UNDERSTANDING PUBLIC SENTIMENT USING SOCIAL MEDIA ACTIVITIES: A MACHINE LEARNING APPROACH

BY

FERDOUS HABIB
ID: 201-15-3374

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

Supervised By

NARAYAN RANJAN CHAKRABORTY
Associate Professor
Department of CSE
Daffodil International University

Co-Supervised By

MD. MIZANUR RAHMAN
Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
DHAKA, BANGLADESH
JANUARY 2024

APPROVAL

This Project titled **"Understanding public sentiment using social media activities: A machine learning approach"**, submitted by **Ferdous Habib,** 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 **23rd January, 2024.**

BOARD OF EXAMINERS

Dr. S.M Aminul Haque
Professor & Associate Head
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Chairman

Nazmun Nessa Moon
Associate Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner 1

Dewan Mamun Raza
Senior Lecturer
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner 2

Dr. Md. Zulfiker Mahmud
Associate Professor
Department of Computer Science & Engineering
Jagannath University

External Examiner

DECLARATION

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

Supervised by:

Narayan Ranjan Chakraborty
Associate Professor
Department of CSE
Daffodil International University

Co-Supervised by:

for 22.1.2024

Md. Mizanur Rahman
Lecturer
Department of CSE
Daffodil International University

Submitted by:

ferdous

Ferdous Habib
ID: 201-15-3374
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

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

I really grateful and wish our profound our indebtedness to **Mr. Narayan Ranjan Chakraborty, Associate Professor**, Department of CSE, Daffodil International University, Dhaka. Deep Knowledge & keen interest of my 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 stage have made it possible to complete this project.

I would like to express our heartiest gratitude to **Dr. Sheak Rashed Haider Noori, Professor and Head,** Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

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

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

ABSTRACT

People use different social media platform to express their thoughts in their profiles or write comments on various aspects of world. Then a huge amount of data is generated and are stored on the internet in an unorganized, unstructured way. In this age of digital technology, million even trillion numbers of data are generated and it has become difficult to analyze those data manually. But sentiment analysis is the scientific way to extract valuable insights from the data, analyzing data and preparing those for using in various purposes. In the world, a large number of people live with different languages as their medium to communicate with others and achieve other objectives. This study was conducted to analyze sentiment of Bangla content from social media data. This study aims to analyze sentiment in two general categories: Positive and Negative. It basically defines content carries what emotions, either positive or negative of people in a particular context. This sentiment analysis approach is conducted using Machine Learning and Natural Language Processing (NLP) techniques. The outcome of this study is to develop a model that can accurately measure emotional states from social media content. This finding can have a bold effect on understanding the public opinion on various issues. From these findings, it can be valuable in real-world events in business, culture, society and more.

Keywords – Sentiment Analysis, Machine Learning, Natural Language Processing (NLP).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

1.1 Introduction

In this era of digital age social media has become integral part of daily life, we cannot think a moment without using Social Media platforms like Facebook, Instagram, Twitter and others. People use Facebook to communicate with others, for sharing thoughts or feelings on various context. Those thoughts, feelings can be defined as Sentiment. Through these things enormous amount of data is generated daily which can present valuable insights. But those data are stored in unstructured way and analyzing those unstructured data became difficult. It needs a structured way for analysis. Machine Learning, Natural Language Processing, Deep Learning techniques creates scope to conduct a structured and scientific way to analysis and extract sentiment from text data.

In the time period of 2021-2022 world's population is seen in increasing with growth rate 0.83%. From year 2022 to 2023 population again increased at 0.88% growth rate and became 8.1 billion in 2023 [1]. As population is increasing the usage of social media is also increasing day by day. From January 2022 to January 2023 the growth rate was 2.1%. There also seen increasing the number of users. Of the total world population around 60% people now use social media and the average usage time is around 2.24 hour daily. In the term of gender and age, Facebook is at the top in most used social media platform list [2]. From the world total population around 234 million people are native Bangla user and other 39 million people use Bangla as their second language. In the list of most spoken language, Bangla holds the 6th position in the world and 5th among the Indo-European language. About 98% Bangla language speaker are from Bangladesh and rest are from states in India and surrounding states [3].

Sentiment Analysis holds the field of huge area of Natural Language Processing (NLP). As enormous research has taken place in various languages, sentiment analysis in Bangla Language is still emerging research problem with number of issues like lacking of resources, complexity of Bangla Language and so on. A few studies were conducted on Multiclass Sentiment analysis but it results low accuracy for using limited dataset.

There also arise limitation with working on the necessary text preparation techniques. For gaining higher accuracy approaches like Stemming, POS tagging, Stopword Removal is must [4]. From the significance and important of sentiment analysis this study is conducted by employing advanced sentiment analysis techniques that includes Machine Learning Algorithms and Natural Language Processing (NLP) techniques. The primary objective of this study is to unveil emotional patterns and discern sentiment in Positive and Negative classes that offers understanding the prevailing attitudes within the online community.

1.2 Motivation

Sentiment analysis is useful in many fields, such as marketing, politics, public opinion research, and customer service. Businesses, governments, and researchers can make informed judgments, modify strategies, and respond effectively to the public psychological state if they learn about the views expressed on social media. In recent years, social media platforms have become part of a virtual ecosystem through which thoughts on various topics are formed, trends are established, and the collective voice of the public finds' significance. The potential for acquiring important insights into prevalent emotion becomes clear to explore people's engagement in social media platforms likes posts, comments. This study is motivated from the desire to extract the depth of information collected from these digital interactions and find the complex patterns of human emotions that characterize our online discourse. This initiative aims to contribute to the expanding field of natural language processing and machine learning, where the integration of technology and human expression opens up new opportunities.

I hope to deliver actionable insights to consumers by understanding the underlying sentiment of social media data, which can affect decision-making processes, shape strategies, and increase our comprehension of the evolving digital environment. In the beginning of journey into the world of digital media the goal is to provide people, businesses, and officials with the tools they need to effectively navigate the numerous the intricate details of digital engagement. The ultimate goal is to bridge the gap between digital and physical emotions, enabling a better understanding of the collective emotions that make up our interconnected society.

1.3 Rationale of the Study

Bangla, now a notable language in the Indian Subcontinent, is used by over 250 million people worldwide and ranks sixth in terms of global popularity. In Bangladesh and India, Bangla is used as the mother tongue and also as the native language. Previous researchers have looked into sentiment analysis in other languages like English, Chinese, and Urdu. However, sentiment analysis in the Bangla language presents challenges that include a lack of resources and the sensitivity of the Bengali language. As a result, sentiment analysis in the Bangla language remains an understudied topic for Bangladeshi researchers. This study offers a chance to tackle these issues by tweaking and improving sentiment analysis models to suit the unique quirks of the Bengali language.

1.4 Research Questions

The objective of this research is to explore sentiment analysis in Bangla text data from social media activities, taking into consideration Bangladesh's unique dialect and cultural aspects. The study is directed by the following research questions:

- What linguistic challenges are there in sentiment analysis in Bengali, given its unique grammar and vocabulary?
- What changes are needed to improve the accuracy of Bengali sentiment analysis models?
- What impact can accurate sentiment analysis have on decision-making processes?
- What amount of data is collected?
- Should I build new models or use well-known and widely used machine learning techniques?

1.5 Expected Output

This paper aims to use contexts to understand human sentiment from their activities, such as posts and shares of the content of different topics on social media. Advanced techniques such as machine learning and natural language processing (NLP) provides us with accurate insights into content. In this study, Bangla text data are used which are collected from social media platforms, like Facebook. In this dataset, 80% data is used for training and rest 20% for testing. A common challenge in labeling sentiment occurs when working with this study. The labeled sentiments may reflect biases in this dataset. For instance, if a dataset predominantly comprises negative sentiments, the machine learning model might skew towards identifying negativity. This could compromise the accuracy and fairness of machine-learning models. Labeled data may also have an imbalanced distribution, with one class significantly outnumbering the others, which can influence models' ability to predict less frequent sentiments accurately. By optimizing all the issues this study will be useful to identifying human sentiment which can become valuable for various purposes.

1.6 Project Management and Finance

Embarking on the sentiment analysis project as a solo endeavor, effective project management becomes paramount for success. The initiation involves personally defining the project scope, encapsulating platforms, languages, and sentiment categories, and my versatile skill set—encompassing natural language processing, machine learning, linguistics, and project management—equips me to navigate diverse perspectives independently. I will meticulously plan a timeline with achievable milestones to ensure consistent progress, and prudent resource allocation, covering personnel, computing, and data resources, will be central to maintaining efficiency without incurring additional costs.

1.7 Report Layout

The following are the contents of this research project:

Chapter 1 delves into the discussion over the research motivation, study rationale, research questions, expected output and Project Management and Finance. In chapter 2 it provides an overview of research's background. It also provides the facts of related works. A thorough description of the methodology used in this research is outlined in Chapter 3. It outlines the data collection process, data preprocessing techniques, and the specific machine learning algorithm used for sentiment analysis. Meanwhile, chapter 4 represents the findings of the research, including the results of the machine learning model and its performance evaluation.

Whereas chapter 5 discussed the impact of sentiment analysis on society and environment, its ethical aspects and also sustainability plan. Lastly, chapter 6 includes the discussion on the summary of this study, conclusion and future work of sentiment analysis.

CHAPTER 2
BACKGROUND STUDY

2.1 Terminologies

Since the early 1990s, the field of sentiment analysis has seen extensive interdisciplinary research. Understanding and evaluating public opinion requires sentiment analysis. Using this information, you can identify whether a sentiment is positive, negative, or neutral, and then use it to analyze that data. Product reviews, social media monitoring, and customer feedback analysis can all be done with this information. However, there are only a few practical Bangla datasets available for research. Moreover, sentiment analysis in Bangla is still in its experimental condition. Researchers are increasingly turning to self-regulated Bangla datasets, as they provide a solution to the lack of practical Bangla datasets.

2.2 Related Works

In the field of sentiment analysis, numerous research efforts have produced a variety of techniques and approaches. In their study [4], the authors proposed a supervised deep learning classifier based on CNN and LSTM for multi-class sentiment analysis. They conducted their study using six machine learning models, where CLSTM improved performance and results with 85.8% accuracy and 0.86% F1 score. In another study [5], the authors achieved 75% accuracy using Skip-Gram with Word2vector and Continuous Bag of Words (CBOW) with a new word-to-index model. The paper [6] proposes an N-gram language model based on contextual similarity for identifying the stems or root forms of Bangla words. The authors implemented a 6-gram model that achieved 40.18% accuracy in stem identification. Authors in [7] presented the development process of Bengali WordNet Affect lists, which already exist in English, showing moderate agreement ranging from 0.44 to 0.56 Kappa Coefficient (k) for six emotion classes. In [8], the paper presents the attention mechanism as effective and efficient in analyzing Bangla sentiment or opinion. Finally, a new technique named PSPWA (Priority Sentence Part weight Assignment) is introduced in [9] to perform aspect category or term extraction on an existing dataset.

While using CNN, author achieves f1-score of 0.59 and 0.67 in two different types of datasets. “BanglaSenti” a lexicon-based corpus for sentiment analysis from textual data has presented in their paper [10] by Ali. The authors conducted a model simulation using the Bangla VADER system, which was generated by modifying the existing English VADER system, as mentioned in [11]. They used multiple translator systems and English VADER to classify sentiment from Bangla text, resulting in improved performance compared to existing models. In [12], the authors experimented with word embedding methods for sentiment analysis, specifically Word2Vec Skip-gram and Continuous Bag of Words. The results showed that the Word2Vec Skip-gram model achieved higher accuracy (83.79%) compared to other models. In [13], the authors conducted a thorough study using supervised machine learning (ML) and supervised deep learning techniques for classification tasks. They employed algorithms such as NB, SVM, and LR in ML approaches, achieving an accuracy of 86.7%. In deep learning approaches, they obtained an accuracy of 72.86%. Researchers [14] studied people's emotions during the COVID-19 period using various deep learning algorithms. CNN achieved an accuracy of 97.24%, while LSTM achieved 95.33% accuracy. Authors [15] developed their own methodology for calculating sentiment from Bangla text using valency analysis. They utilized WorldNet to sense each word based on parts of speech and SentiWordNet to obtain prior valence for each word. Their approach prioritized three emotion classes: Analytical, Depressed, and Angry. In [16], a novel approach for sentiment analysis using Naïve Bayes Classification was presented in the context of Bengali Facebook statuses' language rules detection. For automatically extracting sentiment from Bangla microblog posts and identifying text polarity (positive or negative), researchers [17] explored SVM and MaxEnt algorithms while experimenting with combinations of various feature sets. In another study [18], the potentiality of BERT in sentiment analysis was explored, particularly for aspect-based sentiment analysis. The authors claimed that BERT outperformed other algorithms in terms of effectiveness and generality. The authors [19] conducted a study using linear and non-linear support vector machine and N-gram techniques on diverse web-based data. In their paper, they used N-gram techniques to create vectors containing more than one word instead of single-word vectors. Implicating N-grams, they achieved a better result.

In their paper [20], the authors proposed a framework that includes a classification model of a neural network variance, which is a Convolutional Neural Network (CNN). Instead of the existing Bangla sentiment classifier, CNN obtains an accuracy of 99.87%. The authors used a dataset consisting of Bangla and Roman languages to test sentiment analysis in LSTM, one of the Deep Recurrent models, in their paper [21]. They also demonstrate two types of cross-entropy: binary cross-entropy and categorical cross-entropy. Another work [22] was conducted using classical and deep learning algorithms. SVM, Random Forest, CNN, FastText, and Transformer-based models Ire included among those algorithms. The authors used this model to present a comparison of their performance. This study suggests that the unexplored algorithm performs best for sentiment analysis tasks, particularly transformer-based models. In their study [23], the authors conducted a supervised sentiment analysis using Recurrent Neural Networks (RNN), which is a deep learning model. They classified textual information into three categories and compared deep learning-based algorithms for representing Bangla sentences based on characters. They presented a comprehensive study on sentiment analysis using Bangla conversation data. For their study, they employed common machine learning approaches and initially labeled the data as positive and negative to extract information. Another work [25] on sentiment analysis was conducted in the domain of deep learning. It states that approaches like Recurrent Neural Networks (RNN) with long short-term (LSTM) Ire applied to a cricket-based dataset. This study achieved an accuracy of 95% with LSTM, which surpassed the performance of other methods. A study [26] was conducted based on the Aspect-Based Sentiment Analysis (ABSA) model for e-commerce product reviews. The study also transformed its analysis output into definitive recommendations. In their approach, they classified the dataset into four categories: positive, negative, neutral, and conflicting, and proceeded with further processes. The authors [27] utilize the word2vec model for their sentiment classification study. In their procedure, they introduce a novel approach to extract sentiment from words. They attain an accuracy of 75.5% by combining the word2vec word co-occurrence score and sentiment polarity score. The ABSA-based work in sentiment analysis in the Bangla language domain is conducted in their paper [28]. They conducted their study using the publicly available dataset.

They introduce a baseline approach for their aspect-based sentiment analysis task to extract sentiment from the dataset. The authors proposed a lexicon-based approach for sentiment analysis in their paper [29]. They utilized the Semantic Orientation Calculator (SO-CAL) to perform polarity classification, assigning instances to either positive or negative classes. They suggest that this approach consistently delivers good performance. In their paper [30], the authors conducted a study using the Support Vector Machine (SVM) algorithm. They applied TF-IDF pre-processing methods for data processing and later trained the model with SVM. The outcome of this study showed a very promising score. The authors of the paper [31] proposed an unsupervised method for analyzing consumer reviews. They also used a supervised method. In their work, they applied the association rule of data mining to the unsupervised method, achieving an F1-score of 67%. However, the supervised method outperformed other methods with an F1-score of 84%. The authors [32] discussed the supervised learning algorithm for classifying sentiment from Twitter messages. They used distant supervision with machine learning algorithms for classification. Their study yielded an accuracy above 80% using Naïve Bayes, Maximum Entropy, and SVM classifiers. In their paper [33], the authors present a thorough discussion on analyzing emotional expression, particularly based on Ekman's six emotion classes. Additionally, they also consider three types of intensities for text annotation. They found satisfactory outcomes for each emotion class. The paper [34] finds the use of a combination of unsupervised and supervised techniques for document-level sentiment polarity annotation. They propose a model that leverages both continuous and multi-dimensional sentiment annotation. The results of their study show that their model outperforms previously used methods. The authors conducted a study analyzing sentiment analysis on electric-based products from Twitter text data in their paper [35]. They mentioned that this analysis will have an impressive effect in identifying domain information. To achieve this, they introduced a new feature vector capable of classifying tweets into different classes and extracting people's opinions on certain products. The authors conducted this analysis using machine learning techniques. Another work was also performed for sentiment analysis based on Twitter data. The authors [36] proposed a framework for analyzing sentiment. The study provides step-by-step details about the process. For their work, the authors used machine learning algorithms such as Naïve Bayes and Decision Tree.

2.3 Comparative Analysis and Summary

The above presented discussion on sentiment analysis of different types of research works performed by different research teams shows that in recent years there seen a growing number of research works on Bangla text. It presents that there remain a few good outcomes impacting the sentiment analysis task. But challenges arise in terms of resources, although there is hope that this field will become more resourceful and resulted as a proven and impactful task for analysis public sentiment.

The overview of various research' topics, methodologies and results is provided in Table 2.1 below.

Table 2.1: AN OVERVIEW OF RELATED RESEARCH WORKS

Serial No.	Authors name	Methodologies	Descriptions	Results
1.	R. Haque et al	CNN, LSTM, CLSTM	Using CNN, LSTM model for multiclass sentiment analysis.	85.8% accuracy and 0.86 F1 score by proposed CLSTM model
2.	M. Rahma et al	Word2vector	Using Word2Vec model, three features are extracted	75% accuracy by skip-gram model
3.	T. T. Urmi et al	N-gram model	Using N-gram model detecting similarity in words	40.18% accuracy by 6-gram model
4.	D. Das et al	SentiWordNet	Using SentiWordNet they extract text sentiment to six basic emotions	Kapp coefficient result 0.44 to 0.56 for six emotion classes
5.	F. A. Naim et al	CNN	Using CNN model, performed aspect category or term extraction	0.59 and 0.67 f1-score for two types of datasets by CNN
6.	S. H. Sumit et al	Word2vec, Skip-Gram and CBOW model	Using Word2vec Skip-Gram and CBOW model, exploring Bangla sentiment analysis	83.79% accuracy by Word2vec Skip-Gram model
7.	S. S. Salehin et al	Naïve Bayes, SVM, LR, LSTM model	Using Naïve Bayes, SVM, LR, LSTM model, five features are extracted	86.7% accuracy by SVM
8.	M. S. A. Pran et al	Deep learning model	Using deep learning model, extracted three features	97.24% by CNN and 95.33% by LSTM
9.	M. H. Alam et al	CNN model	Using CNN model, extracted three features	99.87% by CNN

2.4 Scope of the Problem

This research delves into the unexplored territory of sentiment analysis on Bengali social media data, aiming to uncover the complex linguistic and cultural aspects inherent to the language. The scope encompasses challenges such as linguistic diversity, code-switching, cultural sensitivity, the use of symbols of emotion, and the sophisticated handling of sarcasm and irony. The study also takes into account the evolving nature of language, the interaction between multiple dialects, and the evolving role of social media in language use, all of which contribute to the complexity of sentiment analysis. The study, moreover, pays particular attention to temporal dynamics, which play a pivotal role in shaping sentiment expressions, thus influencing the interpretation of data. For instance, the varying expressions of the same sentiment can indicate different emotional states over time, which, if not accounted for, may lead to misleading results. The study also seeks practical applications in real-world scenarios, envisioning a comprehensive understanding of sentiment expression among Bengali social media users. By addressing these challenges, the research contributes not only to the advancement of sentiment analysis but also provides valuable insights into the sentiments of the Bengali-speaking community in the digital era.

2.5 Challenges

Working with the Bangla Dataset may pose a lot of challenges, including data collection, data labeling, Negation handling, data cleaning, and a few more. Manually collecting data from social media platforms like Facebook is time-consuming. Web scrapping is another way, but the availability of proper scrapping tools and their working efficiency raise questions. So, to skip these issues, data is collected from Kaggle and GitHub. These are open-source platforms for any content. Then comes the data labelling issue. Though in the existing dataset, data is assigned to labeling, data labeling is still a major issue in the sentiment analysis task. It may create bias with the data label, which can affect the final output. A sentence that contains a negation word usually denotes a negative sentence, but it can also be a positive sentence based on what the main theme of the sentence is or what the sentence expresses.

CHAPTER 3

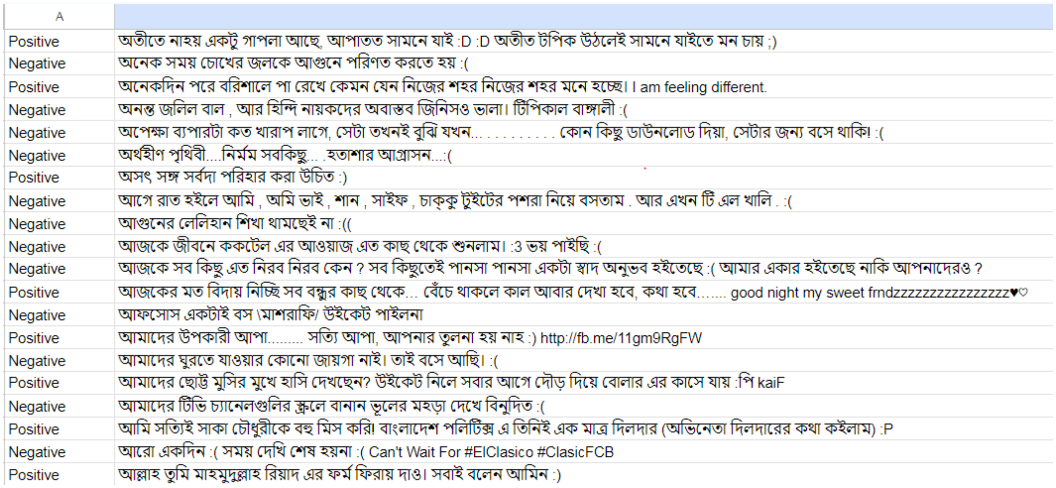
RESEARCH METHODOLOGY

3.1 Research Subject and Instrumentation

The primary focus of this study is the large and growing accumulation of Bengali social media users interacting online. The research topic revolves around various Bengali social media users and their textual expressions and the tool consists of a customized integration of linguistic tools, cultural remarks and machine learning models. This integrated strategy works to uncover the sentiments interconnected into the fabric of Bengali social media, providing significant insight into the complex dynamics of sentiment expression in the digital age.

3.2 Data Collection Procedure

This section describes data collection procedures. Datasets are evaluated based on size, user demographics, and linguistic diversity to ensure they align with research objectives. Existing datasets from Kaggle, an open-source data repository, are used for sentiment analysis. Consequently, the preliminary stage involves finding and choosing Bengali language-based social media datasets on Kaggle. This includes posts, comments, and interactions on Facebook. The amount of data collected is 15,525. Where 13,536 data are used in subsequent processes.



A	
Positive	অতীতে নাহয় একটু গাপলা আছে, আপাতত সামনে যাই :D :D অতীত টপিক উঠলেই সামনে যাইতে মন চায় ;)
Negative	অনেক সময় চোখের জলকে আগুনে পরিণত করতে হয় :(
Positive	অনেকদিন পরে বরিশালে পা রেখে কেমন যেন নিজের শহর নিজের শহর মনে হচ্ছে। I am feeling different.
Negative	অনন্ত জলিল বাল , আর হিন্দি নায়কদের অবাস্তব জিনিসও ভালা। টিপিকাল বাঙ্গালী :(
Negative	অপেক্ষা ব্যাপারটা কত খারাপ লাগে, সেটা তখনই বুঝি যখন... . . . . কোন কিছু ডাউনলোড দিয়া, সেটার জন্য বসে থাকি! :(
Negative	অর্থহীণ পৃথিবী....নির্মম সবকিছু... .হতাশার আগ্রাসন...:(
Positive	অসৎ সঙ্গ সর্বদা পরিহার করা উচিত :)
Negative	আগে রাত হইলে আমি , অমি ভাই , শান , সাইফ , চাক্কু টুইটের পশরা নিয়ে বসতাম . আর এখন টি এল খালি . :(
Negative	আগুনের লেলিহান শিখা থামছেই না :((
Negative	আজকে জীবনে ককটেল এর আওয়াজ এত কাছ থেকে শুনলাম। :3 ভয় পাইছি :(
Negative	আজকে সব কিছু এত নিরব নিরব কেন ? সব কিছুতেই পানসা পানসা একটা স্বাদ অনুভব হইতেছে :(আমার একার হইতেছে নাকি আপনাদেরও ?
Positive	আজকের মত বিদায় নিচ্ছি সব বন্ধুর কাছ থেকে... বেঁচে থাকলে কাল আবার দেখা হবে, কথা হবে....... good night my sweet frndzzzzzzzzzzzzzzzz♥♡
Negative	আফসোস একটাই বস \মাশরাফি/ উঁইকেট পাইলনা
Positive	আমাদের উপকারী আপা......... সত্যি আপা, আপনার তুলনা হয় নাহ :) http://fb.me/11gm9RgFW
Negative	আমাদের ঘুরতে যাওয়ার কোনো জায়গা নাই। তাই বসে আছি। :(
Positive	আমাদের ছোট্ট মুসির মুখে হাসি দেখছেন? উঁইকেট নিলে সবার আগে দৌড় দিয়ে বোলার এর কাসে যায় :পি kaiF
Negative	আমাদের টিভি চ্যানেলগুলির স্ক্রলে বানান ভুলের মহড়া দেখে বিনুদিত :(
Positive	আমি সত্যিই সাকা চৌধুরীকে বহু মিস করি! বাংলাদেশ পলিটিক্স এ তিনিই এক মাত্র দিলদার (অভিনেতা দিলদারের কথা কইলাম) :P
Negative	আরো একদিন :(সময় দেখি শেষ হয়না :(Can't Wait For #ElClasico #ClasicFCB
Positive	আল্লাহ তুমি মাহমুদুল্লাহ রিয়াদ এর ফর্ম ফিরায় দাও। সবাই বলেন আমিন :)

Fig-3.1: Dataset Visualization

3.2.1 Attributes

There were 2 attributes in our raw datasets. Those are Text and Tag. For further approaches more attributes like Text, Label, cleaned are generated. But the main attribute among them is, Label and Cleaned. Moreover, the pre-processed dataset is splits into two different datasets- a training dataset is used to develop the model, and a test dataset is used to evaluate it. In table 3.1 dataset splitting proportion is given below.

Table 3.1: DATA SPLITTING PROPORTION

Train	80%
Test	20%

3.3 Statistical Analysis

As the dataset comprises 13536 data, 6307 are for Negative Sentiment and the remaining 7229 are for Positive Sentiment. Technique for extracting features before model training the TF-IDF Vectorizer is used to transform the text documents into a token count matrix. Several machine learning models are used to discover the best-fitting approaches for this dataset; better accuracy is the ultimate objective here.

3.4 Proposed Methodology

The methodological approach for understanding Bengali social media data must be both comprehensive as well as culturally sensitive to gain a deeper understanding of their sentiments. A detailed description of the research methodology is presented in this section. It describes how sentiment analysis can be explored in the context of Bengali language. The following figure in figure 3.2 presents the necessary steps of the proposed methodology. Later a quick overview of the steps will be given.

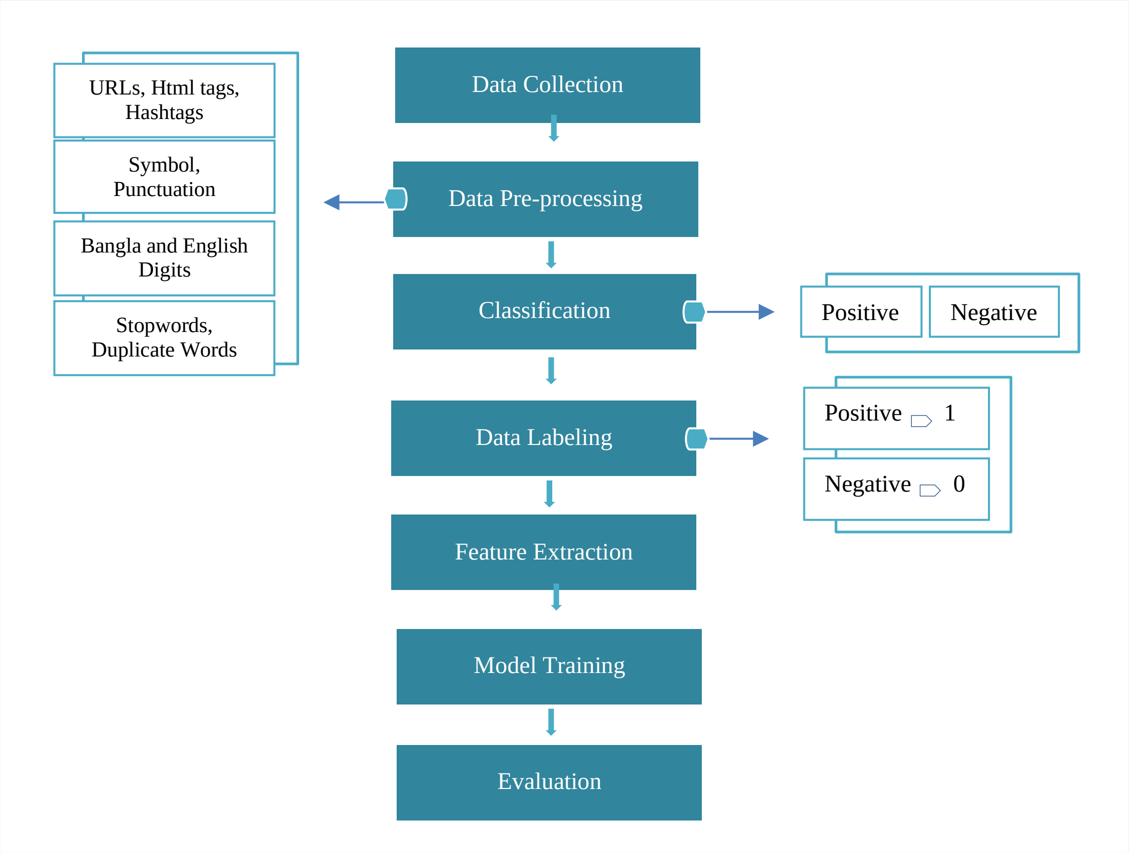


Fig-3.2: Proposed Methodology

3.4.1 Data Pre-processing

In the preprocessing phase, the text is transformed to optimize it for subsequent analysis. The dataset, specifically the data from the 'Text' feature, undergoes cleaning and standardization to ensure optimal performance when integrated into the sentiment analysis model.

The first steps involve tokenization, breaking down each word into individual tokens. This results in structured representations of the text, facilitating easier linguistic analysis. Following tokenization, noise removal is implemented to eliminate extraneous elements such as special characters, symbols, and irrelevant punctuation. This cleanup process ensures that subsequent analyses focus on extracting meaningful patterns from the text.

To further enhance the quality of the data, various text pre-processing techniques are applied in this study:

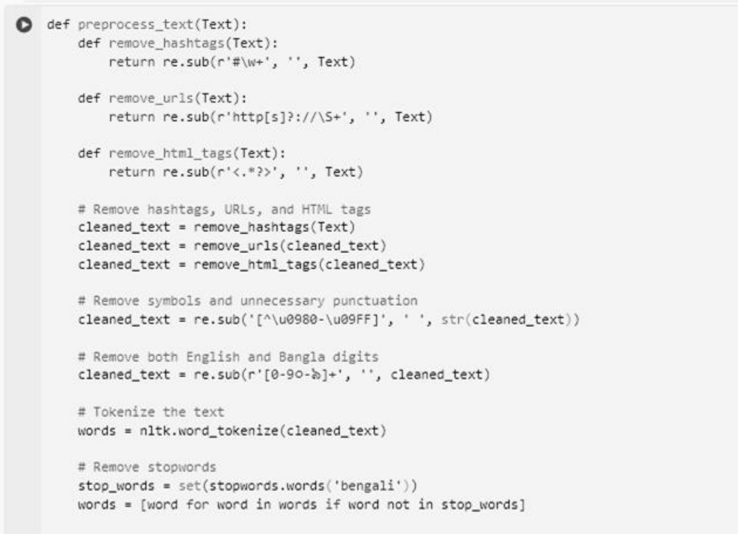
1. Remove hashtags, URL’s, HTML tags, symbols, and punctuation: Extraneous elements that do not contribute to sentiment analysis, such as hashtags, URLs, HTML tags, symbols, and punctuation, are systematically removed. This step simplifies the text and allows the model to concentrate on core linguistic content.

2. Remove both Bangla and English digits: Numerical digits, both in Bengali and English, are removed to reduce numbers that might interfere with sentiment analysis. This ensures a focus on textual patterns without numerical distractions.

3. Tokenization: Tokenization breaks down the text into individual tokens, resulting in a structured and analyzable representation of the textual data.

4. Remove Stopwords and Duplicate Words: Stopwords, common words that do not contribute significantly to sentiment, are removed to streamline the dataset. Additionally, duplicate words are eliminated to speed up the analysis.

By implementing these techniques, the study aims to refine the textual data, making it more conducive to accurate sentiment analysis. These steps in figure 3.3 contribute to the overall optimization of the sentiment analysis model, making sure that it can effectively capture and interpret the sentiments expressed in Bengali social media data.

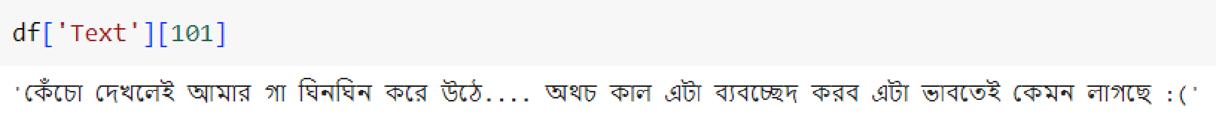


```
def preprocess_text(Text):
    def remove_hashtags(Text):
        return re.sub(r'#\w+', '', Text)

    def remove_urls(Text):
        return re.sub(r'http[s]?://\S+', '', Text)

    def remove_html_tags(Text):
        return re.sub(r'<.*?>', '', Text)

    # Remove hashtags, URLs, and HTML tags
    cleaned_text = remove_hashtags(Text)
    cleaned_text = remove_urls(cleaned_text)
    cleaned_text = remove_html_tags(cleaned_text)

    # Remove symbols and unnecessary punctuation
    cleaned_text = re.sub('[^\u0980-\u09FF]', ' ', str(cleaned_text))

    # Remove both English and Bangla digits
    cleaned_text = re.sub(r'[0-9০-৯]+', '', cleaned_text)

    # Tokenize the text
    words = nltk.word_tokenize(cleaned_text)

    # Remove stopwords
    stop_words = set(stopwords.words('bengali'))
    words = [word for word in words if word not in stop_words]
```

Fig-3.3: Text pre-processing & cleaning

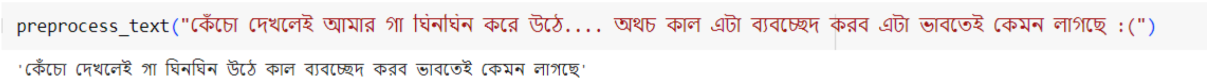
The preprocessing steps described the result given below figure 3.4 in a refined and standardized textual dataset that is ready to be analyzed further.

Before pre-processing:



```
df['Text'][101]
'কেঁচো দেখলেই আমার গা ঘিনঘিন করে উঠে.... অথচ কাল এটা ব্যবচ্ছেদ করব এটা ভাবতেই কেমন লাগছে :('
```

After pre-processing:

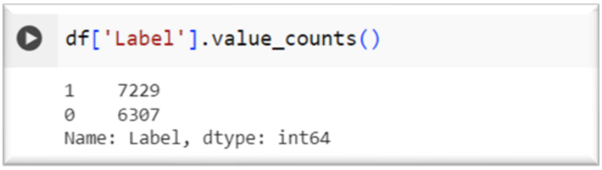


```
preprocess_text("কেঁচো দেখলেই আমার গা ঘিনঘিন করে উঠে.... অথচ কাল এটা ব্যবচ্ছেদ করব এটা ভাবতেই কেমন লাগছে :(")
'কেঁচো দেখলেই গা ঘিনঘিন উঠে কাল ব্যবচ্ছেদ করব ভাবতেই কেমন লাগছে'
```

Fig-3.4: Text pre-processing Result

3.4.2 Classification

Social media user can obtain feeling or emotion when go through content on various context. Analyzing those users' emotion that they express over social media allows to determine whether it has earned a good or bad review. The basic concepts are the same as in English, with a focus on positive, negative, or neutral sentiments. In this dataset contents Ire divided into two categories: Positive and Negative. A content may carry a positive emotion if it conveys good feelings, praise, satisfaction, or favorable judgment. Look for terms like "অসাধারণ"(excellent), "ভালো"(good), "সন্তুষ্ট"(satisfied), or expressions of appreciation. In contrast, if the content indicates dissatisfaction, criticism, disappointment, or adverse opinions, it is most certainly classed as a Negative review. Negative emotions can be communicated by words such as "খারাপ"(poor), "হতাশ"(disappointing), or through expressions of dissatisfaction. The following figures 3.5 and 3.6 depicts how the data is classified. The bar chart shows that the dataset has the right balance to enhance data's accuracy.



```
df['Label'].value_counts()

1    7229
0    6307
Name: Label, dtype: int64
```

Fig-3.5: Classification value count

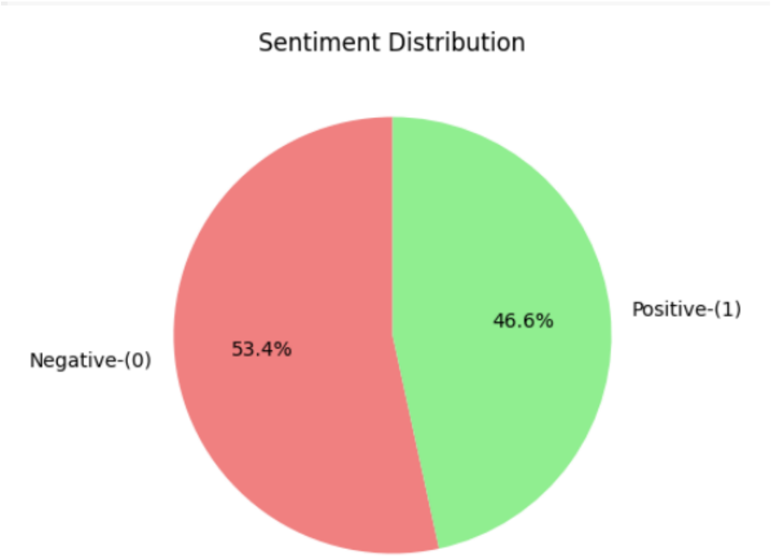
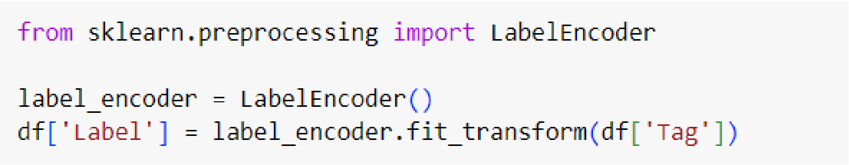


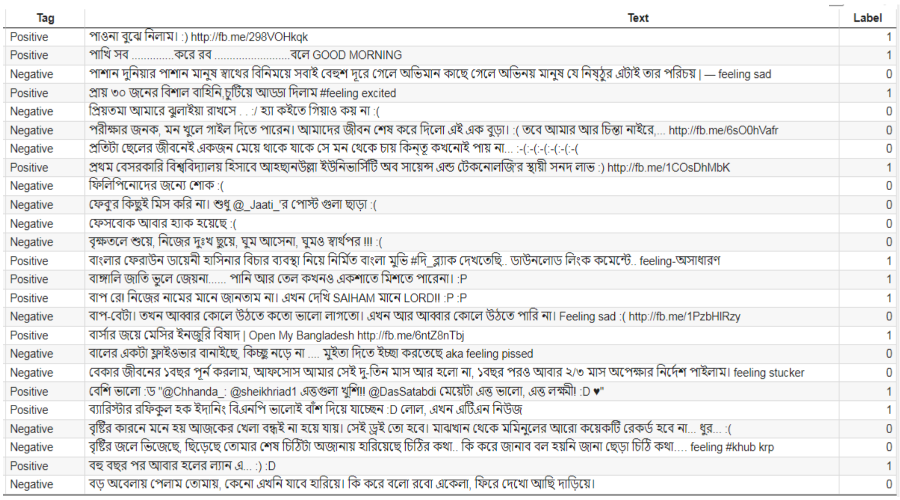
Fig-3.6: Sentiment value distribution

3.4.3 Data Labeling

Data labeling is a pivotal step in the training and evaluation of sentiment analysis models, as this involves assigning sentiment numeric labels to the collected textual data. In this study, ‘Label Encoder’ technique is used for data labelling. The following figure 3.7 shows the label encoding function and dataset after encoding.



```
from sklearn.preprocessing import LabelEncoder

label_encoder = LabelEncoder()
df['Label'] = label_encoder.fit_transform(df['Tag'])
```



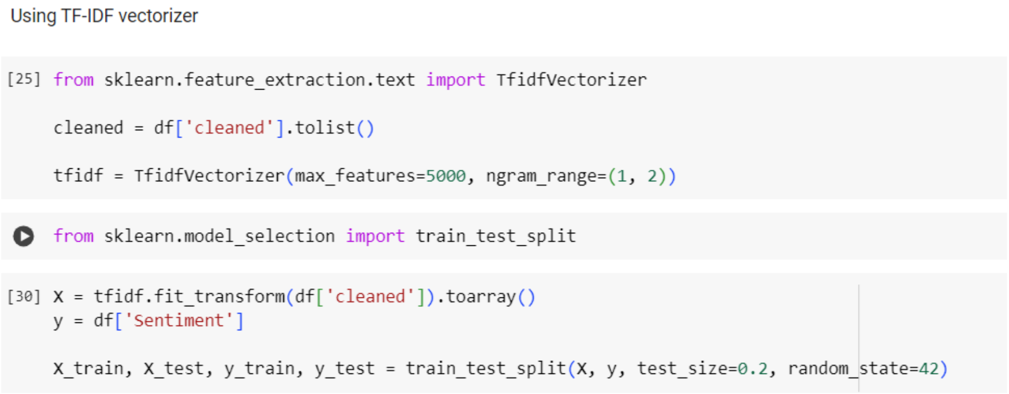
Tag	Text	Label
Positive	পাওনা বুঝে নিলাম। :) http://fb.me/298VOHkqk	1
Positive	পাখি সব ....করে রব ....বলে GOOD MORNING	1
Negative	পাশান দুনিয়ার পাশান মানুষ স্বাথের বিনিময়ে সবাই বেহুশ দূরে গেলে অভিমান কাছে গেলে অভিনয় মানুষ যে নিষ্ঠুর এটাই তার পরিচয় | — feeling sad	0
Positive	প্রায় ৩০ জনের বিশাল বাহিনি,চুটিয়ে আড্ডা দিলাম #feeling excited	1
Negative	প্রিয়তমা আমারে ঝুলাইয়া রাখসে . . :/ হ্যা কইতে গিয়াও কয় না :(	0
Negative	পরীক্ষার জনক, মন খুলে গাইল দিতে পারেন। আমাদের জীবন শেষ করে দিলো এই এক বুড়া। :(তবে আমার আর চিন্তা নাইরে,... http://fb.me/6sO0hVafr	0
Negative	প্রতিটা ছেলের জীবনেই একজন মেয়ে থাকে যাকে সে মন থেকে চায় কিন্তু কখনোই পায় না... :-(:-(:-(:-(:-(:-(	0
Positive	প্রথম বেসরকারি বিশ্ববিদ্যালয় হিসাবে আহছানউল্লা ইউনিভার্সিটি অব সায়েন্স এন্ড টেকনোলজি'র স্থায়ী সনদ লাভ :) http://fb.me/1COsDhMbK	1
Negative	ফিলিপিনোদের জন্যে শোক :(	0
Negative	ফেবু'র কিছুই মিস করি না। শুধু @_Jaati_'র পোস্ট গুলা ছাড়া :(	0
Negative	ফেসবোক আবার হ্যাক হয়েছে :(	0
Negative	বৃক্ষতলে শুয়ে, নিজের দুঃখ ছুয়ে, ঘুম আসেনা, ঘুমও স্বার্থপর !!! :(	0
Positive	বাংলার ফেরাউন ডায়েনী হাসিনার বিচার ব্যবস্থা নিয়ে নির্মিত বাংলা মুভি #দি_ব্ল্যাক দেখতেছি.. ডাউনলোড লিংক কমেন্টে.. feeling-অসাধারণ	1
Positive	বাঙ্গালি জাতি ভুলে জেয়না...... পানি আর তেল কখনও একশাতে মিশতে পারেনা। :P	1
Positive	বাপ রে! নিজের নামের মানে জানতাম না। এখন দেখি SAIHAM মানে LORD!! :P :P	1
Negative	বাপ-বেটা। তখন আব্বার কোলে উঠতে কতো ভালো লাগতো। এখন আর আব্বার কোলে উঠতে পারি না। Feeling sad :(http://fb.me/1PzbHIRzy	0
Positive	বার্সার জয়ে মেসির ইনজুরি বিষাদ | Open My Bangladesh http://fb.me/6ntZ8nTbj	1
Negative	বালের একটা ফ্লাইওভার বানাইছে, কিচ্ছু নড়ে না .... মুইতা দিতে ইচ্ছা করতেছে aka feeling pissed	0
Negative	বেকার জীবনের ১বছর পূর্ন করলাম, আফসোস আমার সেই দু-তিন মাস আর হলো না, ১বছর পরও আবার ২/৩ মাস অপেক্ষার নির্দেশ পাইলাম। feeling stucker	0
Positive	বেশি ভালো :ড "@Chhanda_: @sheikhriad1 এত্তগুলা খুশি!! @DasSatabdi মেয়েটা এত্ত ভালো, এত্ত লক্ষী! :D ♥"	1
Positive	ব্যারিস্টার রফিকুল হক ইদানিং বিএনপি ভালোই বাঁশ দিয়ে যাচ্ছেন :D লোল, এখন এটিএন নিউজ	1
Negative	বৃষ্টির কারনে মনে হয় আজকের খেলা বন্ধই না হয়ে যায়। সেই ড্রই তো হবে। মাঝখান থেকে মমিনুলের আরো কয়েকটি রেকর্ড হবে না... ধুর... :(	0
Negative	বৃষ্টির জলে ভিজেছে, ছিড়েছে তোমার শেষ চিঠিটা অজানায় হারিয়েছে চিঠির কথা.. কি করে জানাব বল হয়নি জানা ছেড়া চিঠি কথা.... feeling #khub krp	0
Positive	বহু বছর পর আবার হলের ল্যান এ... :) :D	1
Negative	বড় অবেলায় পেলাম তোমায়, কেনো এখনি যাবে হারিয়ে। কি করে বলো রবো একেলা, ফিরে দেখো আছি দাড়িয়ে।	0

Fig-3.7: Data Labeling

3.4.4 Feature Extraction

When you're extracting features in sentiment analysis, you're picking out and modifying linguistic features to make things easier to understand for machine learning models. Selecting the right features is crucial because they determine how well a model can identify and understand people's feelings.

So, it's a big deal in sentiment analysis research and applications. In this technique, I've used the 'TF-IDF Vectorizer' to extract features. This vectorizer is particularly useful in sentiment analysis because it helps create a numerical representation of text data that machine learning models can analyze. The below figure 3.8 shows the TF-IDF vectorizer function and 3.9 shows the flowchart of TF-IDF Vectorizer working steps.



Using TF-IDF vectorizer

```
[25] from sklearn.feature_extraction.text import TfidfVectorizer

     cleaned = df['cleaned'].tolist()

     tfidf = TfidfVectorizer(max_features=5000, ngram_range=(1, 2))
```

```
     from sklearn.model_selection import train_test_split
```

```
[30] X = tfidf.fit_transform(df['cleaned']).toarray()
     y = df['Sentiment']

     X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Fig-3.8: Feature Extraction using TF-IDF Vectorizer

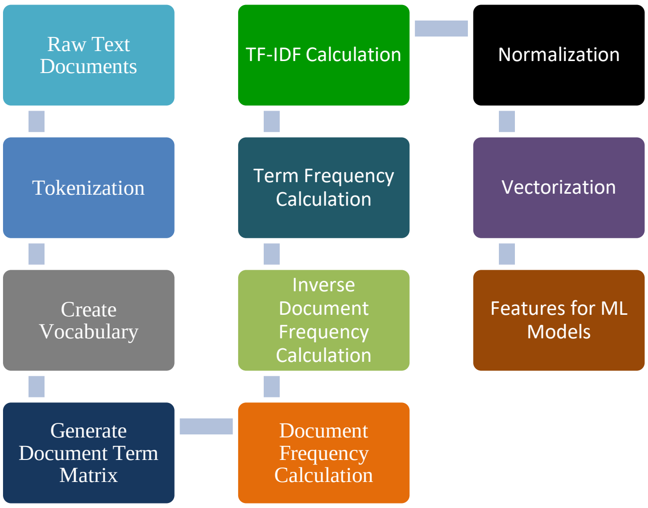

Raw Text Documents
Tokenization
Create Vocabulary
Generate Document Term Matrix
Document Frequency Calculation
Inverse Document Frequency Calculation
Term Frequency Calculation
TF-IDF Calculation
Normalization
Vectorization
Features for ML Models

Fig-3.9: TF-IDF Vectorizer working steps

3.4.5 Model Training

A robust sentiment analysis model tailored to Bengali social media data is developed through the model training process. Regular evaluations, refinements, and adaptations to linguistic nuances improve the model's ability to capture sentiment. This sentiment analysis task uses machine learning and natural language processing (NLP) techniques. Below figure 3.10 shows what machine learning techniques were used and table 3.2 shows fine-tuning parameter used in a few algorithms.

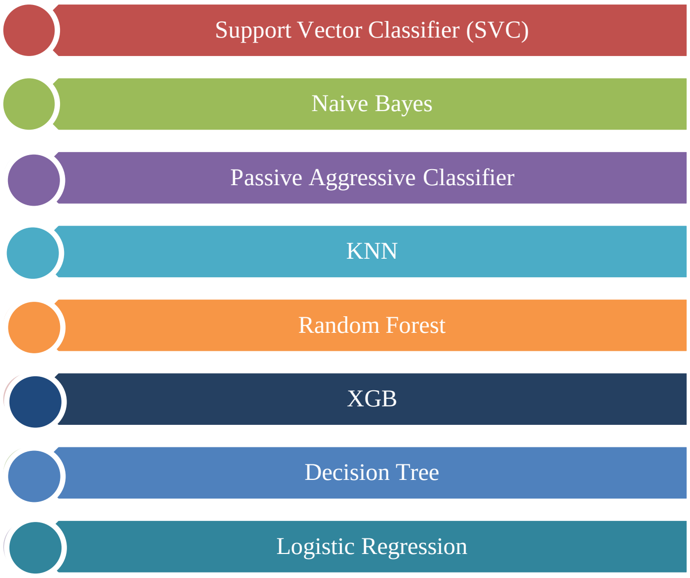


Fig-3.10: Machine Learning Models for training

Table 3.2: PARAMETER USAGE

Algorithms	**Details**
Random Forest	n_estimators=100, random_state=42
XGB	n_estimators=50, random_state=2
Logistic Regression	random_state=42

3.5 Implementation Requirements

Python 3.9

Python 3.9 is the latest version of Python. It's a high-level programming language with a lot of abstractions. Many experts use it for research. It's a fantastic programming language for AI-related work, and it's extremely common among younger programmers because of how easily it can be learned.

Others Requirement:

- Windows or Unix OS (Windows 10 version 21H2 or above / Ubuntu 18.04 or above)
- Web Browser
- Disk space (Minimum 4GB)
- Memory (More than 8GB)

CHAPTER 4
EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Experimental Setup

- For starting sentiment analysis tasks, the initial step is data collection.
- Avoiding all issues regarding data collection, the data collected for this task was from Kaggle, an open-source data resource repository.
- The data is suitable for use after labeling the data.
- Then comes the pre-processing part. After that model training and model evaluation was took place.

4.2 Experimental Results & Analysis

After analyzing the numbers, I used a variety of different calculations to figure out our findings. I had great accuracy, which was over 80%. I also tried to mix real people's details into our predictions, and guess what? The device could actually guess things right based on the information it had. Using multinomial Naive Bayes, support vector machines, and logistic regression, I was able to get an accuracy level of almost 80%. Along with accuracy, I also found out the F1 score and precision. They were also almost 80% accurate. Below figure 4.1 presents accuracy comparison and figure 4.2 presents precision comparison between the all algorithms are presented.

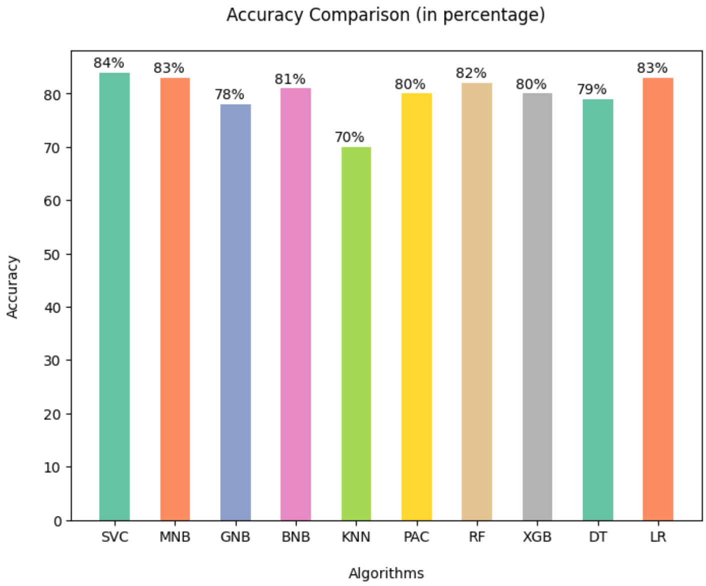


Fig-4.1: Comparison of Accuracy

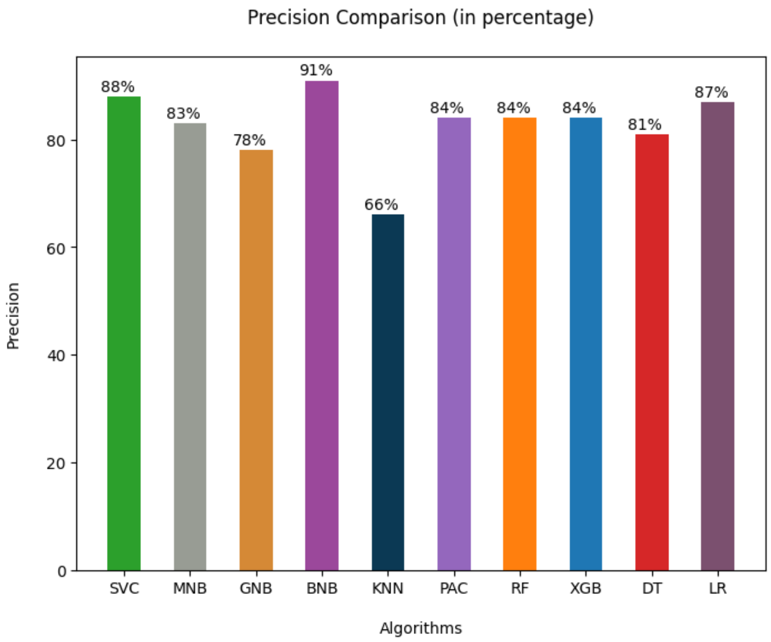


Fig-4.2: Comparison of Precision

Support Vector Machine, a supervised machine learning algorithm, is capable of solving both linear and nonlinear classification and regression problems. Support vector machines are frequently used for classification tasks due to their superior accuracy and minimal computational requirements. It offers an advantage due to its ability to generate trustworthy findings even with a minimal amount of data.

Gaussian Naive Bayes (GNB) refers to Bayes' theorem. It assumes that features follow a Gaussian (normal) distribution. GNB is Ill-suited for continuous data and is commonly used in machine learning tasks where the feature variables are real numbers. Despite its simplicity and the "naive" assumption of feature independence, GNB performs well in various applications, especially when the data distribution aligns with the Gaussian model.

Multinomial Naive Bayes is designed for classification tasks with discrete features, often applied to text classification problems. It assumes that features represent the frequencies with which certain events occur.

Bernoulli Naive Bayes is specifically tailored for binary feature data, where features represent presence or absence. It models each feature as a binary random variable following a Bernoulli distribution. Despite its simplicity and the assumption of feature independence, BNB can be effective in scenarios with binary feature representations.

The Passive Aggressive Classifier is particularly well-suited for scenarios with sequentially arriving data. Its "passive" nature means it maintains its model parameters when predictions are correct, while it becomes "aggressive" and updates parameters when predictions are incorrect. This adaptability makes PAC suitable for real-time applications and streaming data, where it can efficiently adjust to evolving patterns. Its simplicity, efficiency, and ability to handle large datasets make it valuable in various domains.

The K-Nearest Neighbor (K-NN) algorithm, one of the simplest categorization methods, is a supervised machine learning algorithm. It stores all previously observed cases, subsequently classifying new cases based on similarity criteria.

In classification and regression problems, supervised machine learning methods like Random Forest classifiers are often used. It constructs decision trees from various samples and uses their average for categorization and a majority vote for regression.

Among other things, classification and regression tasks are accomplished using a machine learning technique called gradient boosting. It offers a prediction model in the form of a collection of weak prediction models that resemble decision trees. The tree-based ensemble machine learning method Extreme Gradient Boosting is a scalable machine learning system for tree boosting. The main reasons for using XGBoost are speed and model performance.

Decision trees are non-parametric supervised learning techniques used for classification and regression. The goal is to learn simple decision rules based on data characteristics in order to develop a model that predicts the value of a target variable.

Logistic regression is a machine learning algorithm. It is used to address classification problems similar to linear regression. It is a predictive analysis-based approach, with probability serving as its underlying principle.

The below table 4.1 shows accuracy and precision of all algorithms.

Table 4.1: MODEL RESULT

Algorithm	Accuracy	Precision
SVC	84	88
MNB	83	83
GNB	78	78
BNB	81	91
KNN	70	66
PAC	80	84
RF	82	84
XGB	80	84
DT	79	81
LR	83	87

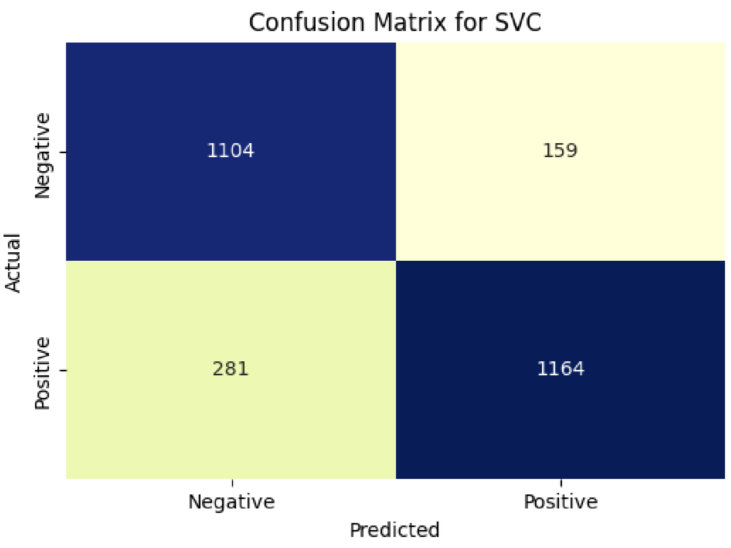
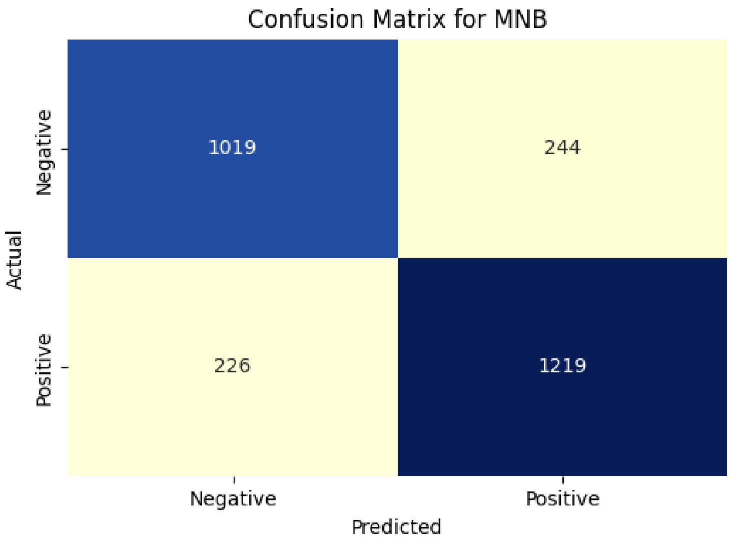
It has been found that the Support Vector Machine classifier is the most accurate in terms of Accuracy and Precision score in the study. This model resulted with the accuracy is 84% and precision 88%, the best value for any classifier that has been tested. Next best count is for Multinomial Naïve Bayes, where accuracy and precision score is 83%.

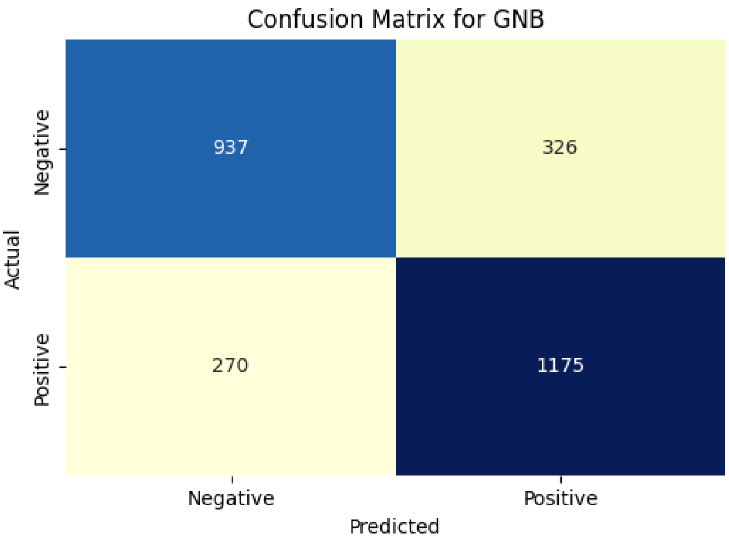
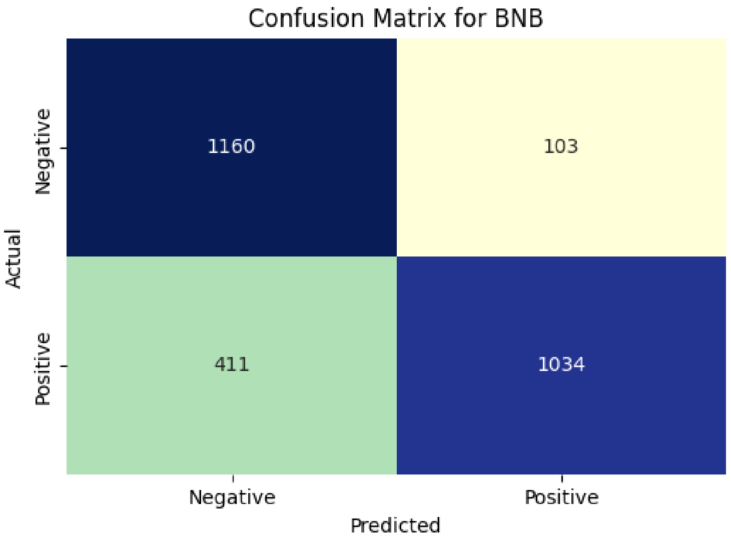
4.2.1 Confusion Matrix Analysis

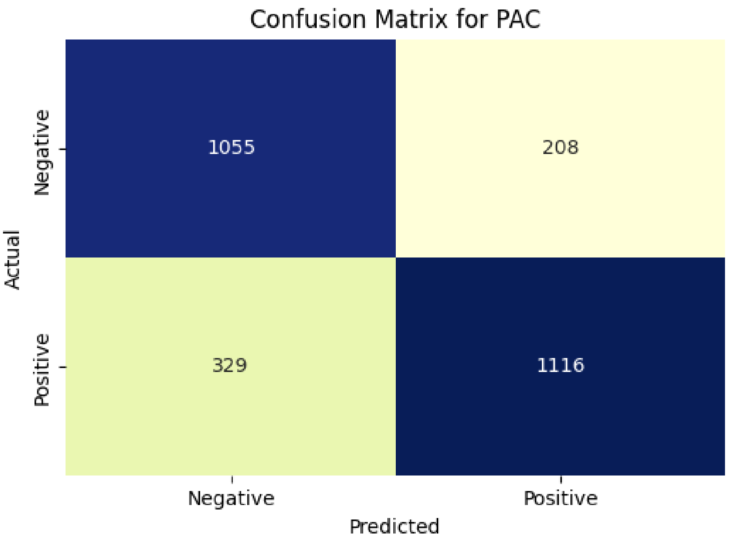
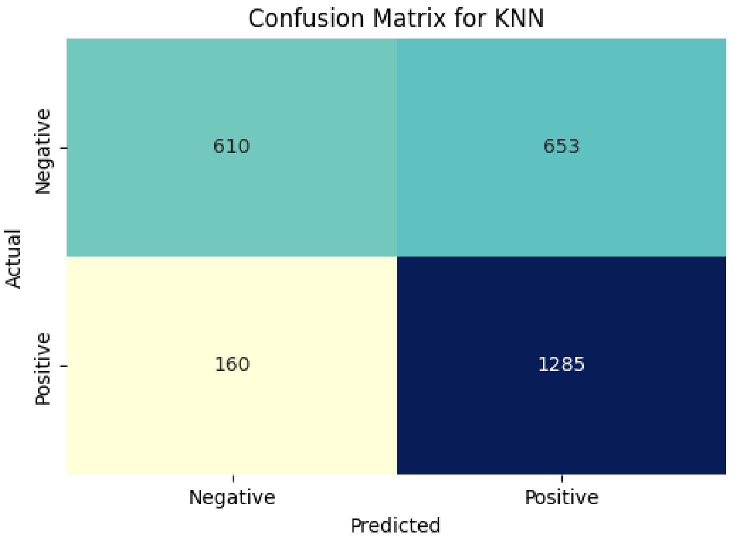
I have constructed confusion matrix to evaluate our model's performance on the test set. Based on the confusion matrix and associated performance metrics to gain valuable insight into the strengths and weaknesses of the model. As a result of these metrics, it may be possible to refine or adjust the model to better align it with the project's goals. Understanding the relationship between true values and predicted values is crucial for evaluating the performance of a machine learning model. The true values represent the actual outcomes, while the predicted values represent the labels that model assigns.

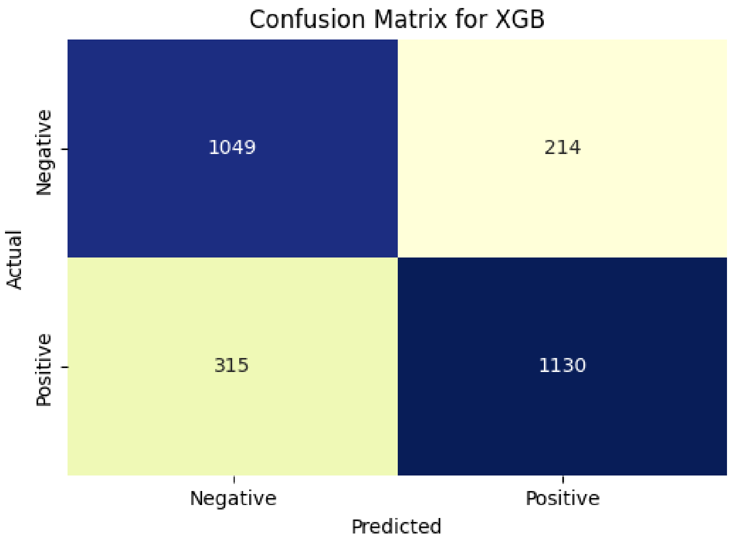
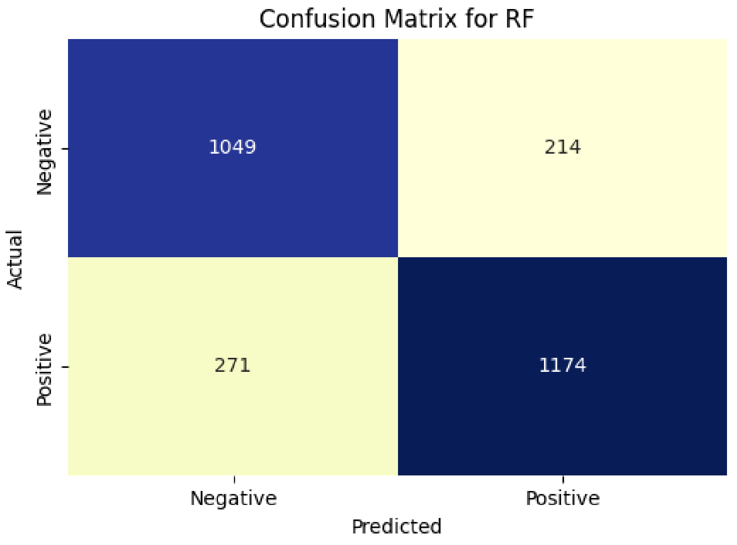
To gain a comprehensive understanding of the model's performance, confusion matrix presented here in the table 4.2 that provides a detailed breakdown of the model's predictions.

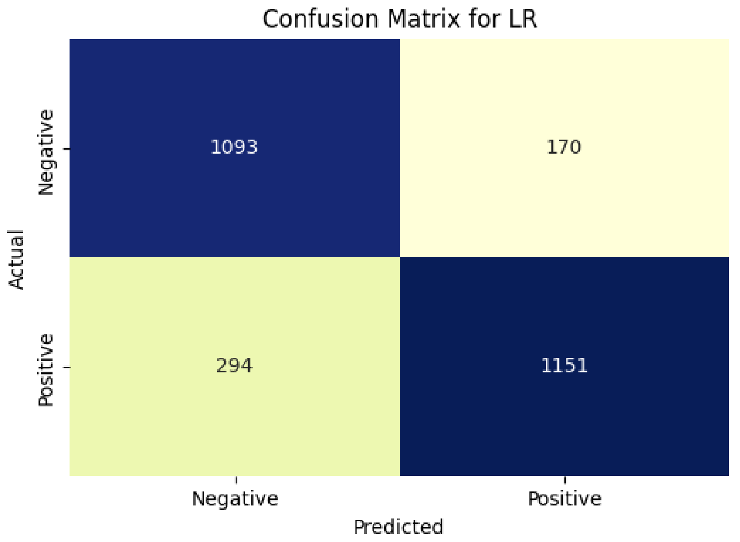
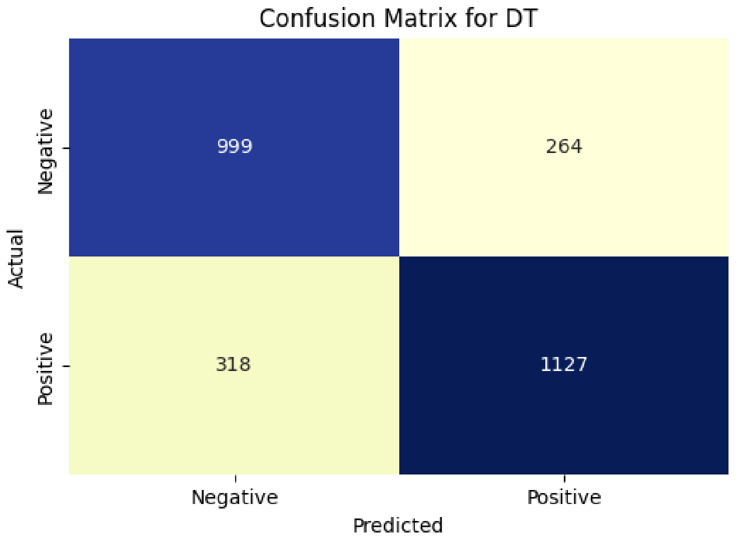
Table-4.2: CONFUSION MATRIX OF ALL ALGORITHMS

Support Vector Machine Classifier (SVC)	
Multinomial Naïve Bayes (MNB)	

Gaussian Naïve Bayes (GNB)	
Bernoulli Naïve Bayes (BNB)	

Passive Aggressive Classifier (PAC)	
K-Nearest Neighbors (KNN)	

XGBoost (XGB)	
Random Forest Classifier (RF)	

Logistic Regression (LR)	
Decision Tree Classifier (DT)	

4.2.2 ROC Curve Score Analysis

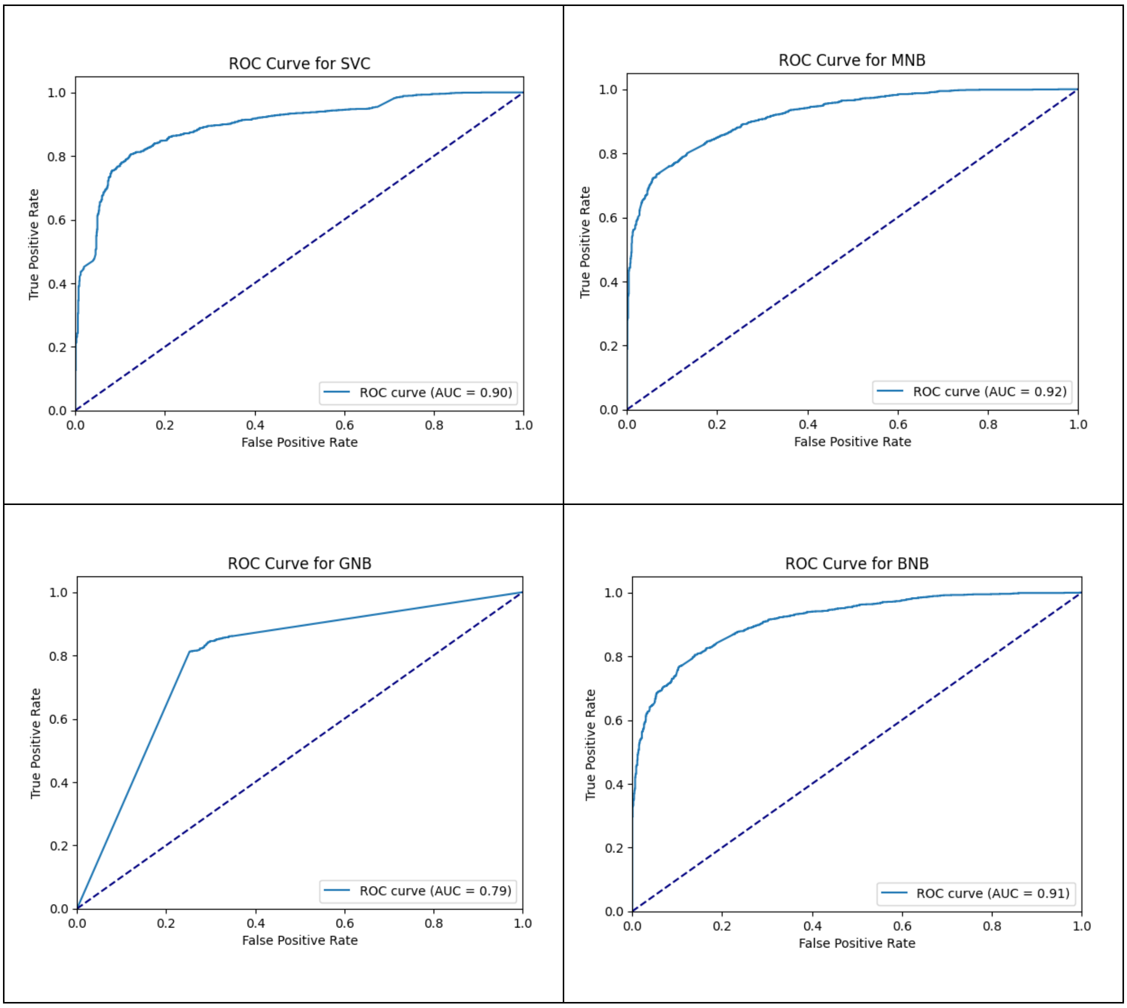
A ROC curve describes the trade-off between sensitivity (true positive rate) and specificity (true negative rate) at different thresholds for a classifier. In this way, you can assess the performance of a classification algorithm at various classification thresholds. There is a diagonal line in the ROC space (from (0,0) to (1,1) that represents random guessing, and a good classifier should have a curve that is higher and to the left of this line. They are useful in assessing the model's ability to discriminate between positive and negative instances across a range of threshold values.

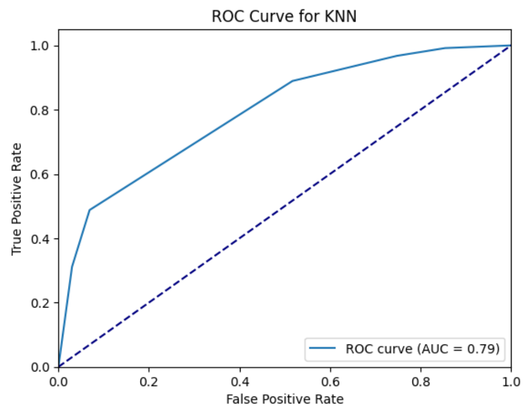
The terms used for ROC are below:

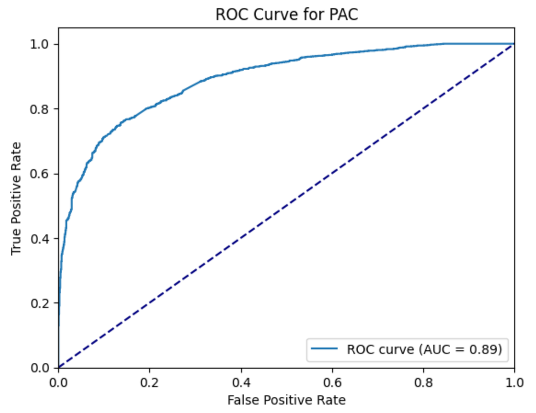
- False Positive Rate (FPR): This represents the false positive rate, which is the ratio of false positives to the total number of negatives (FP / (FP + TN)). It's also called fallout or FPR.
- True Positive Rate (TPR): This rate represents the ratio of true positives to the number of positives (TP / (TP + FN)). TPR is also referred to as sensitivity or recall.

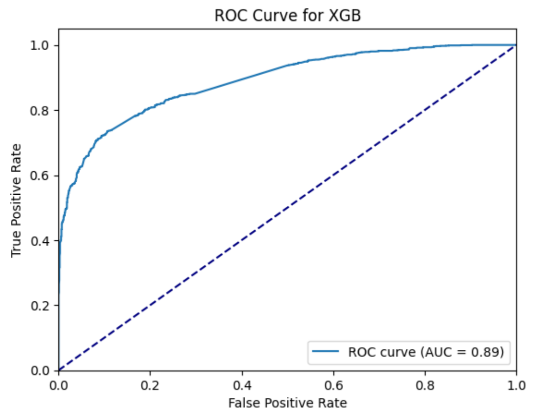
The following table 4.3 presents the ROC curve score of all algorithms used in this study.

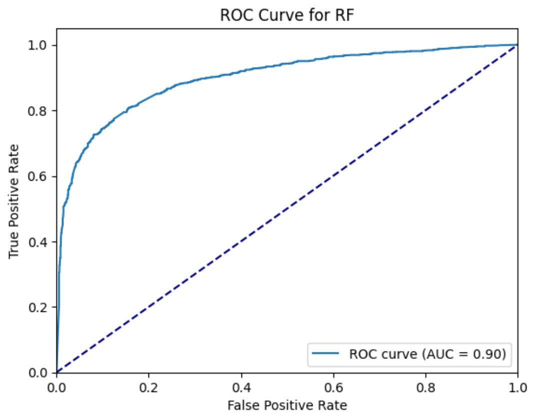
Table 4.3: ROC CURVE OF ALL ALGORITHMS

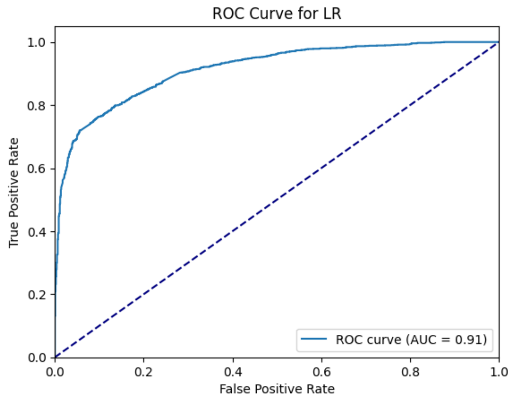


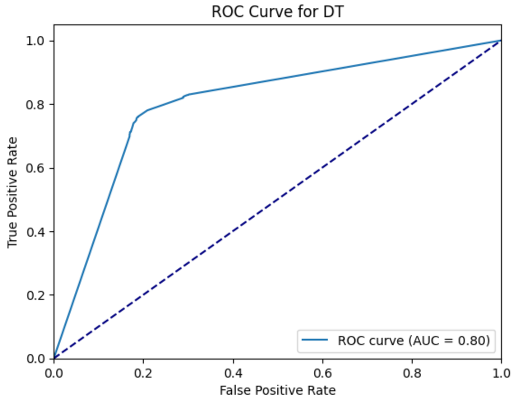

ROC Curve for KNN
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (AUC = 0.79)


ROC Curve for PAC
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (AUC = 0.89)


ROC Curve for XGB
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (AUC = 0.89)


ROC Curve for RF
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (AUC = 0.90)


ROC Curve for LR
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (AUC = 0.91)


ROC Curve for DT
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (AUC = 0.80)

Below figure 4.3 shows the comparison of AUC-ROC curve score of applied algorithms.

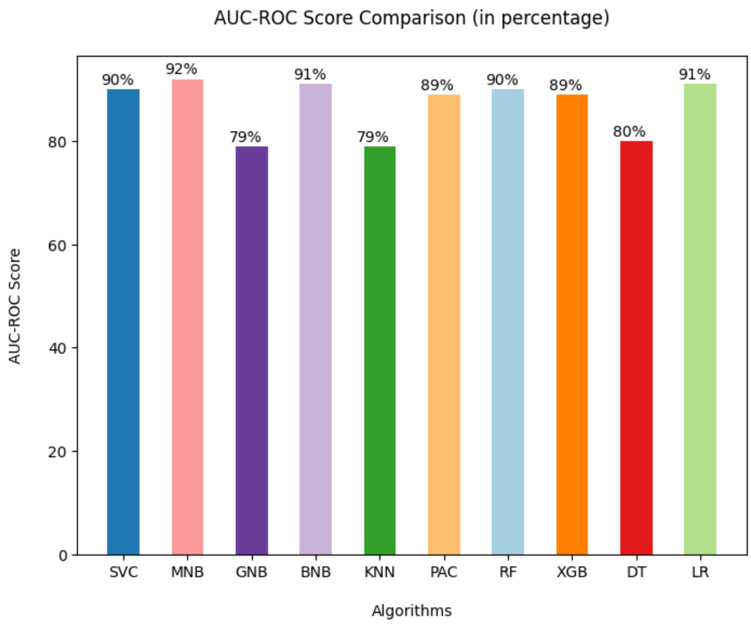


Fig-4.3: Comparison of AUC-ROC score

4.3 Discussion

At achieving this level of accuracy, I became satisfied somehow. However, to improve the level of accuracy, you will need to ensure that the dataset is correctly prepared. There should be maintain consistency for each of the sentiment categories. Currently, there is no option for data cleaning in order to improve the accuracy of predictions. As more data is preprocessed, this classifier's predictions become more precise.

CHAPTER 5

IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

5.1 Impact on Society

Sentiment analysis has had a significant impact on society, affecting a wide range of sectors and aspects of our daily lives. The following figure 5.1 demonstrate the sectors were affects by sentiment analysis.

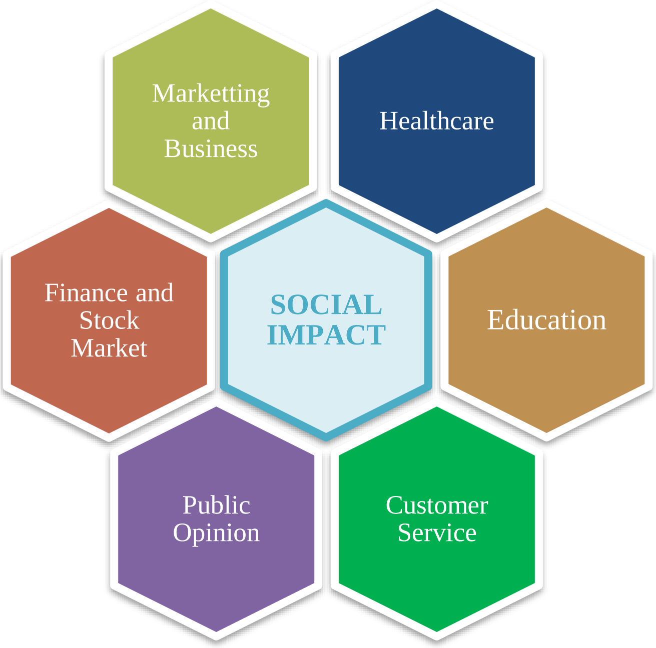


Fig-5.1: Different Sectors Impacting by Sentiment Analysis

The following points highlights the impact on society:

- Businesses can use sentiment analysis to better understand their customers' opinions and feedback.
- A company's marketing campaign can be customized using sentiment analysis to ensure that it resonates positively with the target audience.
- A company can prevent potential problems from escalating by monitoring sentiment trends and addressing emerging issues as they arise.

- Political campaigns and policymakers benefit from sentiment analysis by gauging public opinion on political issues, candidates, and policies.
- To improve the quality of healthcare services, healthcare providers can use sentiment analysis to analyze patient reviews and feedback.
- By analyzing language patterns, sentiment analysis research can help monitor and assess mental health
- Students' feedback and opinions are analyzed using sentiment analysis by educational institutions to improve the curriculum, teaching methods, and overall educational experience.

These points illustrate the multifaceted impact of sentiment analysis research on various aspects of society, ranging from business and politics to healthcare and education.

5.2 Impact on Environment

Sentiment analysis, the art of understanding emotions in text, can have a surprising impact on the environment. By analyzing public opinion on social media and news platforms, it helps gauge public concern about environmental issues like climate change. This data can inform policy decisions, direct resources toward areas of greatest public concern, and even hold companies accountable for environmentally damaging practices. Yet, we must exercise ethical considerations and limitations when trying to capture nuanced emotions. In general, sentiment analysis offers a promising, imperfect, tool for understanding public sentiment and driving positive environmental change. By utilizing the method applied in this study, either favorable or unfavorable feedback can be easily identified, leading to an increase in online purchases compared to the previous period. Individuals who are doubtful about which product to purchase will show interest in it. This will increase online purchases. As our method utilizes internet product reviews and is automated, I do not contribute to environmental damage.

5.3 Ethical Aspects

Sentiment analysis looks at people's emotions, but it has some important ethical issues. One concern is about privacy, which means that people's personal information should be protected. Another concern is about fairness and avoiding discrimination. Sentiment analysis should also be transparent and accountable, meaning people should understand how it works and what it is used for. By being ethical, sentiment analysis can help create a world that values compassion and responsibility.

5.4 Sustainability Plan

The pursuit of knowledge in sentiment analysis, while valuable, often leaves an unseen environmental footprint. To truly embrace sustainability, researchers must shift their focus to environmentally-friendly practices throughout the research lifecycle. Here are some key areas for greening sentiment analysis research:

- **Data collection and storage:** Analyze publicly available data or reuse existing datasets before collecting new data. Design efficient data collection systems to reduce redundancy and energy consumption. Use cloud platforms powered by renewable energy for data storage and computation.

- **Model Training and Analysis:** Sentiment analysis requires less computational power and resources for model selection and training. Distributed computing platforms reduce the burden of individual energy of analysis tasks. Develop tools to measure and communicate the carbon footprint of sentiment analysis research.

- **Dissemination and sharing:** Encourage open publication of research results to reduce printing and distribution costs. Use online platforms and tools for collaboration and communication to reduce travel demand.

By implementing these strategies and constantly seeking new green solutions, sentiment analysis research can reduce its carbon footprint and become a driving force for environmental responsibility.

CHAPTER 6

SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH

6.1 Summary of the Study

The goal of this research project is to predict public emotional states from social media data. Several well-known machine learning models Ire used in this study, along with natural language processing (NLP) approaches. TF-IDF was shown to be less effective than count vectorizer modeling in terms of feature extraction. Data collection from social media sites such as Facebook appears to be time-consuming and challenging; suitable web scraping tools Ire not discovered. As a result, managing each issue dataset for this study was collected from Kaggle, an open-source data repository. However, there might be an issue with data labeling; biases might have been found. With all of these challenges in consideration, the primary objective of this study is to create an effective model that reliably predicts the sentiment of Bengali data from social media.

6.2 Conclusions

Sentiment analysis in Bengali language presents several challenges, including complex grammar, a variety of vocabulary, dialectal differences, slang, colloquialisms, cultural references, and code-mixing with other languages. Pre-trained models struggle with Bengali nuances, whereas models specifically trained in Bengali face challenges due to grammar complexities. Transfer learning with fine-tuned English models using smaller Bengali datasets shows promise for improving Bengali sentiment analysis. Feature selection techniques (e.g., TF-IDF or word embeddings) and hyperparameter optimization can improve Bengali sentiment analysis. Diverse training data from different sources makes sentiment analysis more effective. Sentiment analysis is used in marketing and public policy to develop tailored strategies for businesses The classifier can be trained using a previously learned dataset containing around 15,000 data with two initial features. Although the categorization method used in the study was inaccurate regarding sentiment analysis of Bangla text. Preprocessing can be done directly on raw data, addressing Bangla text concerns. It is hoped that future examiners will find this research valuable for examining Bangla texts or news.

6.3 Implication for Further Study

Sentiment analysis holds great promise for further research, which not only provides greater insight into human emotions but also has specific implications for various fields. Here's a taste of some interesting possibilities:

- Building more complex sentiment analysis algorithms that capture the full range of human emotions and contexts
- Enhancing our ability to reliably assess emotions in local Bengali language and cultural contexts.
- Developing real-time sentiment analysis tools that allow for rapid response to public opinion and crisis situations.
- Combining multiple sources of data such as text, images and video can improve sentiment analysis.
- More transparent and interpretable sentiment analysis techniques for building trust in AI systems, especially in applications with significant social or ethical implications.
- Efforts to reduce bias in sentiment analysis algorithms prioritize fairness, ensuring that the analysis does not reinforce or perpetuate existing biases in the language data.

REFERENCES

[1] “World Population Clock,” [Online]. Available: https://www.worldometers.info/world-population/, [Accessed: 13 November 2023].

[2] “Global social media statistics,” [Online]. Available: https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/, [Accessed: 11 November 2023].

[3] “Bengali language: Wikipedia”, [Online]. Available: https://en.wikipedia.org/wiki/Bengali_languag, [Accessed: 13 November 2023].

[4] Haque, R., Islam, N., Tasneem, M., & Das, A. K. (2023). Multi-class sentiment classification on Bengali social media comments using machine learning. International Journal of Cognitive Computing in Engineering, 4, 21-35.

[5] Rahman, M., Talukder, M. R. A., Setu, L. A., & Das, A. K. (2022). A dynamic strategy for classifying sentiment from Bengali text by utilizing Word2vector model. Journal of Information Technology Research (JITR), 15(1), 1-17.

[6] Urmi, T. T., Jammy, J. J., & Ismail, S. (2016, May). A corpus based unsupervised Bangla word stemming using N-gram language model. In 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV) (pp. 824-828). IEEE.

[7] Das, D., & Bandyopadhyay, S. (2010). Developing bengali wordnet affect for analyzing emotion. In International Conference on the Computer Processing of Oriental Languages (pp. 35-40).

[8] Sharmin, S., & Chakma, D. (2021). Attention-based convolutional neural network for Bangla sentiment analysis. Ai & Society, 36, 381-396.

[9] Naim, F. A. (2021, February). Bangla aspect-based sentiment analysis based on corresponding term extraction. In 2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD) (pp. 65-69). IEEE.

[10] Ali, H., Hossain, M. F., Shuvo, S. B., & Al Marouf, A. (2020, July). Banglasenti: A dataset of bangla words for sentiment analysis. In 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-4). IEEE.

[11] Amin, A., Hossain, I., Akther, A., & Alam, K. M. (2019, February). Bengali vader: A sentiment analysis approach using modified vader. In 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE) (pp. 1-6). IEEE.

[12] Sumit, S. H., Hossan, M. Z., Al Muntasir, T., & Sourov, T. (2018, September). Exploring word embedding for bangla sentiment analysis. In 2018 International Conference on Bangla Speech and Language Processing (ICBSLP) (pp. 1-5). IEEE.

[13] Salehin, S. S., Miah, R., & Islam, M. S. (2020, January). A comparative sentiment analysis on Bengali Facebook posts. In Proceedings of the international conference on computing advancements (pp. 1-8).

[14] Pran, M. S. A., Bhuiyan, M. R., Hossain, S. A., & Abujar, S. (2020, July). Analysis of Bangladeshi people's emotion during COVID-19 in social media using deep learning. In 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-6). IEEE.

[15] Hasan, K. A., & 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.

[16] Islam, M. S., Islam, M. A., Hossain, M. A., & Dey, J. J. (2016, December). Supervised approach of sentimentality extraction from bengali facebook status. In 2016 19th international conference on computer and information technology (ICCIT) (pp. 383-387). IEEE.

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

[18] Song, Y., Wang, J., Liang, Z., Liu, Z., & Jiang, T. (2020). Utilizing BERT intermediate layers for aspect-based sentiment analysis and natural language inference. arXiv preprint arXiv:2002.04815.

[19] Taher, S. A., Akhter, K. A., & Hasan, K. A. (2018, September). N-gram based sentiment mining for bangla text using support vector machine. In 2018 international conference on Bangla speech and language processing (ICBSLP) (pp. 1-5). IEEE.

[20] Alam, M. H., Rahoman, M. M., & Azad, M. A. K. (2017, December). Sentiment analysis for Bangla sentences using convolutional neural network. In 2017 20th International Conference of Computer and Information Technology (ICCIT) (pp. 1-6). IEEE.

[21] Hassan, A., Amin, M. R., Al Azad, A. K., & Mohammed, N. (2016, December). Sentiment analysis on bangla and romanized bangla text using deep recurrent models. In 2016 International Workshop on Computational Intelligence (IWCI) (pp. 51-56). IEEE.

[22] Hasan, M. A., Tajrin, J., Chowdhury, S. A., & Alam, F. (2020, December). Sentiment classification in bangla textual content: A comparative study. In 2020 23rd international conference on computer and information technology (ICCIT) (pp. 1-6). IEEE.

[23] Haydar, M. S., Al Helal, M., & Hossain, S. A. (2018, February). Sentiment extraction from bangla text: A character level supervised recurrent neural network approach. In 2018 international conference on computer, communication, chemical, material and electronic engineering (IC4ME2) (pp. 1-4). IEEE.

[24] Hassan, M., Shakil, S., Moon, N. N., Islam, M. M., Hossain, R. A., Mariam, A., & Nur, F. N. (2022). Sentiment analysis on Bangla conversation using machine learning approach. International Journal of Electrical and Computer Engineering (IJECE), 12(5), 5562-5572.

[25] Wahid, M. F., Hasan, M. J., & Alom, M. S. (2019, September). Cricket sentiment analysis from bangla text using recurrent neural network with long short-term memory model. In 2019 International Conference on Bangla Speech and Language Processing (ICBSLP) (pp. 1-4). IEEE.

[26] Yadav, V., Verma, P., & Katiyar, V. (2021, January). E-commerce product reviews using aspect based Hindi sentiment analysis. In 2021 International Conference on Computer Communication and Informatics (ICCCI) (pp. 1-8). IEEE.

[27] Al-Amin, M., Islam, M. S., & Uzzal, S. D. (2017, February). Sentiment analysis of Bengali comments with Word2Vec and sentiment information of words. In 2017 international conference on electrical, computer and communication engineering (ECCE) (pp. 186-190). IEEE.

[28] Rahman, M. A., & Kumar Dey, E. (2018). Datasets for aspect-based sentiment analysis in bangla and its baseline evaluation. Data, 3(2), 15.

[29] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. Computational linguistics, 37(2), 267-307.

[30] Abubakar, M., Shahzad, A., & Abbasi, H. (2021). Aspect-based sentiment analysis on amazon product reviews. International Journal of Informatics, Information System and Computer Engineering (INJIISCOM), 2(2), 94-99.

[31] Schouten, K., Van Der Iijde, O., Frasincar, F., & Dekker, R. (2017). Supervised and unsupervised aspect category detection for sentiment analysis with co-occurrence data. IEEE transactions on cybernetics, 48(4), 1263-1275.

[32] Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. CS224N project report, Stanford, 1(12), 2009.

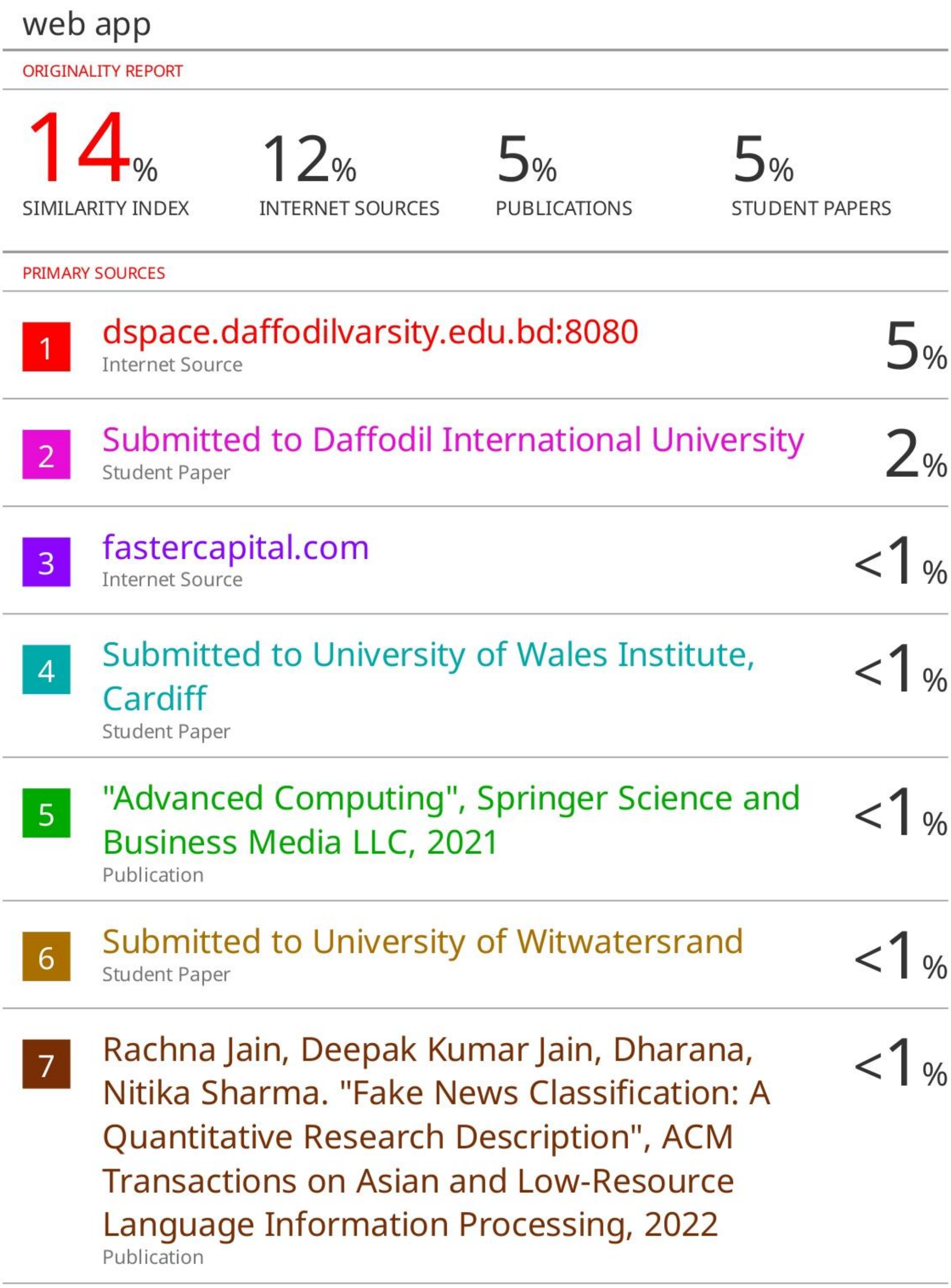
[33] Das, D., & Bandyopadhyay, S. (2010, August). Labeling emotion in Bengali blog corpus–a fine grained tagging at sentence level. In Proceedings of the eighth workshop on Asian language resouces (pp. 47-55).

[34] Maas, A., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011, June). Learning word vectors for sentiment analysis. In Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies (pp. 142-150).

[35] Neethu, M. S., & Rajasree, R. (2013, July). Sentiment analysis in twitter using machine learning techniques. In 2013 fourth international conference on computing, communications and networking technologies (ICCCNT) (pp. 1-5). IEEE.

[36] Jain, A. P., & Dandannavar, P. (2016, July). Application of machine learning techniques to sentiment analysis. In 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT) (pp. 628-632). IEEE.

PLAGIARISM REPORT



web app

ORIGINALITY REPORT

14%	12%	5%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	5%
2	Submitted to Daffodil International University Student Paper	2%
3	fastercapital.com Internet Source	<1%
4	Submitted to University of Wales Institute, Cardiff Student Paper	<1%
5	"Advanced Computing", Springer Science and Business Media LLC, 2021 Publication	<1%
6	Submitted to University of Witwatersrand Student Paper	<1%
7	Rachna Jain, Deepak Kumar Jain, Dharana, Nitika Sharma. "Fake News Classification: A Quantitative Research Description", ACM Transactions on Asian and Low-Resource Language Information Processing, 2022 Publication	<1%