutral

Original:
 শাকিব ভাই
Cleaned:
 শাকিব ভাই
 Sentiment:--  neutral

Original:
 সিরাজগঞ্জের ছেলে আমি সোহেল খান
Cleaned:
 সিরাজগঞ্জের ছেলে আমি সোহেল খান
 Sentiment:--  neutral
```

Fig.3.2.2 shows a fragment of the corpus after cleaning.

- **Data integration:** It is the process of merging data from several sources, including text files, spreadsheets, and databases. One consistent view of the data is what integration aims to produce.

There are two stages to preprocessing: manual and automatic. Typographical errors were removed from the collected data during the manual phase. The text is modified to remove unnecessary or irrelevant words. To train the emotional text classifier, we use the text's main body. The text document is then represented by a list of words and their frequencies. Emojis and punctuation can occasionally lead to confusion about the emotional intensity of the data, according to research. Emojis from the manually processed data were therefore removed during the automatic phase.

C. Data Labeling: The collected data were then manually categorized into eleven emotion categories. Several factors are taken into account in order to discover the distinctive qualities of each emotion in Bengali texts. The factors are mentioned above in the "Properties of Building Bengali Emotion Corpus".

D. Label Verification: The dataset was manually verified each labeling by three experts who were academician and working as senior lecturer on Bangla in a reputed institution.

E. Training and Testing Phase: The classifier is trained using a set of text files as the training sample. A trained machine learning model that has undergone training will be used in the testing phase to identify and detect emotions. Test was used to determine a sample text that extracts the required features that are utilized to train the classifier model. Text that has unclear or undetermined emotional categories is provided as input to the assessment phase. The test text sample's emotion category was determined by the classifier module using features that were extracted. Seven classification methods were used to measure the effectiveness.

1. Linear Regression(LR)
2. Decision Tree(DT)
3. Random Forest(RF)
4. Multinomial Naive Bayes(MNB)
5. K-Nearest Neighbor(KNN)
6. Linear Support Vector Machine (SVM)
7. RBF SVM

Linear Regression(LR): Linear regression is a statistical algorithm used to predict a Y value, given X features. Using machine learning, the data sets are examined to show a relationship. The relationships are then placed along the X/Y axis, with a straight line running through them to predict further relationships.

Linear regression calculates how the X input (words and phrases) relates to the Y output (polarity). This will determine where words and phrases fall on a scale of polarity from "really positive" to "really negative" and everywhere in between.

Decision Tree(DT): Decision trees can be trained on some annotated data as they are supervised algorithms. Therefore, the fundamental concept is the same as for any text classification: given a set of documents (for example, represented as TF IDF vectors along with their labels), the algorithm will determine the degree to which each word correlates with a given label.

For example, it might be discovered that the term "outstanding" frequently appears in documents with a positive label, but the word "awful" typically appears in documents with a negative label. By fusing all of these observations, a model that can label any document is created.

Random Forest(RF): Using the classification group rather than a single classification, the random forest technique for group categorization classifies additional points based on classification predictions. The overfitting problem that occurs in models of decisions is solved by the Random Forest technique, which can handle vast volumes of data with high dimensions.

Multinomial Naive Bayes(MNB): When categorizing texts based on a statistical examination of their contents, the multinomial naive Bayes algorithm is frequently utilized. It offers an alternative to "heavy" AI-based semantic analysis and significantly streamlines the classification of textual material.

By calculating the likelihood that a document belongs to the class of other papers with the same subject, classification attempts to categorize text fragments into different classes.

Each document is made up of several words that help the reader grasp what the paper is about. A class is a tag used to identify one or more papers that are about the same subject. By performing the statistical analysis and evaluating the hypothesis that a document's phrases have already appeared in other documents from a certain class, documents are assigned to one of the categories already in existence.

K-Nearest Neighbor(KNN): The neighborhood of data samples is determined by their proximity and closeness. Depending on the issue at hand, there are various approaches

to determine how close or how far apart two data points are from one another. Most well-known and often used is straight-line distance (Euclidean Distance).
Neighbors typically exhibit comparable traits and behaviors, making it possible to treat them as members of the same social group.
The primary idea behind this straightforward supervised learning classification technique is as follows. Now that we have considered the K-Nearest Neighbors of the unknown data, we may categorize and assign it to the group that appears most frequently in those K neighbors using the KNN technique. When K=1, the unlabeled or unknown data will be given the class of its nearest neighbor.

Linear Support Vector Machine (SVM): Support Vector Machine, generally known as SVM, is a linear model used to solve classification and regression issues. It works well for many real-world issues and can solve both linear and non-linear problems. The SVM concept is straightforward: A line or a hyperplane that divides the data into classes is produced by the algorithm.
Finding a dividing line (or hyperplane) between the data of two classes is, in general, what SVMs perform. The SVM algorithm takes the data as input and, if it is possible, outputs a line that divides the classes.
Let's start with a challenge. Assume you need to separate the red rectangles from the blue ellipses in the dataset below (we'll call them positives and negatives).

RBF SVM: The RBF kernel, which is the default kernel in the sklearn SVM classification algorithm, contains the following formula: where gamma can be manually changed and must be > 0.

3.4 Data Collection Procedure

As we previously mentioned, we have collected data from Facebook comments and YouTube comment sections. We used an extension named “Easy web data scraper” to collect data from Facebook and YouTube.We deleted other information from the datasetlike post ult, owner of the post etcetera. Our data size is 1.181 MB. After processing,

there are mainly two attributes available in our dataset named Comments and Tag. We have converted our dataset to CSV(UTF-8) format.

3.5 Statistical Analysis

For the machine learning process, we focused on the statistical portion. Here, we detect maily classes with the amount of sentiment. Total 5000 data distributed into eleven classes. These data are unique sentences, there are no emojis, punctuations, English words, numeric values and special characters.

Evaluation Measures: We used a variety of evaluation matrices, including the confusion matrix, precision, recall, and F1 score, to determine the effectiveness of our suggested method.

- **Confusion Matrix:** The performance of the classification model is assessed using a tabular representation of the data. Due to the fact that our system uses a multi-class classification model, we used a confusion matrix with the dimensions 7 (row) × 5. (column). The totals for true positives, false positives, true negatives, and false negatives are shown in this matrix, respectively.

- **Precision:** refers to positive predictive value. It calculates the ratio of exactly classified text into a particular class to the total number of classified texts of that emotion class. Precision can be obtained by Eq. (i)

  Precision = TP / TP + FP ….. (i)

- **Recall:** It calculates the ratio of correctly classified text into a particular class to the total number of classified texts of that emotion class.

  Eq.(ii).

  Recall = T P T P + F N …. (ii)

- **F1 score:** It is the weighted mean of recall and precision measures. Eq. (iii) is used to calculate F1 score.

  F1 = 2 × recall × precision recall + precision ……(iii)

• **Accuracy:** Accuracy is used as a statistical evaluation of how well a classification test correctly determines or keeps out a condition. Therefore, the accuracy is the proportion of true results both true positives and true negatives among the total number of test samples. Accuracy can be measured using the Eq. (iv).

$$\text{Accuracy} = TP + TN\,TP + TN + FP + FN \quad \text{.....(iv)}$$

If there are any of them present, there will be some errors. So we carefully handled this. The classifiers are given below:-

1. Happy : 1246
2. Sad : 1195
3. Neutral : 789
4. Funny : 525
5. Praise : 402
6. Surprise : 271
7. Anger : 252
8. Advice : 135
9. Disgust : 71
10. Fear : 70
11. Abusive : 41

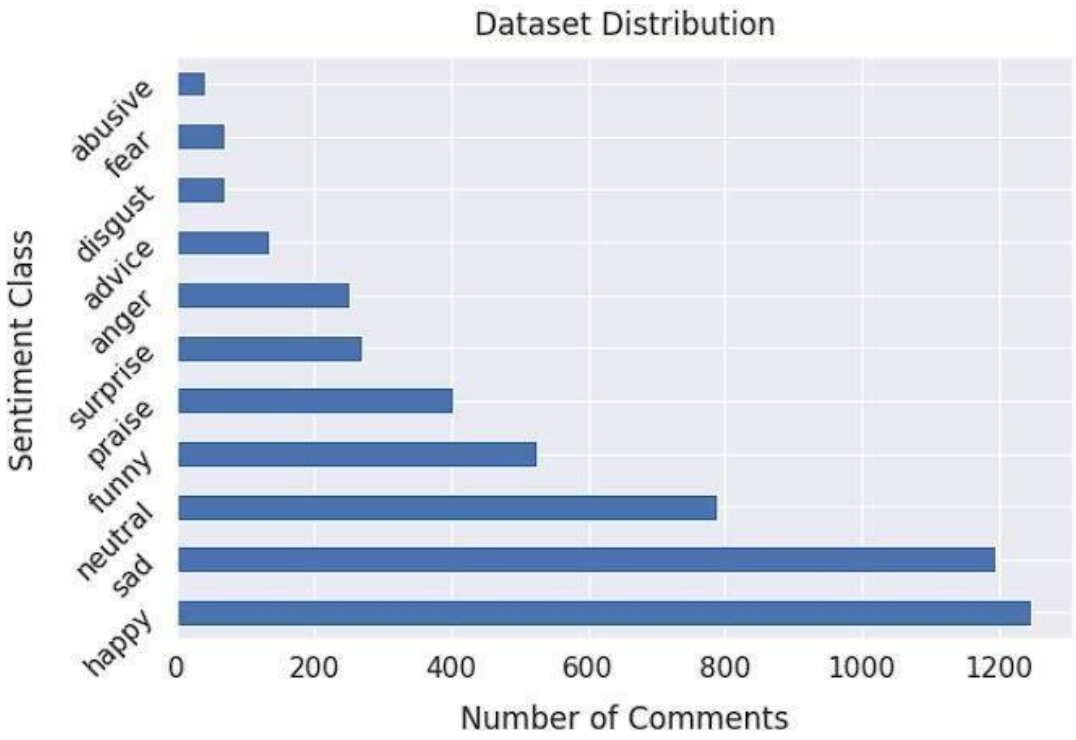
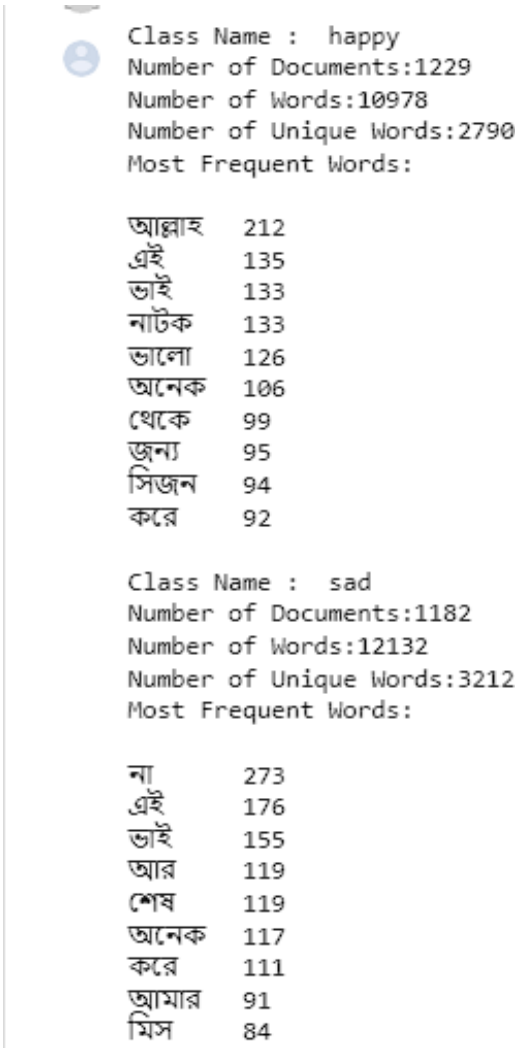


Fig 3.5: Dataset Distribution based on sentiment classification

Now we find out the total number of documents, words, and unique words under the eleven sentimental classes.



```
Class Name :  happy
Number of Documents:1229
Number of Words:10978
Number of Unique Words:2790
Most Frequent Words:

আল্লাহ   212
এই       135
ভাই      133
নাটক     133
ভালো     126
অনেক     106
থেকে     99
জন্য     95
সিজন     94
করে      92

Class Name :  sad
Number of Documents:1182
Number of Words:12132
Number of Unique Words:3212
Most Frequent Words:

না       273
এই       176
ভাই      155
আর       119
শেষ      119
অনেক     117
করে      111
আমার     91
মিস      84
```

Fig 3.5.1: Examples of number of words used

The figure shows the number of documents, words and unique words, frequent words. We can see how much time a word is used. For example, মিস = 84 from sad class means the word is found 84 times from the dataset. They also find out the unique words under every class. Unique word means the word will not be found in any rows except its present position.

SUMMARY OF STATISTICS OF THE BENGALI EMOTION DEVELOPED CORPUS

Class Name	Number of Documents	Number of words	Number of unique words
Anger	250	2250	1194
Fear	69	647	442
Surprise	266	2271	1095
Sad	1182	12132	3212
Happy	1229	10978	2790
Disgust	63	531	400
Funny	517	4127	1845
Abusive	41	413	328
Advice	135	1499	854
Neutral	721	4961	1690
Praise	399	4079	1304
	Total = 4872	Total = 43888	Total = 15154

Table 3.5.2: Statistical overview of the Bengali emotion corpus

Here,

Total Number of Documents = 4,872

Total Number of Words = 43,888

Total Number of Unique Words = 15,154

Data is classified into three categories. They are total documents, total words and total unique words. The Data statistics is given below graphically: -

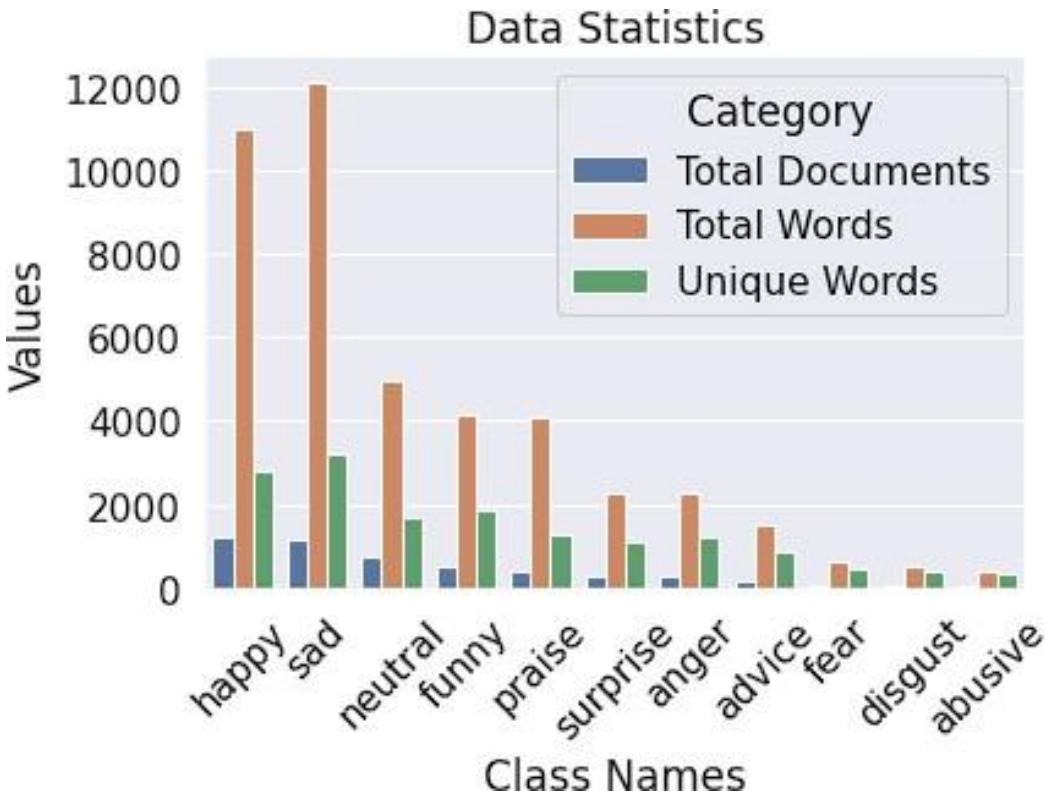


Fig 3.5.3: Graphical representation of data statistics

It’s a graphical representation of a data summary.

Now that we are developing models, we have selected two classes of data: "happy" class (1229 data) and "sad" class (1178 data).
They'll be set to use as training and testing data. In total, 90% of the data will be used for training, and 10% for testing.
The features are grouped into three categories: Unigram, Bigram, and Trigram.

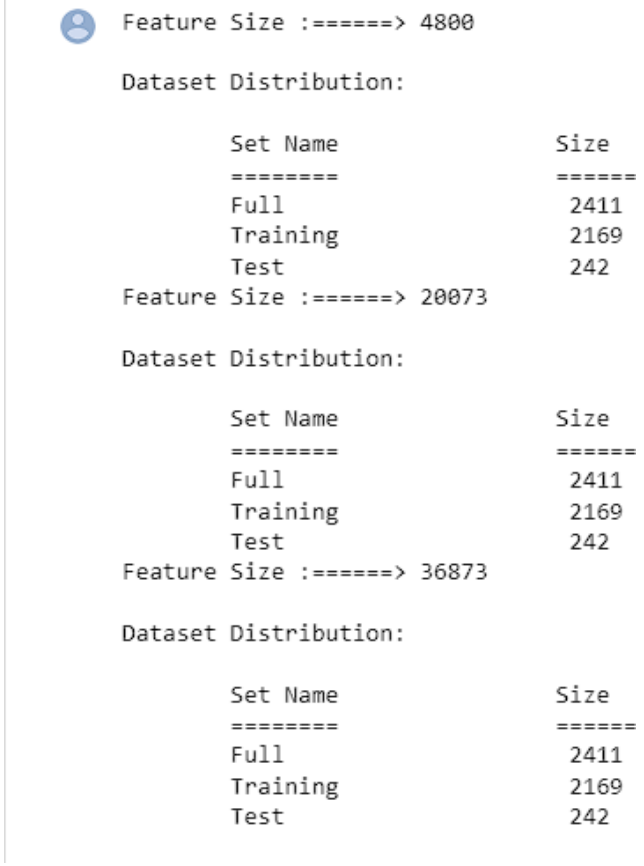


Fig 3.5.4: Training and testing dataset sample

Training and testing data implemented successfully.

3.6 Implementation Requirements

Tools:

3.6.1. NumPy

3.6.2. Pandas

3.6.3. train_test_split

3.6.4. Tokenizer

3.6.5. Dataset: sentiment_analysis_bengali_dataset.csv UTF-8

3.6.6. From the sklearn import all algorithms

3.6.1. NumPy

It is a fundamental library that genuinely assists in our ability to perform scientific computation, particularly in the area of machine learning. It is primarily used in Python for scientific calculations. For improved accuracy, it is a more compact Python list than other lists.

3.6.1. Pandas

To analyze data Pandas is executed. In essence, it serves as a library for data analysis. It serves as a tool for compact and is utilized for packages.

Prior to Panda's Library, it was very challenging to locate the many combine's packages. Pandas essentially found in 2008. The Pandas library has many tools for finding data. Basically, it employs four different types of features.

Basically, pandas use four features,

1. Pivot Table
2. Split Apply Combine
3. Data visualization
4. Working with missing data

3.6.2. train_test_split

Our data is separated into train and test sets using the train test split () function.

3.6.3. Tokenizer

Tokenization is a technique used in natural language processing to break down phrases and paragraphs into simpler language-assignable elements.

3.6.4. Dataset: sentiment_analysis_bengali_dataset.csv UTF-8

We have saved the excel sheet into CSV format and chosen UTF-8 because if we do not choose this, the CSV file will not show us the actual format. That is why it is used.

3.6.5. From the sklearn import all algorithms

In linear regression, a coefficient describes changes in a response variable. The correlation analysis is known as the coefficient of determination. This phrase is used to describe the accuracy or level of fit in a regression.

CHAPTER 4

Experimental Results and Discussion

4.1 Experimental Setup

For this project, we are now utilizing an online experiment. We use Google Colab online to run the experiment. We need a stable internet connection, at least 4 GB of RAM, and other things to run the experiment. RAM is not affected by Colab. So, there is no need for concern.

4.2 Experimental Results & Analysis

For this project, we are using several types of models. Our main motive is to ensure original sentiment of a word or sentence can be found. Although we took eleven classes, it is not easy to not get good accuracy.

4.2.1 Performance Visualization

Accuracy is shown in a bar chart. It will be easy to understand.

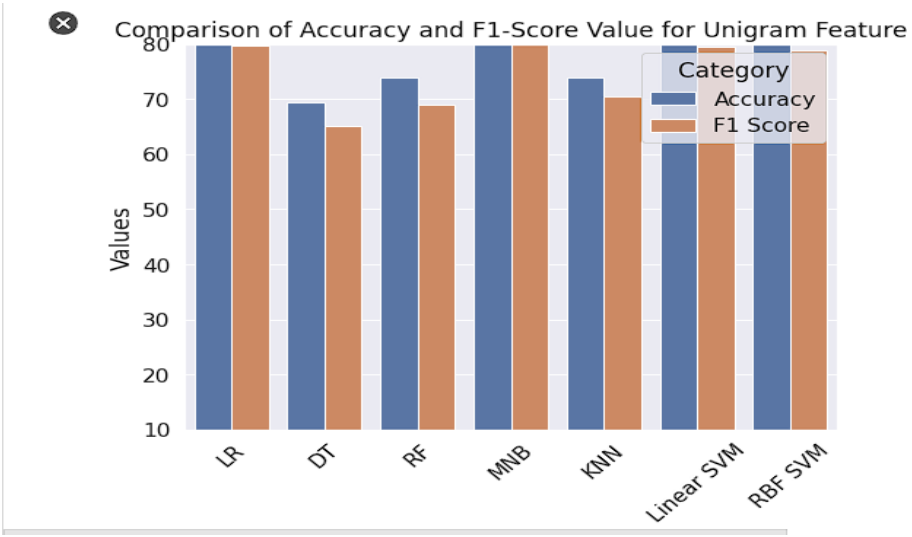


Fig 4.2.1: Data visualization for Unigram Feature

Performance Table (Unigram)

Model	Accuracy	Precision	Recall	F1 Score
LR	80.99	77.59	81.82	79.65
DT	69.42	67.65	62.73	65.09
RF	73.97	75.27	63.64	68.97
MNB	81.40	76.00	86.36	80.85
KNN	73.97	72.82	68.18	70.42
LINEAR SVM	81.40	79.82	79.09	79.45
RBF SVM	80.58	78.38	79.09	78.73

Table: 4.2.1 Performance table for Unigram Feature

Highest Accuracy achieved by MNB at = 81.3999999999999

Highest F1-score achieved by MNB at = 80.85

Highest Precision Score achieved by Linear SVM at = 79.8200000000001

Highest Recall Score achieved by MNB at = 86.36

After implementing all algorithms, highest accuracy is achieved by Multinomial Naive Bayes (MNB) for Unigram. The accuracy is: 81.39%

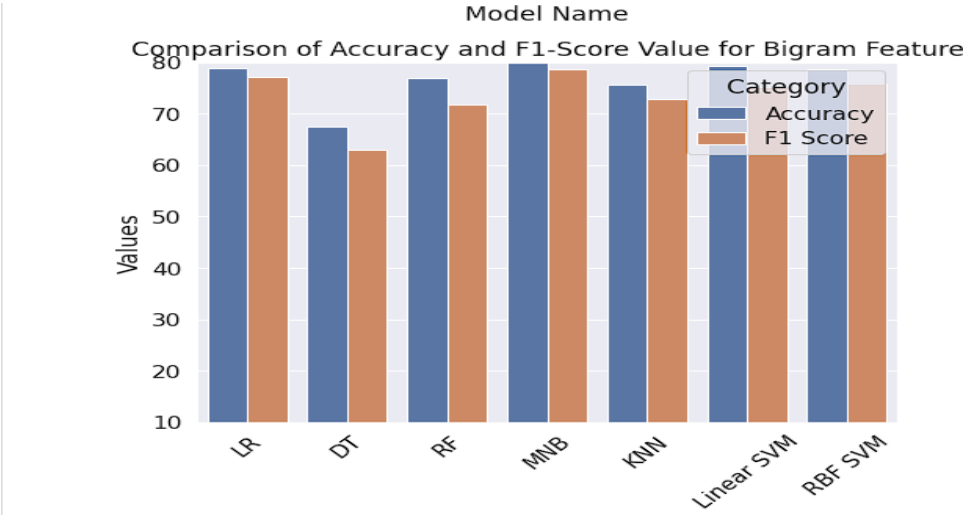


Fig 4.2.2: Data visualization for Bigram Feature

Performance Table (Bigram)

Model	Accuracy	Precision	Recall	F1 Score
LR	78.93	76.11	78.18	77.13
DT	67.36	65.05	60.91	62.91
RF	76.86	80.68	64.55	71.72
MNB	80.17	77.19	80.00	78.57
KNN	75.62	73.83	71.82	72.81
LINEAR SVM	79.34	82.61	69.09	75.25
RBF SVM	78.51	77.36	74.55	75.93

Table 4.2.2: Performance Table for Bigram Feature

Highest Accuracy achieved by MNB at = 80.17

Highest F1-score achieved by MNB at = 78.57

Highest Precision Score achieved by Linear SVM at = 82.61

Highest Recall Score achieved by MNB at = 80.0

Multinomial Naive Bayes (MNB) for Bigram achieves the maximum accuracy after applying all methods. It is 80.17% accurate.

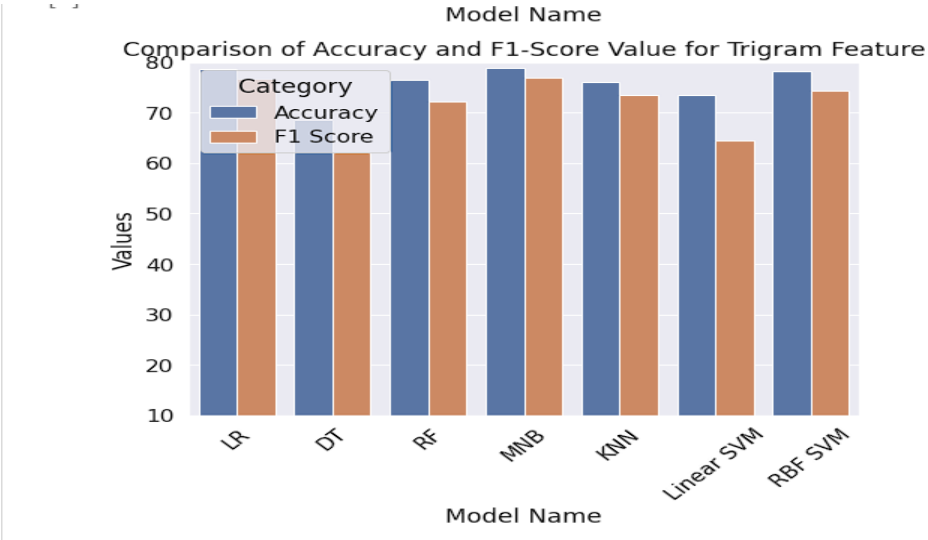


Fig 4.2.3: Data visualization for Trigram Feature

Performance Table (Trigram)

Model	Accuracy	Precision	Recall	F1 Score
LR	78.51	75.44	78.18	76.79
DT	68.60	68.89	56.36	62.00
RF	76.45	77.89	67.27	72.20
MNB	78.93	76.58	77.27	76.92
KNN	76.03	74.07	72.73	73.39
LINEAR SVM	73.55	82.86	52.73	64.44
RBF SVM	78.10	79.38	70.00	74.40

Table 4.2.3: Performance Table for Trigram Feature

Highest Accuracy achieved by MNB at = 78.93

Highest F1-score achieved by MNB at = 76.92

Highest Precision Score achieved by Linear SVM at = 82.86

Highest Recall Score achieved by MNB at = 78.18

Multinomial Naive Bayes (MNB) for Trigram has the highest accuracy after all other algorithms have been implemented. The accuracy is 78.93%.

4.3 Discussion

We analyze Bengali language sentiments in our research. In order to achieve the highest accuracy, we used various approaches. Following the application of the models, we obtained a model named Multinomial Naive Bayes (MNB), which provides us with higher accuracy than the others. The highest accuracy is provided fortrigrams, bigrams, and unigrams.

CHAPTER 5

Impact on Society, Environment and Sustainability

5.1 Impact on Society

It can be observed that MNB gives the highest accuracy. People nowadays often comment on social media platforms. Many companies like restaurants, online shops, e-commerce shops etcetera mainly work on Facebook and YouTube. They can easily find out the sentiment about their product, how people accept the product or how they dislike the product. It will be very for them to know. They can establish their product quality.

5.2 Impact on Environment

Since sentiment analysis is mostly a computational and linguistic work carried out by computers and does not directly alter the physical world, its environmental impact on Bengali is probably insignificant. To lessen their harmful effects on the environment, however, the energy consumption and electronic waste produced by the computers and other devices used for sentiment analysis should be appropriately handled.

5.3 Ethical Aspects

The following are some ethical considerations for sentiment analysis in Bengali:

Privacy: Concerns regarding the privacy and protection of personal information may be raised by the collecting and analysis of text data.

Bias: Sentiment analysis algorithms may reinforce preexisting biases, producing biased results.

Misrepresentation: The results of a sentiment analysis may not adequately reflect the sentiment of a text, which can cause misunderstandings and draw the wrong conclusions.

Freedom of expression: Sentiment analysis tools may be used to limit or keep an eye on freedom of expression.

When creating and implementing sentiment analysis systems, researchers and practitioners should take these ethical implications into account in order to use these systems responsibly and ethically.

5.4 Sustainability Plan

The precise context and objectives of the analysis being conducted will determine the sustainability plan for sentiment analysis on Bengali language. However, there are certain broad factors that might support the longevity of such an analysis, such as:

Data collection: Making sure that the information needed for sentiment analysis is gathered in a manner that respects people's rights to privacy and that is sustainable over time.

Model validation : Involves monitoring the sentiment analysis model's output to make sure it is generating accurate findings and, as necessary, updating and improving it.

Collaboration : It is key to ensuring that sentiment analysis is carried out in a way that is relevant and helpful to those who will benefit from it. Relevant stakeholders include universities, research institutes, and local communities.

Funding: Obtaining enough money to finance the sentiment analysis' continued development and upkeep, as well as any necessary updates or changes.

Ethical considerations: Making sure sentiment analysis is done responsibly and ethically, taking into account any potential effects on people and communities. The specific sustainability plan will rely on the objectives, context, and limitations ofthe analysis being conducted. These are only a few of the variables that could affect the sustainability of sentiment analysis on Bengali language.

CHAPTER 6

Summary, Conclusion, Recommendation and Implication for Future Research

6.1 Summary of the Study

Approximately 210 million individuals globally speak Bengali as a first or second language. To achieve accuracy, researchers applied some models. There weren't many classes or categories, though. Their research was good at understanding the mood of a Bengali sentence. But we are working harder and more effectively. Compared to other researchers, we make this increasingly more effective. Total eleven classes were utilized. After the conclusion, a Bengali sentence or word might be detected using sentiment analysis. MNB provides the most accurate data. Other algorithms that we utilized weren't too inaccurate when we used them.

6.2 Conclusions

In this study, we developed an MNB model sentimental analysis on text data in Bangla gathered from social media (Facebook and YouTube comment section). Our ultimate focus was to make the Bengali language more reliable and accurately recognized. If the system comes across any Bengali sentences, he can quickly determine what the sentences' true meaning is.

Therefore, we categorize the Bengali writings according to eleven fundamental emotions, including anger, fear, surprise, sad, happy, disgust, funny, abusive, advice, neutral and praise. To determine how much accuracy could be acquired, we implemented some models. There are many fruitful impacts of sentimental analysis and a good impact on many aspects.

1. Discovering and Foreseeing Market Trend
2. Monitoring the reputation of the brand
3. Examining polls of both public and political opinion
4. Analysis of customer feedback data is being conducted
5. social media discussion observation and analysis
6. Reduced Employee Turnover

6.3 Implication for Further Study

Here, we actually use a variety of models to achieve our purpose. Our data collection isn't very large. Thus, identifying the true sentiment of Bengali text is our key objective.
The following are the consequences for additional research on sentiment analysis in Bengali:

Accuracy enhancement: Additional study may concentrate on enhancing the precision of Bengali sentiment analysis algorithms.

Reducing bias: It's crucial to continue researching the topic of biases in sentiment analysis algorithms for Bengali language.

Application development: Creation of new sentiment analysis applications in Bengali, such as brand reputation management, opinion mining, and customer feedback analysis.

Sentiment analysis in new domains: Bengali-language sentiment analysis can be used in new domains like social media, news stories, and customer reviews to understand the attitudes and sentiments of the general audience.

Combination with other methods: Data modeling and named entity recognition are two NLP methods that can be used in conjunction with sentiment analysis to produce more thorough analysis and insights.

These are only a few of the numerous opportunities for more study and advancement in the area of sentiment analysis on Bengali language.

Reference

[1] S.M. Samiul Salehin ,Rasel Miah and Md Saiful Islam , “A Comparative Sentiment Analysis On Bengali Facebook Posts ,”In Proceedings of ACM International conference on computing advancement (ICCA 2020). ACM, 8 pages. https://doi.org/10.1145/3377049.337707, January, 2020

[2] Hasan Abid Ruposh, Mohammed Moshiul Hoque, “A Computational Approach of Recognizing Emotion from Bengali Texts”, Proceedings of the 2019 5th International Conference on Advances in Electrical Engineering (ICAEE), 26-28 September, Dhaka, Bangladesh

[3] Nayan Banik, Md. Hasan Hafizur Rahman, “Toxicity Detection on Bengali Social Media Comments using Supervised Models”, International Conference on Innovation in Engineering andTechnology (ICIET) 23-24 December, 2019, See discussions, stats, and author profiles for thispublication at: https://www.researchgate.net/publication/337317824, Preprint · November 2019
DOI: 10.13140/RG.2.2.22214.01608

[4] Abdul Hasib Uddin , Durjoy Bapery , Abu Shamim Mohammad Arif, “Depression Analysis of Bangla Social Media Data using Gated Recurrent Neural Network ” , 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT 2019), profiles for this publication at: https://www.researchgate.net/publication/336853602, Conference Paper · May 2019

[5] Estiak Ahmed Emon, Shihab Rahman, Joti Banarjee, Amit Kumar Das, Tanni Mittra, “A Deep Learning Approach to Detect Abusive Bengali Text”, 2019 7th International Conference on Smart Computing & Communications (ICSCC), 978-1-7281-1557-3/19/$31.00 ©2019 IEEE

[6] MD. Asif Iqbal, Avishek Das,, Omar Sharif, Mohammed Moshiul Hoque, Iqbal H. Sarker, BEmoC: A Corpus for Identifying Emotion in Bengali Texts; Received: 26 November 2020 / Accepted: 6 January 2022 / Published online: 17 January 2022 © The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd 2022 SN Computer Science (2022) 3:135 https://doi.org/10.1007/s42979-022-01028-w

[7] Tanjim Taharat Aurpa, Rifat Sadik, Md Shoaib Ahmed, “Abusive Bangla comments detection on Facebook using transformer-based deep learning models”, Received: 27 June 2021 / Revised: 6 October 2021 / Accepted: 1 November 2021 © The Author(s), under exclusive license to Springer-Verlag GmbH Austria, part of Springer Nature 2021, Social Network Analysis and Mining (2022) 12:24

[8] Puja Chakraborty, Md. Hanif Seddiqui, “Threat and Abusive Language Detection on Social Media in Bengali Language”; 1st International Conference on Advances in Science, Engineering andRobotics Technology 2019 (ICASERT 2019), 978-1-7281-3445-1/19/$31.00
c 2019 IEEE

[9] Md. Saiful Islam, Md. Afjal Hossain, Md. Ashiqul Islam, Jagoth Jyoti Dey, “Supervised Approach of Sentimentality Extraction from Bengali Facebook Status”, 19th International Conference on Computer and Information Technology, December 18-20, 2016, North South University, Dhaka, Bangladesh, 978-1-5090-4090-2/16/$31.00 ©2016 IEEE ISBN 978-1-5090-4089-6

[10] Md. Tazimul Hoque, Public Sentiment Analysis Based on Social Media Reactions for Bangla Natural Language, Hoque, M. T., Islam, A., Ahmed, E., Mamun, K. A., and Huda, M. N. (2019, February). Analyzing Performance of Different Machine Learning Approaches With Doc2vec for Classifying Sentiment of Bengali Natural Language. In 2019 In-ternational Conference on Electrical, Computer and Communication Engineering (ECCE) (pp. 1-5). IEEE.

[11] Debasish Bhattacharjee Victor, Jamil Kawsher, Md Shad Labib, Ms. Subhenur Latif, “Machine Learning Techniques for Depression Analysis on Social Media- Case Study on Bengali Community”, Fourth International Conference on Electronics, Communication and Aerospace Technology (ICECA-2020)IEEE Xplore Part Number: CFP20J88-ART; ISBN: 978-1-7281-6387-1

[12] Nauros Romim, , Mosahed Ahmed, Hriteshwar Talukder and Md Saiful Islam, Hate Speech detection in the Bengali language: A dataset and its baseline evaluation, https://github.com/strohne/Facepager

[13] Manash Sarker, Nazmus Sakib, A Machine Learning Approach to Classify Antisocial Bengali Comments on Social Media, February 2022 DOI:10.1109/ICAEEE54957.2022.9836407 Conference: 2022 International Conference on Advancement in Electrical and Electronic Engineering(ICAEEE)

[14] Md Gulzar Hussain, Tamim Al Mahmud, “Waheda Akthar, An Approach to Detect Abusive Bangla Text”, December 2018, DOI:10.13140/RG.2.2.20631.21920, International Conference on Innovation in Engineering and Technology (ICIET) 27-29 December, 2018

[15] Tanjim Taharat Aurpa, Rifat Sadik, Md Shoaib Ahmed, “Abusive Bangla comments detection on Facebook using transformer-based deep learning models”, Social Network Analysis and Mining (2022) 12:24 https://doi.org/10.1007/s13278-021-00852-x

[16] Shanta Phani,Shibamouli Lahiri, Arindam Biswas, “Sentiment Analysis of Tweets in Three Indian Languages”, Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing, pages 93–102, Osaka, Japan, December 11-17 2016. Licence details: http:// creativecommons.org/licenses/by/4.0/.1 https://en.wikipedia.org/wiki/Languages_of_India

[17] Diana Maynard, Mark A. Greenwood, “Who cares about sarcastic tweets? Investigating the impact of sarcasm on sentiment analysis”, Maynard, D.G. orcid.org/0000-0002-1773-7020 and Greenwood, M.A. (Accepted: 2014)Who cares about sarcastic tweets? Investigating the impact of sarcasm on sentiment analysis. In: LREC 2014 Proceedings. Language Resources and Evaluation Conference (LREC), 26-31 May 2014, Reykjavik, Iceland. ELRA . ISBN 978-2-9517408-8-4, https://eprints.whiterose.ac.uk/130763/

[18] Fabio Del Vigna12, Andrea Cimino23, Felice Dell’Orletta ,Marinella Petrocchi1, and Maurizio Tesconi1, “Hate me, hate me not: Hate speech detection on Facebook”, In Proceedings of the First Italian Conference on Cybersecurity (ITASEC17), Venice, Italy. Copyright c 2017 for this paper by its authors. Copying permitted for private and academic purposes. https://www.researchgate.net/publication/316971988

[19] Eftekhar Hossain , Omar Sharif , Mohammed Moshiul Hoque , and Iqbal H. Sarker, “SentiLSTM: A Deep Learning Approach for Sentiment Analysis of Restaurant Reviews”, https://www.researchgate.net/publication/350937855

[20] Dini Turipanam Alamanda, Abdullah Ramdhani, Ikeu Kania, Wati Susilawati, Egi Septian Hadi, “Sentiment Analysis Using Text Mining of Indonesia Tourism Reviews via Social Media”, International Journal of Humanities, Arts and Social Sciences
volume 5 issue 2 pp. 72-82 doi: https://dx.doi.org/10.20469/ijhss.5.10004-2

[21] Adeep Hande, Ruba Priyadharshini, Bharathi Raja Chakravarthi, KanCMD: Kannada CodeMixed Dataset for Sentiment Analysis and Offensive Language Detection, Proceedings of the Third Workshop on Computational Modeling of PEople’s Opinions, PersonaLity, and Emotions in Social media, pages 54–63 Barcelona, Spain (Online), December 13, 2020. details: http:// creativecommons.org/licenses/by/4.0/.