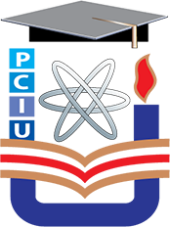
Sentiment Analysis on Bangla OTT Platform Content Using Machine Learning & Deep Learning

# Kutub Uddin

ID: CSE 01806766

# Joy Dey

ID: CSE 01806698



Department of Computer Science and Engineering Port City International University

Chattogram, Bangladesh

Sentiment Analysis on Bangla OTT Platform Content Using Machine Learning & Deep Learning

Submitted by

# Kutub Uddin

ID: CSE 01806766

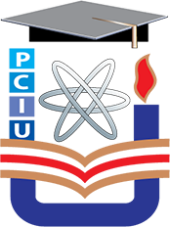
# Joy Dey

ID: CSE 01806698

Supervised by

## Sourav Adhikary

Lecturer



Submitted to the Department of Computer Science and Engineering,

Port City International University

for partial fulfilment of the requirements for the Degree of Bachelor Science in Computer Science and Engineering.

# Dedication

We dedicated this thesis to

our beloved parents who are our role models and source of inspiration.

#### Approval for Submission

This thesis titled “**Lexicon Based Emotion Extraction from Bengali Texts**” by **Amir Hosen**, Student ID: CSE 010 05670, Batch:10 Day has been approved for submission to the Department of Computer Science and Engineering, Port City International University, in partial fulfilment of the requirements for the Degree of Bachelor of science.

#### [MONISHA BISWAS](http://www.portcity.edu.bd/HomePage/SubPageDetailsInfo/23/Teacher/monisha-biswas)

Senior Lecturer

Department of Computer Science and Engineering Port City International University

7-14, Nikunja Housing Society, South Khulshi Chattogram, Bangladesh.

Email: [monisha.cse10@gmail.com](mailto:monisha.cse10@gmail.com) Telephone: 01881075178

#### Declaration

I humbly declare that this thesis entitled, “**Lexicon Based Emotion Extraction from Bengali Texts**” and this work is entirely my own for my undergraduate degree. All section in this thesis has been properly acknowledged.

#### Amir Hosen

ID: CSE 010 05670

Department of Computer Science and Engineering Port City International University

#### Acknowledgement

Alhamdulillah, first of all, I thank my creator who has given me the ability and patience to perform this task. Then from the bottom of my heart, I thank my esteemed supervisors **FARZINA AKTER** and [**MONISHA BISWAS**](http://www.portcity.edu.bd/HomePage/SubPageDetailsInfo/23/Teacher/monisha-biswas)who helped me complete this task with all kinds of support and advice. Special thanks to [**MONISHA**](http://www.portcity.edu.bd/HomePage/SubPageDetailsInfo/23/Teacher/monisha-biswas) ma’am who encouraged me to finish the job with timely instructions after the departure of **FARZINA** ma’am.

Finally, I would like to thank some of the teachers at this university and my friends who have supported and inspired me from the beginning.

#### Amir Hosen

ID: CSE 010 05670

Batch: CSE 10 Day

Department of Computer Science and Engineering Port City International University

#### Abstract

In today's age of the internet, people want to determine almost everything through machines. And the machine is now able to learn everything. In this age of the Internet, people are constantly sharing their emotions on various social media. Sentiment analysis is a term used by the machines to recognize their emotions, which is a part of Natural Language Processing (NLP). This system can bring out emotion. The purpose of this thesis is to bring out the emotions from the shared posts or comments of the people. In this system, I’ve predicted the emotion based on the emotional word. This system has 6 emotion classes (Like, Ha-ha, Wow, Sad, Angry, Love). Our system is a supervised model. In this system, I've used Random Forest (RF), K-Nearest Neighbors (KNN), Decision Tree (DT) as algorithms.

**Keywords:** Sentiment Analysis, Natural Language Processing, Emotional Word, Supervised Model, Random Forest (RF), K-Nearest Neighbors (KNN), Decision Tree (DT).

#### Table of Contents

[Dedication i](#_TOC_250038)

[Approval for Submission ii](#_TOC_250037)

[Declaration iii](#_TOC_250036)

[Acknowledgement iv](#_TOC_250035)

[Abstract v](#_TOC_250034)

[Table of Contents vi](#_TOC_250033)

[List of Tables ix](#_TOC_250032)

[List of Figures x](#_TOC_250031)

[Chapter 1 1](#_TOC_250030)

[Introduction 1](#_TOC_250029)

* 1. [Overview 1](#_TOC_250028)
  2. [Problem Statement 2](#_TOC_250027)
  3. [Motivation. 2](#_TOC_250026)
  4. [Objective 2](#_TOC_250025)

[Chapter 2 3](#_TOC_250024)

Literature Review 3

* 1. [Sentiment Analysis 3](#_TOC_250023)
  2. [Natural Language Processing 3](#_TOC_250022)
  3. Literature Review 4
  4. [Classifying Methodologies 6](#_TOC_250021)
     1. [Random Forest (RF) 6](#_TOC_250020)
     2. [K-Nearest Neighbors (KNN) 6](#_TOC_250019)
     3. [Decision Tree (DT) 7](#_TOC_250018)
  5. [Required Tools 7](#_TOC_250017)
     1. [Python 7](#_TOC_250016)
     2. [Pandas 7](#_TOC_250015)
     3. [NumPy 8](#_TOC_250014)
     4. [Matplotlib 8](#_TOC_250013)
     5. [Google Colab 8](#_TOC_250012)

Chapter 3 9

Methodology 9

* 1. Problem Description 9
  2. System Workflow 10
  3. Data Collection 11
  4. [Data Preprocessing 11](#_TOC_250011)
  5. [Feature Extraction… 14](#_TOC_250010)
     1. [Lexicon Based approach… 14](#_TOC_250009)
  6. [Classification Algorithm 17](#_TOC_250008)
     1. [Random Forest (RF) 17](#_TOC_250007)
     2. [K-Nearest Neighbors (KNN) 17](#_TOC_250006)
     3. [Decision Tree (DT) 17](#_TOC_250005)

Chapter 4 18

Result and Evaluation 18

* 1. Confusion Matrix 18
     1. accuracy 19
     2. Precision 19
  2. Recall 19
  3. F1-Score 19
  4. Sensitivity… 19
  5. Specificity 19
  6. Output 20
  7. Experimental Result 21
     1. [Random Forest Classifier… 21](#_TOC_250004)
     2. K-Nearest Neighbor… 22
     3. [Decision Tree Classifier… 23](#_TOC_250003)

Chapter 5 25

Conclusion and Future Work 25

* 1. [Conclusion… 25](#_TOC_250002)
  2. [Future Work 25](#_TOC_250001)

[References 26](#_TOC_250000)

**List of Tables**

Table 4.1 Precision, Recall, F1-Score of Random Forest Classifier 21

Table 4.2 Precision, Recall, F1-Score of K-Nearest Neighbor 22

Table 4.3 Precision, Recall, F1-Score of Decision Tree Classifier 23

Table 4.4 Table 4.4 Comparison of RF, KNN, DT 24

**List of Figures**

Figure 3.1 System Workflow 10

Figure 3.2 Data Preprocessing 10

Figure 3.3 Feature Extraction 10

Figure 3.4 Result 10

Figure 3.5 URL removal 11

Figure 3.6 Punctuation character removal 12

Figure 3.7 Emoticon removal 12

Figure 3.8 Stop Words removal 13

Figure 3.9 Tokenization 13

Figure 3.10 Lexicon Based approach 14

Figure 3.11 Example Lexicon Based 15

Figure 4.1 Confusion Matrix 18

Figure 4.2 Confusion Matrix of Random Forest Classifier 21

Figure 4.3 Confusion Matrix of K-Neighbors Classifier 22

Figure 4.4 Confusion Matrix of Decision Tree Classifier 23

## CHAPTER 1

**------------------**

**Introduction**

This chapter presents an overview, problem statement, motivation, objectives of this work.

#### Overview

Bengali is one of the most spoken languages in the world. Almost 100% of the people of Bangladesh speak Bengali. This language is also spoken in West Bengal and a few other states in India. The present age is the age of the internet. So, with the passage of time, everyone is leaning towards the internet. Currently the number of internet users in the world 4,648,228,067 people. The number of Bangladeshi internet users is 96,199,000. Day by day people are engaging themselves through social media and they constantly share their emotions through various posts or comments.

Nowadays people stay at home or with family but their focus is on mobile and the internet. They reveal the events of their every moment in front of others. They share their feelings and emotions on different social media platforms and respond to different emotions by calculating what other people's posts are pointing to. Nowadays people are very thoughtful about everything. They easily present their feelings to others and others respond by understanding what he means.

The current topic is that with the blessings of science, people now want everything through machines. They also want to know their reactions through the machine. They are making everything automatic. It can also be called an addiction.

Here I have created a model of Machine Learning with a term Natural Language Processing which will automatically extract emotions from Bangla text. I’m prepared our model using some classifying algorithms.

#### Problem Statement

The number of internet users in the world is increasing day by day and the competition to automate everything in the world has also started. Part of this is sentiment analysis through which various organizations bring out the emotions of their users through machines. At present, with the help of natural language processing, many researchers are processing different languages through machines. Has already created systems to express emotions from many languages of the world.

But the Bengali language lags behind all other languages in the world and the purpose of this thesis is to express feelings from the Bengali language through machine learning.

#### Motivation

Sentiment analysis is currently occupying a leading position in the field of research. In many cases, it uses are bringing benefits. It is helpful for getting results without wasting time and brain. Results are being obtained by applying this method in almost all languages and it is also being done in Bengali to express emotions with some methods.

#### Objective

* + - The purpose of this thesis is to extract emotions from Bengali sentences through a lexicon- based method of 6 categories (Ha-Ha, Wow, Love, Angry, Sad, Like).
    - lexicon-based emotion extraction is rare in Bengali sentences.

## CHAPTER 2

**-------------------------**

**Literature Review**

This chapter discusses previous research on sentiment analysis and natural language processing on Bengali text.

#### Sentiment Analysis

Sentiment analysis is contextual mining of text which identifies and extracts subjective information in the source material and helping a business to understand the social sentiment of their brand, product, or service while monitoring online conversations. However, analysis of social media streams is usually restricted to just basic sentiment and count-based metrics. This is akin to just scratching the surface and missing out on those high-value insights that are waiting to be discovered [10].

#### Natural Language Processing

Natural Language Processing, usually shortened as NLP, is a branch of artificial intelligence that deals with the interaction between computers and humans using the natural language. The ultimate objective of NLP is to read, decipher, understand, and make sense of the human languages in a valuable manner. Most NLP techniques rely on machine learning to derive meaning from human languages [11].

#### Related work

In [1], the authors of this paper tried to extract sentiment in three-class (Positive, Negative, Neutral) from Bengali text using a Character Level Supervised Recurrent Neural Network Approach. They collected total- 34271 data from Facebook using graph API where Positive-8271, Negative-14000, and Neutral-12000. By this model, they got 80% accuracy by using the LSTM technique.

By using Multinomial Naïve Bayes (MNB) and Support Vector Machines (SVM) in [2], the author’s tries to Detect Sentiment Polarity in Bengali Tweets. They got the highest accuracy 44.20% in MNV by (Unigram + bigram + SentiWordNet) and 45% in SVM by (Unigram

+SentiWordNet) then compare this accuracy with some top systems of SAIL 2015 contest.

In this paper [3], the author’s presented a model to extract positive and negative emotions. Here they applied some methods like 1) Parts of speech ratio. 2) Cosine similarity using TF-IDF. 3) Cosine similarity using custom TF-IDF. 4) Naïve Bayes model using Uni-gram & stammer. 5) Naïve Bayes model using Bi-gram stammer & normalizer. 6) Word Embedding with Hellinger PCA for generate good accuracy. They got better accuracy of 83.20% from the Naïve Bayes model using a Bi-gram stammer & normalizer.

In [4], the authors built a system that can predict emotion from text by using a machine learning approach. They focused on Ekman’s six basic emotions. Used TF-IDF for feature extraction and Multinomial Naïve Bayes, Support Vector Machine, Decision Trees, and K-Nearest Neighbors algorithms for classification. They got the highest accuracy of 64.08% from Multinomial Naïve Bayes. They used ISEAR (The International Survey on Emotion Antecedents and Reactions) dataset for this classification.

Here the authors used n-gram for mining sentiment from Bangla text using Support Vector Machine in [5], and approach this model for classifying in two-classes positive and negative. They collected 9500 comments from different social media sites. In this system, they applied linear SVM and non-linear SVM and found that linear SVM is better than non-linear SVM by giving 91.684% accuracy.

This authors in [6] used youtube API for collect Bangla, English, and Romanized Bangla youtube comments. Then tried to detect multilabel sentiment and emotions like 3 classes (positive, negative, neutral), 5 classes (very positive, positive, neutral, negative, very negative) and five category emotions (anger, disgust, fear, joy, sadness, and surprise). They implemented Continuous Bag of Words (CBOW) and Skip Gram (SG) for a feature and used LSTM, CNN, SVM, NB for classification. The highest accuracy for 3, 5 class, and five category emotions are 65.97%, 54.24%, and 59.23% respectively.

Using Naïve Bayes Classifier, the authors tried to detect Emotion from Bangla Text Corpus in [7]. They used features such as stemmer, parts-of-speech (POS) tagger, n-grams, term frequency- inverse document frequency (tf-idf). They did this job for three emotion classes (happy, sad, and angry). They collected 4200 comments from different Facebook groups and some public posts of popular bloggers. By building this model they got 78.6% accuracy.

In [8], the authors tried to build an automated system of sentiment analysis from Bangla text using Supervised Learning Techniques. They did this for 6 emotion classes (Happy, Sad, Tender, Angry, Excited, Scared). They created 7,500 Bangla sentences as a dataset and applied the Naïve Bayes Classification Algorithm and Topical Approach. They got above 90% accuracy from the Topical Approach.

A Lexicon Based Approach was proposed in [9], Here the authors analyses sentiment from Bangla text. Here they build this model for positive, negative classes. They collected 5200 sentences from some sites and social media (Facebook groups, Twitter, Blogs, online newspaper sites) where 2600 sentences are positive and 2600 sentences are negative. Here they proposed a method by creating a Bengali Sentiment Words Dictionary which contains more than 5100 words, Boost Word Check, Negation Check for their feature extraction. Finally, they got better accuracy 92% form their proposed method.

#### Classifying Methodologies

This section presents an introduction to the algorithms I used in my thesis work. I used Random Forest (RF), K-Nearest Neighbors (KNN), Decision Tree (DT) as algorithms for the experiment in my system.

#### Random Forest (RF)

Random forest could be a versatile, simple to use machine learning algorithmic program that produces, even while not hyper-parameter standardization, a good result most of the time. it's conjointly one among the foremost used algorithms, attributable to its simplicity and variety (it is often used for each classification and regression tasks) Random forest could be a supervised learning algorithmic program. The "forest" it builds, is Associate in the Nursing ensemble of decision trees, typically trained with the "bagging" technique the overall plan of the textile technique is that a mix of learning models will increase the general result Let's look into the random forest in classification since classification is typically consist building block of machine learning. In random forest, solely a random set of the options is taken into thought by the algorithmic program for ripping a node. we will even create trees additional random by to boot victimization random thresholds for every feature instead of finding out the simplest attainable thresholds sort of a traditional decision tree does.

#### K-Nearest Neighbors (KNN)

K-nearest neighbors (KNN) algorithm is a type of supervised Machine Learning algorithm which is mainly used for classification predictive problem. KNN algorithm uses feature similarity to predict the values of new data points which further means that the new data point will be assigned a value based on how closely it matches the points in the training set. KNN captures the idea of proximity with some mathematics such as calculating the distance between points on a graph. There are many ways of calculating distance, and one way might be preferable depending on the problem we are solving. The straight-line distance also called the Euclidean distance is a popular and familiar choice for calculating distance.

#### Decision Tree (DT)

Decision Tree (DT) is one of the classification techniques in data mining that uses branches method to depict each feasible result of decision making in each possible outcome. DT comprises three kinds of the node that frame an established a tree which a tree required to have "root node, internal node', and 'leaf". DT breaks down a set of data into smaller and smaller subsets while at the same time an associated decision tree is incrementally built. The root node is known as the initial attribute or the topmost decision node in a tree which corresponds to the best predictor for a tree to make decision making that has zero incoming and outgoing edges. While internal nodes have both incoming and outgoing edges at least one. Followed by a leaf node that has no outgoing edges represents a classification or decision.

#### Required Tools

#### Python

Python is AN understood, object-oriented programing language just like PERL, that has gained quality due to its clear syntax and readability. Python is claimed to be comparatively straightforward to find out and moveable, which means its statements are understood in a very variety of operative systems, as well as UNIX-based systems, Mac OS, MS-DOS, OS/2, and numerous versions of Microsoft Windows ninety-eight. Python was created by Guido van Rossum, a former resident of the Kingdom of The Netherlands, whose favorite comedy cluster at the time was Monty Python's Flying Circus. The ASCII text file is freely accessible and open for modification and utilization. Python encompasses an important variety of users

#### Pandas

Pandas could be a high-level knowledge manipulation tool developed by Wes McKinney. it's engineered on the NumPy package and its key organization is termed the Data Frame. Data Frames permit you to store and manipulate tabular knowledge in rows of observations and columns of variables

#### NumPy

NumPy is a python library used for operating with arrays. It conjointly has functions for working in the domain of algebra, Fourier rework, and matrices. NumPy was created in 2005 by Travis Oliphant. It is an open-source project and you'll be able to use it freely. NumPy stands for Numerical Python.

#### Matplotlib

Matplotlib is a comprehensive library for making static, animated, and interactive visualizations in Python.

#### Google Colab

Colab is a free Jupyter notebook environment that runs entirely in the cloud. Most importantly, it does not require setup. Colab supports many popular machine learning

libraries.

**CHAPTER 3**

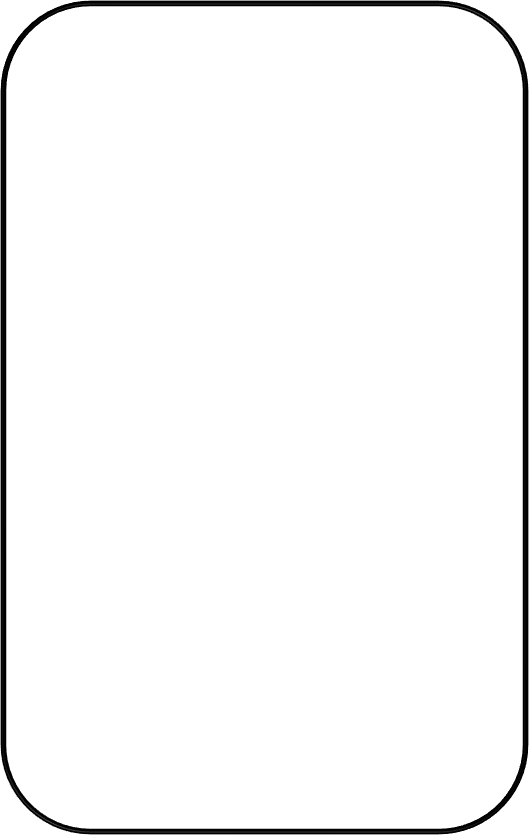
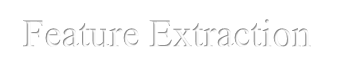
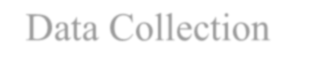
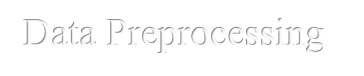
**--------------------**

**Methodology**

This section describes my thesis methodologies like problem description, system workflow, feature extraction, tools and technologies

**3.1 Problem Description**

In this research, I focus on extract emotion from Bengali text data. I did this by applying a lexicon- based approach. I applied a lexicon dictionary for feature extraction. I also applied other machine learning techniques like Random Forest (RF), K-Nearest Neighbors (KNN), Decision Tree (DT) for classify. I create a lexicon dictionary that contains six classes (Like, Ha-ha, Wow, Love, Sad, Angry) and each individual class contain related lexicon. Here I created a dataset from Facebook groups and label them based on my classes. Firstly, I have processed the data for cleaning. Then I extract features from that clean data and fit them into classifiers. I divide my dataset into 80% for training and 20% for testing after that I calculate accuracy, precision, recall, etc.



**3.2 System Workflow**

URL removal

Data Collection

Punctuation & Special

Character removal

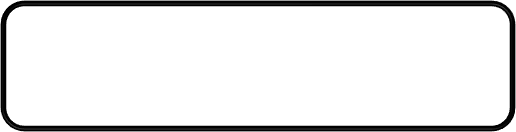
Data Preprocessing

Emoticon removal

Feature Extraction

Different word Removal

Tokenization

Figure 3.2 Data Preprocessing



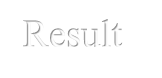
Train



Test

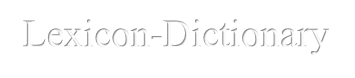
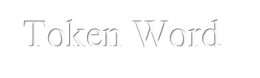


Classification Algorithm



Result

Figure 3.1 System Workflow



Token Word

Lexicon-Dictionary

Figure 3.3 Feature Extraction



Like

Love

Wow

Sad

Angry

Ha-ha

Figure 3.4 Result

**3.3 Data Collection**

It is difficult to get the Bengali dataset as there are fewer works on Bangla language. So, I have

created a dataset of 1288 Bangla sentences by myself and leveled the data in 6 classes. It contains Ha-ha-102 sentences, Angry-169 sentences, Love-160 sentences, Wow-255 sentences, Sad-310 sentences, Like- 292 sentences. The data in this dataset is taken from posts and comments on different groups on Facebook.

#### Data Preprocessing

In addition to original data, I have some unnecessary data in this dataset that I do not need. In the beginning, I have to remove that unnecessary thing. The processes will be discussed below.

* + - **URL removal**: Some comments or post contain URL which is not bear any emotion. That’s why I need to remove the URL.

আমার ফেইসবুক আইডি ড িংক [www.facebook.com/amirhosen](http://www.facebook.com/amirhosen)।আমার ফেইজ টি বুস্ট করর 10$ চাই

😖😖।follow this Id



আমার ফেইসবুক আইডি ড িংক ।আমার ফেইজ টি বুস্ট করর 10$ চাই 😖😖 ।follow this Id

Figure 3.5 URL removal

* + - **Punctuation character removal:** This dataset contains some unnecessary punctuation. That’s why I remove punctuation.



আমার ফেইসবুক আইডি ড িংক ।আমার ফেইজ টি বুস্ট করর 10$ চাই 😖😖।follow this Id

আমার ফেইসবুক আইডি ড িংক আমার ফেইজ টি বুস্ট করর 10 চাই 😖😖 follow this Id

Figure 3.6 Punctuation character removal

* + - **Emoticon removal:** Some comments have emoticon but, in my system, I focus only on Bangla text that’s why I strip emoticon.



আমার ফেইসবুক আইডি ড িংক আমার ফেইজ টি বুস্ট করর 10 চাই 😖😖 follow this Id

আমার ফেইসবুক আইডি ড িংক আমার ফেইজ টি বুস্ট করর 10 চাই follow this Id

Figure 3.7 Emoticon removal

* + - **Different words removal:** Different words like non-Bengali words, non-Bengali numbers, Bengali numbers. In this section, I removed all those Different words.

আমার ফেইসবুক আইডি ডगিংক আমার ফেইজ P বুস্ট করর 10 চাইग follow this Id



আমার ফেইসবুক আইডি ডगিংক আমার ফেইজ P বুস্ট করর চাইग

Figure 3.8 Stop Words removal

* + - **Tokenization:** Tokenization is the method of breaking large corpus to single corpus /word or single line corpus. Here I did tokenize the data for a single corpus/word.



আমার ফেইসবুক আইডি ডगিংক আমার ফেইজ P বুস্ট করর চাইग

“আমার”,“রেইসবুক”,“আইডি”,“ডगিংক”, “আমার”,“রেইজ”,“P”,“বুস্ট”,“করর”,“চাইग”

Figure 3.9 Tokenization

#### Feature Extraction

The feature is to represent text data as numerical. We know that machines do not understand anything except 0 and 1. Similarly, machine learning cannot understand without numerical data. But in my Dataset contains only string data. So, I have to move this string data in a numerical format and how to do it will be discussed in this section.

#### 3.5.1 Lexicon Based approach

I am bringing out the features of this work through a Lexicon based model. I am matching the token data with my Lexicon dictionary. By doing this it will send a count in my code from the class, when they will match. Thus, after checking a whole comment or post in the lexicon dictionary, my system will generate a label based on the counts. I am getting and those labels will go to classifier later.



Token Word



Ha-ha

Love

Wow

Sad

Like

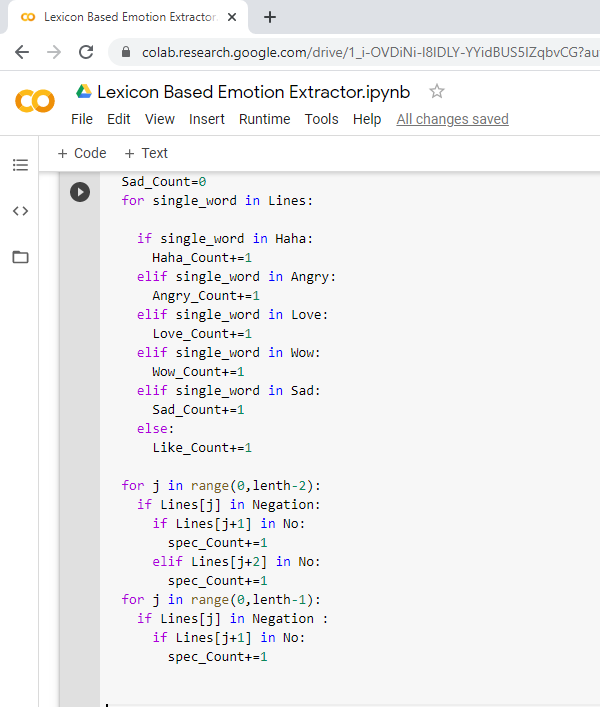
Angry



Label

Figure 3.10 Lexicon Based approach

**CODE SNAPSHOT- Feature Extraction**



**Example:**

If matched

“েু ग”, “P”, “সুন্দর”

Wow



Dictionary

If not matched

Figure 3.11 Example Lexicon Based

#### Classification Algorithm

##### Random Forest (RF)

Classifier=RandomForestClassifier() X\_train=X\_train.reshape(-1,1)

text\_clf=Pipeline([('Classifier',RandomForestClassifier())]) text\_clf.fit(X\_train,y\_train)

##### K-Nearest Neighbors (KNN)

Classifier=KNeighborsClassifier() X\_train=X\_train.reshape(-1,1)

text\_clf=Pipeline([('Classifier', KNeighborsClassifier())]) text\_clf.fit(X\_train,y\_train)

##### Decision Tree (DT)

Classifier=tree.DecisionTreeClassifier() X\_train=X\_train.reshape(-1,1)

text\_clf=Pipeline([('Classifier', tree.DecisionTreeClassifier())]) text\_clf.fit(X\_train,y\_train)

**CHAPTER 4**

**--------------------**

**Results and Evaluation**

This section describes my thesis outcomes like performance metrics, experimental results.

**4.1 Performance Metrics**

Metrics for Evaluating Machine Learning Algorithms. Different performance metrics are used to evaluate different Machine Learning Algorithms.

#### Confusion Matrix

The Confusion matrix is one of the most intuitive and easiest (unless of course, you are not confused) metrics used for finding the correctness and accuracy of the model.

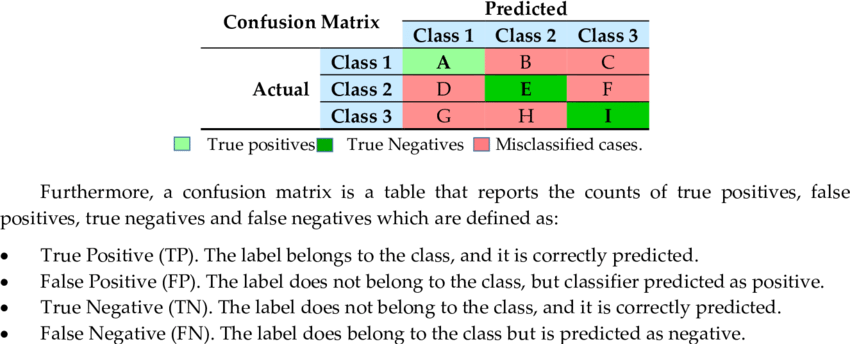


Figure 4.1 Confusion Matrix

#### Accuracy

Accuracy in classification problems is the number of correct predictions made by the model over all kind’s predictions made

Accuracy=(TP+TN)/(TP+FP+FN+TN)

#### Precision

It is the fraction of relevant instances among the retrieved instances.

#### Recall

Precision=TP/(TP+FP)

It is the fraction of the total amount of relevant instances that were actually retrieved

#### F1- Score

Recall=TP/(TP+FN)

The **F1 score** is a measure of a test's accuracy. It considers both the [precision](https://en.wikipedia.org/wiki/Precision_(information_retrieval)) *p* and the [recall](https://en.wikipedia.org/wiki/Recall_(information_retrieval)) *r* of the test to compute the score

F1 Score = 2 \* Precision \* Recall / (Precision + Recall)

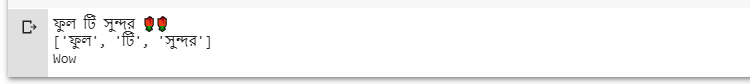
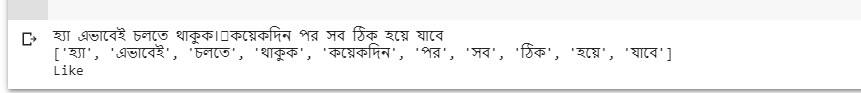
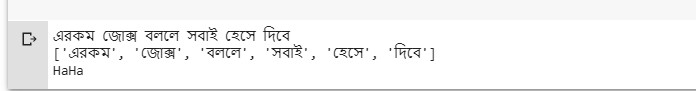
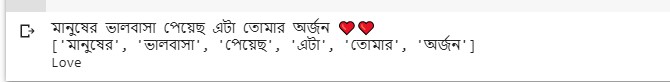
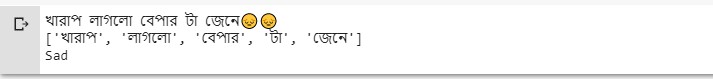
#### Sensitivity

* + 1. **Specificity**

Sensitivity=TP/(TP+FN)

Specificity=TN/(TN+FP)

#### Output



* 1. **Experimental Results**

#### Random Forest Classifier

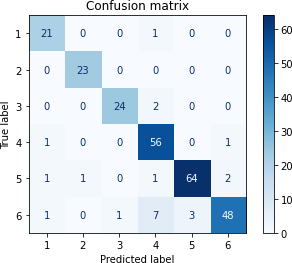


Figure 4.2 Confusion Matrix of Random Forest Classifier

precision recall f1-score support

1

2

3

4

5

6

0.88

0.96

0.96

0.84

0.96

0.94

0.95

1.00

0.92

0.97

0.93

0.80

0.91

0.98

0.94

0.90

0.94

0.86

22

23

26

58

69

60

Table 4.1 Precision, Recall, F1-Score of Random Forest Classifier

### K-Neighbors Classifier

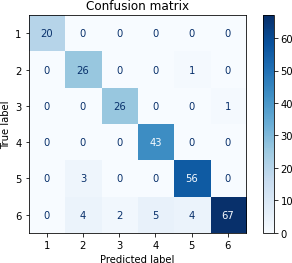


Figure 4.3 Confusion Matrix of K-Neighbors Classifier

Precision recall f1-score support

1

2

3

4

5

6

0.93

0.94

0.77

0.87

0.92

0.96

0.93

1.00

0.96

0.97

0.98

0.68

0.93

0.97

0.86

0.92

0.95

0.80

15

30

25

64

58

66

Table 4.2 Precision, Recall, F1-Score of K-Neighbors Classifier

### Decision Tree Classifier

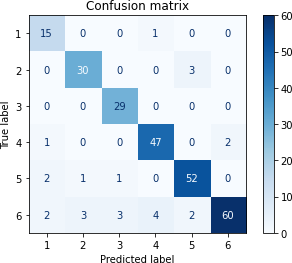


Figure 4.4 Confusion Matrix of Decision Tree Classifier

precision recall f1-score support

1

2

3

4

5

6

0.75

0.88

0.88

0.90

0.91

0.97

0.94

0.91

1.00

0.94

0.93

0.81

0.83

0.90

0.94

0.92

0.92

0.88

16

33

29

50

56

74

Table 4.3 Precision, Recall, F1-Score of Decision Tree Classifier

**Performance Comparison of different algorithm given below:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Algorithm | Accuracy | Precision | Recall | F1-Score |
| RF | 91.47% | 92% | 93% | 92% |
| KNN | 89.92% | 90% | 92% | 90% |
| DT | 90.31% | 88% | 92% | 90% |

Table 4.4 Comparison of RF, KNN, DT

**CHAPTER 5**

**--------------------**

**Conclusion and Future Work**

#### Conclusion

In this experiment, emotions are extracted from Bangla posts and comments in different social medias using ML classifiers (RF, KNN, DT) with lexicon-based feature extraction techniques. Here I have created a dataset of 1288 Bangla sentences where emotions are labeled as Like, Ha- ha, Wow, Love, Sad, Angry. Through this research work, 91.47% accuracy is achieved using the RF algorithm.

#### Future Work

This work can be extended in the following ways in future:

* + - Increase Emotion Class.
    - Increase Emotion Classes Data.
    - Adding more feature.

### ---------------------------------------------------------------------------------------

**References**

1. Mohammad Salman Haydar, Mustakim Al Helal, Syed Akhter Hossain. “**Sentiment Extraction from Bangla Text: A Character Level Supervised Recurrent Neural Network Approach.**” In [2018 International Conference on Computer, Communication, Chemical,](https://ieeexplore.ieee.org/xpl/conhome/8433310/proceeding) [Material and Electronic Engineering (IC4ME2),](https://ieeexplore.ieee.org/xpl/conhome/8433310/proceeding) IEEE,2018.
2. Kamal Sarkar1, Mandira Bhowmick2. “**Sentiment Polarity Detection in Bengali Tweets Using Multinomial Naïve Bayes and Support Vector Machines.**” In 2017 IEEE Calcutta Conference (CALCON), IEEE,2017.
3. Md. Al- Amin, Md. Saiful Islam, Shapan Das Uzzal. “**A Comprehensive Study on Sentiment of Bengali Text.**” In International Conference on Electrical, Computer and Communication Engineering (ECCE), February 16-18, 2017, Cox’s Bazar, Bangladesh.
4. 1Ahmad Fakhri Ab. Nasir, 1Eng Seok Nee, 1Chun Sern Choong, 1Ahmad Shahrizan Abdul Ghani, 1Anwar P. P. Abdul Majeed, 1Asrul Adam and 2Mhd. Furqan. “**Text-based emotion prediction system using machine learning approach.**” The 6th International Conference on Software Engineering & Computer Systems. IOP Conf. Series: Materials Science and Engineering769 (2020) 012022.
5. SM Abu Taher, Kazi Afsana Akhter, K.M. Azharul Hasan. “**N-gram Based Sentiment Mining for Bangla Text Using Support Vector Machine.**” In International Conference on Bangla Speech and Language Processing (ICBSLP), 21-22 September, 2018.
6. Nafis Irtiza Tripto, Mohammed Eunus Ali. “**Detecting Multilabel Sentiment and Emotions from Bangla YouTube Comments.**” In International Conference on Bangla Speech and Language Processing (ICBSLP), 21-22 September, 2018.
7. Sara Azmin, Kingshuk Dhar. “**Emotion Detection from Bangla Text Corpus Using Naïve Bayes Classifier.**” In 4th International Conference on Electrical Information and Communication Technology (EICT), 20-22 December 2019, Khulna, Bangladesh.
8. Rashedul Amin Tuhin, Bechitra Kumar Paul, Faria Nawrine, Mahbuba Akter, Amit Kumar Das. “**An Automated System of Sentiment Analysis from Bangla Text using Supervised Learning Techniques.**” In 2019 IEEE 4th International Conference on Computer and Communication Systems, IEEE, 2019.
9. Rajib Chandra Dey, Orvila Sarker. “**Sentiment Analysis on Bengali Text using Lexicon Based Approach.**” In 2019 22nd International Conference on Computer and Information Technology (ICCIT), 18-20 December, 2019.
10. **“Sentiment Analysis,”** towardsdatascience**,**

<https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17>

1. **“Natural Language Processing,”** becominghuman.ai,

<https://becominghuman.ai/a-simple-introduction-to-natural-language-processing-ea66a1747b32>