Emotion Analysis on Bengali Song Lyrics

Ashiqul Islam	**170204070**
Alam Khan	**170204084**
Mehedi Hasan	**170204096**

Project Report

Course ID: CSE 4214

Course Name: Pattern Recognition Lab

Semester: Spring 2021

Department of Computer Science and Engineering

Ahsanullah University of Science and Technology

Dhaka, Bangladesh

March 2022

Emotion Analysis on Bengali Song Lyrics

Submitted by

Ashiqul Islam	**170204070**
Alam Khan	**170204084**
Mehedi Hasan	**170204096**

Submitted To

Faisal Muhammad Shah, Associate Professor

Md. Tanvir Rouf Shawon, Lecturer

Department of Computer Science and Engineering

Ahsanullah University of Science and Technology

Department of Computer Science and Engineering

Ahsanullah University of Science and Technology

Dhaka, Bangladesh

March 2022

ABSTRACT

Music influences its audiences. It creates emotions- makes us happy, sad, or calms us down. Different cultures can have very different approaches to songs, lyrics, and melody. Music and songs represent the culture of a country. Bangla folk song, Bhatiali song, Robingra sangeet, Nazrul Songeet etc are represents our culture world wide. So, song is important in daily part of our life. Here the purpose of our project is to analyze the Bengali song lyrics and find the emotion of the lyrics. We applied six machine learning classifying algorithms like Multinomial Naive Bayes , K-Nearest Neighbours, Support Vector Machine, Logistic Regression, Random Forest, Decision Tree Classifier and one boosting classifier XGBoost. We have seen that Support Vector Machine(SVM) model perform better than all along with 73% accuracy with tri-gram and word stemming.

Contents

List of Figures

Chapter 1

Introduction

Songs are a way to express our emotion and harmony through vocal or instrumental sounds or both. We human beings are singing a song or hearing songs when we are happy or sad. Music and songs control our emotions and make our minds cheerful and joyous and reduce loneliness. Songs are meant to be the way of expressing emotion. Different cultures can have very different approaches to songs, lyrics, and melody. Songs usually have a meter or beat. Whether we sing or speak the lyrics, we can feel a pattern or pulse in the way the words move the song forward. Here lyrics play a major role to make a song cheerful or emotional to the audience. In this project, we will analyze emotion from Bengali songs lyrics. Here we will analyze songs lyrics and detect the lyrics as sad songs or not sad song, etc.

1.1 Motivation

Music influences its audiences. It creates emotions- makes us happy, sad, or calms us down. But one of the most-known aspects is the role music played in setting the tempo for the War of Liberation of Bangladesh. When we are depressed or sad music reduces it calms us down. Again listening to music is a benefit for our health also. It contributes to reducing pain and favors the production of endorphins. Lyrics is the key factor to making a song great. Here in our project, we will analyze the Bengali song's lyrics and analysis their emotion of it. This will help us to hear music according to our moods and emotions.

1.2 Uniqueness

In this project, we create a unique dataset. Our dataset has 1000 song lyrics and label the lyrics as sad or not sad. We collect these lyrics from different websites which lyrics are available publicly.

Chapter 2

Literature Reviews

Emotion or sentiment analysis of songs is a topic which is not that explored for low resource languages as such. And when it comes for bengali emotion analysis on song lyrics there resources is not available enough. Though a not so new field of research, most of the works have been done for the highly resourced languages only. Chen and Tang [1] proposed a method based on computational analysis of the lingual part of song lyrics. They constructed a composite emotion point matrix for each song which can then be used to further classify songs based on its inherent emotion and make recommendation accordingly through extracting and combining the term frequency and inverse document frequency (tf-idf) from song lyrics.

Though in emotion analysis on song there are indeed some work done in Bengali. However, Textual Lyrics Based Emotion Analysis of Bengali Songs by Devjyoti Nath et al [2]. They proposed methods to classify the Bengali songs into two classes - positive emotions and negative emotions. They used CNN with Bag-of-words , Linear Kernel SVM classifier and tf-idf for unigrams.

Chapter 3

Data Collection & Preprocessing

3.1 Data Collection

Our project dataset was collected from various open source Website of bengali song lyrics which are publicly available. After collecting 1000 lyrics we manually labeled the dataset as Sad = 1 and not sad = 0. In our dataset there are 567 not sad song and 433 sad song.

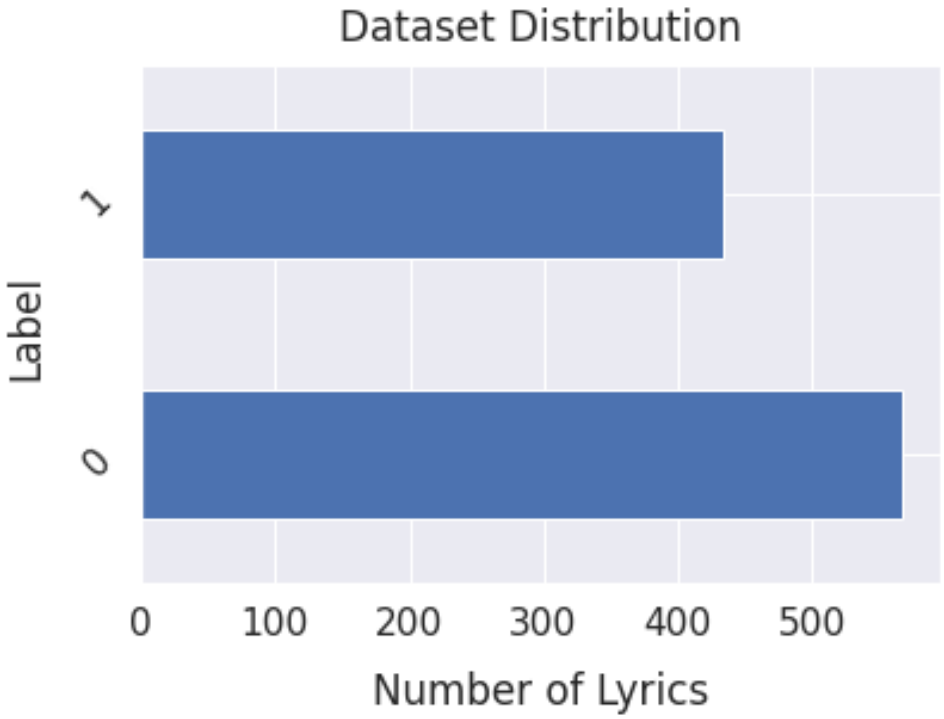

Figure 3.1: Data Distribution

3.2 Data Preprocessing

In preprocessing part we remove white space, comment, quotation symbol or other unwanted character from the dataset.
Then we remove stopwords which are available online to remove irrelevent words from the dataset.
We also use word stemming to get the root words.
In feature extraction part we use count vectorizer, tfidf transformer, bigram and trigram model.

	lyrics	label	cleaned	cleaned_stopwords	cleaned_stem
0	জনস্রোত By Warfaze\nসবাই বলে, আর তুমিও বলো\nআর...	0	সবাই বলে আর তুমিও বলো আর তুমি কি বলো আর তুম...	সবাই তুমিও বলো বলো বলো বলো পথ জনস্রোত অবরোধ পথ...	সবা বলে আর তুমি বলো আর তুমি কি বলো আর তুমি কি ...
1	প্রজন্ম-২০১২ By Warfaze\nনা শুনি তোমার কানে\nন...	0	না শুনি তোমার কানে না তাকাই তোমার পানে না দেখ...	শুনি কানে তাকাই পানে দেখি চোখে চোখ রাঙিয়ে তাক...	না শুনি তোম কানে না তাকা তোম পানে না দেখি তোম ...
2	পূর্ণতা By Warfaze\nসেদিন ভোরে, বুকের গভীরে\nশ...	1	সেদিন ভোরে বুকের গভীরে শুনেছি জমে থাকা নীল ব...	সেদিন ভোরে বুকের গভীরে শুনেছি জমে নীল বেদনার ড...	সেদিন ভোরে বুক গভীরে শুনে জমে থাকা নীল বেদন ডা...
3	রূপকথা By Warfaze\nশক্তি দাও বিধাতা, অনন্তকাল ...	1	শক্তি দাও বিধাতা অনন্তকাল ধরে জ্বলছে হৃদয় হ...	শক্তি দাও বিধাতা অনন্তকাল জ্বলছে হৃদয় হয়তো ব...	শক্তি দাও বিধাতা অনন্তকাল ধরে জ্বল হৃদয় হয় ক...
4	না By Warfaze\nআর চার দেয়ালে কেন একা ডুবে থাক...	0	আর চার দেয়ালে কেন একা ডুবে থাকা এই বর্তমানকে...	চার দেয়ালে একা ডুবে বর্তমানকে দূরে ঠেলে অতীতে...	আর চার দেয়ালে কেন একা ডুবে থাকা এই বর্তমান দূ...
...	...	...	...	...	...
995	আগুন – অসুর\nএকমুঠো ছাই উড়লো কোথায়\nপুড়ছে ক...	0	একমুঠো ছাই উড়লো কোথায় পুড়ছে কে আজ বলে দে আ...	একমুঠো ছাই উড়লো কোথায় পুড়ছে দে আমায় আগুন ম...	একমুঠো ছাই উড়লো কোথায় পুড় কে আজ বলে দে আমায়...

Figure 3.2: Pre Processing the dataset

Chapter 4

Methodology

For our project we used six machine learning classifying algorithms like Multinomial Naive Bayes , K-Nearest Neighbours, Support Vector Machine, Logistic Regression, Random Forest, Decision Tree Classifier. Then we use one boosting classifier XGBoost in our project.
For this we preprocessed as described in chapter 3, the data and split the data for training and testing the model. For training we used 70% of the data and rest of the 30% for testing purpose as that is ideal. After training the models, we evaluated the models with the test data based on Accuracy, Precision, Recall and F1 scores.

Dataset Distribution:

Set Name	Size
Full	1000
Training	700
Test	300

Figure 4.1: Dataset Distribution

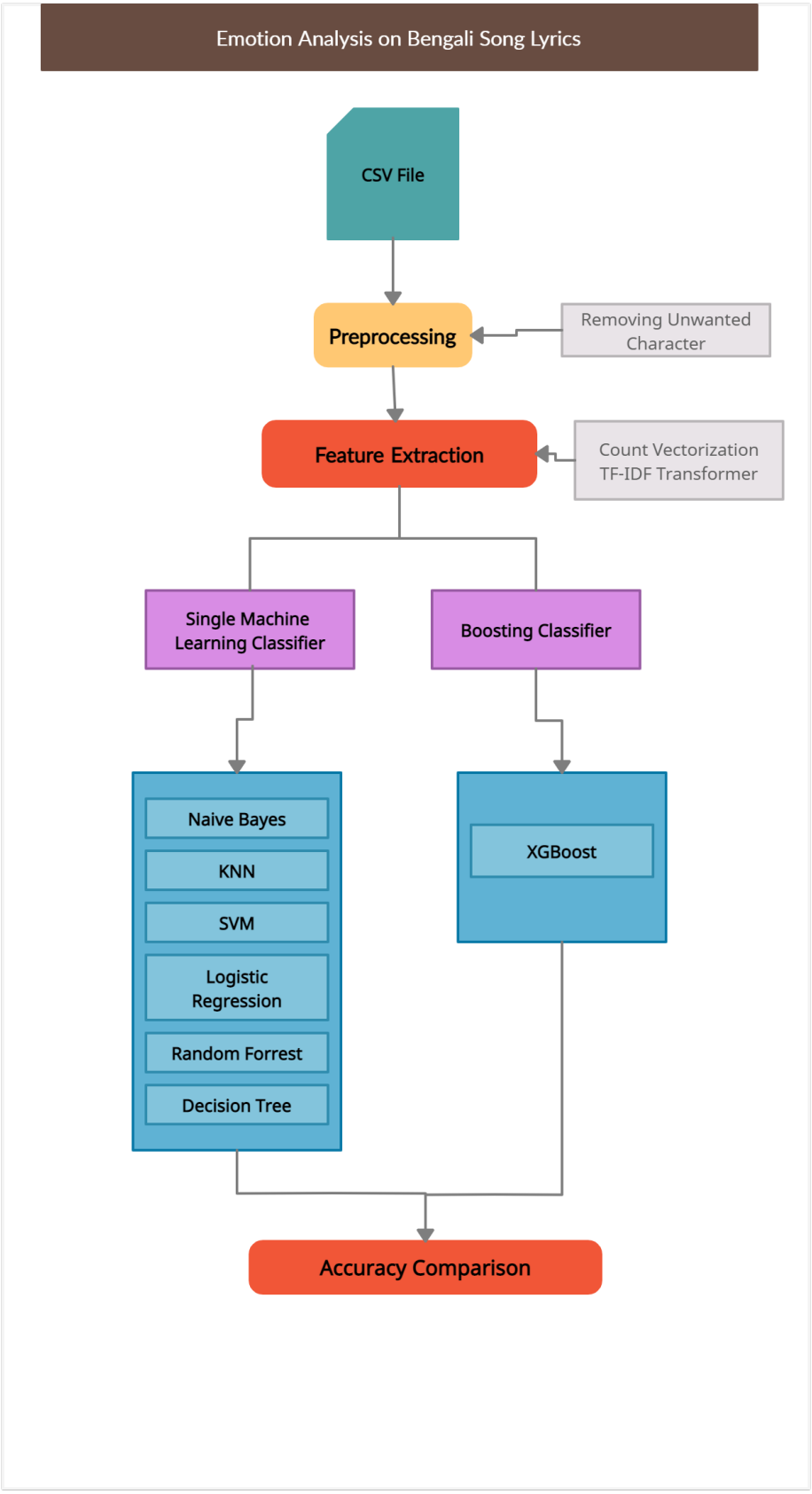

Figure 4.2: Methodology

Chapter 5

Experiments and Results

5.1 Experiment

For this project, we experimented with various machine learning models and fed them data once by Stemming and by Removing stop words using tfidf vectorizer feature extraction model with ngrams, keeping the other factors constant.

5.1.1 Experiment 1

In experiment 1 we work with TfIdf Vectorizer. Then the process data is sent to the machine learning models. The table shows the comparision of the ML models.

Metrics	NB	KNN	SVM	LR	RM	XGBoost	DT
Accuracy	0.68	0.63	0.69	0.65	0.64	0.63	0.52
Precison	0.68	0.62	0.69	0.65	0.64	0.62	0.54
Recall	0.68	0.63	0.69	0.65	0.64	0.63	0.52
F1 Score	0.65	0.61	0.69	0.64	0.64	0.62	0.53

Then we remove stopwords and apply the same model,keeping the other factors constant.

Metrics	NB	KNN	SVM	LR	RM	XGBoost	DT
Accuracy	0.71	0.61	0.72	0.72	0.62	0.62	0.59
Precison	0.71	0.59	0.72	0.72	0.61	0.61	0.60
Recall	0.71	0.61	0.72	0.72	0.62	0.62	0.59
F1 Score	0.69	0.59	0.72	0.71	0.60	0.61	0.60

We also use word stemming to get the root word and check the results.

Metrics	NB	KNN	SVM	LR	RM	XGBoost	DT
Accuracy	0.67	0.62	0.66	0.68	0.59	0.63	0.57
Precison	0.68	0.61	0.66	0.68	0.58	0.63	0.58
Recall	0.67	0.62	0.66	0.68	0.59	0.63	0.57
F1 Score	0.65	0.60	0.66	0.67	0.58	0.63	0.57

5.1.2 Experiment 2

In experiment 2 we work with TfIdf Vectorizer with bigram model. Then the process data is sent to the machine learning models. The table shows the comparision of the ML models.

Metrics	NB	KNN	SVM	LR	RM	XGBoost	DT
Accuracy	0.64	0.59	0.70	0.73	0.64	0.66	0.57
Precison	0.70	0.57	0.70	0.73	0.64	0.66	0.58
Recall	0.64	0.59	0.70	0.73	0.64	0.66	0.57
F1 Score	0.56	0.57	0.70	0.72	0.63	0.66	0.57

Then we remove stopwords and apply the same model,keeping the other factors constant.

Metrics	NB	KNN	SVM	LR	RM	XGBoost	DT
Accuracy	0.66	0.58	0.68	0.69	0.62	0.62	0.57
Precison	0.70	0.56	0.67	0.69	0.61	0.61	0.58
Recall	0.66	0.58	0.68	0.69	0.62	0.62	0.57
F1 Score	0.61	0.56	0.67	0.69	0.61	0.62	0.57

We also use word stemming to get the root word and check the results.

Metrics	NB	KNN	SVM	LR	RM	XGBoost	DT
Accuracy	0.67	0.60	0.72	0.73	0.66	0.64	0.58
Precison	0.72	0.59	0.72	0.72	0.66	0.65	0.58
Recall	0.67	0.60	0.72	0.73	0.66	0.65	0.58
F1 Score	0.61	0.58	0.72	0.72	0.65	0.65	0.58

5.1.3 Experiment 3

In experiment 3, we work with TfIdf Vectorizer with trigram model. Then the process data is sent to the machine learning models. The table shows the comparision of the ML models.

Metrics	NB	KNN	SVM	LR	RM	XGBoost	DT
Accuracy	0.68	0.63	0.72	0.70	0.60	0.68	0.57
Precison	0.70	0.62	0.72	0.70	0.59	0.68	0.57
Recall	0.68	0.63	0.72	0.70	0.60	0.68	0.57
F1 Score	0.64	0.61	0.72	0.69	0.59	0.68	0.57

Here, we remove stopwords and apply the same model,keeping the other factors constant.

Metrics	NB	KNN	SVM	LR	RM	XGBoost	DT
Accuracy	0.70	0.61	0.70	0.72	0.67	0.67	0.57
Precison	0.73	0.60	0.70	0.72	0.67	0.66	0.55
Recall	0.70	0.61	0.70	0.72	0.67	0.67	0.55
F1 Score	0.68	0.59	0.70	0.72	0.66	0.66	0.55

Again, we use word stemming to get the root word and check the results.

Metrics	NB	KNN	SVM	LR	RM	XGBoost	DT
Accuracy	0.67	0.62	0.73	0.71	0.62	0.67	0.57
Precison	0.72	0.61	0.73	0.71	0.61	0.67	0.57
Recall	0.67	0.62	0.73	0.71	0.62	0.67	0.57
F1 Score	0.61	0.60	0.73	0.70	0.61	0.67	0.57

5.2 Results

From the above illustrated data, we can conclude that Support Vector Machine (SVM) performed the best with an accuracy of 73%, Precision 73%, Recall 73%, F1 Score 73% on Experiment 3 i.e. with word stemming and trigram feature extraction. The second best result is shown by Logistic Regression.

SVM is an algorithm that determines the best decision boundary between vectors that belong to a given category and vectors that do not belong to it. It can easily search a classification hyperplane in feature space and for the generalisation capability of the classifier as well. That is a main reason why the SVM can achieves very good results.

Logistic regression is easy to implement, interpret, and very efficient to train. It is very fast at classifying unknown records. It performs well when the dataset is linearly separable. It can interpret model coefficients as indicators of feature importance.

Stemming is the process of reducing inflection in words to their root forms such as mapping a group of words to the same stem even if the stem itself is not a valid word in the Language. It extracts the root words from a document, that's why we get good results here.

Chapter 6

Future Work and Conclusion

6.1 Future Work

In this project we applied six machine learning classifying algorithms like Multinomial Naive Bayes , K-Nearest Neighbours, Support Vector Machine, Logistic Regression, Random Forest, Decision Tree Classifier. We have 1000 dataset of Bengali song lyrics. In future we will web scrapping lyrics from different website and make a large dataset suitable for applying Neural Network. As we know sequence is important for analyzing lyrics so Recurrent Neural Network model might be more suitable here.

6.2 Conclusion

As future works, we intent to emply other feature extraction methods, such as the different approaches of Stemming. We are also intend to apply a variety of preprocessing techniques to boost the accuracy and see the differences between various models. we will utilize a different enriched dataset and intend to employ alternative feature extraction methods.

References

[1] V. X. Chen and T. Y. Tang, "Combining content and sentiment analysis on lyrics for a lightweight emotion-aware chinese song recommendation system," *2018 10th International Conference on Machine Learning and Computing*, pp. 85–89, 2018.

[2] S. K. S. A. G. S. P. Devjyoti Nath, Anirban Roy, "Textual lyrics based emotion analysis of bengali songs," *2020 International Conference on Data Mining Workshops (ICDMW)*, 16 February 2021.

Generated using Undegraduate Thesis LaTeX Template, Version 1.4. Department of Computer Science and Engineering, Ahsanullah University of Science and Technology, Dhaka, Bangladesh.

This project report was generated on Monday 14th March, 2022 at 6:21am.