Shahjalal University of Science and Technology,

Department of Computer Science and Engineering.

**CSE 408**



Sentiment Analysis - Detect the sentiment of Bangla text

Name: Shoaib Alam Reg. no: 2013331064 Year/Semester: 4/1 Dept: CSE

Name: Fowzia Yesmin Reg. no: 2013331037 Year/Semester: 4/1 Dept: CSE

Supervisor:

Name: Md. Masum Designation: Associate Professor Dept: CSE

12th September , 2018

Recommendation Letter from Thesis Supervisor

This report is for the completion of first part (4/1 semester) of our thesis on “Sentiment Analysis- Detect the sentiment of Bangla text” which is assigned to us by CSE dept. of SUST for the fulfillment of Course No. 408.

Group Members

Shoaib Alam

2013331064

Fowzia Yesmin Munni

2013331037

Md. Masum

Associate Professor

Dept. of Computer Science and Engineering, SUST

Qualification Form of Bachelor Degree

Student Name: Shoaib Alam, Fowzia Yesmin Munni,

Thesis Title: Sentiment Analysis – Detect the sentiment of Bangla text.

This is to certify that the thesis submitted by the student named above in March, 2018. It is qualified and approved by the Thesis Examination Committee.

|  |  |  |
| --- | --- | --- |
| **Acting Head of the Dept.** | **Chairman, Exam. Committee** | **Supervisor** |

Dr. Mohammad Reza Selim Dr. MD. Jahirul Islam MD. Masum

Professor & Head Professor Associate Professor

Department of Computer Department of Computer Department of Computer

Science and Engineering. Science and Engineering Science and Engineering

**ABSTRACT**

Sentiment Analysis is an area of important research over the last decade. The basic task in sentiment analysis is classifying a given text whether the expressed opinion in the text. Sentiment analysis is an important type of text analysis process that aims to measure how much positive or negative a text is. Though Bangla is one of the top ten most spoken languages in the world and is spoken by more than 200 million people, it still lacks significant research efforts in the area of Bangla Sentiment Analysis. In the first phase of our thesis work we built a custom Bangla POS tagger that automatically tags Bangla documents with different parts of speech.

In the next phase, we use different machine learning techniques to determine whether the given query is positive or negative.

**Keywords：**

Parts of Speech (POS) Tagger, Stammer, Supervised Learning, N-gram model, Cosine similarity, Term frequency, Document frequency, Inverse document frequency, Naïve Bayes, Smoothing.

**Acknowledgments**

We wish to express my profound sense of gratitude to our supervisor Assistant Professor Md. Masum for introducing us to this research topic and providing their valuable guidance and unfailing encouragement throughout the course of the work. Besides, we would like to thank our teacher, staffs, Juniors, friends, surroundings specially parents for their beloved support, inspiration and motivation.

We are immensely grateful to them for their constant advice and support for successful completion of this work.

TABLE OF CONTENTS

Page

[LIST OF TABLES](#_Toc82246438) 8

[LIST OF pictures](#_Toc82246438) 9

[IllUStration of symbols](#_Toc82246438) 10

1 [INTRODUCTION](#_Toc82246440) 11

2 [BACKGROUND STUDY](#_Toc82246456) 12

[2.1 ACTIVITIES ON SENTIMENT ANALYSIS FOR FOR ENGLISH](#_Toc82246454) 12

[2.2 ACTIVITIES ON SENTIMENT ANALYSIS FOR FOR BANGLA](#_Toc82246454) 13

[2.3 SENTIMENT ANALYSIS IN NEWSPAPER & BLOGS](#_Toc82246454) 14

[2.4 SENTIMENT ANALYSIS IN SOCIAL MEDIA](#_Toc82246454) 14

3 [WHY SENTIMENT ANALYSIS](#_Toc82246456) 16

4 [DATA COLLECTION](#_Toc82246456) 17

5 [METHODOLOGY](#_Toc82246448) 18

[5.1 PARTS OF SPEECH RATIO](#_Toc82246449) 18

5.2 COSINE SIMILARITY USING TF-IDF 19

[5.3 COSINE SIMILARITY USING CUSTOM TF-IDF](#_Toc82246449) 20

[5.4 NAÏVE BAYES MODEL WITH UNI-GRAM & STAMMER](#_Toc82246449) ..21

5.4.1  [NAÏVE BAYES MODEL WITH BI-GRAM, STAMMER & NORMALIZER](#_Toc82246454) 21

* 1. [NAÏVE BAYES MODEL WITH BI-GRAM, STAMMER & NORMALIZER](#_Toc82246454) 25

[5.6 DETERMINNING THE SENTIMENT OF AN INDIVIDUAL FROM HIS SOCIAL MEDIA POSTS AND COMMENTS…………………………………………………27](#_Toc82246454)

6 PERFORMANCE ANALYSIS27

7 CONCLUSION 29

[8.1 FUTURE SCOPE](#_Toc82246454) 29

[8.2 CONCLUSION](#_Toc82246454) 29

8 REFERENCES 30

**List of Tables**

**Page**

|  |  |  |
| --- | --- | --- |
| Table 1 | Activities On Sentiment Analysis | 12 |
| Table 2 | Example of Sentiment Analysis | 13 |
| Table 3 | Personal pronoun for nominative case | 20 |
| Table 4 | Personal pronoun for objective case | 20 |
| Table 5 | Verb | 21 |
| Table 6 | After normalizing sentence | 23 |
| Table 7 | Performance analysis | 25 |

**List of Pictures**

**Page**

|  |  |  |
| --- | --- | --- |
| Picture 1 | Parts of speech ratio | 15 |
| Picture 2 | Cosine similarity using TF-IDF | 17 |
| Picture 3 | Cosine similarity using custom TF-IDF | 18 |
| Picture 4 | Naive Bayes Model with Uni-gram & stammer | 19 |
| Picture 5 | Naive Bayes Model with Bi-gram, stammer & normalizer | 22 |
| Picture 6 | Performance comparison bar | 26 |
|  |  |  |

**Illustration of Symbols**

POS: Parts Of Speech

PV: Positive Vector

NV: Negative Vector

QV: Query Vector

TF: Term Frequency

DF: Document Frequency

IDF: Inverse Document Frequency

## Chapter 1

## INTRODUCTION

Sentiment analysis is the most interesting and newly emerged research topic in the field of Natural Language Processing (NLP). Basically sentiment analysis is an automated system which uses various machine learning techniques on a user’s or author’s opinion to detect the sentiment. A user may express his/her emotion, excitement, confession, feelings, boredom through a text and a system needs to detect it. Now a day’s social networking sites are getting popular exponentially. Most of the business minded people use social networking sites for advertising their product and getting review from user’s or customers. Besides that, people write blogs, advertise movie and many other things. One can easily know how people are reacting about their works through sentiment analysis.

Sentiment analysis can be categorized into two major parts positive and negative. In this respect, a sentiment analysis task can be interpreted as a classification task where each category represents a sentiment. Sentiment analysis for a company with a means to estimate the product acceptance and to determine the strategy to improve the product quality. To analysis a sentiment we need to build a classifier or patterns to classify a specific type of sentiment. If one pattern or classifier to fail to classify one document then it will check the other patterns until it finds a suitable classifier. In this thesis we work on sentiment analysis in order to detect the sentiment of a Bangla text.

Chapter 2

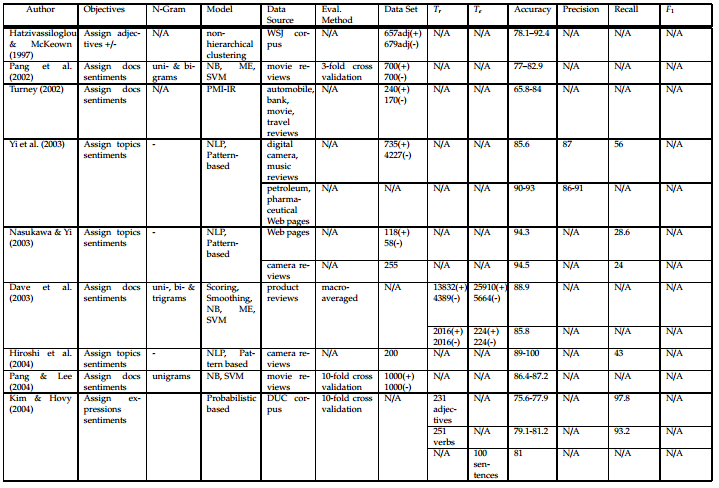
Background study

## 2.1 Activities on Sentiment Analysis for english: [1]

While most researchers focus on assigning sentiments to documents, others focus on more specific tasks: finding the sentiments of words (Hatzivassiloglou & McKeown 1997), subjective expressions (Wilsonet al. 2005; Kim & Hovy 2004), subjective sentences (Pang & Lee 2004) and topics (Yi et al. 2003; Nasukawa & Yi 2003; Hiroshi et al. 2004). These tasks analyze sentiment at a fine-grained level and can be used to improve the effectiveness of a sentiment classification, as shown in Pang & Lee (2004). Instead of carrying out a sentiment classification or an opinion extraction, Choi et al. (2005) focus on extracting the sources of opinions, e.g., the persons or organizations who play a crucial role in influencing other individuals’ opinions. Various data sources have been used, ranging from product reviews, customer feedback, the Document Understanding Conference (DUC) corpus, the Multi-Perspective Question Answering (MPQA) corpus and the Wall Street Journal (WSJ) corpus. To automate sentiment analysis, different approaches have been applied to predict the sentiments of Words, expressions or documents.

These are Natural Language Processing (NLP) and pattern-based (Yiet al. 2003; Nasukawa & Yi 2003; Hiroshi et al. 2004; K¨onig & Brill 2006), machine learning algorithms, such as Naive Bayes (NB), Maximum Entropy (ME), Support Vector Machine (SVM) (Joachims 1998),and unsupervised learning (Turney 2002).

Table 2 lists some existing work in this area, and shows different types of objectives along with the associated models used and the experimental results produced.

****

**TABLE 1[1]**

## 2.2 Activities on Sentiment Analysis for Bangla:

Some work has already done for Bangla language. Such as, by using the WorldNet to get the senses of each word according to its parts of speech and SentiWordNet to get the prior valence (i.e. polarity) of each word. We calculate the total positivity, negativity and neutrality of sentence or document with respect to total sense **[3].**

Used SentiWordNet, WorldNet API for lexical analysis and implement support vector machine and maximum entropy model to detect the sentiment of microblog. Besides, they gave preference of emoticons **[4].**

# **2.3 Sentimental analysis in newspaper & Blogs:**

Newspapers and blogs express opinion of news entities (people, places, things) while reporting on recent events. We present a system that assigns scores indicating positive or negative opinion to each distinct entity in the text corpus.

News can be good or bad, but it is seldom neutral. Although full comprehension of natural language text remains well beyond the power of machines, the statistical analysis of relatively simple sentiment cues can provide a surprisingly meaningful sense of how the latest news impacts important entities.

Recent years have brought a significant growth in the volume of research in sentiment analysis, mostly on highly subjective text types (movie or product reviews). The main difference these texts have with news articles is that their target is clearly defined and unique across the text. Following different annotation efforts and the analysis of the issues encountered, we realized that news opinion mining is different from that of other text types. We identified three subtasks that need to be addressed: definition of the target; separation of the good and bad news content from the good and bad sentiment expressed on the target; and analysis of clearly marked opinion that is expressed explicitly, not needing interpretation or the use of world knowledge. Furthermore, we distinguish three different possible views on newspaper articles - author, reader and text, which have to be addressed differently at the time of analysing sentiment.

**2.4** **Sentimental analysis in Social media:**

This thesis presents a comparison of different machine learning techniques applied to the case of sentiment analysis in social media. Social media sites like Facebook and Twitter have become extremely popular since their appearance. Today millions of people share their impressions about the world with their friends and acquaintances in social media. In today’s connected world, users can send messages in any time. However, social media is not only used as a casual tool for messaging and sharing private things and thoughts; it is also used by journalists, politicians and public figures, series of companies and universities who want to be more open to the public, share their thoughts and take an interest in opinion of persons. The active growth of the audience of social media on the Internet led to the formation of these resources as a new source of the people’s mood and opinion. The tracking of citizens’ reactions in social media during crises has attracted an increasing level of interest in the research community .

Social media can be referred to as the ”group of internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user-generated content”, as defined by Kaplan and Haenlein [40]. In recent years in addition to the leaders of **the World Wide Web** such as **Facebook, Google+, LinkedIn and Twitter**, there are new services for different groups of users: social network for students, the network for specific groups of professionals, communities of ethnic minorities, and even a special network for all the world’s drinkers. This extends the scope to very different kinds of research from consumer preferences to psychological characteristics.

Facebook dominates the global social media landscape, claiming 1.366 billion active users in January 2015. Meanwhile, instant messenger services and chat apps continue to grow, with **WhatsApp**, **WeChat**, **Facebook Messenger** and **Viber** all reporting more than 100 million new monthly active users over the past 2014. Instant messenger services and chat apps now account for 3 of the top 5 global social platforms, and 8 instant messenger brands now claim more than 100 million monthly active users. In Twitter, the number of monthly active users is 284 million in 2015. In 2016 the number of monthly active users exceeded 320 million.

The simplest and most common polarity scheme assumes two categories, positive and negative. These two categories constitute the extreme ends of a discrete or continuous scale. This definition covers most voting schemes used in practice, such as

• thumbs up/down (e.g., Facebook, YouTube)

• positive, neutral, negative (e.g., eBay)

• star ratings (e.g., Amazon, IMDb).

## chapter 3

**WHY SENTIMENT ANALYSIS:**

Sentiment Analysis is a very new topic in Bangla Language. No work has been done in that topic so far. Today’s world is based on statistic. So, automated sentiment analyzer can build up a basement to predict a specific product or topic acceptance to others. 24% people in Bangladesh pay income tax is more reliable information than few people in Bangladesh pay income tax. Statistical report or analysis is more reliable and believable to others than assumption. In our thesis, our required destination is to polarize human’s document or statement and built a statistical report.

For an example;

|  |  |  |
| --- | --- | --- |
| Comment | Positive | Negative |
| রুবেল আজ হ্যাপি করে দিল । | 1 | 0 |
| বাংলদেশ আজ ভাল খেলেও ম্যাচটি হেরে গেল। | 0 | 1 |

**Table 2**

So,

Total comment = 2

Positive comment = 1

Negative comment = 1

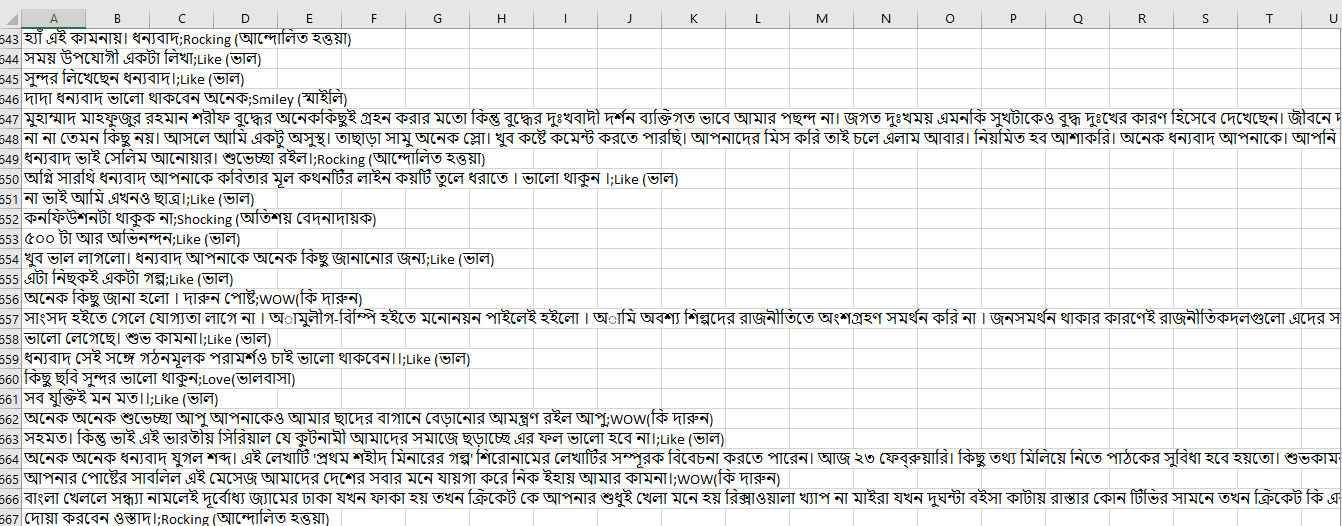
Percentage of positive comment = 50%

Percentage of Negative comment = 50%

So, we can easily say that 50% people comment positively about today’s match and 50% people comment negatively.

## chapter 4

**DATA COLLECTION**



This figure is a snapshot of our collected corpus. The size of our corpus is about 50 Thousands sentences. In our corpus, there are sentences of various types. There are not only positive and negative sentences in our corpus, but also, neutral and other kinds of sentimental sentences available in our corpus. So this corpus can be considered as a general and perfect corpus for running the training models and other algorithm.

## 4.1 Sentiment Analysis on social networking site

Two most popular and common social networking sites are Facebook and Twitter. Generally, Facebook is more popular and common than Twitter. Facebook also contain more Bangla comments and review than others. We are working on Bangla language so Facebook is the best option to retrieve data.

Many work has already done to analysis the sentiment of Twitter data. And the data set on Twitter is English. But sentiment analysis on Facebook Bangla data is rare.

**4.2 FACEBOOK USER CLASSIFICATION**

If we analyze Facebook we found that there are 5 types of user. They are picture people, commenters, likers, statusers and the nothings.

**4.3 FACEBOOK COMMENT CLASSIFICATION [2]**

According to **[2]** we can categorize Facebook comments into 14 parts. They are angry, evil, frown, rock-on, shock, smiley, wink, angel, blush, fail, neutral, shades, skeptical, and tongue. For binary classification, we chose to classify into positive and negative sentiment labels. The labels are represented by:

Positive (Smiley, Wink, Tongue, Angel, Shades, Blush, Rock on, Favor)

Negative (Frown, Shock, Skeptical, Evil, Angry, Fail, Against)

**4.4 DATA PROCESSING**

We collected user’s comment from Facebook manually. And give them a structural format and tagged the data set either positive or negative. We collect above 1000 positive comment and 1000 negative comment for training purpose and approximately 500 comment for test purpose. All the comment we collected is public.

## chapter 5

## Methodology

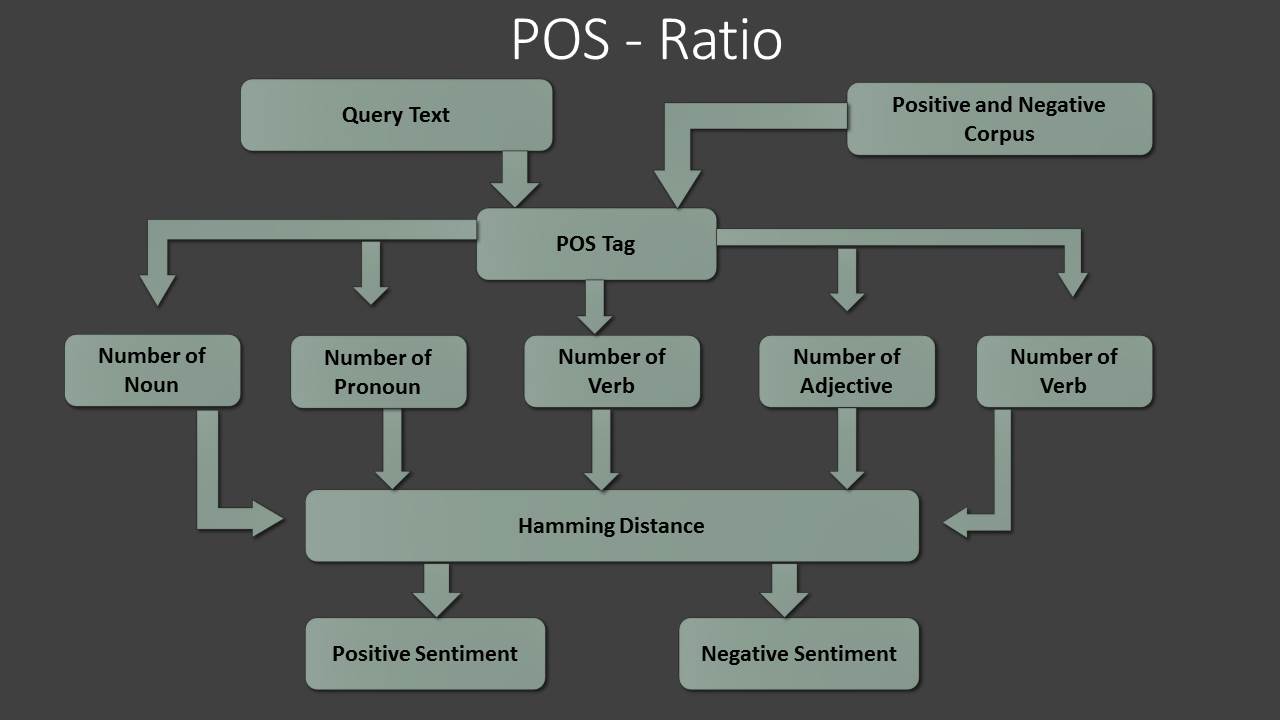
We use five different methodologies. They are:

* + Parts of speech ratio.
  + Cosine similarity using TF-IDF
  + Cosine similarity using custom TF-IDF
  + Naïve Bayes model using Uni-gram & stammer
  + Naïve Bayes model using Bi-gram stammer and normalizer
  + Determining the sentiment of an individual from his social media posts and comments.

**5.1 PARTS OF SPEECH RATIO**

First of all we consider positive and negative data set as a classifier. We tagged all our positive and negative data set with our own custom POS tagger. Then we count the number of noun,

adjective, verb, pronoun, conjunction for both classifier. We also compute the POS ratio for both positive and negative classifier. When a query comes we also POS tagged the query and compute the POS ratio. Then we calculate the hamming distance between positive classifier and query and negative classifier and query. Then minimum distance defines the classifier.



**Picture: 1**

**5.2 COSINE SIMILARITY USNG TF-IDF**

Here we tried to find the cosine similarity between documents. First of all we consider all positive data set as DOC1 and all the negative data set as DOC2 and query data set as DOC3. We will find maximum similarity between DOC1 and DOC3 and DOC2 and DOC2. To find cosine similarity we need to build two vector. We make DOC1 as PV (Positive Vector) DOC2 as NV (Negative Vector) and DOC3 as QV (Query Vector).

**Term Frequency (TF):** We listed all the unique word contains in the document and count the frequency how many time each word occurs.

**Document Frequency (DF):** check how many documents contains each unique words. Suppose, if মানুষ contains in both DOC1 and DOC2 then the DF of মানুষ is 2. If it contain only DOC1 or DOC2 then DF of মানুষ is 1.

**Inverse Document Frequency (IDF):**

Inverse document frequency (x) =

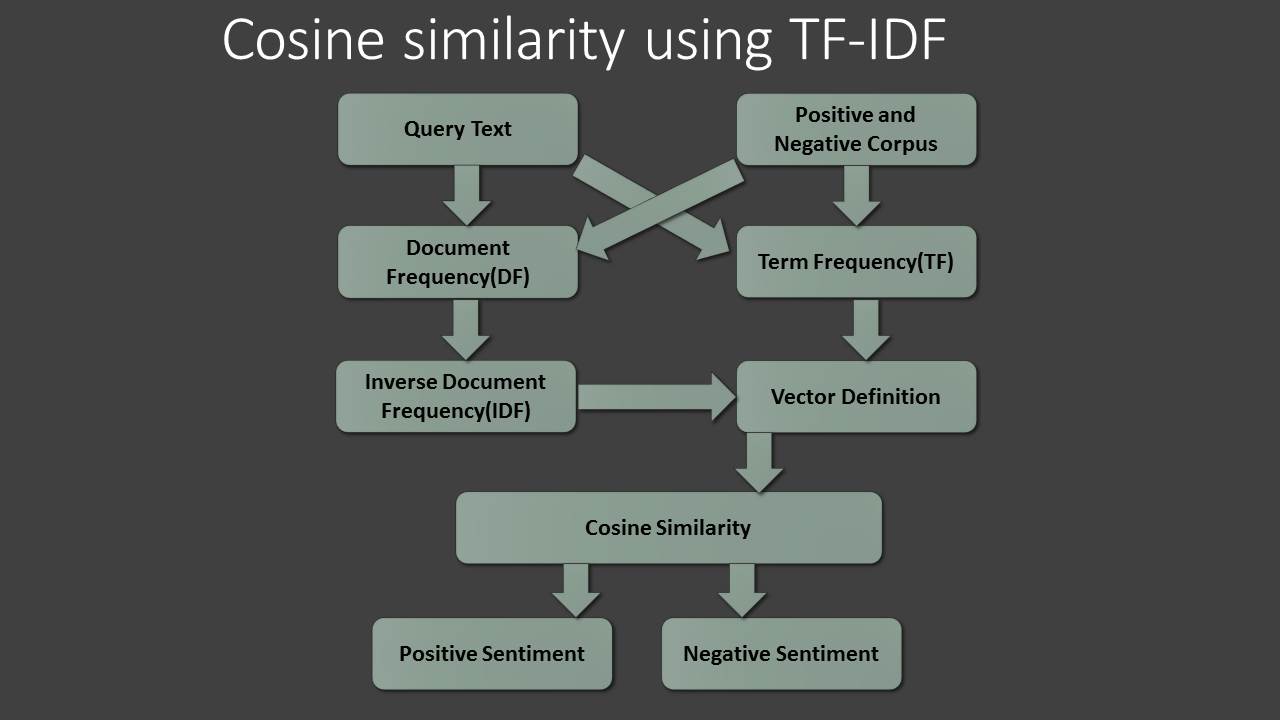
We have TF, DF, and IDF for both the document and the query. So, we can make the vectors. The vectors are:

**Positive Vector (PV) = (TF1\*IDF1) i1+ (TF2\*IDF2) i2+…………..+ (TFn\*IDFn) in**

**Negative Vector (NV) = (TF1\*IDF1) i1+ (TF2\*IDF2) i2+…………..+ (TFn\*IDFn) in**

**Query Vector (QV) = (TF1\*IDF1) i1+ (TF2\*IDF2) i2+…………..+ (TFn\*IDFn) in**

Now we find the cos between PV and QV and NV and QV. The maximum defines its classification. If the cos value between PV and QV is greater than value between NV and QV then the query is positive otherwise negative.

****

**Picture: 2**

**5.3 COSINE SIMILARITY USNG CUSTOM TF-IDF**

In Bangla language two word can make different meaning as well as the different parts of speech. It can change the sentiment. Suppose,

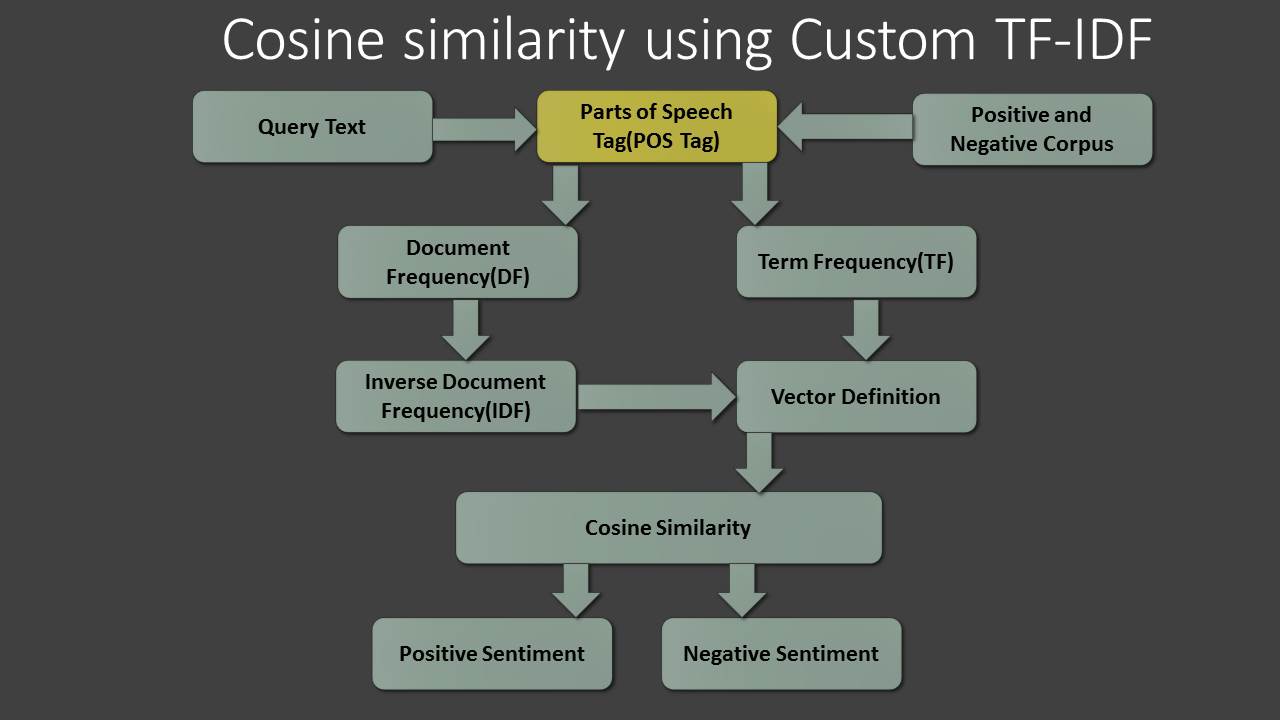
* + - তারা can be used as Noun.
    - তারা can be used as Pronoun.

Here same word তারা creating different parts of speech and meaning of a statement.

তারা মাঠে ফুটবল খেলছে।

আকাশের তারা অনেক সুন্দর।

To solve we use our own custom POS tagger which detects the parts of speech of a text. To find the TF and DF we consider তারা N and তারা P as completely different word. Then follow the same procedure as we used in **5.2**.

****

**Picture: 3**

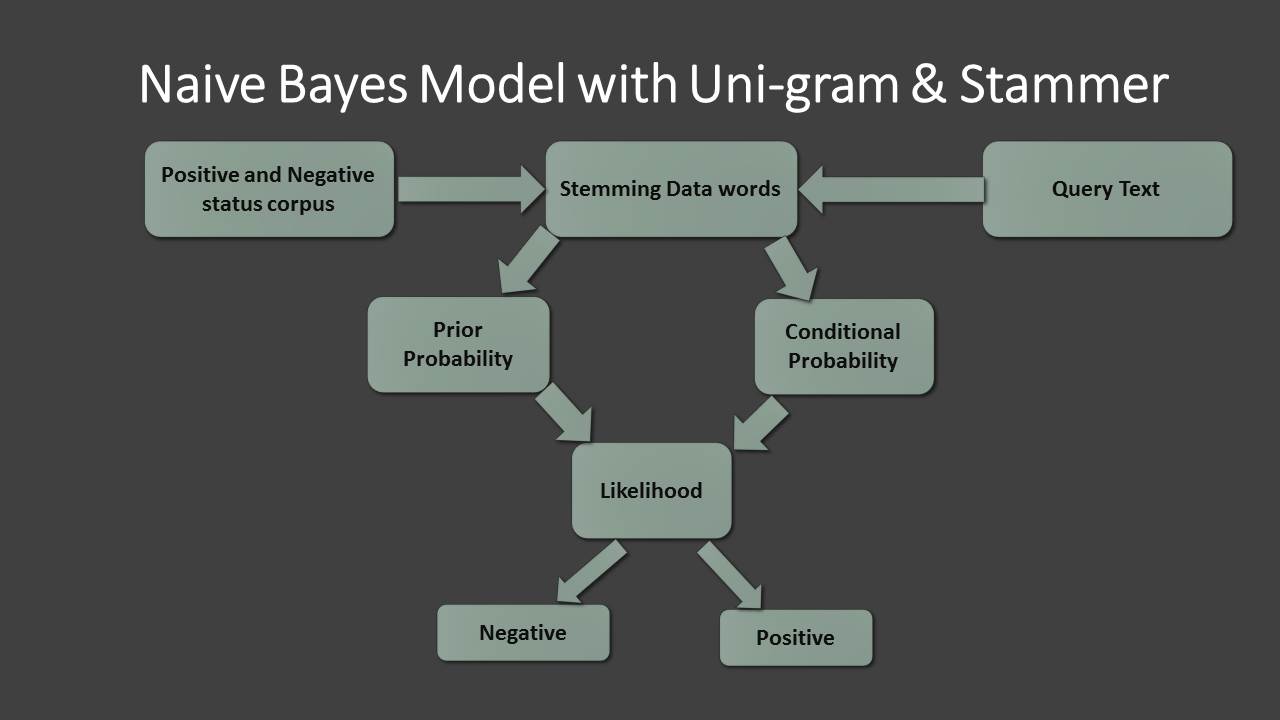
**5.4 NAÏVE BAYES MODEL WITH UNI-GRAM & STAMMER**

In NLP our sentiment detection is a Binary-Classification supervised learning problem. That’s why we try to implement supervised classification machine learning approach. There is lot of classification model in machine learning such as:

* Naive Bayes classification(NB)
* Support Vector Machine (SVM)
* Maximum Entropy Model(ME)

For documents classification Naïve Bayes Model well performed (Sida Wang and Christopher D. Manning, Department of Computer Science in Stanford University)

Model View given below:



**Picture: 4**

**5.4.1 Positive and Negative Corpus**

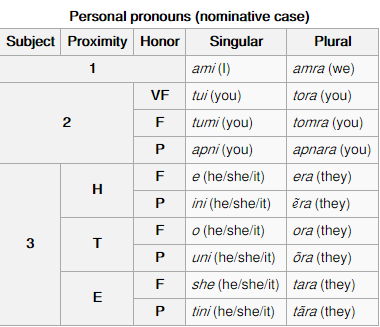
For our thesis purpose we collect 1000 positive and 1000 negative Facebook comments.

**5.4.2 Stemming the data**

Bangla text structure is different from other languages like English, Arabic, Chinese, and Japanese.

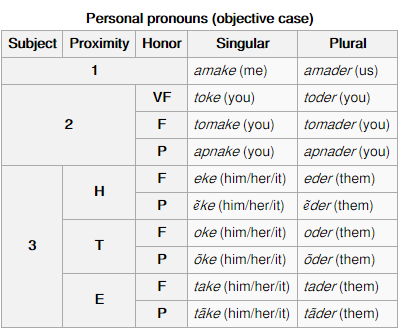
Some example given below that is taken from Wikipedia that claim the variation of Bangla text structure.

**For Personal pronouns (Nominative case):**

****

**Table 3**

**For Personal pronouns (Objective Case):**

****

**Table 4**

**For Verb**

****

**Table 5**

For our Kernel (8.3.4) we have to calculate prior and conditional probability distribution. In probability distribution we try to find out the root word of a word because it’s smooth our Naive Bayes parameter definition.

**5.4.3 Kernel**

Bayes' Theorem is a theorem of probability theory originally stated by the Reverend Thomas Bayes. We use ‘Add Ones’ normalize form of Bayes Theorem used by Stanford NLP-Dan Jurafsky and Chris Manning.

Three part of Kernel is:

1. Prior Probability
2. Conditional Probability
3. Likelihood

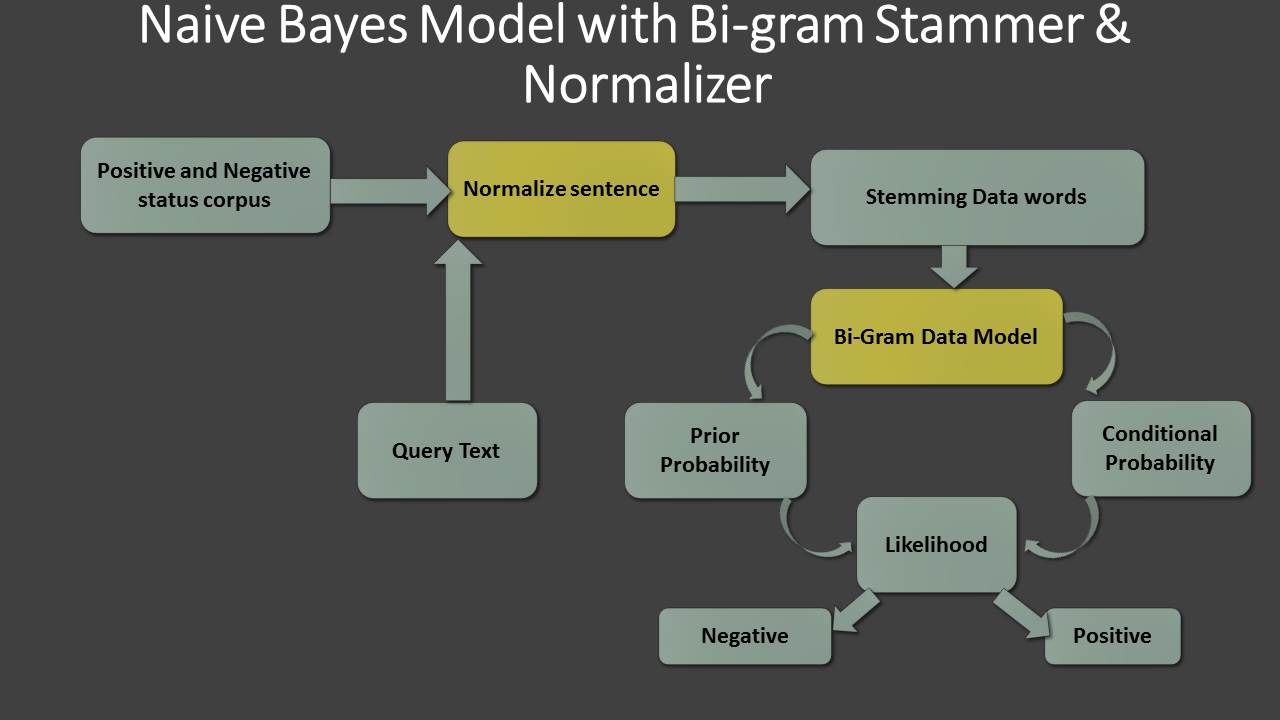
Prior Probability, Conditional Probability and Likelihood (Math):

We have two classes:

* Positive class = c1
* Negative class = c2
* Prior Probability(c1) = (Total comments in c1 ) / (Total comments in c1+c2)
* Conditional Probability:
  + CP(w|c1) = count(w,c1)+1/count(c1)+|V|
  + CP(w|c2) = count(w,c2)+1/count(c2)+|V|
  + Where |V| = Total unique words in corpus.
* The best class in NB classification is the most likely or *maximum a posteriori* (MAP) class C. Class chooser : Cmap = arg-max ( P(c1|Qd),P(c2|Qd))

**5.5 NAÏVE BAYES MODEL WITH BI-GRAM & STAMMER & NORMALIZER**

We design another model by using Stammer and Naïve Bayes. Model figure given below:

****

**Picture: 5**

Normalize sentence and Bi-gram Data is the key different of our previous model (8.3).

Positive and Negative corpus (8.3.1), Stemming the data word (8.3.2) and Kernel (8.3.3) is as same as given before.

**5.5.1 Normalize Sentence**

Valance shifter word example are given below

Ex: না, নি, নাই, নও, নয়

This valance shifter word play an important role in our sentiment detection. Some example given below:

আমি ভাল ফুটবল খেলি। - positive

আমি ভাল ফুটবল খেলি না। -negative(One “না” change the polarity)

আমি মন্দ ছেলে না। -positive

আমি মন্দ ছেলে। - negative (“absence of “না” change the polarity”)

One valance shifter word can change the polarity of a sentence. Valance shifter word can appear in positive sentiment corpus and Negative sentiment corpus. To normalize this fact we implement one important rule.

* Detect simple sentence with one adjective.
* Find “না, নি, নাই…” valance shifter word placed at the right side of a simple sentence.
* Build corpus of “**বিপরীতার্থক শব্দ**” that contain 900 unique words.
* অস্থিবাচক শব্দ থেকে নেতিবাচক শব্দ রূপান্তর…………
  + বিশেষন পদের বিপরীথ শব্দ ব্যবহার করে অস্থিবাচক শব্দ থেকে নেতিবাচক শব্দে রূপান্তর করা যায়
    - ড হায়াৎ মাহমুদ

After normalize us have our sentence like that:

|  |  |
| --- | --- |
| আমি ভাল ফুটবল খেলি না। - negative | আমি খারাপ ফুটবল খেলি । -negative |
| আমি মন্দ ছেলে না। - positive | আমি ভাল ছেলে । - positive |

**Table 6**

**5.5.2 Bi-gram Model**

Some observation comes when we introduce normalize in our sentence. What happen to our decision when comes like given below sentence?

না না এসব কথা বলিস না। এসব কথা ভাল ছেলেরা বলে না।

তুমি যাতটা না দুষ্ট তার চেয়ে বেশি ভাল।

For this case observation, this kind of complex and compound sentence play an important role in our sentiment detection. When valance shifter word comes, the polarity of a sentence not only depends on adjective or other POS tag. It makes a tradeoff observation between adjective or other POS tag. To solve this kind of tradeoff we implement Forward Bi-Gram Model and Backward Bi-Gram model to normalize the impact of valance shifter word.

* 1. **: Determining the sentiment of an individual from his social media posts and comments.**

Most of the individuals in the current world use social media as a way to express their emotions and feelings regarding any situations. Those social media posts were considered as corpus in our thesis. Now we will use different machine learning based approaches to correctly determine the sentiment behind a social media post. The algorithms which we are going to use are Naïve bayes, Recurrent Neural Network, Artificial Neural Network and Convolutional Neural Network. We have made a corpus which have 50 thousands sentences. By using the above mentioned machine learning models on our corpus we will determine the sentiment of any individuals from his social media posts and comments.

## chapter 6

**PERFORMANCE ANALYSIS**

**Performance of parts of speech ratio method:**

The performance of POS ratio method is very low. Accuracy is very low more precisely less than 40.12%. Failed to detect the sentiment.

**Performance of cosine similarity using TF-IDF method:**

Performance of cosine similarity using TF-IDF is hopeful. Here, it give every unique individual word a point or value to make vector. It also normalize the very frequent word. So, it produce much better performance. The accuracy of detect comment as positive or negative is 73.54%.

**Performance of cosine similarity using custom TF-IDF method:**

Cosine similarity using custom TF-IDF create difference among the same words having different parts of speech or meaning. It is more accurate than the normal TF-IDF method and able to detect the sentiment of a text. Accuracy of detect comment as positive or negative is 78.12%. Here, there are some ambiguous comments that is really hard to detect. Suppose,

কালকে কয়েকজনের ড্রপের পরীক্ষা আছে, তাই পরশু।

রাজনের বিষয়ে কিছু একটা লেখার জন্য আমি হাতে কলম এবং কিছু কাগজ নিয়ে গত কয়েক ঘণ্টা চুপচাপ বসে আছি।

কাজটি ভাল না খারাপ জানি না।

একটা কথা মনে রেখো আমি ভালর ভাল আর খারাপের খারাপ।

**Performance of Naïve Bayes model with Uni-gram & stammer method:**

Accuracy = (total correct answer / total test comments) \* 100. Performance of Naive Bayes uni-gram model produce 76.72% accuracy.

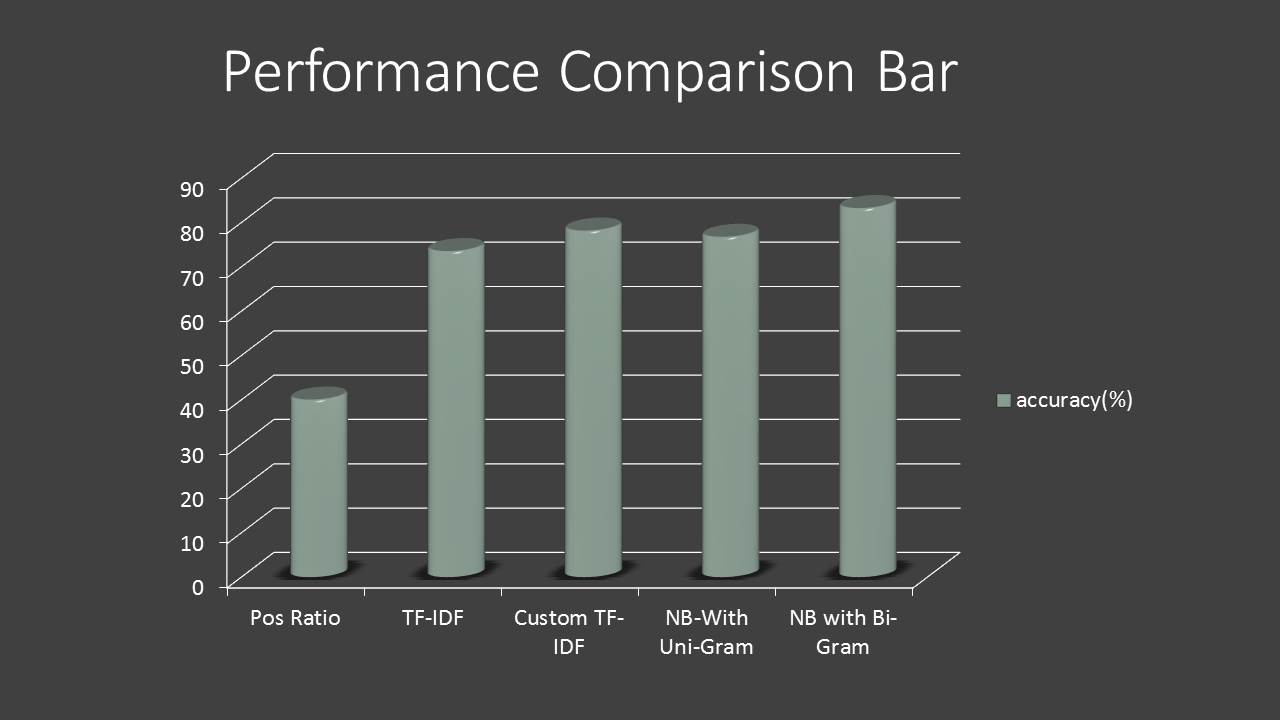
**Performance of Naïve Bayes model with Bi-gram, stammer & normalizer method:**

Model Accuracy = 83.20%. Naive Bayes Model with Bi-gram, Stammer and Normalizer is batter.

|  |  |
| --- | --- |
| **Methodology** | **Accuracy (%)** |
| Parts of speech ratio | 40.12 |
| Cosine similarity using TF-IDF | 73.54 |
| Cosine similarity using custom TF-IDF | 78.12 |
| Naïve Bayes model with Uni-gram & stammer | 76.72 |
| Naïve Bayes model with Bi-gram, stammer & normalizer | 83.20 |

**Table**

**Performance comparison bar:**

****

**Picture: 6**

##### Besides that we implement our methodologies in Professor **Dr.** [**Muhammed Zafar Iqbal**](https://www.facebook.com/our.zafar.sir?fref=nf)**’s blog** **“লেখা পড়া নিয়ে কিছু কথা”** on Facebook published in August 28. We find that 70.73% people give positive review or comment against the blog and the remaining 29.27% people gives negative comment.

## chapter 7

**CONCLUSION**

**7.1 Future Scope**

Sentiment analysis is the most interesting and newly emerged research topic. It will open a new door for the writers, bloggers, and businessman. One can easily know the percentage of product acceptance and make their strategy to improve the product quality. Our model can also be used to accurately predict stock market and other business strategical decisions. We think different type of Machine learning model will enforce the result to a significant level.

**7.2 Conclusion**

In this thesis report, we have designed Sentiment Analyzer which detects the sentiment of a text or statement. We used several supervised machine learning methods. It gives us approximately satisfactory accuracy. The approach presented here is flexible and suggests promising avenues for further research. So we conclude our hypothesis as one of the most flexible and accurate implemented model in the field of sentiment analysis.

## chapter 8

**REFERENCES**

**[1]**Rudy Prabowo, Mike Thelwall Sentiment Analysis: A Combined Approach. School of Computing and Information Technology University of Wolverhampton

**[2]**Julie Kane Ahkter (kanej), Steven Soria (ssoriajr) Sentiment Analysis: Facebook Status Messages Final Project CS224N

**[3]** K. M. Azharul Hasan Mosiur Rahman, Badiuzzaman Sentiment Detection from Bangla Text using Contextual Valency Analysis

**[4]** Shaika Chowdhury Wasifa Chowdhury sentiment Analysis for Bangla Microblog Posts.**[2]**Stone, P. J., Dunphy, D. C., Smith, M. S., & Ogilvie, D. M. (1966). The general inquirer: a computer approach to content analysis. The MIT Press. Swan, R. & Allan, J. (2000). Automatic generation of overview timelines.

**[3]**Thelwall, M. (2000). Web impact factors and search engine coverage. Journal of Documentation, 56(2), 185–189.

**[4]**Thelwall, M. (2008). Fk yea I swear: Cursing and gender in a corpus of MySpace pages. Corpora, 3(1), 83–107.

**[5]**Turney, P. D. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40 annual meeting of the Association for Computational Linguistics (ACL), July 6–12, 2002 (pp. 417–424). Philadelphia, PA, USA.

**[6]** Apoorv Agarwal Boyi Xie Ilia Vovsha Owen Rambow Rebecca Passonneau Sentiment Analysis of Twitter Data Department of Computer Science Columbia University New York, NY 10027 USA

**[7]** Sabir Ismail, Lecturer Shahjalal University of Science and Technology, Sylhet.

**[8]** [www.facebook.com](http://www.facebook.com) for data.

**[9]** [www.wikipedia.com](http://www.wikipedia.com) for data.

**[10]** Merialdo. 1994. Tagging English text with a probabilistic model. Computational Linguistics, 20(2):155-171

**[11]** Luciano Barbosa and Junlan Feng. 2010. Robust sentiment detection on twitter from biased and noisy data. Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pages 36–44.