Bachelor of Science in Computer Science & Engineering

Sentiment Analysis of Bangla Texts Using Machine Learning Techniques

by

Md. Mashiur Rahaman Mamun

ID: 1504071

Department of Computer Science & Engineering

Chittagong University of Engineering & Technology (CUET)

Chattogram-4349, Bangladesh.

April, 2021

Sentiment Analysis of Bangla Texts Using Machine Learning Techniques

Submitted in partial fulfilment of the requirements for
Degree of Bachelor of Science
in Computer Science & Engineering

by

Md. Mashiur Rahaman Mamun
ID: 1504071

Supervised by

Dr. Mohammed Moshiul Hoque
Professor
Department of Computer Science & Engineering

Chittagong University of Engineering & Technology (CUET)
Chattogram-4349, Bangladesh.

The thesis titled '**Sentiment Analysis of Bangla Texts Using Machine Learning Techniques'** submitted by ID: 1504071, Session 2015-16 has been accepted as satisfactory in fulfilment of the requirement for the degree of Bachelor of Science in Computer Science & Engineering to be awarded by the Chittagong University of Engineering & Technology (CUET).

Board of Examiners

Supervisor

Dr. Mohammed Moshiul Hoque
Professor
Department of Computer Science & Engineering
Chittagong University of Engineering & Technology (CUET)

Head of Dept

Dr. Md. Mokammel Haque
Professor & Head
Department of Computer Science & Engineering
Chittagong University of Engineering & Technology (CUET)

External

Dr. Mohammad Samsul Arefin
Professor
Department of Computer Science & Engineering
Chittagong University of Engineering & Technology (CUET)

Declaration of Originality

This is to certify that I am the sole author of this thesis and that neither any part of this thesis nor the whole of the thesis has been submitted for a degree to any other institution.

I certify that, to the best of my knowledge, my thesis does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my thesis, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. I am also aware that if any infringement of anyone's copyright is found, whether intentional or otherwise, I may be subject to legal and disciplinary action determined by Dept. of CSE, CUET.

I hereby assign every rights in the copyright of this thesis work to Dept. of CSE, CUET, who shall be the owner of the copyright of this work and any reproduction or use in any form or by any means whatsoever is prohibited without the consent of Dept. of CSE, CUET.

Mashin

Signature of the candidate

Date: 24 May, 2021

Acknowledgements

This thesis would not have been possible without the help of a number of people. First and foremost, I would want to express my gratitude to my supervisor, Professor Dr. Mohammed Moshiul Hoque, for his unwavering support and assistance. His inspiration for doing research on the realm of Natural Language Processing, infinite patience, understanding and willingness to let me find my own wings, have all contributed to my humble development of this thesis and the trining dataset. I am gratful to him for his constant encouragement and his stimulating ideas.

I'd like to thank the members of my defense committee for their valuable advice and feedback. I am eternally grateful to Prof. Dr. Assaduzzaman, Head of the Department of Computer Science and Engineering at CUET, for his unwavering support during my undergraduate studies.

I'd also like thank Omar Sharif, Lecturer, Dept of CSE, CUET, for his invaluable help, time, and criticism. I am grateful to the CUET NLP Lab for their assistance during this process. I'm also appreciative to my friends for their assistance in validating the traing dataset.

Abstract

Owing to the widespread use of the Internet has resulted in a revolutionary way for people to share their feelings or sentiment on blogs, social media, e-commerce site and online platforms in recent years. Most of the feelings expressed on the online platforms are in textual forms (such as status, tweets, comments and reviews). Most of these textual expressions are unstructured, extremely hard and time-consuming to organize, manipulate, or efficient storage due to their messy forms. Textual sentiment analysis refers to the automatic process of assigning an expression or text to an appropriate polarity (positive, negative, and neutral). This paper proposes a machine learning-based framework to classify Bengali textual sentiment into two categories: positive and negative. Due to the unavailability of the Bengali sentiment corpus, this work also developed a dataset (called 'BSaD') containing 8,122 text expressions to perform the sentiment classification. This work investigates the performance of eight popular ML techniques (such as LR, NB, RF, KNN, SVM, MNB, AdaBoost and SGD) and an ensemble technique with TF-IDF and BoW features for sentiment classification task on the developed corpus. Experimental results show that tf-idf + ensemble approach with uni-gram + bi-gram + tri-gram combination outperformed the other classifier models achieving the highest accuracy of 82%.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

Sentiment analysis is closely connected to natural language processing and textual expression of human can be determined by this. Peoples opinion can be used to extract information about their interest on something. What people want, how they like it or which improvement needed can be known by sentiment analysis. Review based sentiment analysis can be helpful to provide important information to the commercial, marketing, strategies companies. Sentiment analysis from Bangla text will provide us information about the texts polarity and information on certain matters. This chapter contains some introductory information on sentiment analysis, motivation of the work, challenges of the work and our contribution to this work.

1.1.1 Problem Statement

Sentiment is defined to be peoples opinion towards a specific interaction [1] Sentiment Analysis or opinion mining can be defined as the process of understanding the opinion of people. Sentiment analysis, also known as opinion mining, is the systematic identification, extraction, quantification, and study of emotional states and subjective opinions using natural language processing and text analysis. Customer reviews and survey answers, internet and social media, and healthcare applications ranging from marketing to customer service to clinical medicine are all examples of where sentiment analysis is used.Several subtasks within opinion mining can be identified; all need tagging at the document / sentence / phrase / word level based on the opinion stated. One such subtask is focused on a particular opinion piece of text on a single problem or item in order to classify the viewpoint as falling below one of two opposing sentimental polarities or finding its location on a continuum between these two polarities. In general, sentiment analysis seeks to ascertain a speaker's or writer's attitude toward a topic, as well as their emotional reaction to a document, encounter, or event. The attitude could be a judgment or evaluation, or it may be an emotional expression. High resources language such as English, Japanese have made progress in analyzing while Bangla is lacking due to limited standard dataset and resources. We can see, resources like standard datasets are not available in Bangla language. Though Bangla being spoken widely, still it is a low resourced language. Approximately 280 million people speak it as a first language in Bangladesh, West Bangel and in some other parts of India [2]. Nowadays, the majority of Bangla speakers are interested in online activities. As a result, sentiment analysis is very much needed in order to understand consumer interests and their reviews based on it. However, analyzing this huge amount of data in a manual way is very much difficult and sometimes impossible. In these circumstances, sentiment analysis can be quite useful in determining people's interests and opinions. Online news websites provide timely and up-to-date information, as opposed to traditional newspapers, which need readers to wait until the next morning to read the latest international news. In other words, if one is seeking for speedy and up-to-date news, Internet news is a better option than a traditional newspaper. Bangladesh is a small nation, but it has a big number of newspapers, TV, online newspapers and magazines, news portals and blog pages. There are over 10,000 online media in our country [3].So, it is easily understandable how important role sentiment analysis can play in this area. Sentiment polarity classification or polarity classification refers to the binary classification problem of categorizing an opinionated material as conveying either an overall positive or overall negative viewpoint. Much research has been done on sentiment polarity classification in the context of reviews in the English language (e.g., Sentiment Polarity Classification in Twitter posts) [4]. While positive and negative opinions are frequently evaluative in this context (e.g., like vs. dislike), there are other issues where the understanding of positive and negative is subtly different.

Works on sentiment analysis have not been that successful in Bangla language because of lacking in standard resources and datasets. For this reason, it will be tough to collect datasets and other resources which is needed for this work. However, such works are hardly observed in Bangla language.

1.1.2 Objectives

The primary goal of this work is to create a sentiment analysis system for the Bangla language. The following are the main goals and potential outcomes of this thesis:

- To design and develop a sentiment analysis system that will determine the polarity of Bangla texts as positive or negative.
- To develop Bangla dataset containing 8,122 labeled sentiment documents.
- To develop a supervised machine learning model using LR, DT, RF, KNN, MNB, SVM, SGD and AdaBoost to categorize the sentiment from Bangla texts into positive or negative polarities.
- To develop an ensemble learning classifier to classify the sentiment polarity.
- To analyze the effectiveness of our proposed technique on different distributions of the developed dataset.
- To evaluate the system performance in real environment.

9

1.2 Framework/Design Overview

The general framework of the system consists of some components as shown in Figure 1.1 This general framework shows the simple working procedure of sentiment analysis from Bangla texts. Most challenging part is to create a system

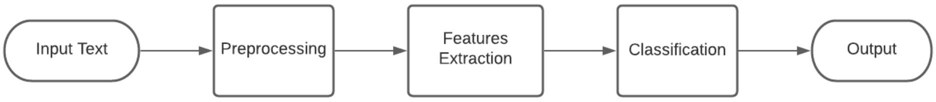


Figure 1.1: General Framework of Sentiment Analysis.

that can detect the sentiment from textual data and recognize it. At first the sentiment is recognized from the input which is a Bangla text. Then the category of the expression for that sentiment is displayed at the output interface.

1.3 Difficulties

Lacking of standard resources and datasets are the main obstacle of this work. This type of work hardly seen in Bangla language. So, it is tough to collect datasets and works through it. For the implementation of the system was to develop a dataset which can be used by our learning algorithm. We know that for any machine learning algorithm a well-furnished dataset a key. If the dataset is accurate then the accuracy of the output becomes perfect. Bangla is low resource language.We strive to solve this problem by gathering a vast volume of text from many Bangla sources. Another problem is labeling data into positive and negative. As we using supervised learning algorithm, we need to label the data. As using wrong data for training does not give correct output. As a result, we will obtain poor result from our system. The performance will decrease too. So, we have to collect data from various sites for training and testing purposes. Another challenge we had to face was lack of literature resources. There has been a very few works in this area so it was tough for us to get a proper understanding regarding this topic. There are limitations of linguistic resources also.

1.4 Applications

Nowadays sentiment analysis is becoming more popular. With the help of this tool we can do a lots of great things like Social networking monitoring, customer service management, and customer feedback analysis are all popular sentiment analysis applications.
Here are some of the most common real-world uses of sentiment analysis:

1. Social media monitoring
2. Customer support
3. Customer feedback

4. Market research and competitive research
5. Voice of employee
6. Voice of customer (VoC)
7. Product analysis
8. Brand monitoring and reputation management etc.

1.5 Motivation

Sentiment Analysis has two-fold motivation. Both consumers and Customers' views on goods and services are highly regarded by manufacturers, so Sentiment Analysis has become popular. There has been a major push from both business and academia. It is important for us to know the opinions of those around us before making a decision. Previously, this community was small, consisting of only a few trustworthy family members and friends. People who are looking for information on a certain entity are now regularly reading these (product, movie etc.). As a result, there are various views available on the Internet. Extracting views on a single individual is crucial from the standpoint of users. The overwhelming number of data makes it difficult for people to sift through such a massive quantity of data in order to grasp the general consensus. As a result, a framework that distinguishes between positive and negative feedback is needed. Furthermore, marking these documents with their sentiment will offer a concise description of the public feeling of a person to the readers. Because of the growth of Web 2.0 platforms such as blogs, comments sections, and so on, consumers now have a platform to express their brand impressions and ideas, favourable or unfavourable, on any product or service. Since customers began to use the Internet to broaden their horizons, there has been an explosion in review pages and blogs where people can assess the benefits and drawbacks of a product or service. As a result, the future of the product or service is shaped by these viewpoints. Vendors need a framework that can detect patterns in consumer feedback and use them to enhance their product or service while also identifying potential needs.

To check multiple reviews or news at a time, tedious and time-consuming and there are huge number of datasets have to check to gain knowledge about overall reaction. The emergence of social networks and blogs over the last decade has prompted a deluge of interest in sentiment analysis. With reviews, ratings, suggestions, and other kinds of online expression, internet opinion has become useful information for organizations wanting to advertise their products, uncover new prospects, and manage their reputations. Many people are turning to sentiment analysis to help them locate relevant content and take necessary actions. Sentiment Analysis and its application to business analytics has seen a steady growth in interest from organizations and researchers in recent years.Today's corporate world, as in many data analytics streams, is looking for "business insight." As more written content is written and shared online, such as through Social Channels, Blogs, and Review Sites, we are becoming more vocal and open about our online experiences. This, once again, demonstrates the rising necessity and desire for firms to mine this information in order to derive commercial insight from it. Businesses are attempting to unlock the hidden value of text in order to better understand the attitudes and needs of their target audiences and make better, more informed strategy decisions.Traditionally, businesses relied on surveys, workshops, and focus groups to gain insight into their audiences' opinions and feelings, but with modern technology, we can harness the power of Machine Learning and Artificial Intelligence to extract meaning from text and dive into audience opinions and see them outside of the often controlled environment of a survey. Machine learning algorithms are best to do this kind of job perfectly. We used machine learning classifier along with an ensemble technique to do this job. As mentioned earlier, this work is focused on the polarity of Bangla texts as positive and negative.

1.6 Contribution of the thesis

Contributions of this work are given bellow:

- We design and develop a sentiment analysis system that determines the polarity of Bangla texts as positive or negative.

- We developed a Bangla dataset containing 8,122 labeled sentiment documents.

- We developed a supervised machine learning model using LR, DT, RF, KNN, MNB, SVM, SGD and AdaBoost to differentiate the sentiment from Bangla texts into positive or negative polarities.

- We implement both BoW and tf-idf feature extraction technique considering different combination of n-gram words.

- We also develop an ensemble learning classifier to classify the sentiment polarity more accurately.

1.7 Thesis Organization

The entire report is represented in six different chapters. We can outline the report structure as follows:

1. Chapter 1 gives an overview of sentiment analysis of Bangla Texts and discuss about motivation behind this project. It also lists the objectives of the project and the challenges we may face in this work.

2. Chapter 2 gives an overview on literature review as well as classification history and related works and implementation challenges.

3. Chapter 3 gives an overview of proposed methodology, Data processing, classifier algorithm, training set, testing set and Implementation.

4. Chapter 4 gives an overview on the dataset used to train the model as well as the impact of the thesis and evaluation of performance and framework. It also gives an overview of experimental analysis, result and evaluation.

5. Chapter 5 gives conclusion to this work and how it should be improved further in future.

1.8 Conclusion

From the above discussion we have got to know about sentiment analysis and our objectives and opportunities of sentiment analysis of Bangla texts. We have also discussed the limitations and the challenges we are going to face ahead. We will try our best to overcome those challenges and fulfil our objectives.

Chapter 2

Literature Review

2.1 Introduction

In this chapter we will shortly describe history of text classification and learn about Logistic Regression(LR), Decision Tree(DT), Multinomial Naïve bayes (MNB), Random Forest (RF), SGD Classifier, KNN Classifier method which are really useful for classifying text which are important to understand. Also the ensemble technique that is used will also be explained. This chapter also contains brief discussion on related previous works that is already implemented, their limitations.

2.1.1 Sentiment Analysis

Sentiment Analysis also known as Opinion Mining is a field within Natural Language Processing (NLP) that builds systems that try to identify and extract opinions within text. Usually, besides identifying the opinion, these systems extract attributes of the expression e.g.:

1. Polarity: if the speaker expresses a positive or negative opinion,

2. Subject: the thing that is being talked about

3. Opinion holder: the person, or entity that expresses the opinion.

Sentiment analysis is currently a popular field of research and development due to its numerous practical applications. Because the amount of publicly and privately available information on the Internet is always rising, there are a great number of texts expressing opinions available on review sites, forums, blogs, and social media. This unstructured information might be automatically turned into structured data on public opinions about products, services, brands, politics, or any other issue on which individuals may express their ideas using sentiment analysis tools.This information can be used for a variety of business purposes, including marketing analysis, public relations, product reviews, net promoter scores, product feedback, and customer service. We considered and worked with polarity in Bangla texts in our project.

2.1.2 Opinion

Before we get into the specifics, let's first define opinion. Text information can be divided into two categories: facts and opinions. Facts are unbiased statements about anything. Opinions are typically subjective expressions that represent people's opinions, assessments, and opinions about a subject or topic. Sentiment analysis, like many other NLP issues, can be described as a classification problem with two sub-problems: Subjectivity classification is the process of determining whether a sentence is subjective or objective.Polarity categorization is the process of categorizing a sentence as positive, negative, or neutral. In an opinion, the entity the text talks about can be an item, its components, its aspects, its attributes, or its features. It could also be a product, a service, a person, a company, an event, or a topic.

2.1.3 Natural Language Processing (NLP)

Natural language processing (NLP) is a field of computer science, artificial intelligence, and computational linguistics concerned with computer interactions with human (natural) languages. As a result, NLP is linked to the field of humancomputer interaction. Many NLP issues include: natural language understanding, letting computers to draw meaning from human or natural language input, and natural language generation. Natural language processing (NLP) is an important medium of communication. For any language, it is an empirical field of work.To create a program that understands spoken language, we need to grasp all of the facilities of a written language as well as enough additional information to handle all types of ambiguity. Natural language processing encompasses activities like comprehension and generation, among others.

2.1.4 Text Classification

Text classification is the task of automatically categorizing a text into a set of predetermined classes. Because of the increasing proliferation of online content, text classification has grown more difficult as well as more crucial. A massive amount of textual data is available online, and text classification is required to extract the expected and desired information from that data. Text-based classification is being used by researchers all around the world to extract information. Text Classification categorizes a document by assigning it to one or more classes based on its content. Classes are chosen from a pre-existing taxonomy (a hierarchy of categories or classes). Text Ccassification API for English, Hindi, and other languages are available as open source on the Internet to assist academics with their relevant work. All preparation operations (text extraction, tokenization, stop word removal, and lemmatization) required for automatic classification are handled by the API.

2.1.4.1 History of Text Classification

Text classification for document classification) first emerged in late 1990 as "Text Data Mining" and has been actively implemented by many researchers. Text classification is a preprocessing technology which can be used to filter out irrelevant documents from a large scale corpus.A text source is handled as a bag-of-words in early techniques. The bag-of-words is a simple representation used in natural language processing in which a sentence or document is expressed as a set of words. But there was no ability to understand the semantics of a document in earlier approaches. After early ages, researchers try to find out hidden relationships and another complex pattern within data-sets. Several techniques such as clustering, classification, decision trees, and link analysis are used to find out these complex relations. This technique along with machine learning algorithms enables us to find deeper linguistics that enables us to understand the semantics meaning of a document or a sentence. It has now become one of the most accredited research topics. If a system can classify texts accurately, we can use it to predict events which will happen in the future from our present data. The modern world is evaluating towards Industry 4.0 whose main concerns are data and information.

The company which or person who has a collection of huge information is considered to be more powerful than others. Text classification detects emotion by which we can predict the opinion of people about new products or events. Text classification can be described into two categories.

2.1.4.2 Supervised Classification

In supervised classification of text classification categories are defined. It works on training and testing principle. During training phase. the machine learning algorithm works on the label data.The classifier is trained on labeled data and produces the expected results. During testing phase, unobserved data are fed into algorithm and classifier classifies them based on the knowledge of training phase.

2.1.4.3 Unsupervised Classification

In unsupervised classification of text classification categories are not defined. Here machine learning algorithm try to discover natural structure in data. The algorithm searches the data points for similar patterns and structures and groups them into clusters. The data is classified depending on the clusters that form. One can also apply some other ways to classify text such as Custom Text Classification, Semi-Supervised Text Classification etc.

2.1.5 Text Classification Methods

There are many classification algorithm like Logistic Regression (LR), Decision Tree (DT), Multinomial Naïve Bayes (MNB), Random Forest (RF), Stochastic Gradient Descent (SGD), Support Vector Machine (SVM), AdaBoost Classifier, Kth Nearest Neighbour (KNN) etc. We have used all eight of those ML algorithms to train our model and then shown a comparison among these models. We have also applied an ensemble technique using the combination of SVM, LR, and DT.

2.1.5.1 Logistic Regression (LR)

Using a logic function [5] LR's predictions are transformed. If we apply log to hypothesis (predicted) we get some values (cost) which is useful to estimate the overall error. The outcome and the cost functions are determined by Eq. 2.1 and Eq. 2.2 respectively.

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}} \tag{2.1}$$

$$J(\Theta) = \frac{1}{m} \sum_{i=1}^{m} Cost(h_\theta(x^{(i)}), y^{(i)}) \tag{2.2}$$

$$Cost(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if}: \ y = 1 \\ -\log(1 - h_\theta(x)) & \text{if}: \ y = 0 \end{cases}$$

where, Eq. 2.1 is also known as Sigmoid function, $h_\theta(x)$ denotes the hypothesis function of the ith instances, number of training instances is denoted by m and y^i indicates the input label of ith training instances.

2.1.5.2 Multinomial Naïve Bayes (MNB)

The Naive Bayes classifier is a probabilistic classifier that is extensively used in machine learning. Bayesian classifiers are statistical and have the ability to learn. For processing large data-set multinomial model of Naive Bayes is used. By searching the dependencies among attributes, the performance of Naive Bayes could be enhanced. Because of its ease of computation, it is primarily employed in data preparation applications. In order to predict the target class, Bayesian reasoning and probability inference are used. When employing a probabilistic model for classification, attributes are crucial. As a result, attribute weight values play an essential role in improving the model's performance. Because of its versatility, it has been used in both the preparation and classification stages. It's a probabilistic classifier that can find out how to analyze a collection of records. It has been classified. It compares the contents to the word list to delegate records to the relevant group [6].

The conditional probability for Naive Bayes can be defined as

$$P(X|y_j) = \prod_{i}^{n} P(x_i|y_i)$$

where i=1 and X is the feature vevtor defined as $X = x_1, x_2, x_3, ...X_n$ and y_i is the class label. The main limitation of Bayesian networks is that the time complexity increases when high dimensional text data is processed using these networks. Moreover, ill Bayesian networks interaction between features cannot be achieved and the probabilities calculated are not accurate but relative probability.

2.1.5.3 Support Vector Machine (SVM)

SVM plots data points as points in an n-dimensional space (n denotes the number of features), with the value of each feature representing the value of a given coordinate. Finding the ideal hyperplane that best separates the two groups is utilized to categorize them. Using Eq. 2.3 we can obtain a hyperplain.

$$f(x) = w^T x + b \tag{2.3}$$

where b is the bias and n is a normal to the line.The hypothesis function h is defined as:

$$h(x_i) = \begin{cases} +1 \text{ if } w.x + b \geq 0 \\ -1 \text{ if } w.x + b < 0 \end{cases} \tag{2.4}$$

2.1.5.4 Decision Tree

The Decision Tree algorithm is a member of the supervised learning algorithm family. The decision tree approach, unlike other supervised learning algorithms, may also be utilized to solve regression and classification issues. The purpose of employing a Choice Tree is to build a training model that can predict the class or value of the target variable by learning basic decision rules from prior data (training data). The Decision Tree algorithm use information gain to compute the amount of information that a feature provides about a class when constructing a Decision Tree.

Expected gained information= Entropy before - Entropy after a decision Here, the amount of disorder is known as entropy. In Decision Trees, we begin at the root of the tree to forecast a class label for a record. The values of the root attribute are compared to the values of the record's attribute. Based on the comparison, we follow the branch corresponding to that value and proceed to the next node.

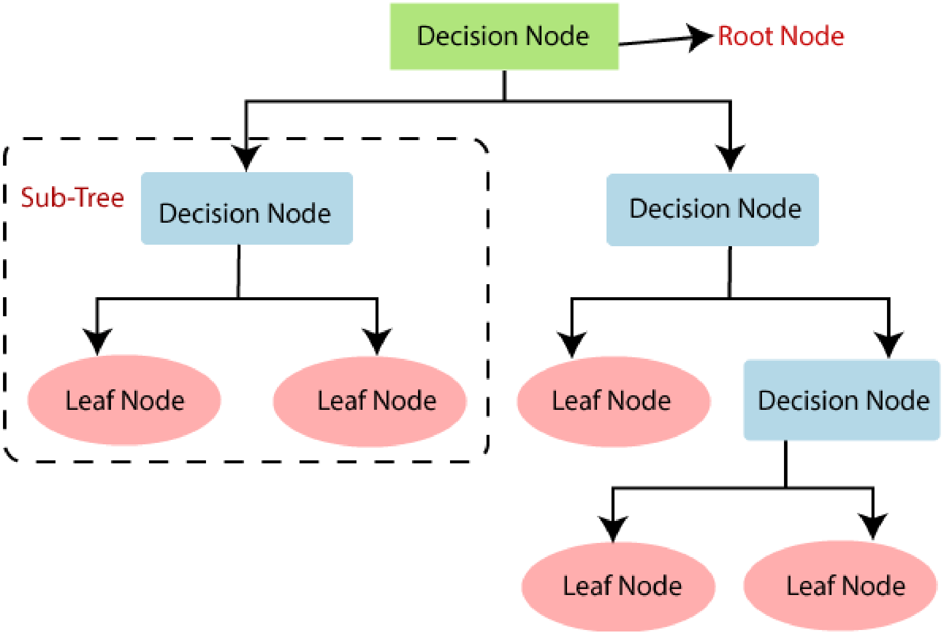


Figure 2.1: Decision Tree Classification Algorithm.

2.1.5.5 Random Forest (RF)

It is theoretically a decision tree ensemble method created on a randomly divided dataset (based on the divide-and-conquer approach). A group of decision tree classifiers is referred to as the forest. Using an attribute selection indicator such as information gain, gain ratio, or Gini index, specific decision trees are generated for each attribute. The Gini index is calculated by:

$$gini = 1 - \sum_{i=1}^{C} (y_i)^2 \tag{2.5}$$

where, C is denoted by total number of class and y_i denotes the probability of ith class.

2.1.5.6 Stochastic Gradient Descent (SGD) method

For large training dataset this method is used.Instead of calculating the gradient, each iteration of the SGD method calculates the value of the gradient based on a single randomly selected case [7]

$$W_{t+1} = W_t - \gamma_t \delta_w Q(z_t, w_t)$$

Table 2.1: Ensemble Pseudocode

```
PSEUDOCODE 1
Majority Vote Based Ensemble Classifier
Step 1: Apply three classifiers (RF, SVM, LR) on the training data.
Step 2: Compare the performances of the three classes.
Step 3: Performing majority voting for each observation.
Step 4: Compare the performance of the majority voting
          with the RF, LR, and SVM.
Step 5: Aggregate vote to the ensemble.
```

At each iteration this process $w_t, t = 1, 2, 3, ...$ depends on randomly picked examples, where $Q(z_t, w_t)$ minimizes the risk and the learning rate is γ_t. By noisy approximation of the gradient the convergence gets affected.The parameter estimate w_t decreases equally slowly if the learning rate decreases slowly, but the parameter estimate takes a significant amount of time if to reach it's optimum point if the rate decreases too slowly.

2.1.5.7 Ensemble

Ensemble methods are conceptual that combine multiple machine learning techniques into a single predictive model in order to reduce uncertainty, bias, or improve prediction accuracy. By incorporating the advantages of each individual classifier, the Majority Vote Based Ensemble Classifier method improves accuracy. LR, RF and SVM are taken as the base classifier to implement the ensemble. We experimented with different combinations to find the best base classifiers for the ensemble, and found that LR, RF, and SVM produced the best results. Table 2.1 is the Majority Voting pseudocode used in the proposed work:

2.2 Related Literature Review

The term sentiment analysis is first appeared in Nasukawa and Yi [8]. In their work, they proposed their approach towards word level analysis. By analyzing sentiment to identify the semantic relationships between the subject and the sentiment expressions. They applied semantic analysis with sentiment lexicon and a syntactic parser. On the other hand, Prabowo and Thelwall, they combined supervised learning, rule-based classification and machine learning [9]. They also tested their new method on movie reviews. However, in their method they used thousands of rules and it is hard to maintain and it becomes complex. V.K. Singh et al. [10] proposed a new method and it is feature based domain specific heuristic for aspect level sentiment analysis for movie reviews. He used SentiWordNet with verbs, adjectives and adverbs. Singh et al. [10] approached a SentiWordNet that having two linguistic features. This SentiWordNet approaches sentiment analysis of movie reviews in document level. Here he used two techniques: SVM (Support Vector Machine) and Naïve Bayes for the sentiment analysis. In order to find the emotional polarity, T. Prasad et al [11] and P. Bhoir et al [12] performed sentiment analysis using SentiWordNet to gain overall reaction. As we can see, preprocessing is not used in these papers which can be necessary in increasing the accuracy. G. Gautam D. Yadav [13] used Naive Bayes, Maximum Entropy, and SVM to perform sentiment analysis on Twitter data, as well as the Semantic Orientation-based WordNet, which extracts synonyms and similarities for the content characteristic of Twitter data in English.

Sentiment analysis in Bangla languages have not been at light due to lack of standard resources. At present some researchers have working in Bangla sentiment analysis. In order to gain more update, I went through a few works of them. For example, Hossain et al. [14] worked in Bangla Book Review using Multi Nominal Naïve Bayes and as it is only on book reviews and used only 2000 Bangla dataset. Rahman et al. [**rahman2018datasets**] worked on Aspect-Based Sentiment Analysis; used cricket(2900) & restaurant(2800) and got maximum of 42% f1-score only. Sumit et al. [**sumit2018exploring**] worked on word embedding for Bangla sentiment analysis, they have implemented Skip-Gram, Word2vec, and Continuous Bag of Words with Word to Index model and found that Word2vec Skip-Gram model outperformed other models and achieved 83.79% accuracy. Tabassum et al. [**tabassum2019design**] performed sentiment analysis using Random Forest classifier and used negation, POS tagging, and unigram; they used only 850 texts to traing the model and got 87% accuracy. Tripto et al. [**tripto2018detecting**] worked on detecting multilabel sentiment and emotions from Bangla YouTube comments and for three-class (positive, negative, neutral) they found 65.97% accuracy for their dataset. Taher et al. [**taher2018n**] implemented a n-gram based sentiment analysis using Support Vector Machine and found 91.68% accuracy.

Chowdhury et al. [**chowdhury2019analyzing**] worked on analyzing the sentimet of Bangla movie reviews and found maximum of 88.90% accuracy for Support Vector Machine algorithm. From the above discussion and to our knowledge it is clear that there has been no good work on sentiment analysis. All of those work considered single feature extraction techniques.Furthermore, to the best of our knowledge, no work on sentiment analysis utilizing Ensemble Learning has been done. So, we used eight different machine learning models along with ensemble learning techniques. And used both Bag of Words and Tf-Idf feature extraction techniques. In contrast to these, it is intended to design a system to analysis Bangla texts in order to find out the polarity based on Bangla sentiment analysis.

2.3 Conclusion

From the above discussion we can conclude that for Bangla language sentiment analysis is still in its beginning process. As most of the work has been done in other languages. So there are lots of opportunities and scope of work in this field.

2.3.1 Implementation Challenges

The system's implementation required the development of a dataset that could be used by our learning algorithm. We all know that a well-equipped dataset is necessary for every machine learning algorithm. If the dataset is accurate, the output's accuracy is flawless as well. Bangla is a language with limited resources. We are attempting to solve this problem by obtaining a vast volume of text from numerous Bangla sources. And we have managed to create a corpus of 8122 Bangla dataset but still this isnt enough. Besides for a huge dataset we need a very good computational system and the best resource we had was Google Colab. Another issue is categorizing data into positive and negative categories. We must label the data since we are using a supervised learning algorithm. Since inaccurate data is used for instruction, the results are incorrect. As a result, our method would deliver poor results. Output can deteriorate as well. As a result, for preparation and research purposes, we must gather data from different sites with more accurately.

Chapter 3

Methodology

3.1 Introduction

In this chapter, we will go through our proposed methodology as well as learn about each module of our system. We will gain a little mathematical understanding of the feature extraction techniques used in our system. Finally, our system's total implementation.

3.2 Diagram/Overview of Framework

Development of this kind of system involves the development of classification problem that will evaluate different basic emotions from textual input. Figure 3.1 shows a Schematic Diagram of our Proposed System.

3.3 Detailed Explanation

In the implementation the raw data is processed into several steps. It includes data preprocessing, feature extraction and some other steps. In this chapter we will give a brief explanation of all those steps.

3.3.1 Implementation

Sentiment analysis in Bangla text is a difficult task. We made every effort to efficiently implement this System. In this chapter, we have supplied some pictures of our project together with the essential explanations that will clearly convey the project's results. We've also included our project's experimental setup.

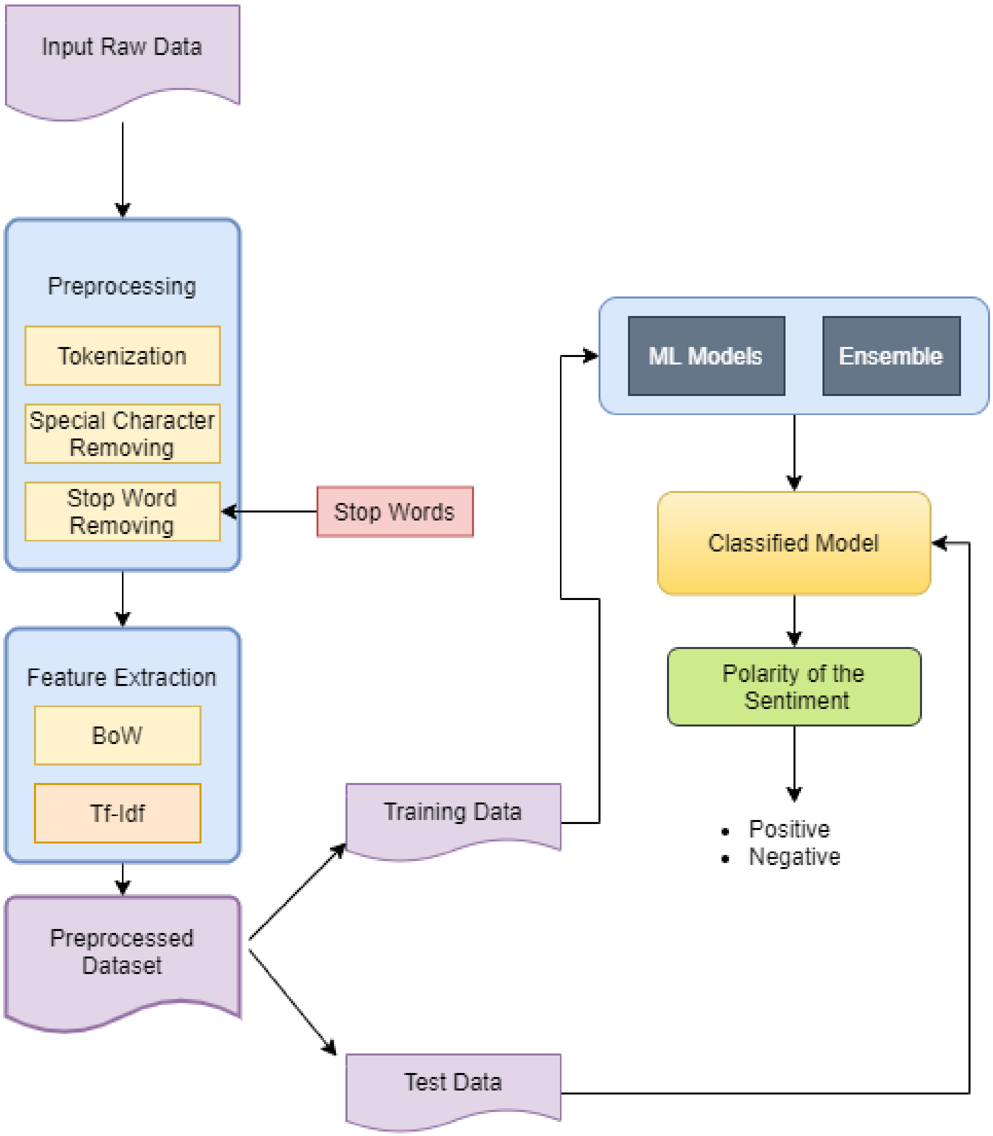


Figure 3.1: Schematic Diagram of the Proposed System.

3.3.1.1 System Requirements

Our system takes Bangla text as input and classifies it according to its expressing polarity such as negative, positive. Some hardware and software tools are required to implement this system. The hardware and software tools required are given below.

3.3.1.2 Hardware Requirements

- Personal Computer
- Intel Core i5 CPU
- Physical Memory 4GB

3.3.1.3 Software Requirements

- Operating System: Microsoft Windows 10
- Software Libraries: Python 3.6.6
- IDE: Sublime3
- Other: notepad++
- Jupiter Notebook
- Google Collaboratory

3.3.2 Implementation Details

To implement the system first the dataset is collected. Since the datasets are labeled datasets, they are sorted in a folder depending on their label. Then passed through preprocessing steps and classifier to train the system. Each step is described below.

3.3.2.1 Training Text

Supervised learning involves using a set of training examples that make up the training data. A text corpus is a big, organised collection of texts. For a machine learning algorithm, a well established corpus is a must for obtaining performance according to expectation. In the text corpus, there consists input Bangla sentence and desired output value. An algorithm is then applied to the data to produce a classifier which will determine the correct output for any further valid input values. In Bengali language processing there is no quality full corpus available which contains subjective emotional text. But as ours is a supervised classification system, our emotional text detector needs a training corpus which consist of text and labels indicating whether a text expresses positive or negative. To implement this system, we collect 8122 Bangla text documents and labeled then as positive or negative. Training corpus is divided into two categories. The accuracy of the learning algorithms depends on uniqueness of training examples.

3.3.2.2 Data collection

We have collected data from Facebook, Youtube, newspapers, blogs and various offline sources.. Lets consider the following text collected from a newspaper “মধ্যপ্রাচ্যে যুদ্ধ নিয়ে মার্কিন প্রেসিডেন্ট ডোনাল ট্রাম্পকে সতর্ক করেছে ইরান। ইরান জানিয়ে দিয়েছে, যে কোন আক্রমণাত্মক পদক্ষেপের বিরুদ্ধে চূড়ান্ত প্রতিক্রিয়া দেখানো হবে।” (Iran warns US President Donald Trump about war in the Middle East, Iran has said that they will take "final action" against any offensive) This sample dataset contains two sentences and expressing a negative sentiment.

3.3.2.3 Data Cleaning

To ensure accuracy, we had to go through data cleaning step. It is a vital step for the system. Data cleaning is done in two steps given below:

3.3.2.3.1 Special Character removal We have performed post collection procedure to remove unwanted or unnecessary data such as coma(,), fullstop (.) extra space etc. This helped us to get rid of junks. For example, after cleaning above sample review we will get two sentences-

1. মধ্যপ্রাচ্যে যুদ্ধ নিয়ে মার্কিন প্রেসিডেন্ট ডোনাল ট্রাম্পকে সতর্ক করেছে ইরান

2. ইরান জানিয়ে দিয়েছে যে কোন আক্রমণাত্মক পদক্ষেপের বিরুদ্ধে চূড়ান্ত প্রতিক্রিয়া দেখানো হবে

3.3.2.3.2 Data preprocessing The raw data obtained through different resources are noisy and hence are preprocessed. Preprocessing has been performed through some steps Tokenization, Stop word removal, Extracting polar signs. First the sentences are tokenized into tokens. Table 3.1 shows the tokenization of the sample dataset:

Table 3.1: Tokenization of sample review

Sentence	**Tokens**
মধ্যপ্রাচ্যে যুদ্ধ নিয়ে মার্কিন প্রেসিডেন্ট ডোনাল ট্রাম্পকে সতর্ক করেছেইরান	“মধ্যপ্রাচ্যে”, “যুদ্ধ”, “নিয়ে”, “মার্কিন “,“প্রেসিডেন্ট”, “ডোনাল”, “ট্রাম্পকে”, “সতর্ক”, “করেছে”, “ইরান”
ইরান জানিয়ে দিয়েছে যে কোন আক্রমণাত্মক পদক্ষেপের বিরুদ্ধে চূড়ান্ত প্রতিক্রিয়া দেখানোহবে	“ইরান”, “জানিয়ে”, “দিয়েছে”, “যে”, “কোন”,“আক্রমণাত্মক”, “পদক্ষেপের”, “বিরুদ্ধে”, “চূড়ান্ত”, “প্রতিক্রিয়া”, “দেখানো”,“হবে”

3.3.2.4 Feature Extraction

The modified dataset after pre-processing has a number of distinguishing characteristics. The aspect (adjective) is extracted from the dataset using the feature extraction approach. Bag of Words (BoW) and Term Frequency and Inverse Document Frequency were the two feature extraction techniques applied (Tf-Idf). One of the most major challenges with text is that it's really noisy and unstructured; machine learning algorithms prefer organized, well-defined fixed-length inputs, and we can use the Bag-of-Words technique to convert variable-length texts into fixed-length vectors. The frequencies of the words is considered as features in BoW [15]. The weights/importance of context-related terms that are less common can be lower than the weights/importance of insignificant words with a high frequency. To overcome this problem tf-idf technique [16] was used as contextual words get more importance in this method.

3.3.2.4.1 Bag of Words A bag-of-words model is a method of extracting textual features for use in modeling, such as with machine learning techniques. The method is quite adaptable, and it is a text representation that represents the frequency of words within a text document. It involves two things:

- A vocabulary of known words.
- A measure of the presence of known words.

It is called a "bag" of words, because any information about the order or structure of words in the document is discarded. The model is only concerned with whether known words occur in the document, not where in the document. The intuition is that documents are similar if they have similar content. Further, that from the content alone we can learn something about the meaning of the document. The complexity of BOW comes both in deciding how to design the vocabulary of known words or tokens and how to score the presence of known words. In this approach, we look at the histogram of the words within the text, i.e. considering each word count as a feature. The objective is to turn each document of free text into a binary vector that we can use as input or output for a machine learning model. All ordering of the words is nominally discarded and we have a consistent way of extracting features from any document in our corpus, ready for use in modeling. New documents that overlap with the vocabulary of known words, but may contain words outside of the vocabulary, can still be encoded, where only the occurrence of known words is scored. Accuracy increases overall as the number of words used increases, and as the number of dimensions increases. But we should be careful about associated complexities. In our work number of maximum features taken is 15,000.

3.3.2.4.2 TF-IDF The majority of machine learning algorithms work with data in the form of a matrix of numbers. The sentiment data, on the other hand, is always in text format. As a result, it must be transformed to a number matrix. TF or TF-IDF have been suggested by different scholars as a tool for transforming text into a numerical matrix. However, in this thesis, the combination of TF-IDF and CountVectorize is used to transform text data into a matrix of numbers. The number of times a word appears in a document divded by the total number of words in the document. Every document has its own term frequency.

$$tf_{i,j} = \frac{n_{i,j}}{\sum_{k}, n_{i,j}}$$

The total number of documents divided by the total number of documents containing the word w is the log of the total number of documents. Inverse data frequency is used to determine the weight of unusual words across all documents in the corpus.

$$idf(w) = \log(\frac{N}{df_t})$$

The IDF is computed once for all documents. Lastly, the TF-IDF is simply the TF multiplied by IDF.

$$W_{i,j} = tf_{i,j} * \log(\frac{N}{df_t})$$

We took into action the combination of uni-gram, uni-gram + bi-gram, uni-gram + bi-gram + tri-gram features.

3.3.2.4.3 Preprocessed Dataset The dataset will be used in this work is already labeled. Labeled dataset has a negative and positive polarity and thus the analysis of the data becomes easy. The raw data having polarity is highly

Table 3.2: Summary of Parameters

Classifiers	Parameters
LR	class_weight='balanced', max_iter=400,random_state = 123
DT	criterion = 'entropy', random_state = 0,splitter='best'
RF	min_samples_split=2, n_estimators=100, criterion='gini'
KNN	n_neighbors=15, weights='uniform', metric='minkowski', p=2
SVM	probability=True, gamma = 0.0001, random_state = 0,C=1.0, kernel='rbf'
MNB	class_prior='None', additive_smoothing=1.0, ft_prior='true'
SGD	loss="log", penalty="l2", max_iter=5
AdaBoost	random_state=0, learning_rate=1, n_estimators=50

susceptible to inconsistency and redundancy. The quality of the data affects the results and therefore in order to improve the quality, the raw data is preprocessed. It deals with the preparation that removes the repeated words and punctuations and improves the efficiency. It is a difficult task to select dataset for this work because it is tough to cover most of the words in one dataset. As there is shortage of data in Bangla language. We had to make a whole dataset for this thesis. To make the result more accurate and perfect, we have to use as many dataset as possible. In dataset preparation stage the preprocessed datasets have to divide into training datasets and the testing datasets. To improve the systems performance, it is needed to enrich words and phrases.

3.3.2.4.4 Classification Algorithms After the preprocessing the and extracting the features using tf-idf and BoW the ML classifiers was implemented. In this work, we used eignt classifiers, namely,Multinomial Naïve Bayes (MNB), Kth Nearest Neighbours (KNN), Logistic Regression (LR), Decision Tree (DT), Stochastic Gradient Descent (SGD), Random Forest (RF), AdaBoost Classifier and Support Vector Machine (SVM). An ensemble technique is applied on the dataset to get more accuracy. Table 3.2 shows all the parameters used by the classifiers.

3.4 Conclusion

In this chapter we have given a descriptive explanation of our implementation. During the implementation we had faced many difficulties and was able to overcome all of those.

Chapter 4

Results and Discussions

4.1 Introduction

In this chapter we will give an overview on the dataset used to train the model as well as the impact i.e. social, environmental and ethical impact of the thesis and evaluation of performance and framework. We will also discuss about the experimental analysis, result and evaluation.

4.2 Dataset Description

Due to the unavailability of benchmark textual sentiment dataset in Bengali, this work developed a dataset (i.e., BSaD: Bengali Sentiment Analysis Dataset) to perform the sentiment analysis task. We followed the directions suggested by Das et al. [6] for developing dataset. Following steps are carried out to prepare the BSaD:

4.2.1 Text Accumulation and Preprocessing

We started collecting data manually on August 2019 and finished on December 2020. Our sentiment texts are collected from online newspapers, Facebook, Youtube, blogs, and offline (storybooks, novels, conversations). This data collection process was done by five human crawlers and total 8815 text documents was collected over this period. This collected raw data needed some correction and preprocessing before annotation. We had to remove duplicate data, non-Bengali words and also for getting a distinctive sentiment meaning by discarding data that is less than three words long. After preprocessing the dataset holds 8535 text documents.This processed texts are then feasible for manual annotation.

4.2.2 Text Annotation and Quality Measure

Initially five undergraduate NLP enthusiast were assigned to annotate the dataset. Using majority voting technique [17] the initial label was choosen. An expert corrected the labelling if any initial annotation was performed improperly. The expert discard 413 texts as they had neutral and mix sentiment. Cohen Kappa score [18] is calculated to measure the quality of annotation. There is a lot of consensus among the annotators as we get % of agreement: 97.34% and Cohens k: 76.58% which is Substantial agreement. Final corpus contains 8122 documents.

4.2.3 Dataset Statistics

Finally, after the preprocessing and annotation process BSaD contains 8250 text documents. BSaD consists data from various sources. Among online sources Facebook contributes 2796, online newspaper 2306, Youtube 610 and blogs contributes with 483 text documents. Substantial amount of data was also collected from offline sources (2084 text documents). Table 4.1 shows the summary of the dataset. Here we can see almost all of our data's word frequency is more that 50 words. Fig. 4.1 shows the length of the texts.

Table 4.1: Dataset Summary

Dataset attributes	**Quantity**
Number of documents	8,122
Number of positive documents	2,421
Number of negative documents	5,702
Number of words	2,20,988
Total unique words	35,748
Size (in bytes)	3,654,967

4.3 Impact Analysis

The power of sentiment analysis can be enormous. It can have a significant impact on society or the environment. Because of its potential to extract useful information from social data, it is a practice that is widely used by enterprises all over the world.

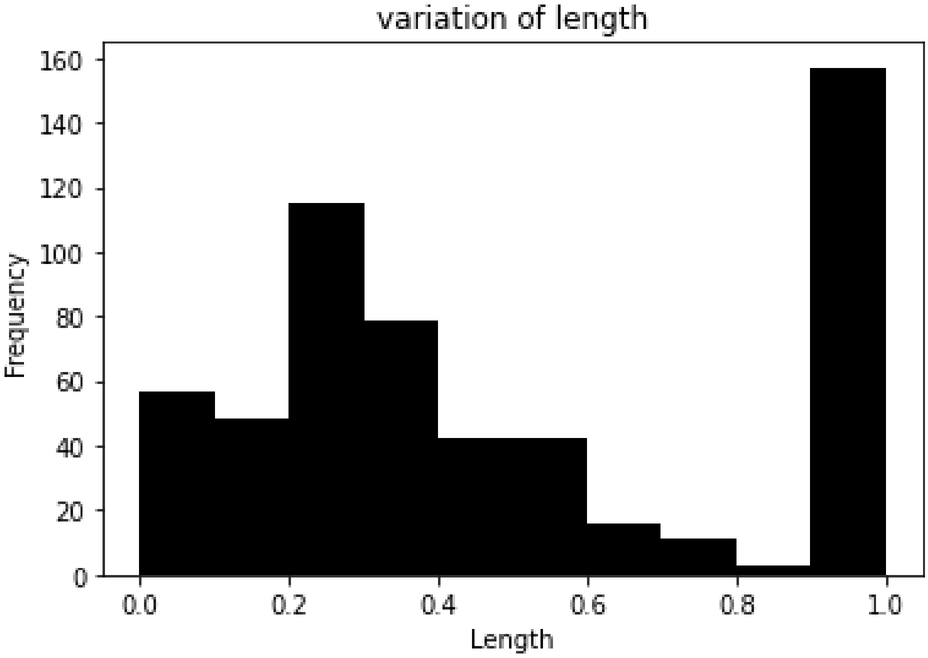


Figure 4.1: Variation of length of BSaD.

4.3.1 Social and Environmental Impact

We can accomplish a great deal with the assistance of sentiment analysis like Social networking monitoring, customer service management, and customer feedback analysis and so on. As sentiment analysis detects whether a text is positive or negative so we can easily filter the negative news, articles, comments and status and can spread more positivity throughout our society. We can easily determine or choose a good book or movie based on the sentiment analysis of previous customers reviews.

4.3.2 Ethical Impact

Obviously, sentiment analysis has some ethical impacts also. When we use the word ethics first thing comes to our mind is that whether the thing is good or bad. And its fully related to sentiment analysis as in this terminology we measure emotions as positive or negative. If we are able to implement this method in every sector of our digital life we can eliminate negativity from out timeline as well as our society.

4.4 Evaluation of Performance

We employed several evaluation matrices to evaluate our suggested system. We will learn about this evaluation matrices in this section. Some of this evaluation matrices are:

- Confusion Matrix
- Precision
- Recall
- F1 Score
- Accuracy

Confusion Matrix: A confusion matrix is a table that is used to evaluate the performance of a classification model. As ours is a multi-class classification model, the confusion matrix of our system has seven rows and seven columns. This matrix reports the number of false positives, false negatives; true positives, and true negatives. Table 4.2 shows the confusion matrix.

Table 4.2: Confusion Matrix

	Correct Labels	
	Positive	Negative
Positive	TP (True Positive)	FP (False Positive)
Negative	FN (False Negative)	TN (True Negative)

- **True Positive (TP):** Number of documents that is positive and also classified as positive.
- **True Negative (TN):** Number of documents that is negative and also classified as negative.
- **False Negative (FN):** The number of documents that are positive yet are labeled as negative.
- **False Positive (FP):** The number of documents that are negative yet are labeled as positive.

This numbers are used to calculate other evaluation.

Precision: Precision refers as positive predicted value. That is the ratio of correctly classified positive instances to the total number of instances classified as positive. Precision can be obtained from the following equation.

$$Precision = \frac{TP}{TP + FP}$$

Recall: The ratio of accurately identified positive instances to the total number of positive cases is known as recall. It is also known as the true positive rate. The following equation can be used to calculate recall.

$$Precision = \frac{TP}{TP + FN}$$

A model must have a high level of precision and recall. Unfortunately, precision and recall have a trade-off. In other words, increasing precision often decreases recall and vice versa.

- If threshold of a classifier is increased then it causes high precision and low recall
- If threshold of a classifier is decreased then it causes low precision and high recall.

To sort out this problem we need another measure that is F1 measure.

F1 score: The F1 score is calculated as the weighted average of Precision and Recall. As a result, this score considers both false positives and false negatives. In most cases, F1 is more useful than precision. Accuracy works best when the cost of false positives and false negatives is comparable. If the cost of false positives and false negatives is considerably different, it is best to consider Precision and Recall. F1 score can be obtained from the following equation:

$$F_1 Score = \frac{2 * Precision * Recall}{Precision * Recall}$$

Accuracy: Accuracy is the most obvious performance metric, because it is simply the ratio of properly predicted observations to total observations. For our system accuracy can be obtained by following equation:

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

Table 4.3: Performance (weighted) of ML classifiers with BoW

Feature Extraction	**Classifiers**	**Auc. (%)**	**Pr. (%)**	**Re. (%)**	**F1. (%)**
BoW	LR	77	78	77	78
	DT	72	72	72	72
	RF	78	79	78	76
	KNN	71	71	71	63
	SVM	75	74	75	73
	SGD	76	76	76	76
	MNB	79	78	79	78
	AdaBoost	76	75	76	74
	Ensemble	**80**	**80**	**80**	**80**

Error of our system can be obtained easily after calculating accuracy,

$$Error = 1 - Accuracy$$

4.4.1 Experimental Results

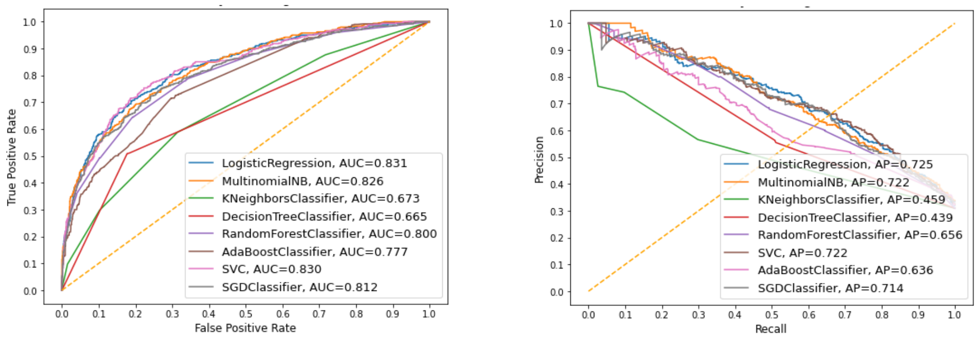
Table 4.3 shows the weighted performance of nine ML classifiers including ensemble using BoW feature extraction technique. Where Auc. Pr. Re. and F1 stands for Accuracy, Precision, Recall and F1-score. Here we can see that ensemble outperforms all other classifier with highest accuracy (80%) and f1-score (80%).

In Table 4.4 the weighted performance of ML classifiers including ensemble technique with the combination of uni-gram, unigram + bi-gram, and uni-gram + bi-gram + tri-gram features has been shown. Here we can see that for uni-gram features ensemble outperforms all other classifier with 81% accuracy and 80% f1-score. For the combination of uni-gram + bi-gram features SGD performs best with 82% accuracy and 81% f1-score. Finally, for the combination of uni-gram + bi-gram + tri-gram features ensemble outperforms all other classifiers with 82% accuracy and 82% f1-score. In contrast, we get best f1-score of 82% with tf-idf (uni-gram + bi-gram + tri-gram) + ensemble approach.

Fig. 4.2 shows a graphical comparison of the classifiers for BoW feature extraction technique. Fig. 4.3 Fig. 4.4 Fig. 4.5 also shows the graphical comparison of uni-gram, bi-gram, and tri-gram combination using ROC and PR curve analysis.

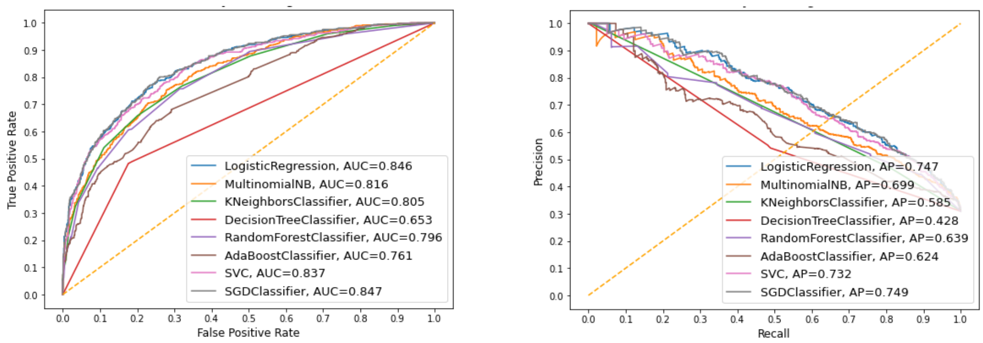
Table 4.4: Performance (weighted) of ML classifiers with tf-idf

Feature Extraction		**Classifiers**	**Auc. (%)**	**Pr. (%)**	**Re. (%)**	**F1 (%)**
Tf-Idf	Uni-gram	LR	79	79	79	77
		DT	72	71	72	71
		RF	79	80	79	75
		KNN	77	77	77	75
		SVM	79	79	79	79
		SGD	77	80	77	72
		MNB	75	80	75	68
		AdaBoost	76	75	76	74
		Ensemble	**81**	80	81	**80**
	Uni-gram + Bi-gram	LR	80	81	80	77
		DT	72	71	72	72
		RF	78	80	78	75
		KNN	78	77	78	77
		SVM	80	80	80	80
		SGD	**82**	81	82	**81**
		MNB	77	81	77	72
		AdaBoost	77	76	77	75
		Ensemble	80	80	80	78
	Uni-gram +Bi-gram+ Tri-gram	LR	80	82	80	78
		DT	73	73	73	73
		RF	78	80	78	75
		KNN	77	76	77	76
		SVM	81	80	81	80
		SGD	82	82	82	81
		MNB	76	81	76	71
		AdaBoost	76	75	76	74
		Ensemble	**82**	82	82	**82**



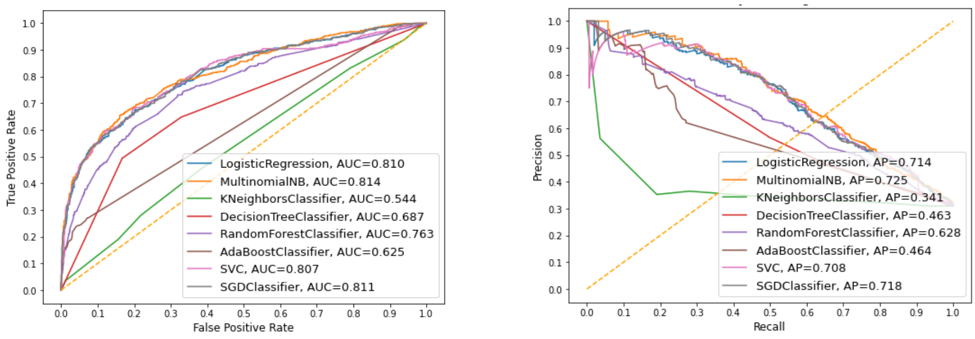
(a) ROC curve analysis for BoW features (b) PR curve analysis for BoW features

Figure 4.2: Classifiers performance on BoW features.



(a) ROC curve analysis for uni-gram features (b) PR curve analysis for uni-gram features

Figure 4.3: Classifiers performance on Uni-gram features.

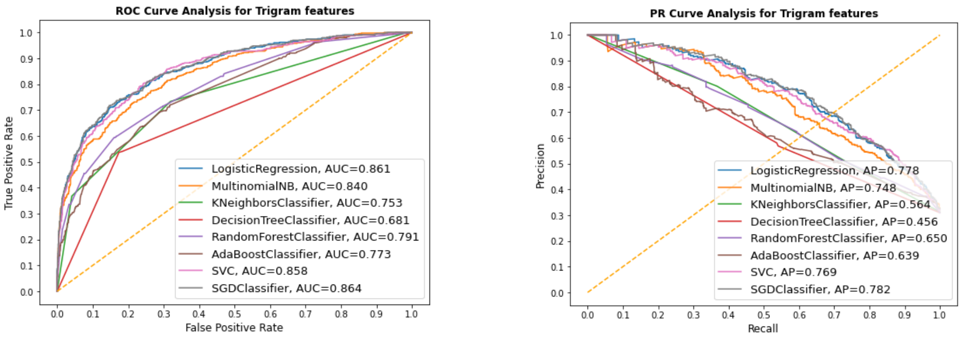


(a) ROC curve analysis for bi-gram features (b) PR curve analysis for bi-gram features

Figure 4.4: Classifiers performance on Bi-gram features.

4.4.2 Comparative Analysis with Existing Datasets

To evaluate our model we used two other dataset, dataset1 (DS1) [19] and dataset2 (DS2) [20] that are available online. DS1 has 3,500 positive and 3,500 negative text documents on the other hand DS2 has 5,000 positive and 5,000 negative text documents. Table 4.5 shows that ourdataset (BSaD) outperformed the other datsets with 80% f1-score. In Table 4.6 we can that our dataset BSaD showed best f1-score for evey combination of tf-idf. For uni-gram DS1 shows maximum



(a) ROC curve analysis for tri-gram features

(b) PR curve analysis for tri-gram features

Figure 4.5: Classifiers performance on Tri-gram features.

Table 4.5: Performance comparison with BoW features

Dataset	**Classifier**	**Pr. (%)**	**Re. (%)**	**F1. (%)**
DS1	LR	75	75	75
	RF	78	78	78
	SVM	75	75	75
	SGD	71	70	69
	Ensemble	77	76	76
DS2	LR	69	69	69
	RF	71	71	71
	SVM	69	69	69
	SGD	67	66	66
	Ensemble	70	70	70
BSaD	LR	78	77	78
	RF	79	78	76
	SVM	74	75	73
	SGD	76	76	76
	Ensemble	**80**	80	**80**

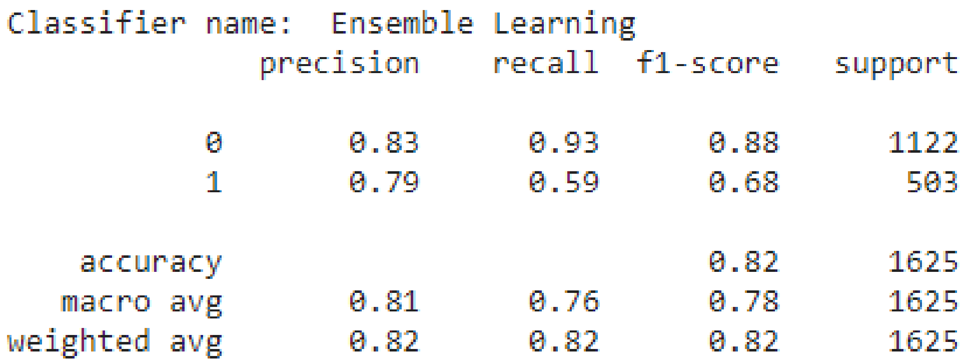
79% f1-score for RF while DS2 shows maximum of 72% for RF also. But in BSaD our proposed ensemble technique outperformed DS1 and DS2 with 80% f1-score. For uni-gram+bi-gram combination ensemble shows best f1-score of 80% and 73% respectively for DS1 and DS2 while in BSaD we get f1-score of 81% for SGD classifiers. Finally for uni-gram + bi-gram + tri-gram combination our peoposed method ensemble technique shows best performance for all dataset. That is 81% for DS1, 74% for DS2 and 82% for BSaD.

4.4.3 Error Analysis

From Table 4.4 we can see that tf-idf+ensemble with uni-gram +bi-gram + tri-gram features performs best for our dataset. Using confusion matrix a details error analysis is explained. Figure 4.6 depicts a class-wise distribution of the number of expected labels. For example, among 1078 negative samples 78 of

Table 4.6: Comparison among datasets for tf-idf features

Dataset	**Feature**	**Classifiers**	**Pr. (%)**	**Re. (%)**	**F1. (%)**
DS1	Uni-gram	LR	75	75	75
		RF	80	79	79
		SVM	76	76	76
		SGD	75	74	74
		Ensemble	77	77	77
	Uni-gram+ Bigram	LR	78	78	78
		RF	80	79	79
		SVM	78	78	78
		SGD	76	76	76
		Ensemble	81	80	80
	Uni-gram +Bi-gram+ Trigram	LR	78	78	78
		RF	80	79	79
		SVM	80	79	79
		SGD	79	78	78
		Ensemble	81	81	81
DS2	Uni-gram	LR	71	71	71
		RF	72	72	72
		SVM	71	71	71
		SGD	70	69	68
		Ensemble	72	71	71
	Uni-gram+ Bi-gram	LR	72	72	72
		RF	71	71	71
		SVM	72	72	72
		SGD	70	65	64
		Ensemble	73	73	73
	Uni-gram +Bi-gram+ Tri-gram	LR	73	73	73
		RF	72	71	71
		SVM	73	73	73
		SGD	73	72	72
		Ensemble	74	74	74
BSaD	Uni-gram	LR	79	79	77
		RF	80	79	75
		SVM	79	79	79
		SGD	80	77	72
		Ensemble	80	81	**80**
	Uni-gram+ Bi-gram	LR	81	80	77
		RF	80	78	75
		SVM	80	80	80
		SGD	81	82	**81**
		Ensemble	80	80	78
	Uni-gram +Bi-gram+ Tri-gram	LR	82	80	78
		RF	80	78	75
		SVM	80	81	80
		SGD	82	82	81
		Ensemble	**82**	**82**	**82**



```
Classifier name:  Ensemble Learning
              precision    recall  f1-score   support

           0       0.83      0.93      0.88      1122
           1       0.79      0.59      0.68       503

    accuracy                           0.82      1625
   macro avg       0.81      0.76      0.78      1625
weighted avg       0.82      0.82      0.82      1625
```

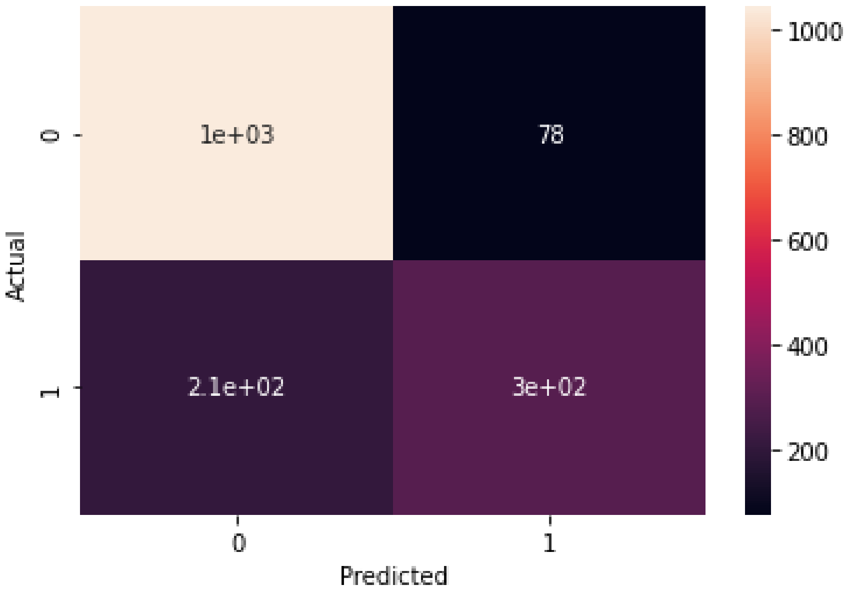


Figure 4.6: Confusion matrix of the proposed model.

them are misclassified as positive. And for the positive, among 510 samples 210 of them are classified as negative. For positive data our model performed not that well because we used very few positive data to train our model than negative data. This can be easily overcomed by adding more positive data to train the model. As a result, creating a balanced dataset with diverse data can help to reduce incorrect predictions to some degree.

4.5 Conclusion

In this chapter we got an overview on the dataset used to train the model as well as the impact i.e. social, environmental and ethical impact of the thesis and evaluation of performance and framework. Finally we can conclude that with 82% accuracy our proposed tf-idf+ensemble classifier with the combination of uni-gram + bi-gram + tri-gram features outperforms all other classifiers.

Chapter 5

Conclusion

5.1 Conclusion

To recognise emotions from text, various Machine Learning techniques along with ensemble learning are used. Symbolic methods are more complex and time-consuming than Machine Learning techniques. So we used Machine Learning techniques and apply 8 different methods to classify the sentiment. We implemented both if-idf and BoW featue extraction techniques.And discovered that the integration of tf-idf+ensemble with the conjunction of uni-gram + bi-gram + tri-gram features is the best feature of the proposed framework. We also took account in uni-gram and uni-gram + bi-gram to extract features and got the best accuracy of 82% for tf-idf+ensemble with uni-gram + bi-gram + tri-gram features. Ensemble technique also performs best for uni-gram features with 80% f1-score. But for the combination of uni-gram + bi-gram features SGD performs best with 82% accuracy and 81% f1-score. For Bow ensemble outperformed all other classifier with 80% accuracy.

5.2 Future Work

The main goal of our project was to use supervised machine learning to create a system for detecting polar texts. We can use this to determine whether a text is positive or negative. Our system can be improved in different areas

- Enrich the dataset by including more data at corpus.
- We will use neural network approach to find semantic relation between text which will help us to predict more accurately.

- Our future efforts will be the focus of implementing more feature extraction techniques.
- We can improve our performance by implementing more pre-processing methods like POS tagging, stemming etc.

References

[1] W. Medhat, A. Hassan and H. Korashy, ‘Sentiment analysis algorithms and applications: A survey,’ *Ain Shams engineering journal*, vol. 5, no. 4, pp. 1093–1113, 2014 (cit. on p. 1).

[2] ‘Banglapidia. bangla language.available from: http://en.banglapedia.org/index.php?title=bangla$_{l}anguage$,’ 2019 (cit. on p. 2).

[3] ‘Online bangla newpaper list. http://onlinebanglanewspaperlist.com/ article$_{d}etails.php?id = 18$,’ (cit. on p. 2).

[4] A. Montejo-Ráez, E. Martnez-Cámara, M. T. Martn-Valdivia and L. A. Ureña-López, ‘Ranked wordnet graph for sentiment polarity classification in twitter,’ *Computer Speech & Language*, vol. 28, no. 1, pp. 93–107, 2014 (cit. on p. 2).

[5] T. Pranckeviius and V. Marcinkeviius, ‘Application of logistic regression with part-of-the-speech tagging for multi-class text classification,’ in *2016 IEEE 4th Workshop on Advances in Information, Electronic and Electrical Engineering (AIEEE)*, IEEE, 2016, pp. 1–5 (cit. on p. 12).

[6] B. Ren and L. Cheng, ‘Research of classification system based on naive bayes and metaclass,’ in *2009 Second International Conference on Information and Computing Science*, IEEE, vol. 3, 2009, pp. 154–156 (cit. on p. 13).

[7] L. Bottou, ‘Stochastic gradient descent tricks,’ in *Neural networks: Tricks of the trade*, Springer, 2012, pp. 421–436 (cit. on p. 15).

[8] T. Nasukawa and J. Yi, ‘Sentiment analysis: Capturing favorability using natural language processing,’ in *Proceedings of the 2nd international conference on Knowledge capture*, 2003, pp. 70–77 (cit. on p. 16).

[9] R. Prabowo and M. Thelwall, ‘Sentiment analysis: A combined approach,’ *Journal of Informetrics*, vol. 3, no. 2, pp. 143–157, 2009 (cit. on p. 16).

[10] V. K. Singh, R. Piryani, A. Uddin and P. Waila, ‘Sentiment analysis of movie reviews: A new feature-based heuristic for aspect-level sentiment classification,’ in *2013 International Mutli-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s)*, IEEE, 2013, pp. 712–717 (cit. on p. 17).

[11] T. P. Sahu and S. Ahuja, ‘Sentiment analysis of movie reviews: A study on feature selection & classification algorithms,’ in *2016 International Conference on Microelectronics, Computing and Communications (MicroCom)*, IEEE, 2016, pp. 1–6 (cit. on p. 17).

[12] P. Bhoir and S. Kolte, 'Sentiment analysis of movie reviews using lexicon approach,' in *2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, IEEE, 2015, pp. 1–6 (cit. on p. 17).

[13] G. Gautam and D. Yadav, 'Sentiment analysis of twitter data using machine learning approaches and semantic analysis,' in *2014 Seventh International Conference on Contemporary Computing (IC3)*, IEEE, 2014, pp. 437–442 (cit. on p. 17).

[14] E. Hossain, O. Sharif and M. M. Hoque, 'Sentiment polarity detection on bengali book reviews using multinomial naive bayes,' *arXiv preprint arXiv:2007.02758*, 2020 (cit. on p. 17).

[15] Y. Zhang, R. Jin and Z.-H. Zhou, 'Understanding bag-of-words model: A statistical framework,' *International Journal of Machine Learning and Cybernetics*, vol. 1, no. 1-4, pp. 43–52, 2010 (cit. on p. 23).

[16] T. Tokunaga and I. Makoto, 'Text categorization based on weighted inverse document frequency,' in *Special Interest Groups and Information Process Society of Japan (SIG-IPSJ*, Citeseer, 1994 (cit. on p. 23).

[17] D. Magatti, S. Calegari, D. Ciucci and F. Stella, 'Automatic labeling of topics,' in *2009 Ninth International Conference on Intelligent Systems Design and Applications*, IEEE, 2009, pp. 1227–1232 (cit. on p. 27).

[18] J. Cohen, 'A coefficient of agreement for nominal scales,' *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960 (cit. on p. 27).

[19] S. Taher, K. Akhter and K. M. Hasan, *Bangla dataset for opinionmining*, Sep. 2018. DOI: 10.13140/RG.2.2.20214.96327 (cit. on p. 33).

[20] 'Banglapidia. bangla language.available from: Https://www.kaggle.com/tazimhoque/bengali-sentiment-text,' 2019 (cit. on p. 33).