SENTIMENT ANALYSIS OF BANGLA SONG COMMENTS REVIEW: A MACHINE LEARNING APPROACH

BY

Sabuj Chandra Das
ID: 191-15-12525

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

Supervised By

Professor Dr. Md. Ismail Jabiullah
Professor
Department of CSE
Daffodil International University

Co-Supervised By

Md. Sadekur Rahman
Assistant Professor
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

NOVEMBER 2023

APPROVAL

This Project titled "**Sentiment Analysis of Bangla Song Comments Review: A Machine Learning Approach**", submitted by Sabuj Chandra Das, ID No: 191-15-12525 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 13 November 2023.

BOARD OF EXAMINERS

Chairman

Narayan Ranjan Chakraborty (NRC)
Associate Professor and Associate Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Md. Sadekur Rahman (SR)
Assistant Professor
Department of Computer Science and Engineering
Daffodil International University

Internal Examiner

Md. Abbas Ali Khan (AAK)
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

External Examiner

Dr. Mohammed Nasir Uddin (DNU)
Professor
Department of Computer Science and Engineering
Jagannath University

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Dr. Md. Ismail Jabiullah, 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.

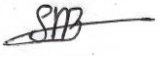
Supervised by:

Dr. Md. Ismail Jabiullah
Professor
Department of CSE
Daffodil International University

Co-Supervised by:

Md. Sadekur Rahman
Assistant Professor
Department of CSE
Daffodil International University

Submitted by:



Sabuj Chandra Das
ID: 191-15-12525
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First I express our 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 **Dr. Md. Ismail Jabiullah, Professor, Department of CSE**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "*Deep Learning and 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 draft and correcting them at all stage have made it possible to complete this thesis.

I would also like to thank my co-supervisor **Md. Sadekur Rahman, Assistant Professor, Department of CSE**, Daffodil International University, Dhaka. When I face any problem, he helped me with valuable ideas and suggestions. He motivated me and help me to complete this work.

I would like to express our heartiest gratitude **to Dr. Touhid Bhuiyan, 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 our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, I am very thankful to my parents and friends who were always motivate me and criticize my work in a manner to improve my work. At least I thank all of them form the core of my heart.

ABSTRACT

Sentiment analysis represents a contemporary approach in Natural Language Processing used to determine the sentiment of a user. Bangla music, with its unique rhythms, melodies, and lyrical depth, stands as a musical treasure that reflects the soul of the Bengali culture. This music serves as a powerful storyteller, narrating tales of love, longing, joy, and resilience. From the soulful tunes of Rabindra Sangeet to the electrifying beats of contemporary Bangla pop, this music genre captures the essence of Bengali life and spirit. It transcends borders, touching the hearts of listeners worldwide with its timeless beauty and profound lyrics, making it a cherished part of global music heritage. This research study has attempted to investigate on Bangla music using different machine learning classification algorithms from over 2224 data points. People of all ages are considered as data collection. After pre-processing and feature engineering the collected data, Random Forest, Decision Tree, Multinomial Naive Bayes, XGBoost, K-Nearest Neighbor and Support Vector Machine are used to train the model. These classifiers predict with high accuracy, with Random Forest having the highest accuracy of 63.68%.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1 Introduction

Regarding the modern age, interaction over the internet is so simple that anyone, anywhere, at any time, can communicate their feelings with anyone else. Nowadays, people spend more time online engaging in activities like online gaming, purchasing and selling goods, running businesses, and using services like Telegram, WhatsApp, Skype, Facebook and YouTube. YouTube has become one of the most addicting music streaming services, where users spend the most time. Bangla song tells tales of love, longing, joy, and resiliency and is a potent storyteller. One of the Natural Language Processing (NLP) methods is sentiment analysis, which is formed from the words sentiment and analysis. The concept of "sentiment" and the range of human emotions that might be experienced are closely related. Analysis is the process of dissecting a complex material into smaller pieces to provide better results, and sentiment analysis is the process of dissecting text utilizing various statistical and language processing approaches. An excel sheet has been created, it holds comments that are collected from social media, where various types of people's comments are classified, in order to carry out the specified research. We gather data and make predictions using machine learning techniques [11], [12].

Sentiment analysis primarily supported rule-based, automated, and hybrid approaches. Popular methodologies including lexicons, parsing, stemming, parts of speech tagging, tokenization, and others are all included in the rule-based approach. The proposed model's input has undergone automatic training, or the automatic approach. The hybrid strategy will be the best choice if a model needs to incorporate both systems. These techniques are employed to provide precise results [13].

Researchers have tried to investigate and identify suitable solutions to a variety of real-time difficulties while also using sentiment analysis for diverse purposes. The online actions can be examined using sentiment analysis. Numerous public attention items, such as the outcomes of the election are taken into consideration when determining the level of violence. Additionally, this technology is capable of

identifying the general public's preferences in food, attire, and other aspects. Sentiment analysis is crucial in figuring out things like human psychology, rates of depression, suicide risk, and other things. This can also be used to forecast a movie's box office earnings and analyze the state of the market.

All age groups, especially teenagers and aged people, have shown an increased interest in listening to Bangla songs. They spend the majority of their precious time listening to music that perfectly depicts the character and way of life of Bengalis, from the electric beats of modern Bangla pop to the soulful melodies of Rabindra Sangeet. The primary problem, though, is that after a few days of listening to music, they seem to have become dependent on it. This study's main objective is to examine how Bangla songs are now used in our culture and how they relate to mental health. In order to gather data, the planned study has made an effort to contact people of different ages. The initial dataset consists of a huge number of comments kept in an Excel sheet. Unwanted character, tokenization, and applied contraction are removed during the pre-processing of the data.After feature engineering, the proposed dataset is trained using six classification algorithms viz. Random Forest (RF), Decision Tree (DT), Multinomial Naive Bayes (MNB), Extreme Gradient Boosting (XGB), K-Nearest Neighbor (KNN), and Support Vector Machines (SVM). The dataset is trained through all the six algorithms individually with good accuracy. Finally, an accuracy graph and classification report has been developed to easily and clearly analyze the results.

1.2 Motivation

This research is driven by the goal of harnessing machine learning to gain a deeper understanding of the sentiments conveyed in Bengali song comments. Such an analysis holds practical significance for the music industry, furnishes cultural insights, and tackles the complexities of sentiment analysis in a less-explored language, thereby making meaningful contributions to both academic inquiry and practical applications. The driving force behind this research project is the aspiration to employ advanced machine learning methodologies in addressing the unique challenges of sentiment analysis in Bengali song comments, ultimately shedding light on cultural responses to music and enriching the broader domain of natural language processing.

1.3 Objective

This research is driven by the desire to harness machine learning for a deeper comprehension of the sentiments conveyed in Bengali song comments. Such an analysis holds pragmatic significance for the music industry, provides valuable cultural insights, and tackles the complexities of sentiment analysis in a language that has received less attention, thereby making noteworthy contributions to both academic scholarship and practical applications. The overarching goal of this study is to close the gap in sentiment analysis for underrepresented languages, delivering advantages to the music industry, fostering cultural understanding, and enriching the wider landscape of natural language processing.

1.4 Research Questions

- What are the sentiments expressed in Bangla song comments, and how can machine learning techniques be applied to accurately classify these sentiments as positive, negative, or neutral?

- What cultural and emotional insights can be gained from the sentiment analysis of Bangla song comments, and how might these findings impact the Bangla music industry?

- Which machine learning algorithms and natural language processing techniques are most effective for sentiment analysis in the Bangla language?

1.5 Expected Outcome

The expected outcome of this research is the development of accurate machine learning models for sentiment analysis of Bangla song comments, providing valuable insights into audience emotions, potential impacts on the music industry, and enhanced cross-cultural understanding. The anticipated results of my experiment are outlined as follows:

- Successfully identify and distinguish objects.
- Investigate the training and testing image datasets.
- Achieve a higher level of accuracy using the proposed algorithm.

CHAPTER 2

BACKGROUND STUDY

2.1 Terminology System

A few numbers of existing research literatures are analysed in order to move forward with the suggested research project. This section will introduce a few pieces of work associated with the proposed study project.

2.2 Literature Review

Hasmot et al. highlight the "BanglaSenti" dataset, a valuable resource for Bengali sentiment analysis, particularly in user-generated social media content. This dataset, containing 61,582 Bengali words categorized by sentiment, advances Bangla Natural Language Processing (BNLP) and finds versatile applications beyond sentiment analysis, including emotion detection and opinion mining [1].

Rashedul et al. delve into Bengali text emotion detection, with wide-reaching applications in product analysis, social media monitoring, and market research. The paper introduces two effective methods, though it acknowledges data-related challenges. The study aims to enhance performance using hybrid approaches and other strategies, paving the way for practical user applications in the future [2].

Tapasy Rabeya and et al. [3] primary focus is on performing sentiment analysis of Bengali song reviews from a specific YouTube channel, aiming to assess the popularity of a rising young artist. The paper introduces a backtracking algorithm with a sentiment lexicon, achieving a commendable accuracy rate exceeding 70% in capturing public sentiment. This experimental study analyzes 288 unprocessed comments, and the algorithm demonstrates strong performance in sentiment detection.

In their paper [4], Mst. Eshita Khatun and Tapasy Rabeya conduct sentiment analysis on Bengali book reviews, aiming to detect positive and negative sentiments. They analyze 5500 user-generated reviews, using five algorithms, and

achieve remarkable accuracy, with Random Forest reaching 98.39%. The study emphasizes the significance of sentiment analysis in online shopping during the COVID-19 pandemic and suggests improvements for future predictions in the Bengali language.

Deebha Mumtaza and Bindiya Ahujab et al. [5] focus lies in sentiment analysis of Twitter movie reviews using the "Senti-lexicon" algorithm. It categorizes reviews as positive, negative, or neutral, tackling challenges like sarcasm and complexity. This approach proves simpler and more practical than conventional machine learning methods. Future work will prioritize improving the algorithm's performance and addressing diverse sentiment analysis challenges.

This paper [6], Nagamma P. and et al. emphasis is on employing sentiment analysis and machine learning to forecast box office earnings based on online movie reviews. It introduces a streamlined sentiment-aware model and uses data mining methods like clustering to enhance prediction precision. The study proposes future directions, including the detection of double negative phrases and expanding the model's utility to broader applications.

In paper [7], V.K. Singh and et al. fresh method is presented for aspect-level sentiment analysis in movie reviews. It assesses reviews, attributes sentiment labels to distinct aspects, and constructs in-depth sentiment profiles for each movie. The study utilizes linguistic features and contrasts findings with document-level sentiment analysis, showcasing its precision and precision in capturing nuanced sentiments.

In their paper [8], Abdur Nur Tusher et al. delve into online gaming addiction, particularly among youth in the COVID-19 era. They gather data from 401 individuals, including students, and employ six machine learning classification algorithms to assess the situation. The study categorizes data into positive, negative, and neutral groups, with Multinomial Naive Bayes (MNB) attaining the highest accuracy of 73.27%. Nevertheless, challenges in data collection and hesitancy to express opinions are acknowledged, signaling potential avenues for future research.

In this paper [9], Namita Mittal and et al. main thrust is Hindi sentiment analysis, introducing an algorithm to enhance sentiment classification. It attains an 80.21% accuracy, with 82.89% for positive and 76.59% for negative reviews. The paper

underscores the importance of Hindi sentiment analysis and hints at future extensions to boost accuracy.

In their paper [10], Rasika Wankhede and Prof. A.N.Thakare focus is on sentiment analysis, particularly in movie reviews from the "Times of India" database. It assesses sentiment using the Random Forest classifier, achieving a remarkable 90% accuracy. The research underscores the significance of sentiment analysis in online communication and delves into alternative classifiers for improved outcomes.

2.3 Comparative Assessment and Overview

The primary goal of the current research work is to reflect various user viewpoints on the Bangla song. The suggested method has also made an effort to identify the song-related actions that have the poorest impacts over time.

2.4 Challenges

Engaging in sentiment analysis of Bengali song comments presents a range of complex challenges. These encompass the limited availability of high-quality annotated data for training models, the intricacies of the Bengali language with its nuanced expressions, the scarcity of dedicated natural language processing resources tailored for Bengali, and the inherent difficulty in deciphering subtle emotional nuances and idiomatic expressions. Moreover, cultural factors loom large, as sentiments often draw deeply from cultural references and historical contexts. These challenges are further compounded by the domain-specific nature of music-related terminology and references in song comments, underscoring the importance of developing specialized models capable of accommodating these intricacies while being sensitive to cross-cultural considerations. Surmounting these obstacles necessitates innovative strategies in data collection, preprocessing, model development, and assessment, as well as the incorporation of cultural expertise to ensure the precise analysis of sentiments within Bengali song comments.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Proposed Methodology in Diagram

This section defines the many Bengali language categories. Here, a pair of separate parameters are taken into consider. The first is "Class," and the second is "Opinions." The suggested approach is primarily concerned with collecting user comments on Bangla songs that are available globally, analyzing user data, and identifying the appropriate types of datasets. This critical process encompasses the sequential steps as shown in Figure 1.

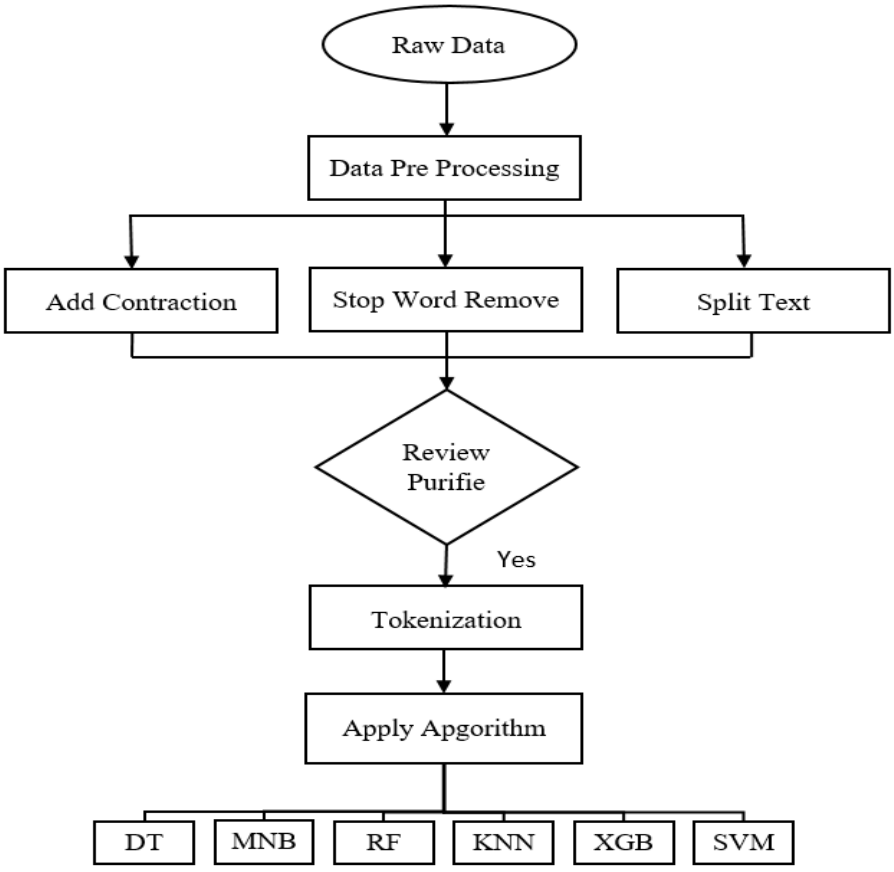


Figure 3.1: Proposed Methodology

3.2 Input Materials

The fundamental input for this study is a thorough dataset of Bengali song comments, which should include comments from diverse sources, genres, and historical periods, ensuring that it aligns with the research goals in terms of diversity and relevance. Additionally, access to linguistic resources like Bengali dictionaries, lists of stop words, and sentiment lexicons is crucial for the preprocessing of data and extracting relevant features. These resources are essential for the development of precise sentiment analysis models that are customized to the unique characteristics of the Bengali language.

3.3 Data Preprocessing

Data pre-processing stands as a pivotal phase within the realm of machine learning systems dedicated to Natural Language Processing (NLP).

STEP_1: A contraction should first be inserted to the text. Various contractions exist, they are ‘ শিক্ষ.’, ‘ছা.’, ‘ডা.’, ‘কৃষ.’ etc. This will be replaced with their full from like ‘শিক্ষক’, ‘ছাত্র/ছাত্রী’, ‘ডাক্তার’, ‘কৃষক’.

STEP_2: The data will then be manually checked to see if it is ordered, if there are any typographical errors, and if anything, else, it will be manually replaced.

STEP_3: After that, the data will be cleaned by deleting the compressed text or paragraph. Excessive symbolic characters, punctuation problems, and components such as [‘*’, ‘৩১’, ‘ ‘, ‘|’, etc.

STEP_4: Then, several types of Bengali stop words, such as 'সুতরাাং', cause issues when evaluating the data. The suggested dataset will be free of any stop words.

STEP_5: Finally, the necessary data will be acquired, and the sentiment-based outcome will be displayed.

3.6 Train and Test Data

In the context of data for training and testing, I collect information about them using an epoch, which represents all the training datasets combined.

The training period involves the following steps:

- Read the dataset and assign it as the training set.
- Preprocess the dataset to prepare it for further analysis.
- Implement a KNN model to process and analyze the data.

During the testing period, the following actions are performed:

- Read the dataset and designate it as the testing set.
- Include both the training and testing datasets for analysis.
- Perform classification to identify the valid Comments.

CHAPTER 4

EXPERIMENTAL DETAILS

4.1 Information about the Dataset

The amount of data required by each machine learning method is rather minimal. The ‘datasets' quality and size have an impact on the outcomes. Due to the large number of social media users, it might be challenging to gather enough data for a certain domain. For these reasons, just the necessary information should be gathered from the viewpoints of different social media users. Because of this, the suggested dataset is limited to 2224 text documents and two columns. A section or text form is one, while classification is another. On the other hand, it denotes the user's "Opinion" of Bangla songs in the form of a paragraph or text that could be either positive, neutral, or negative.

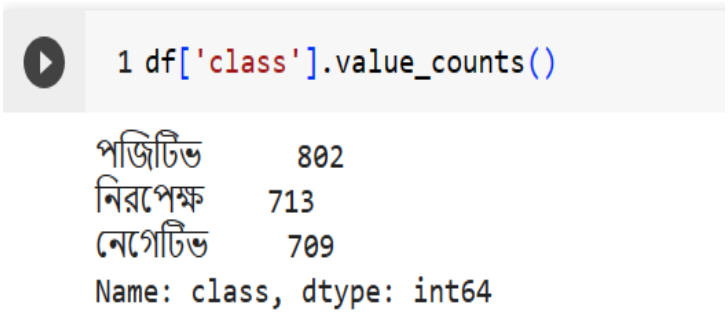


Figure 4.1: Data Set Count

4.2 Essential Tools and Equipment

Hardware and Software:

- Web IDE: Google Colab
- Hard Disk
- High Speed Internet Connection

Advanced Libraries:

- Python
- Pandas
- NumPy
- Matplotlib

Required Tools:

- Windows 10
- Dataset- Raw Data
- Algorithm- K-Nearest Neighbor (KNN), Decision Tree, Random Forest, Multinomial Naive Bayes, XGBoost, and Support Vector Machines.

4.3 Data Tokenization

Throughout the data preprocessing, the text will be tokenized. Tokenization is the process of converting speech into acceptable words. It is also known as 1 hot encoding. Nowadays, count vectorizers are a popular method for converting material into numerical values. Because the same dimensional data is used to estimate the dictionary, and if the data shape coordinates with the same vocabulary term, it will be taken into account. Finally, while it is evident that we will raise 1s or leave 0s in total for this term, we did not follow through.

4.4 KNN Algorithm

K-Nearest Neighbors (KNN) is a simple but effective supervised machine learning algorithm used for classification and regression. It relies on the proximity of data points in a feature space to make predictions. In classification, it assigns the most common class among the k nearest neighbors, while in regression, it calculates an average or other statistic from their values. The choice of 'k' is crucial and requires optimization through methods like cross-validation. KNN works well for non-linear and multi-class problems but can be sensitive to the choice of distance metric and may struggle with high-dimensional data.

4.5 KNN Architecture

KNN algorithm departs from the traditional architectures seen in neural networks or decision trees, instead embracing a straightforward and intuitive principle rooted in proximity. Utilizing KNN necessitates a dataset comprising labeled examples, each mapped within a feature space. When confronted with the task of predicting or classifying a new, unlabeled data point, KNN identifies the 'k' closest data points from the training dataset, employing a designated distance metric such as the Euclidean distance. Subsequently, the labels or values of these 'k' neighbors are employed to ascertain the prediction or classification for the new data point. Essentially, KNN's approach is devoid of intricate structures and operates on a case-by-case basis, relying on the local density of data points in the feature space to inform its predictions or classifications.

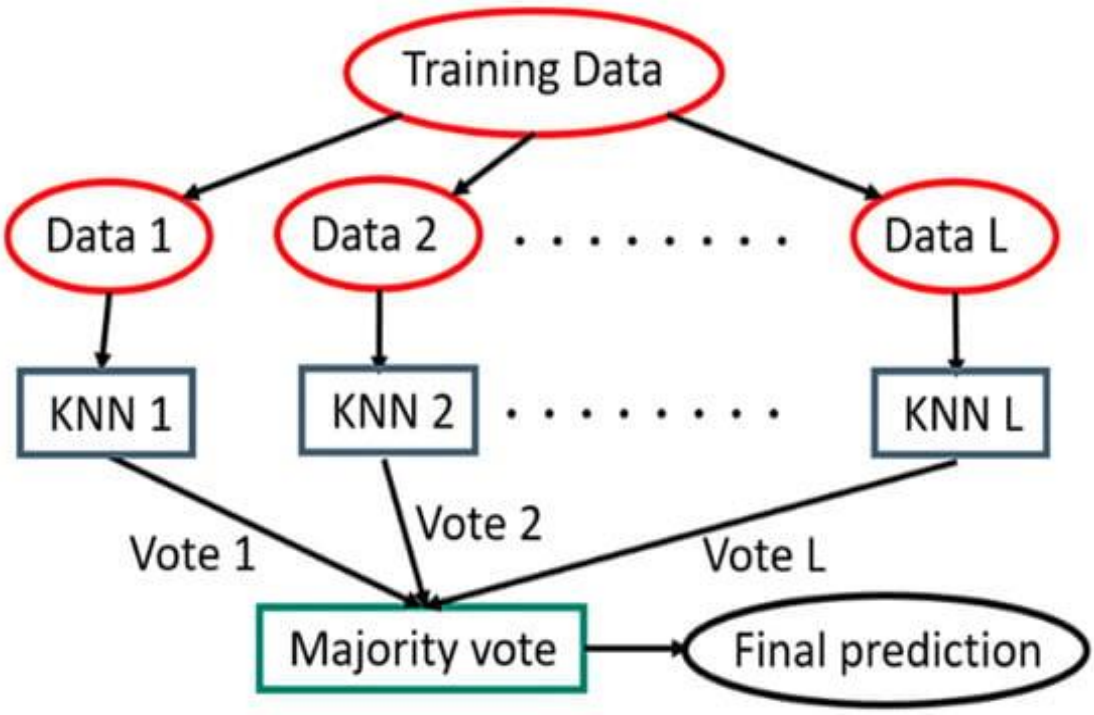


Figure 4.6: KNN Architecture

CHAPTER 5

RESULT AND DISCUSSION

5.1 Result Analysis:

The collected data are divided into three categories, first one is the neutral, then negative, and last one is positive. The proposed data is put to the test, and the performance of the system is assessed in terms of accuracy, precision, recall, F1-score, and support. To train the dataset, six classification approaches are used. The outcomes of each machine learning classification model are included in Fig. 3. From the accuracy graph, the accuracy of Decision Tree (DT) is 57.4%, Multinomial Naïve Bayes (MNB) is 52.02% and Random Forest (RF) classifier is 63.68%. The accuracy of K Nearest Neighbor (KNN), Support Vector Machine (SVM)and Extreme Gradient Boosting (Xgboost) classifiers are 54.71%, 60.09% and 61.43% respectively.

5.2 Accuracy Analysis

Additional metrics for assessing accuracy in the model include precision, recall, and the F1 score. Precision and recall measure the TP and TN in the model. The F1 score, commonly referred to as the F1 measure, combines recall and precision into a single metric and gives each equal weight.

TABLE 5.2.1: ACCURACY MATRICES

Name	Equation
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$
Precision	$\frac{TP}{TP + FP}$
Recall	$\frac{TP}{TP + FN}$

5.3 Model Accuracy

For measuring accuracy, Fig. 2: shows in comparison to other methods, RF accuracy has reached its highest level. With the exception of MNB, which achieved a somewhat lower accuracy, DT, KNN, SVM, and XGB also came close to the most significant accuracy.

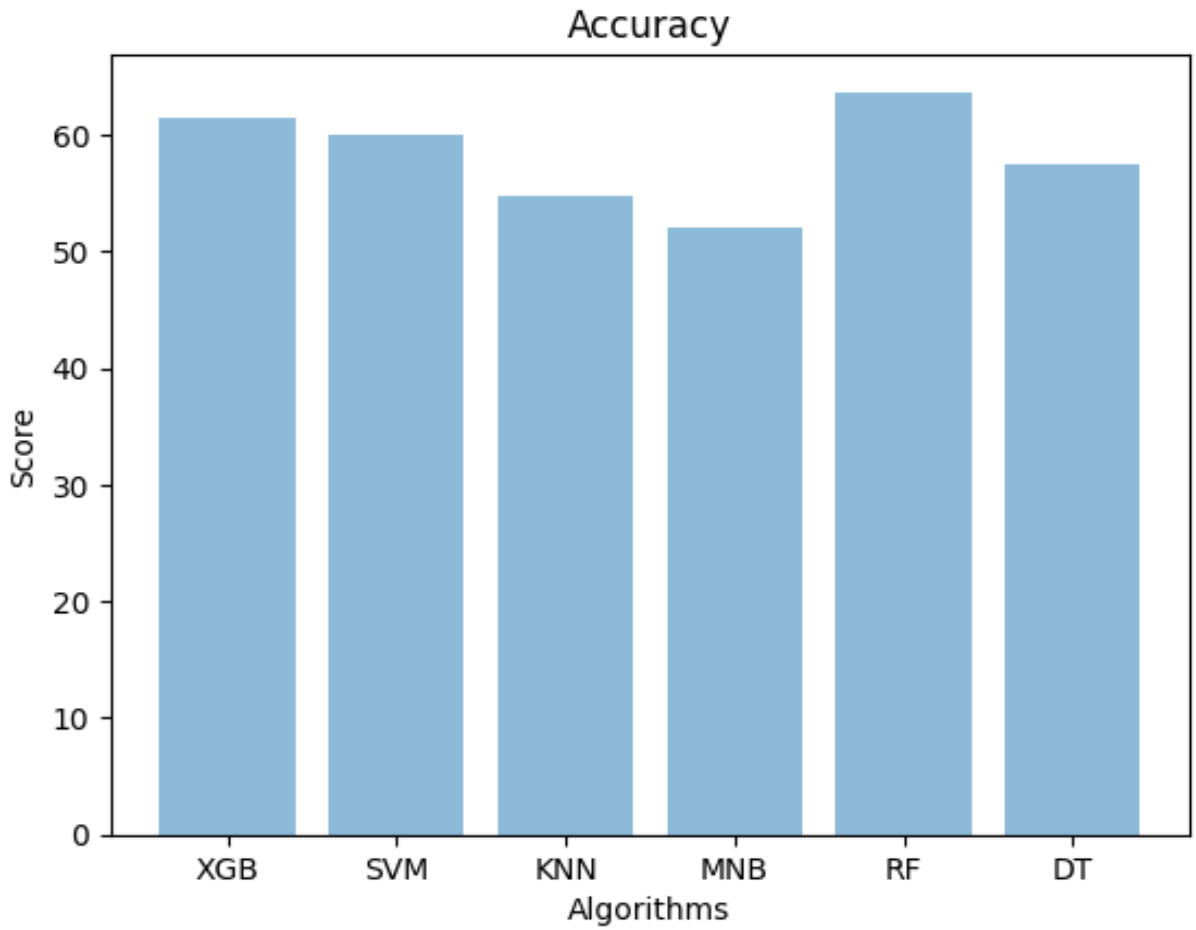


Figure 5.3: Model Accuracy

5.4 Classification Reports

TABLE 5.4.1: CLASSIFICATION REPORT OF XGB

Algoritm	Class	precision	recall	f1-score	support
XGB	Positive	0.49	0.60	0.54	70
	Neutral	0.62	0.52	0.57	73
	Negative	0.74	0.71	0.73	80
accuracy				0.61	223
macro avg		0.62	0.61	0.61	223
weighted avg		0.62	0.61	0.62	223

TABLE 5.4.2: CLASSIFICATION REPORT OF SVM

Algoritm	Class	precision	recall	f1-score	support
SVM	Positive	0.48	0.57	0.52	70
	Neutral	0.56	0.55	0.55	73
	Negative	0.79	0.68	0.73	80
accuracy				0.60	223
macro avg		0.61	0.60	0.60	223
weighted avg		0.62	0.60	0.61	223

TABLE 5.4.3: CLASSIFICATION REPORT OF KNN

Algoritm	Class	precision	recall	f1-score	support
KNN	Positive	0.44	0.44	0.44	70
	Neutral	0.54	0.60	0.57	73
	Negative	0.65	0.59	0.62	80
accuracy				0.55	223
macro avg		0.55	0.54	0.54	223
weighted avg		0.55	0.55	0.55	223

TABLE 5.4.4: CLASSIFICATION REPORT MNB

Algoritm	Class	precision	recall	f1-score	support
MNB	Positive	0.57	0.56	0.56	70
	Neutral	0.58	0.49	0.53	73
	Negative	0.38	0.51	0.44	80
accuracy				0.52	223
macro avg		0.51	0.52	0.51	223
weighted avg		0.53	0.52	0.52	223

TABLE 5.4.5: CLASSIFICATION REPORT OF RF

Algoritm	Class	precision	recall	f1-score	support
RF	Positive	0.56	0.53	0.54	70
	Neutral	0.58	0.64	0.61	73
	Negative	0.76	0.72	0.74	80
accuracy				0.64	223
macro avg		0.63	0.63	0.63	223
weighted avg		0.64	0.64	0.64	223

TABLE 5.4.6: CLASSIFICATION REPORT OF DT

Algoritm	Class	precision	recall	f1-score	support
DT	Positive	0.60	0.49	0.54	70
	Neutral	0.50	0.59	0.54	73
	Negative	0.64	0.64	0.64	80
accuracy				0.57	223
macro avg		0.58	0.57	0.57	223
weighted avg		0.58	0.57	0.57	223

TABLE 5.4.7: PREDICITION FOR POSSITIVE SENTENCE

Raw Data	গানটা যত শুনি ততই ভালো লাগে।
Actual Sentence Type	পজিটিভ
Actual Predicition Sentence Type	পজিটিভ
Preprocess Data	গানটা যত শুনি ততই ভালো লাগে
Applied Algorithm	Predicted Value
XGB	পজিটিভ
SVM	পজিটিভ
KNN	পজিটিভ
MNB	পজিটিভ
RF	পজিটিভ
DT	পজিটিভ

TABLE 5.4.8: PREDICITION FOR NEUTRAL SENTENCE

Raw Data	গানটি সম্পর্কে আমার মতামত নেই।
Actual Sentence Type	নিরপেক্ষ
Actual Predicition Sentence Type	নিরপেক্ষ
Preprocess Data	গানটি সম্পর্কে আমার মতামত নেই
Applied Algorithm	Predicted Value
XGB	নিরপেক্ষ
SVM	নিরপেক্ষ
KNN	নিরপেক্ষ
MNB	নিরপেক্ষ
RF	নিরপেক্ষ
DT	নিরপেক্ষ

TABLE 5.4.9: PREDICITION FOR NEGATIVE SENTENCE

Raw Data	এক কথায় বাজে গান।
Actual Sentence Type	নেগেটিভ
Actual Predicition Sentence Type	নেগেটিভ
Preprocess Data	এক কথায় বাজে গান
Applied Algorithm	Predicted Value
XGB	নেগেটিভ
SVM	নেগেটিভ
KNN	নেগেটিভ
MNB	নেগেটিভ
RF	নেগেটিভ
DT	নেগেটিভ

CHAPTER 6
COMPARATIVE STUDY AND ADVANTAGES

6.1 Comparative Study

The objective is to evaluate and compare different sentiment analysis models, encompassing both conventional machine learning and deep learning methods, to gauge their efficiency in accurately recognizing sentiments conveyed in comments about Bengali songs. Through a systematic assessment of these models based on criteria such as accuracy, precision, recall, F1-score, and computational efficiency, this research aims to pinpoint the most resilient and culturally attuned approach for comprehending audience emotions in Bengali music comments. This endeavor not only provides valuable insights for the music industry, cultural comprehension, and the wider domain of natural language processing in the Bengali language but also serves as a means to identify the most effective method for this specific context."

TABLE 6.1: COMPARATIVE STUDY

Existing Work	This Work
The existing model primarily focuses on identifying errors or mistakes in previous studies.	Model emphasizes on improving precision and accuracy.
Most of the research conducted so far has been based on the less powerful ANN algorithm, as compared to KNN.	Model is based on the more powerful KNN algorithm, which surpasses the capabilities of ANN.
The accuracy achieved by the current model ranges from 50% to 60%.	Have achieved an accuracy of 63% with this model.

6.2 Advantages

The primary advantage of conducting for sentiment analysis of Bangla song comments within a machine learning framework is the ability to pinpoint the most effective sentiment analysis model or approach tailored to the unique linguistic and cultural nuances of the Bangla language. This approach allows for a systematic assessment of different techniques, including traditional and deep learning models, leading to the selection of the best-performing model. By identifying the model that excels in accurately classifying sentiments in Bangla song comments, this research not only provides valuable insights into audience emotions and cultural context but also offers practical applications in the music industry, marketing, and social sciences, ultimately contributing to improved cross-cultural understanding and informed decision-making.

CHAPTER 7

FUTURE SCOPE, LIMITATION AND CONCLUSION

7.1 Future Scope

The future prospects for research in the field of sentiment analysis of Bengali song comments offer exciting possibilities for further exploration and progress. Upcoming studies may explore the development of hybrid models that combine the strengths of traditional machine learning and deep learning approaches, with the goal of enhancing both accuracy and computational efficiency. Furthermore, research could expand its horizons to encompass sentiment analysis across a wide array of musical genres, facilitating a deeper understanding of how sentiments vary across different musical styles and themes. Additionally, there is potential for investigating the evolution of sentiment over time in Bengali song comments, which could yield valuable insights into changing cultural trends and preferences. Collaborative efforts with the music industry and marketing sectors may yield practical applications, such as automated feedback analysis for artists and personalized music recommendations. Ultimately, future research in this field has the potential to contribute to a greater appreciation of Bengali music culture and the advancement of natural language processing techniques for languages that have received less attention.

7.2 Limitation

One significant challenge when conducting sentiment analysis on comments about Bengali songs is the difficulty of acquiring a sufficiently large and diverse dataset that accurately reflects the various genres, time periods, and cultural backgrounds within the Bengali music landscape. The availability of annotated data for training and evaluating models may be limited, potentially introducing dataset bias and hindering model generalization. Moreover, ensuring that comments are culturally relevant and contextually understood across different regions and dialects within the Bengali-speaking world can be quite intricate, as interpretations of sentiments

can vary widely. Additionally, although machine learning models can offer valuable insights, they may struggle to capture the full spectrum of emotional subtleties and cultural references found in song comments, underscoring the ongoing need for manual verification and cultural expertise in the analysis process.

7.3 Conclusion

The number of who enjoy Bangla songs has risen dramatically. We collected data from many audiences in order to determine the true state of Bangla Songs and the opinions of various people. six different classification algorithms are used to predict and analyze the given data. To achieve the best results, the acquired data is divided into three categories: positive, negative, and neutral. The Random Forest algorithm produced the best accuracy (63.68%) among six machine learning-based classification algorithms, whereas Xgboos provided the lowest accuracy (54.71%). Other models' accuracy in the spectrum of greatest and lowest levels. Despite the fact that we successfully constructed a model for predicting and interpreting data, several obstacles remain. The biggest difficulty arises during the data collection phase. Furthermore, most individuals are uninterested in giving their thoughts about Bangla songs and music. Some people are afraid to express themselves. People have also raised other types of queries while refusing to share their opinions. Another issue is that, while people express their opinions, they are not valid. The proposed study effort was effectively performed and the intended result was attained by overcoming the aforementioned hurdles.

REFERENCES

[1] M. H. Imam Bijoy, M. Hasan, A. N. Tusher, M. M. Rahman, M. J. Mia and M. Rabbani, "An Automated Approach for Bangla Sentence Classification Using Supervised Algorithms," 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2021, pp. 1-6, doi: 10.1109/ICCCNT51525.2021.9579940.

[2] Khan, M.R.H., Afroz, U.S., Masum, A.K.M., Abujar, S. and Hossain, S.A., 2020, July. Sentiment analysis from bengali depression dataset using machine learning. In 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-5). IEEE.

[3] Ardianto, Rian, et al. "Sentiment analysis on E-sports for education curriculum using naive Bayes and support vector machine." Jurnal Ilmu Komputer dan Informasi 13.2 (2020): 109-122.

[4] Hassan, Rakibul, and Md Rabiul Islam. "Impact of sentiment analysis in fake online review detection." 2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD). IEEE, 2021.

[5] Tariqul Islam, M., Tusher, A.N. (2022). Automatic Detection of Grape, Potato and Strawberry Leaf Diseases Using CNN and Image Processing. In: Nanda, P., Verma, V.K., Srivastava, S., Gupta, R.K., Mazumdar, A.P. (eds) Data Engineering for Smart Systems. Lecture Notes in Networks and Systems, vol 238. Springer, Singapore. https://doi.org/10.1007/978-981-16-2641-8_20

[6] Dhawan, Sanjeev, Kulvinder Singh, and Priyanka Chauhan. "Sentiment analysis of Twitter data in online social network." 2019 5th International Conference on Signal Processing, Computing and Control (ISPCC). IEEE, 2019.

[7] Choudhary, M. and Choudhary, P.K., 2018, December. Sentiment analysis of text reviewing algorithm using data mining. In 2018 International Conference on Smart Systems and Inventive Technology (ICSSIT) (pp. 532-538). IEEE.

[8] Tanna, D., Dudhane, M., Sardar, A., Deshpande, K. and Deshmukh, N., 2020, May. Sentiment analysis on social media for emotion classification. In 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS) (pp. 911-915). IEEE.

[9] Tusher, A.N., Islam, M.T., Sammy, M.S.R., Hasna, S.A., Chakraborty, N.R. (2022). Automatic Recognition of Plant Leaf Diseases Using Deep Learning (Multilayer CNN) and Image Processing. In: Chen, J.IZ., Tavares, J.M.R.S., Shi, F. (eds) Third International Conference on Image Processing and Capsule Networks. ICIPCN 2022. Lecture Notes in Networks and Systems, vol 514. Springer, Cham. https://doi.org/10.1007/978-3-031-12413-6_11

[10] Kim, E.J., Namkoong, K., Ku, T. and Kim, S.J., 2008. The relationship between online game addiction and aggression, self-control and narcissistic personality traits. European psychiatry, 23(3), pp.212-218.

[11] Ko, Y., & Seo, J. (2000). Automatic text categorization by unsupervised learning. In COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics.

[12] Ortigosa A, Martín JM, Carro RM. Sentiment analysis in Facebook and its application to e-learning. Computers in human behavior. 2014 Feb 1;31:527-41.

[13] Hodson, Hal. "Twitter hashtags predict rising tension in Egypt." (2013): 22.

[14] Talukder, M.A.I., Abujar, S., Masum, A.K.M., Faisal, F. and Hossain, S.A., 2019, July. Bengali abstractive text summarization using sequence to sequence RNNs. In 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-5). IEEE.

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

[16] Noble, W.S., 2006. What is a support vector machine?. Nature biotechnology, 24(12), pp.1565-1567.

[17] Rabbani, M., Alam, K.M.R. and Islam, M., 2014, May. A new verb based approach for English to Bangla machine translation. In 2014 International Conference on Informatics, Electronics & Vision (ICIEV) (pp. 1-6). IEEE.

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

APPENDICES

```
1 import itertools
2 import os
3 import numpy as np
4 %matplotlib inline
5 import matplotlib.pyplot as plt
6 import numpy as np
7 import pandas as pd
```

```
[ ]  1 df = pd.read_excel("/content/drive/MyDrive/Dataset.xlsx")
```

```
[ ]  1 len(df)
```

```
2224
```

```
[ ]  1 df.head(10)
```

	opinion	class
0	সুন্দর গান।	পজিটিভ
1	এই গানের জনপ্রিয়তা পরিমাপ করার জন্য ওই কয়েকট...	পজিটিভ
2	৪৯ মিলিওন এর অপেক্ষায় গানটা।	পজিটিভ
3	গানটাতে দেওরা মানে কী কেউ বলবেন?	নিরপেক্ষ
4	গানটা শোনার পর এলাকার সকল ভাবি আর বৌদির হাত ধর...	নেগেটিভ
5	আসলেই অসাধারন গান।	পজিটিভ
6	বাজে গান ।	নেগেটিভ
7	বাঙালি হয়ে গর্বিত গানটা শুনে।	পজিটিভ
8	বাংলা গানের সুদিন আসতেছে।	পজিটিভ
9	এই গানটি আগে শোনেছি এতোটা ভালো লাগেনি, কোক স্ট...	নেগেটিভ

```
[ ]  1 df.length.describe()
```

```
count     2224.000000
mean        12.026529
std         18.084938
min          1.000000
25%          5.000000
50%          8.000000
75%         13.000000
max        375.000000
Name: length, dtype: float64
```

```
  1 df.isna().sum()
```

```
opinion    0
class      0
length     0
dtype: int64
```

```
  1 df['class'].value_counts()
```

```
পজিটিভ      802
নিরপেক্ষ    713
নেগেটিভ     709
Name: class, dtype: int64
```

```
[ ]  1 from sklearn.model_selection import train_test_split, cross_val_score
```

```
[ ]  1 x_train, x_test, y1_train, y1_test = train_test_split(x, y1, test_size = 0.10, random_state=42)
```

```
[ ]  1 print("Total Data: "+str(len(x)))
     2 print("Train Size: "+str(len(x_train)))
     3 print("Test Size: "+str(len(x_test)))
```

```
Total Data: 2224
Train Size: 2001
Test Size: 223
```

```
Confusion Matrix for Multinomial Naive Bayes:
[[50 21 19]
 [24 42 20]
 [14  9 24]]
Score: 52.02
Classification Report:
              precision    recall  f1-score   support

           0       0.57      0.56      0.56        90
           1       0.58      0.49      0.53        86
           2       0.38      0.51      0.44        47

    accuracy                           0.52       223
   macro avg       0.51      0.52      0.51       223
weighted avg       0.53      0.52      0.52       223
```

```
[ ]
Confusion Matrix for Decision Tree:
[[33 23 14]
 [17 40 16]
 [10 20 50]]
Score: 55.16
Classification Report:
              precision    recall  f1-score   support

           0       0.55      0.47      0.51        70
           1       0.48      0.55      0.51        73
           2       0.62      0.62      0.62        80

    accuracy                           0.55       223
   macro avg       0.55      0.55      0.55       223
weighted avg       0.55      0.55      0.55       223
```

```
]  1 acc1=round(accuracy_score(y1_test,predmnb)*100,2)
   2 print("Accuracy:",acc1,"%")
```

```
Accuracy: 52.02 %
```

```
[ ]  1 acc2=round(accuracy_score(y1_test,preddt)*100,2)
     2 print("Accuracy:",acc2,"%")
```

```
Accuracy: 55.16 %
```

```
Confusion Matrix for Random Forest Classifier:
[[37 23 10]
 [18 47  8]
 [11 11 58]]
Score: 63.68
Classification Report:
              precision    recall  f1-score   support

           0       0.56      0.53      0.54        70
           1       0.58      0.64      0.61        73
           2       0.76      0.72      0.74        80

    accuracy                           0.64       223
   macro avg       0.63      0.63      0.63       223
weighted avg       0.64      0.64      0.64       223
```

```
]
Confusion Matrix for Support Vector Machines:
[[40 20 10]
 [29 40  4]
 [14 12 54]]
Score: 60.09
Classification Report:
              precision    recall  f1-score   support

           0       0.48      0.57      0.52        70
           1       0.56      0.55      0.55        73
           2       0.79      0.68      0.73        80

    accuracy                           0.60       223
   macro avg       0.61      0.60      0.60       223
weighted avg       0.62      0.60      0.61       223
```

```
]  1 acc3=round(accuracy_score(y1_test,predrmfr)*100,2)
   2 print("Accuracy:",acc3,"%")
```

```
Accuracy: 63.68 %
```

```
]  1 acc4=round(accuracy_score(y1_test,predsvm)*100,2)
   2 print("Accuracy:",acc4,"%")
```

```
Accuracy: 60.09 %
```

```
1 import matplotlib.pyplot as plt; plt.rcdefaults()
2 import numpy as np
3 import matplotlib.pyplot as plt
4
5 objects = ('XGB', 'SVM', 'KNN', 'MNB', 'RF', 'DT')
6 y_pos = np.arange(len(objects))
7 performance = [acc6,acc4,acc5,acc1,acc3,acc2]
8
9 plt.bar(y_pos, performance, align='center', alpha=0.5)
10 plt.xticks(y_pos, objects)
11 plt.ylabel('Score')
12 plt.title('Accuracy')
13 plt.xlabel('Algorithms')
14 plt.show()
```

PLAGIARISM REPORT

Sentiment Analysis of Bangla Song Comments Review: A Machine Learning Approach

ORIGINALITY REPORT

16%	12%	12%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.researchgate.net Internet Source	6%
2	Abdur Nur Tusher, Saiful Islam, Md. Tariqul Islam, Mst. Sakira Rezowana Sammy, Md. Sadekur Rahman, Md. Shibli Sadik. "User Perspective Bangla Sentiment Analysis for Online Gaming Addiction using Machine Learning", 2022 Sixth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2022 Publication	5%
3	dspace.daffodilvarsity.edu.bd:8080 Internet Source	1%
4	Submitted to Leeds Beckett University Student Paper	1%
5	Ton Duc Thang University Publication	1%
6	www.ijcnis.org Internet Source	1%

THANK YOU!