

**Bangla Sentiment Analysis on Cricket**

**Comments BY**

**Papya Sultana**

**ID:162-15-**

**8207**

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

Supervised By

**Mr. Md. Jueal Mia**

Lecturer

(Senior Scale)

Department of CSE

Daffodil International University

Co-Supervised By

**Mr. Abdus Sattar**

Assistant Professor & Coordinator

M.Sc Department of CSE

Daffodil International University



**DAFFODIL INTERNATIONAL UNIVERSITY**

**DHAKA, BANGLADESH**


**October 20**

## **APPROVAL**

This Project/internship titled “**Bangla Sentiment Analysis on Cricket Comments**”, submitted by Papya Sultana, ID No: 162-15-8207 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 6/1/2022.

## **BOARD OF EXAMINERS**

**Chairman**



---

**Dr. Touhid Bhuiyan**  
**Professor and Head**

Department of Computer Science and Engineering  
Faculty of Science & Information Technology  
Daffodil International University



---

**Zahid Hasan (ZH)**  
**Associate Professor**

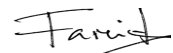
Department of Computer Science and Engineering  
Faculty of Science & Information Technology  
Daffodil International University



---

**Mohammad Monirul Islam (MMI)**  
**Senior Lecturer**

Department of Computer Science and Engineering  
Faculty of Science & Information Technology  
Daffodil International University



---

**Dr. Dewan Md. Farid**  
**Professor**

Department of Computer Science and Engineering  
United International University

**Internal Examiner**

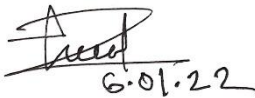
**Internal Examiner**

**External Examiner**

## DECLARATION

We hereby declare that this project has been done by us under the supervision of **Mr. Md. Jueal Mia, Lecturer, and Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

### Supervised by:



6.01.22

---

**Mr. Md. Jueal Mia**  
Lecturer  
(Senior Scale)  
Department of CSE  
Daffodil International University

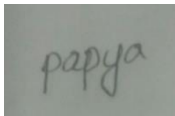
### Co-Supervised By:



---

**Mr. Abdus Sattar**  
Assistant Professor & Coordinator M.Sc  
Department of CSE  
Daffodil International University

### Submitted by:



---

**Papya Sultana**  
**ID:162-15-**  
**8207**  
Department of CSE  
Daffodil International University

## ACKNOWLEDGEMENT

First we express our heartiest thanks and gratefulness to almighty Allah for His divine blessing makes us possible to complete the final year project successfully.

I would like to express my sincere gratitude to my honorable project supervisor **Mr. Md. Jueal Mia, Lecturer**, Department of CSE Daffodil International University, Dhaka, for his valuable advices, constructive suggestions and sincere guidance with all the necessary facilities for assimilation, research and preparation for the project.

We would like to express our heartiest gratitude to **Mr. Md. Jueal Mia, Lecturer**, Department of CSE, **Mr. Abdus Sattar, Assistant Professor & Coordinator M.Sc** , Department of CSE, and **Professor Dr. Touhid Bhuiyan, Head**, Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

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

I would like to thank my family for their constant love and support. Finally, I would like to take this opportunity to express my gratitude to one and all, who directly or indirectly, have lent their hand in this venture.

## **ABSTRACT**

What is the space between the countries of the world from south to north, east to west of the world? If you get the answer to this query from a modern point of view, you will see that there is no distance at all. Right now, somebody is receiving all kinds of news reports from around the world in just a few seconds. And it has only happened because of the visual media. Online news sites indeed publish news live, but it is disappointing that users do not like all kinds of news published on the news site. At the time, there was a need for a platform that could easily identify user preferences in the news and publish only according to their priority. Dividing stories by user choice requires researching news text. A lot has been done in English news at the moment but there are very limited functions in Bangla news. Because of this, even though Bangla is one of the world's eight most frequently spoken languages, we should continue our investigation. We're using Bangla news articles pulled from a database for our study. From furtherance of storytelling, we are trying to implement all kinds of text-separating processes using the Linear SVM and Random Forest Reading Machines. Finally, we create an interface to capture news stories and show the category of those stories.

## TABLE OF CONTENTS

<b>CONTENTS</b>	<b>PAGE</b>
Board of examiners	i
Declaration	ii
Acknowledgements	iii
Abstract	iv
<b>CHAPTER</b>	
<b>CHAPTER1: INTRODUCTION</b>	<b>1-4</b>
1.1 Introduction	1
1.2 Objectives	2
1.3 Motivation	2
1.3 Rationale of the Study	3
1.4 Research Questions	3
1.5 Expected Output	3
1.6 Report Layout	4
<b>CHAPTER 2: BACKGROUND</b>	<b>5-6</b>
2.1 Introduction	5
2.2 Related Works	5
2.3 Research Summary	6
2.4 Challenges	6

<b>CHAPTER 3: RESEARCH METHODOLOGY</b>	<b>7-11</b>
3.1 Introduction	7
3.2 Research Subject and Instrumentation	7
3.3 Data Collection Procedure	7
3.4 Data Pre Processing	7
3.5 Work Flow of Identifying News Category	8
3.6 Implementation Requirements	10
<b>CHAPTER 4: EXPERIMENTAL RESULTS AND DISCUSSION</b>	<b>12-17</b>
4.1 Introduction	12
4.2 Raw Data	12
4.3 Cleaning Raw Data	12
4.4 Creating Input File	13
4.5 Excluded Words Removal	13
4.6 Features Selection and Extraction	13
4.7 Building Model and Fit dataset for classifier	14
4.8 Expected Result	15
4.9 Accuracy of Model	15
4.10 Summary	16
<b>CHAPTER 5: SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH</b>	<b>18-18</b>
5.1 Summary of the Study	18
5.2 Conclusions	18
5.3 Recommendations	18

5.4 Implication for Further Study	18
<b>REFERENCES</b>	<b>19-20</b>
<b>APPENDIX</b>	<b>21</b>
<b>Plagiarism Report Screenshot</b>	<b>22</b>



## **LIST OF FIGURES**

<b>FIGURES</b>	<b>PAGE NO</b>
Figure 3.5.1: Show the excluded Bangla word.	8
Figure 3.5.2: Proposed Working Flow chart for classification.	10
Figure 4.2.1: Experimental raw data.	12
Figure 4.2.2: Tab separated Bangla text.	13
Figure 4.7.1: Dataset chart ratio.	14
Figure 4.8.2: Experimental output of Bangla Sentiment Analysis.	15

## **LIST OF TABLES**

### **FIGURES**

### **PAGE NO**

Table 4.9.1: Precision, Recall, F1-Score for Support Vector Machine.	15
Table 4.9.2: Precision, Recall, F1-Score for Random Forest.	16
Table 4.9.3: Compare Precision of all classifier.	16
Table 4.9.4: Compare Recall of all classifier.	17
Table 4.9.5: Compare Recall of all classifier.	17

# CHAPTER 1

## INTRODUCTIO

### N

#### 1.1 Introduction

The emotional analysis is a field of study to analyze the views of people, feelings, and emotions of various sources such as developments, organizations, benefits, events, colonial issues. When we know what individuals enjoy and hate, we may better understand what we can do to improve. Any kind of user feedback, whether it comes from social networks, websites, or contact center employees, is a goldmine of information. It's not enough to understand what other people are saying, though. We need to know what they're thinking. These sensations can be obtained through an emotional analysis. Taking apart and comparing ideas stated in a piece of writing, especially in order to evaluate if a person's standing in connection to a certain topic, product, etc. is a type of analysis. We are good, not bad, or we are neutral. While mechanical power cannot give a logical split of emotions, mathematical analysis can provide a thorough grasp of natural language.

Cricket is a religion in our nation. In other words, their perspectives on this game are very different. As in world of cricket, there is a constant exchange of ideas and opinions. As a result, we've had a lot of fun exploring the thoughts and sentiments of actual cricket fans in this area. All in all, our data is structured in Bengali emotions since individuals express their feelings about cricket with their own language. However, the absence of Bengali Language Processing tools makes it difficult to operate in the Bengali language.

Cricket is an emotion in our nation. As a result, they had quite varied opinions on the game in general. Cricketers frequently display a wide range of emotions during the course of a match in a variety of ways. We had a lot of fun analyzing the sentiments of genuine cricket fans throughout this period. Overall, our data is grouped in Bengali sentiments as people who express their sentiments on cricket in their local tongues, which is a strength of our study. However, the paucity of resources for Bengali Language Processing makes working in Bengali a problem.

#### 1.2 Objectives:

- To investigate the process of classifying or categorizing Bangla news using a classifier algorithm.
- To design a platform capable of classifying provided Bangla news.
- To visualize some analytical analysis of Bangla News classification classified by classifier algorithms.

#### 1.3 Motivation

We see that the news portals publish all sorts of news. That means all absolute news is being published in

these news portals. But all people do not prefer all sorts of news. Some people prefer to read sports news most than political news. Some people like to read political news than other news. Some people like to read enjoyment news. It depends on one's choice. But sometimes, it has become so much boring to see the news that is not preferred by the user. The news portal becomes the most efficient if it shows the news according to the specific user's choice. But, for this, the first task is to identify the news variety. We find lots of tasks in the news category in English. But there is very poor work on Bangla. If Bangla news classification gets some research works on it, it can be used in many real-life applications.

Cricket is an emotion in our nation. As a result, they had various feelings about the game. Cricketers frequently display a wide range of emotions during the course of a match in a variety of ways. We had a lot of fun analyzing the sentiments of genuine cricket fans throughout this period. When asked in Bengali how they felt about cricket, the vast majority of our respondents responded in Bengali emotions. However, due to a lack of resources for Bengali Language Processing, working in Bengali proved difficult. Our interest in this type of research-based activity was piqued by this. Machine learning and data mining are two of the main approaches we use in our work.

#### **1.4 Rationale of the Study**

Natural Language Processing (NLP) is a wide sector in research specially in English and these approaches or the processes are being used in many automated system as well as robotics system. But, Natural Language Processing on Bangla is very rare. To develop more automated application or make much more efficient of Machine Learning approaches in Bangla, there has no alternative to work with Bangla text. This made us to be interested to work with this Bangla Sentiment Analysis on Cricket Comments. In the present time, we see that the notepad editors are much more intellectual. These have

some features like auto corrections, grammar checking, auto suggestion etc. These features are the outcome of the blessing of Natural Language Processing. But these features are mostly seen for English. Such kind of features is very rare for Bangla text. These, actually, take us to work with Bangla news as well as Text.

### **1.5 Research Question**

- Can we have collected raw data of Bangla News?
- How to pre-process the raw data for the Machine Learning approaches?
- Can Linear SVM Classifier algorithm be use on the pre-processed data?
- Is the Machine Learning technique capable of appropriately detecting or classifying the provided Bangla dataset?

### **1.6 Expected Output**

It is expected that this research-based project would result in the creation of an algorithm or a comprehensive efficient technique for categorizing provided Bangla news based on the created model of training dataset that was developed.

### **1.7 Report Layout**

The report will be followed as follows-

Chapter 1 provides the summary of this research-based project. Introductory discussion is the key term of this first chapter. Apart from, what motivated us to do such a research-based project is explained well in this chapter to. The most important part of this chapter is the Rationale of the Study. Then, what are the research questions and what is the expected outcome is discussed in the last section of this chapter.

Chapter 2 summarizes the previous discussion on this report. The extent of this category is shown in the final portion of this second chapter. It's time to discuss the barriers or challenges that this study faces.

For the most part, Chapter 3 is a theoretical exploration of the findings presented in Chapter 2. This chapter covers the mathematic methodologies used in this study in order to discuss a portion of the research hypothesis. In addition, the Linear SVM machine learning technique is described in detail in this chapter. A complicated matrix analysis is provided at the end of this chapter to validate the model and illustrate its precision label. Chapter 4 summarizes the findings of the whole study. This chapter includes a few unique photographs to help draw attention to the project.

Chapter 5 is based on the project title conclusion. This chapter is responsible for showing the entire project report in line with the recommendation. The chapter concludes with an overview of the potential of our work, which

may be the future of others who want to work in this field.

## CHAPTER 2

### BACKGROUND

#### D

##### 2.1 Introduction

This chapter is based on past study in this topic and represents the findings of previous researchers. The limitations of these works will also be illustrated in this chapter, which also serves as a good representation of the extent of our investigation and the difficulties it presented.

##### 2.2 Related Works

Some of our work is based on previous research in these fields, and some of it is for our benefit. From [10], where they attempted to define additional classes such as coarseness and emotion, in our view has indeed been greatly revived. As well as, one of the most impressive activities we've witnessed [18]. Text messages are a source of inspiration for them. TF-IDF and SVM segmentation were used to improve class accuracy. It's the same approach that we used in our first research effort. As both a text representation model, those who applied the Vector Space Model (VSM). Empty Bayes and genteel distinctions have been proposed in this paper [9] to differentiate tweets from positive, negative, or neutral behavior. By focusing on a text's overall feeling rather than a specific topic, [12] explores the challenge of classifying texts. Upon that identification of insults and flames, they carried out their emotional analysis in [8]. Having this in mind, we've chosen and taken action on feedback from our Bangladesh Cricket class in our database. Bengali's language position on Twitter is being used to assess whether or not Bengali text is large enough. For POS Tagging and Support Vector Machine and Maximum Entropy, they have created the Bangla Pos-Tagger Package for POS Tagging and experimented with various feature sets. We're excited to continue working with POS-Tagger in the future. SentiWordNet can be created in Bengali, Hindi, or Telugu using a variety of computer techniques, including those based on WordNet, dictionaries, or other organizational methods [2]. And in the report [16], they aimed to automatically remove emotions or bonds using HMM to perform POS tagging and SVM editing. And in the paper [16], they used HMM to do POS tagging and SVM editing to eliminate emotions or relationships. As well as working [17] to elicit emotions like positive and negative movie reviews based on data from 2000. They employed the TF-IDF and the weka tool's Support Vector Machine to classify their data. This feature set includes unigrams, bigrams, POS tags of words, and words

with function. Mixed Bengali English gold data related to coding and linguistic and a valence tag for emotional analysis was developed in paper [14] and described ideas they had to gather and filter raw Twitter data. Another recognized the the paper followed [7] where they did the digging of ideas and the expressing of sentiments in which they distinguished the unity of the text as the good, the bad, and the neutral. [5] completed research on emotional analysis that analyzes the classification of text in an opinion mine. Apart from [19], textual data analysis was performed for emotional analysis. [3] In this paper, researchers compared the attitudes expressed in English and Bengali literature and drew similar conclusions. We've covered a lot of ground when it comes to Bengali linguistics in these volumes. In addition, [11] worked on Twitter microblogging data and identified good, negative, and neutral data from the tweets they retrieved. They also divided the font size in [6] to focus on café reviews. [13] employed a variety of machine learning approaches, including Naive Bayes and Highest Entropy Models, to analyze Twitter micro-blogging. In addition, we combed through social media and the website Prothom-Alo to see what the general public thinks about Bangladesh cricket. As a result of our research, we've decided to separate our data using TF-IDF Vectorizer and SVM.

### **2.3 Research Summary**

The above discussion done on various types of research works from different research teams, it is being appeared to us that recently, research work on Bangla text is increasing day by day. Some good outcomes already prove this statement well. Though, enough resources are not present, but hope is that this field is becoming more resourceful each after passing a single day.

### **2.4 Challenges**

The main challenges of this work are dealing with the datasets. To clean the dataset, we need some efficient approaches to perform it but there are not enough recognized approaches to do it. Another challenge of this work is not having enough resources regarding this topic.

## CHAPTER 3

### RESEARCH METHODOLOGY

#### 3.1 Introduction

This chapter focuses on the research's forward-looking insights. It'll help you have a better grasp on the idea of labor. Research Topic and Instrumentation are specified succinctly initially to make things more obvious. When we look at the data acquisition process or machine learning, we see that data are at the core. Because of this, this section explains how data is gathered. A description of our project's statistical methodologies and a clear picture of the performance requirements round up this chapter.

#### 3.2 Research Subject and Instrumentation

In other words, a research topic is anything that is being investigated in depth in order to get a deeper knowledge of. The analysis subject is responsible for more than just producing a clear performance; they must also deliver the correct learning of many inquiry parameters. Instrumentation, on the other hand, refers to the tools and equipment that researchers need to carry out their work.

#### 3.3 Data Collection Procedure

Data is the best and fastest way to conduct research in a certain topic. The essence of machine learning is, in fact, the collection of data. And data is the only option available to us in our investigation. As a result, it became our most difficult research challenge. Prothom Alo, Bangladesh's most popular Bangla-language news site, serves as our primary source of information. Our Bangla news is collected from this site by using corpus. We collected almost 4 years news from them. And the news is stored as text document format.

#### 3.4 Data Pre-Processing

The pre-processed data is critical when working with row data. When data is pre-processed more effectively, the end result is more accurate. When it comes to research-based work, this is the first hurdle to overcome. There are certain HTML tag names in our table data. Because of this, it must be published in print. This was our first response to remove all HTML tag names from the news text. Then, we have to maintain some for bathing the excessive space from the document. Then it removes an all-new line to place it into a line. That means, after gathering any news file, every line will be treated as a piece of news. Then, lastly, for every individual news, this script allocates a number for determining a category.

Finally, after allocating a specific number for each news, this script produces a tsv file formatted file that is tab-separated. This tsv file is, actually, our pre-processed data with its category. Thus, all six actual news





### **Building Model:**

Finally, we're ready to create our model. And we're able to do this because we've been working with our machine to teach it. We divided our data set in thirds, resulting in a final ratio of 3:1. All of our data set is used for training purposes, and the remainder is used for testing purposes. There are 75% of datasets utilized for training, while the remaining 25% are used for testing purposes.

### **Classifier Fitting:**

At this point, our machine is classifier-ready. To identify our news content, we employ Support Vector Machine (SVM) and Random Forest (RF). All we have to do is import it and put it together.

### **Predict the Category**

This is the last step in our categorization process for breaking news. Bangla text input is being readied for testing in this stage of our model's development. This model is able to categorize the provided input text using two classifiers, such as SVM and Random Forest, based on the supplied input text.

## Flow Chart:

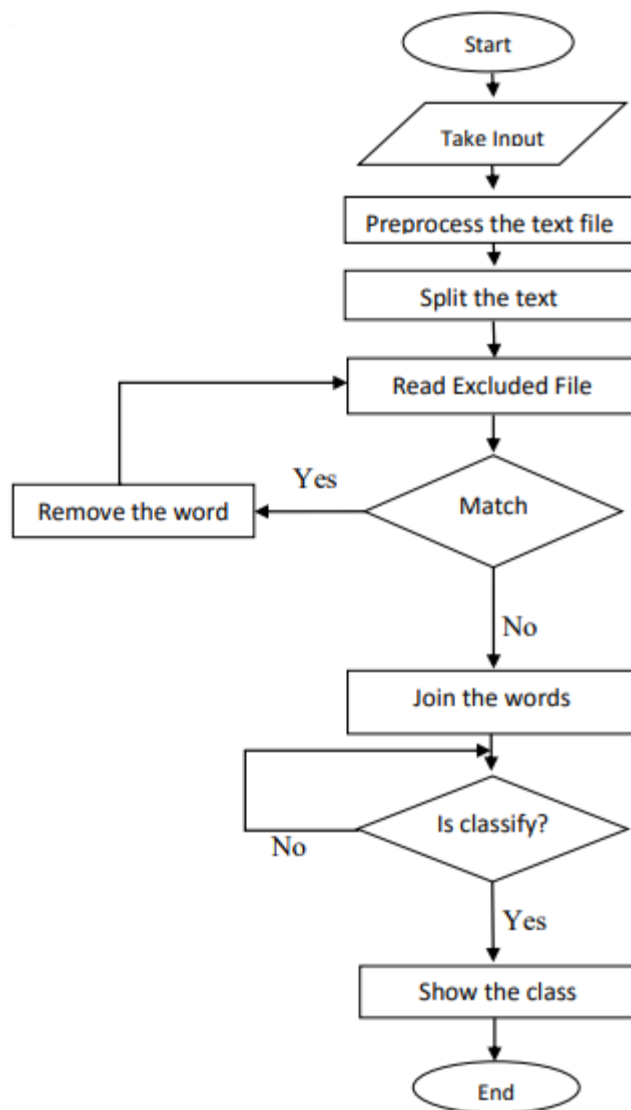


Figure 3.5.2: Proposed Working Flow chart for classification.

### 3.6 Implementation Requirements

A list of requirements for a Bangla News Classification has been compiled after an in-depth examination of all relevant statistical and theoretical ideas and approaches. The following are likely necessities:

## **Hardware/Software Requirements**

- Operating System (Windows 7 or above)
- Hard Disk (minimum 4 GB)
- Ram (more than 1 GB)
- Web Browser (preferably chrome)

## **Developing Tools**

- Python Environment
- Jupyter (Anaconda3)
- Notepad++

# CHAPTER 4

## EXPERIMENTAL RESULTS AND DISCUSSION

### 4.1 Introduction

Throughout this chapter, we'll be analyzing both the data we utilized in our research and our project's findings.

### 4.2 Raw Data

Prothom Alo, Bangladesh's most popular news site, provides the raw data for our analysis. Corpus is the tool we use to gather all of our data. Once the information has been gathered, it is then saved as a text document file. These files include data tagged with an html tag name. Our raw data looks like:



Figure 4.2.1: Experimental raw data.

Therefore, it has become obvious to clean the data. That means pre-processed the row data for preparing for the model.

### 4.3 Cleaning Raw Data

Our data preprocessing activity is made easier by the usage of a script file. The following are the responsibilities of the python script:

- i. Remove all instances of the html tag name.
- ii. Eliminate extraneous spaces from the text.
- iii. Remove all new lines from each news item and group them together in a line.
- iv. Assign an integer number for pre defining the category of each news.

This script result in a Tab Separated Value (tsv) formatted file and it looks like:



Figure 4.2.2: Tab separated Bangla text.

Actually, by this process, we can get all our categorical news in individuals file but the outputted file data are pre-processed and categorical.

#### 4.4 Creating Input File

After the data cleaning phase, we get six absolute tsv files as we are working on this investigation on these classes. Hence, after successfully preprocessing process, there has this categorical news file in our hand. Then, to perform Natural Language Process on Bangla news, we must join all these files into a file. For this, we use another python script named join.py. This file takes the folder name that includes all tsv files as an input and delivers only a file where all news possessed individually is merged..

#### 4.5 Excluded Words Removal

Code written in Python is used to categorize a news item into several types. To begin developing a prototype, we've consolidated all the relevant information in one place. This necessitates some preparatory work. We compile a list of Bangla terms unrelated to the story's subject matter. The list was labeled "Banned words" and given that moniker. Our input file is being checked to see whether any of the terms we've omitted are there. It is imperative that any such thing be ruled out.

#### 4.6 Feature Selection and Extraction

The selection and extraction of features is the most important aspect of the classifying strategy in this step. When it comes to categorization, it really makes the final decision. Our feature selection is based on word count and we build it.

## 4.7 Building Model and Fit Dataset for Classifier

We split our dataset in two to begin the modeling process.

- Training Dataset
- Testing Dataset

When constructing our model, we adopt a 3:1 ratio. The three-part data set is used for training purposes, while the remaining component will be used for testing purposes.

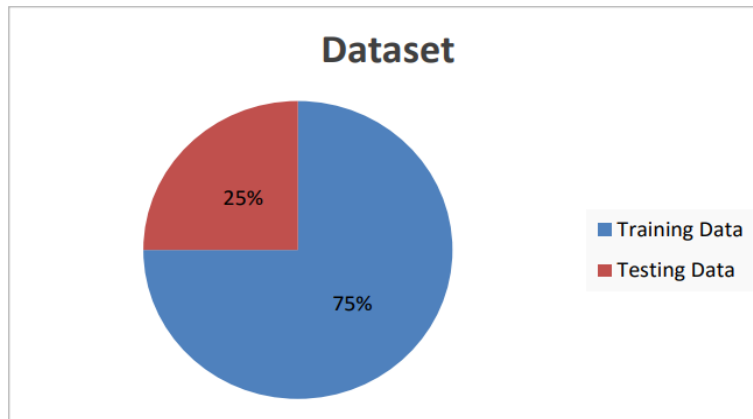


Figure 4.7.1: Dataset chart ratio.

In the concept of percentage, 75% data will be for training and 25% will be for testing. And this will make our expected model, we import the sklearn package because we're working with many classifiers. This classifier is capable of generating an integer indicating the predicted news category.

## 4.8 Experimental Result

After completing the Bangla Sentiment Analysis on Cricket Comments, User Interface shown in figure 4.8.1 It is an experimental input field where user can produce any kind of Bangla news text.

```
Test set:      Acc  Precision      Recall  F1
71.58  65.55  71.58  66.70

          precision    recall  f1-score   support

   negative      0.75      0.92      0.83      273
   neutral      0.25      0.03      0.05       34
   positive      0.49      0.29      0.36       73

 accuracy
macro avg      0.50      0.41      0.41      380
weighted avg   0.66      0.72      0.67      380

time taken:
0.20356082916259766
```

```
# ইমরুল বাদে বাকি তিনজনের আউট মেনেনিতে পারছিনা ? negative
test = ["ইমরুল বাদে বাকি তিনজনের আউট মেনেনিতে পারছিনা ?"]
test_tf = tfidf_vectorize.transform(test)
```

Figure 4.8.2: Experimental output of Bangla Sentiment Analysis.

#### 4.8 Accuracy of Model

We've created a Confusion Matrix for our model, which is a way to summarize the performance of the classifier. Just looking at classification accuracy might be deceptive if your dataset has an unequal number of observations for each class or if there are more than two classifications.

#### For Support Vector Machine

Table 4.9.1: Precision, Recall, F1-Score for Support Vector Machine.

Class Name	Precision	Recall	F1-Score
negative	0.75	0.92	0.83
neutral	0.25	0.03	0.05
positive	0.49	0.29	0.36



Accuracy of Model:

Total News = 6530

Testing News (25%) = 1632.5

Accuracy of this model =  $(1175 / 1632.5) * 100$   
= 71.58%

### For Random Forest

Table 4.9.2: Precision, Recall, F1-Score for Random Forest.

Class Name	Precision	Recall	F1-Score
negative	0.73	0.93	0.82
neutral	0.00	0.00	0.00
positive	0.41	0.18	0.25

Accuracy of Model:

Total News = 6530

Testing News (25%) = 1632.5

Accuracy of this model =  $(1145 / 1632.5) * 100$   
= 70.26%

### Compare Algorithms

Table 4.9.3: Compare Precision of all classifier.

Algorithms	Accuracy
Support Vector Machine	0.75
Random Forest	0.73

Table 4.9.4: Compare Recall of all classifier.

Algorithms	Accuracy
Support Vector Machine	0.92
Random Forest	0.93

Table 4.9.5: Compare Recall of all classifier.

Algorithms	Accuracy
Support Vector Machine	0.83
Random Forest	0.82

In terms of Precision, Recall, f1-score, and accuracy, we can observe that the Support Vector Machine classifier is the most accurate. The precision, recall, and f1-score of this model are all above average, at 0.75, 0.92, and 0.83, respectively, and the accuracy is 72.58 percent, the best of any classifier tested.

#### **4.9 Summary**

Since we've already achieved this degree of accuracy, we're pleased, but if you want to enhance the level of accuracy, you'll need to correctly prepare the dataset. There should be an equal number of stories for each of the several categories. At this point, there is no option to data cleansing in order to improve the accuracy of the results. This classifier's predictions get more precise as more data is preprocessed.

## **CHAPTER 5**

### **SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH**

#### **5.1 Summary of the Study**

On the subject of natural language processing, there is no doubt that there is a great deal of interest. Such study is being extended as a result of the significant changes in our digital lives that have resulted from such work recently. As a result of such research, we've seen some very remarkable results in the real world. However, the absence of comparable research into the Bangla language is deeply regrettable. However, the fact that several scholars from various nations have begun to investigate this area gives us reason for optimism. As part of our research, we employ a variety of techniques from our Bangla News to classify it.

#### **5.2 Conclusion**

No matter how inaccurate the classification method we utilized in our study was, we still gained enough from this investigation. The Bangla Text may now be dealt with. It is possible to do preprocessing on the raw data directly. And we may use our previously trained dataset to train the classifier. I am hopeful that this type of research on Bangla Text or Bangla news would be extremely beneficial to future experimenters.

#### **5.3 Recommendations**

The following are a few of the best ways to do this:

- In order to get the most out of this study, it's important to streamline the data collection process.

#### **5.4 Implication for Further Study**

- This project might be made more effective by include other categories.
- As additional classifiers are applied to this dataset, it becomes clearer which classifier is best suited for this task.

## References

- [1] S. Chowdhury and W. Chowdhury. Performing sentiment analysis in bangla microblog posts. In 2014 International Conference on Informatics, Electronics & Vision (ICIEV), pages 1–6. IEEE, 2014.
- [2] A. Das and S. Bandyopadhyay. Sentiwordnet for indian languages. In Proceedings of the Eighth Workshop on Asian Language Resources, pages 56–63, 2010.
- [3] D. Das. Analysis and tracking of emotions in english and bengali texts: a computational approach. In Proceedings of the 20th international conference companion on World wide web, pages 343–348. ACM, 2011.
- [4] G. Diaz. (2018, Oct.) Bengali stopwords. [Online]. Available: <https://github.com/stopwords-iso/stopwords-bn>.
- [5] D. M. E. D. M. Hussein. A survey on sentiment analysis challenges. Journal of King Saud University-Engineering Sciences, 2016.
- [6] H. Kang, S. J. Yoo, and D. Han. Senti-lexicon and improved naïve bayes algorithms for sentiment analysis of restaurant reviews. Expert Systems with Applications, 39(5):6000–6010, 2012.
- [7] A. Kaur and V. Gupta. A survey on sentiment analysis and opinion mining techniques. Journal of Emerging Technologies in Web Intelligence, 5(4):367–371, 2013.
- [8] A. Mahmud, K. Z. Ahmed, and M. Khan. Detecting flames and insults in text. 2008.
- [9] R. Mehra, M. K. Bedi, G. Singh, R. Arora, T. Bala, and S. Saxena. Sentimental analysis using fuzzy and naive bayes. In Computing Methodologies and Communication (ICCMC), 2017 International Conference on, pages 945–950. IEEE, 2017.
- [10] S. M. Mohammad. Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In Emotion measurement, pages 201–237. Elsevier, 2016.
- [11] A. Pak and P. Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In LREC, volume 10, pages 1320–1326, 2010.
- [12] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL02 conference on Empirical methods in natural language processing Volume 10, pages 79–86. Association for Computational Linguistics, 2002.
- [13] R. Parikh and M. Movassate. Sentiment analysis of user-generated twitter updates using various classification techniques. CS224N Final Report, 118, 2009.
- [14] B. G. Patra, D. Das, and A. Das. Sentiment analysis of code-mixed indian languages: An overview of sail code-mixed shared task@ icon2017. arXiv preprint arXiv:1803.06745, 2018.
- [15] A. Rahman. (2018, Oct.) Bengali ABSA dataset. [Online]. Available: <https://github.com/AtikRahman/Bangla-Datasets-ABSA>.
- [16] A. Roy and A. A. Singh. (2018, Oct.) Sentiment Analysis ANLP Research Report. [Online]. Available: <https://github.com/abhie19/Sentiment-Analysis-Bangla-Language/>.

- [17] P. H. Shahana and B. Omman. Evaluation of features on sentimental analysis. *Procedia Computer Science*, 46:1585–1592, 2015.
- [18] J. D. Silva and P. S. Haddela. A term weighting method for identifying emotions from text content. In *Industrial and Information Systems (ICIIS)*, 2013 8th IEEE International Conference on, pages 381–386. IEEE, 2013.
- [19] H. Tang, S. Tan, and X. Cheng. A survey on sentiment detection of reviews. *Expert Systems with Applications*, 36(7):10760–10773, 2009.

## **Appendix**

### **Project Reflection**

To complete the project we faced so many problem, first one was to determine the methodological approach for our project. It was not traditional work it was a research based project, more over there were not much work done before on this area. So we could not get that much help from anywhere. Another problem was that, collection of data, it was big challenge for us. There was no available source where we could get Bangla news text data, that's why we were develop a corpus for data collection. Also we started collect data manually. After a long time with hard work we could do that.

Plagiarism Report Screenshot:

## Bangla Sentiment Analysis on Cricket Comments

### ORIGINALITY REPORT

19%	13%	14%	10%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

### PRIMARY SOURCES

1	Shamsul Arafin Mahtab, Nazmul Islam, Md Mahfuzur Rahaman. "Sentiment Analysis on Bangladesh Cricket with Support Vector Machine", 2018 International Conference on Bangla Speech and Language Processing (ICBSLP), 2018 Publication	11%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	3%
3	Submitted to Daffodil International University Student Paper	2%
4	Keith Fraser, Dylan M. Bruckner, Jonathan S. Dordick. "Advancing Predictive Hepatotoxicity at the Intersection of Experimental, In Silico and Artificial Intelligence Technologies", Chemical Research in Toxicology, 2018 Publication	<1%
5	Sara Azmin, Kingshuk Dhar. "Emotion Detection from Bangla Text Corpus Using Naïve Bayes Classifier", 2019 4th International	<1%