UNIVERSITY OF CHITTAGONG

BACHELOR THESIS

Bangla Text Sentiment Analysis

Author:
Md. Sajib Hosen

Supervisor:
Prof. Dr. Md. Hanif Seddiqui

A thesis submitted in fulfillment of the requirements
for the degree of B.Sc. (Engg.)

in the

Department of Computer Science & Engineering

September 25, 2017

Declaration of Authorship

I, Md. Sajib Hosen, declare that this thesis titled, "Bangla Text Sentiment Analysis" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

_____

Date:

_____

University of Chittagong

Abstract

Faculty of Engineering

Department of Computer Science & Engineering

B.Sc. (Engg.)

Bangla Text Sentiment Analysis

by Md. Sajib Hosen

Sentiment Analysis or Opinion mining is a field of where we can extract peoples opinions, attitudes something like that. Increase in user-generated content provide an important aspect for the researchers, industries and government to mine this information. The user-generated content is one important source for various organizations to know/learn/identify the general expression/sentiment of different users on the product.

In this manuscript we tried to extract sentiment or opinion from the text. We mainly developed a system that classify the text as either POSITIVE or NEGATIVE.

We have implemented a Machine Learning based approach which uses Multinomial Naive Bays classifier to classify the text. Before Applying classifier ,we have done some preprocessing both the training and test data.

Acknowledgements

At first, I want to express gratitude to the Almighty Allah for His endless kindness for keeping me mentally and physically fit to complete this sophisticated task.

I would like to express my sincerest gratitude to my supervisor Prof. Dr. Md. Hanif Seddiqui, for giving me the inspiration and flexibility to explore my ideas and research interests. Without his continuous guidance, support and encouragement, this journey would not have been possible.

I would also like to thank my classmates for helping me whenever I was in any problem.

Md. Sajib Hosen

Id:12205055

Session:2011-2012

Department of Computer Science and Engineering

University of Chittagong

...

Contents

To my Family and Teachers...

Chapter 1

Introduction

1.1 Sentiment Analysis - An Introduction

Sentiment analysis or Emotion recognition from the text is the popular part of Natural language processing(NLP). It has drawn a considerable attention for NLP researchers. Analyzing sentiment of a language is a challenging task in NLP.In general, the motive of sentiment analysis is determine opinion, feelings, attitude or comments of the writer which indicate emotion of the writer.

"Sentiment analysis (also known as opinion mining) refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials. Sentiment analysis is widely applied to reviews and social media for a variety of applications, ranging from marketing to customer service."(Source: Wikipedia)

Emotion or Feelings are hidden in text.This field of computer science deals with analyszing and predicting the hidden information stored in the text. Sentiment analysis focus on categorizing the text.such as subjective and objectives. for example.

subjectives: Sakib al Hasan is a good player. (this sentence has an opinion.it's talk about the player of sakib al hasan good or bad.)

Objectives: Sakib al hasan play in Bangladesh national cricket team. (this sentence does not contain any opinion or a view. It is only fact.

1.2 Bangla text sentiment analysis:Overview

Bangla (or Bengali)is an Indo-Iranian language spoken in the Indian Subcontinent.With over 250 million speakers, Bengali is the seventh most spoken native language in the world.It is the primary language in Bangladesh and second language in India.Over the internet world,

a lot of pepole write or read bangla text. many e-commerce site also in bangla text. a lot of pepole, in many social media site, such as facebook,twitter e.t.c also use bangali user.

People may joke that others spend too much time on the internet, but this intricate series of tubes has become an important part of everyday life so much. so that its become a human rights violation to take it away.

These social networking sites and other platforms leads to the generation of petabytes of data per week. Some of the popular and widely used social networking platforms with its brief overview-

Facebook - The total monthly active users on facebook are around 850 million. Everyday about 250 million photos are uploaded and 2.7 billions likes are made on it. Facebook stores, accesses, and analyzes 30+ Petabytes of user-generated data.

Twitter - Twitter has about 465 million accounts and about 175 millions tweets are done per day.

Google+ - It has about 90 millions users and about 675,000 users are been added everyday.

Common bangla online news paper view per day is given bellow:

TABLE 1.1: Online Bangla newsPaper View

Name Of The News Paper	Total view per day
Prothom-alo	1.75M/ Day
bdnews24.com	197.2K/ Day
Bangladesh pratidin	198.99K/ Day
Anandabazar Patrika	303.03K/ Day
somoyer konthosor	40.37K/ Day

The statistics mentioned above gives an idea about the rate at which the web has been increasing. With such vast data generated regularly, it provides enormous business opportunities to handle this data safely and precisely.

1.3 Main Challenges

Because of the relatively new field, there are many challenges to be faced. According to the [1] current techniques are just identification emotional expression and Topic Identification. Mainly these challenges are related to the authenticity of the extracted data and the methods used in it. Reference [2] also discusses some issues of opinion mining. A summary of challenges of opinion mining is as follows:

Unstructured Data- Unstructured Sentiments are an informal and free text format, the writer does not follow any constraints. The data available on the internet is very unstructured, there are different forms of the data talking about the same entities, persons, places, things and events. The web contains data from different sources varying from books, journals, web documents, health records, Companies logs, internal files of an organization and even data from multimedia platforms comprising of texts, images, audios, videos etc. The diverse sources of the data makes the analysis more complex as the information is coming in different formats.

Noise (slangs, abbreviations)- The web content available is very noisy. In todays era of 140 characters texting, for their ease people use various abbreviations, slangs, emoticons in normal text which makes the analysis more complex and difficult.Now a day people write their text in small format. For example If a people want to write awesome the he write it asm or osam e.t.c.

Foreign Word: There are many foreign word mixing in bangla text. For exam I miss you (). The are no exact bangla meaning . some time its a challenging work identifying emotion or opinion from this type of text.

Contextual Information- Identifying the context of the text becomes an important challenge to address. Based on the context the behavior/use of the word changes in a great aspect.

Ex-1 The movie was long.

Ex-2 Lecture was long.

Ex-3 Battery life of samsung galaxy-2 is long.

In all the above 3 examples, meaning of long is same- indicating the duration or passage of time. In ex-1 and ex-2 "long" indicates boredness hence a Negative expression Where as in ex- 3 "long" indicates efficiency hence a Positive expression. With the help of above examples, it's clear that same word with same meaning can have multiple usage depending on the context. So, it becomes important to detect the context to find the subjective information in a text

1.4 motivation

Understanding emotions, analyzing situations and the sentiments associated with it is the natural ability of a human being. But how efficiently can we train a machine to exhibit the same phenomenon becomes an important and vital question to be explored and answered.

Sentiment Analysis provides an effective mechanism for understanding individuals attitude, behavior, likes and dislikes of a user.

Small Story- *Its believed that when a child is small the mother knows very well what and when he/she is going to need, at what time the child drinks, eats or even the time of difficulty when it cries and the possible causes for same. She works towards rectifying same. She very well knows the difference in the cry for food and cry for getting the diapers changed, thus the mother can analyze and take the necessary action very well*

Analogous to this small story, how well our lives would be if what we want can be automatically analyzed, suggested and provided to us without putting much efforts? Sentiment analysis provides us with the services and products we want of our taste at our ease. With e-commerce business spreading at a great speed the task of mining opinions on various products becomes an useful resource to guide and help people in making choices and decisions. Mining sentiments and subjective information helps to provide products and services in a personalized fashion and as per individuals taste and likings. With more emphasis laid on personalized information it becomes necessary and important to go about catering information to an individual, based on his likings and taste. The study of sentiment analysis also provide enough information about how human beings perceive and express information in the form of text to express their feelings and emotions. This wide multi-dimension aspects discussed above, motivated me to take this problem as my Research Problem.

Today the Web has become an excellent source of opinions with the explosive growth of social media (e.g., reviews, forum discussions, blogs, micro-blogs, Twitter, comments, and postings in social network sites). Individuals and organizations are increasingly using the content in these media for decision making. Nowadays, if one wants to buy a consumer product, one is no longer limited to asking ones friends and family for opinions because there are many user reviews and discussions in public forums on the Web about the product. For an organization, it may no longer be necessary to conduct surveys, opinion polls, and focus groups in order to gather public opinions because there is an abundance of such information publicly available.

,

1.5 Application

The Internet has become a rich platform for people to express their opinion, attitude, feeling, and emotion. From this point of view, Web is an important source of product reviews,

news reviews, blog reviews, movie review, stock market reviews, travel advice, social issue discussions, consumer complaints, etc. Nowadays, Bangla has been using widely in the Web. Automatic sentiment classification will become very useful in above applications. Sentiment analysis is now a great interest to the social networking media such as Twitter, Facebook, Google+ as well.

Sentiment analysis is widely used for understanding subjective nature of text or writer opinion. there are few area where sentiment analysis are applied :-

1.5.1 Decision making System

What other people think has always been an important piece of information for most of us during the decision-making process. Long before awareness of the World Wide Web became widespread, many of us asked our friends to recommend an auto mechanic or to explain who they were planning to vote for in local elections, requested reference letters regarding job applicants from colleagues, or consulted Consumer Reports to decide what dishwasher to buy. But the Internet and the Web have now (among other things) made it possible to find out about the opinions and experiences of those in the vast pool of people that are neither our personal acquaintances nor well-known professional critics that is, people we have never heard of. And conversely, more and more people are making their opinions available to strangers via the Internet.

The interest that individual users show in online opinions about products and services, and the potential influence such opinions wield, is something that vendors of these items are paying more and more attention to. Thus, aside from individuals, an additional audience for systems capable of automatically analyzing consumer sentiment, as expressed in no small part in online venues, companies are anxious to understand how their products and services are perceived.

1.5.2 Recommendation Systems

Most of the websites, we visit have a recommendation system built to assist us, online media, entertainment, music, film, industry to other forms of art. this system uses our personal information, previous history, likes and dislikes and our friend's information to make a suggestion.

1.5.3 Products Analysis

With the help of sentiment analysis,it has become easier to analyze different products and make the choices accordingly. This kind of analysis also helps to select a product based on its feature specifications. The comparison between two products has also been made quite easier.

1.6 Proposal

In this manuscript, we tried to focus on extracting opinion from the text. there are a lot of method for extracting sentiment or opinion analysis. we tried to classifying data as either "POSITIVE" or "NEGATIVE" though the "Naive Bays classifier". Before classifying we have done some pre-processing on this data.

1.7 Chapter Outline

The thesis has been organized as follows- Chapter 2 discuss the related work done in the area of sentiment analysis which is spitted into five subsections. Chapter 3 contain two subsections 3.1 General approaches and 3.2 is System architecture. Chapter 4 focus Data collection and pre-processing. Chapter 5 discuss the work done towards sentiment classification. it shows the main processing .chapter 6 focus the experiment and Evaluation. We present the conclusions and the possible future extensions of this research work in Chapter 7.

Chapter 2

Related Works on Sentiment Analysis

Sentiment Analysis is the new topics for researcher. Sentiment Analysis has been the focus of research community from last decade. There has been a large amount of work or research done for English language. But unfortunately the work for bangla language is not rich. it has just began. The initial lexicon and dictionary based approaches for extracting sentiment or opinion. Machine learning approach is for using the syntactic and semantic feature. In this chapter , we discuss the work done in the past in the area of sentiment Analysis.

Identifying the sentiment polarity is a complex task, to address the task of sentiment classification various methodologies have been applied earlier. Most common, widely used approaches for identifying sentiments for a given piece of text are as follows

2.1 Sentiment polarity classification

This is the area that has been researched the most in academia. Sentiment polar ity classification assumes that the given document is opinionated and aims to find the general opinion of the author in the text [47]. For example, given a product review, it determines whether the review is positive or negative. Sentiment classification, in contrast to subjectivity analysis, does not usually need manual effort for annotating training data. Training data used in sentiment classification are mostly online product reviews that have already been labeled by reviewers with the assigned overall ratings (usually in the range form 1 to 5). Typically a review with 4-5 stars are considered positive, and a review with 1-2 stars are considered negative [2].

2.2 Syntactic Approaches

Syntactic approach towards sentiment classification using N-Grams have been used by Bo Pang, Lillian Lee and Vaithyanathan[35]. They used the traditional n-gram approach along with POS information as a feature to perform machine learning for determining the polarity. They used Naive Bayes classification, Maximum Entropy and Support Vector Machines on a three fold cross validation. In their experiment, they tried different variations of N-Gram approach like unigams presence, unigrams with frequency, unigrams+bigrams, bigrams, unigrams + POS, adjectives, most frequent unigrams, unigrams + positions. They concluded from their work that incorporating the frequency of matched n-gram might be a feature which could decay the accuracy. Maximum accuracy achieved by them among all the experiments they performed was 82.9% which was obtained in unigrams presence approach on SVM.

2.3 Semantic/Pattern Mining

Semantic approaches using part of speech learning has also been used quite popularly for identifying sentiments in a text. Turney and Benamara used this approach for binary classification in [40] and [5]. Much work has also been done in the field of extracting sentiment expressions using various NLP techniques. Nasukawa and Yi[33], Bloom, Garg and Argamon[7] used techniques like word sense disambiguation, chunking, n-gram to perform binary polarity classification . Ohana and Tierney [34], Saggion and Funk[38] used sentiwordnet to perform opinion classification. They calculated positive and negative score for a review and based on the maximum score, the polarity of the review was assigned. They also extracted features and used machine learning algorithms to perform sentiment classification. Turney [40] also worked on part of speech (POS) information. He used tag patterns with a window of maximum three words (i.e) till trigrams. In his experiments, he considered JJ, RB, NN, NNS POS-tags with some set of rules for classification. His work is extension to the work done on adjectives alone by Hatzivassiloglou and McKeown[20] they consider patterns like RB, NN/NNS. Given a phrase he calculates the PMI (Point-wise Mutual Information) from the strong positive word excellent and also from the strong negative word poor, and the difference gives the semantic orientation of the phrase. Dave, Lawrence and Pennock devised their own scoring function which was probability based [13]. They performed lexical substitutions for negation handling and used rainbow classifiers to decide the class of the review.

2.4 Features/Machine Learning

Much of the work has also been done towards using machine learning approach for identifying the sentiment expressions. Bo Pang, Lillian Lee and Vaithyanathan [35], Zhang [45], Go, Bhayani and Huang [18] deduced features to perform supervised machine learning . The feature based learning has proved to perform better in comparison to the traditional approaches of syntactic and semantic approaches. The features learned from the N-Grams models along with that of subjective lexicon with little bit of fine tuning perform better as compared to the normal N-Gram and Subjective Lexicon scoring mechanism.

2.5 Various levels of sentiment analysis

Research in the field of sentiment analysis is done at various levels which are as follows

2.5.1 Document Level

The document level analysis deals with classifying the whole document as a single polarity positive, negative or objective. Bo Pang, Lilling Lee and Vaithyanathan[35], Turney [40] performed document level classification .

2.5.2 Sentence Level

The sentence level analysis focus on analyzing the documents at sentence level. The sentences are analyzed individually and classified as objective, negative or positive. The overall document thus has a set of sentences with each sentence being marked with its corresponding polarity. There has been significant work done by Wiebe, Bruce and OHara[42], Yu and Hatzivassiloglou[44], Theresa Wilson [23], Hu and Liu[22] and Kim [30] with respect to the sentence level classification .

2.5.3 Phrase Level

This analysis involves going much deeper and deals with identifying the phrases in a sentence for a given document and analyze the phrases and classify them accordingly as positive, negative or objective. The phrase level analysis is also known as fine grained analysis going deep into the text to identify the subjective items/entities and classify same as done by Wilson et al. [43], Agarwal, Biadsky and Mckeown[1] .

The work done in the past in the area of sentiment analysis can be categorized into various genres such as Reviews, News, Blogs analysis etc. Blogs level analysis- Work done

at blog level can be attributed to Chesley [9], Ku, Liang and Chen[27], Zhang [46], Ben He, Macdonald, Jiyin He and Ounis[21], Melville, Gryc and Lawrence [28], Draya, Plantie, Harb, Poncelet, Roch and Trousset[16], Godbole, Srinivasaiah and Skiena[19] Reviews level analysis- Some of the work done at reviews level for mostly English Language is as follows- Wiebe, Bruce and OHara[42], Bo Pang, Lillian Lee and Vaithyanathan[35], P Turney [40], Yu and Hatzivassiloglou[44], Theresa Wilson [23], Hu and Liu [22], Blitzer, Dredze and Pereira[6]. News level analysis- Godbole, Srinivasaiah and Skiena[19], Alexander Balahur[4] worked on News.

All the above categories deal with large text, in addition to this there has been much focus given towards micro-blogs analysis which includes analyzing tweets, forums and chats by Go, Bhayani and Huang[18], Nicholas A and David A[15], Apoorv, Boyi, Ilia, Owen and Rebecca[2], Dmitry, Oren and Ari[14] Draya, Plantie, Harb, Poncelet, Roche and Trousset[16] tried to identify domain specific adjectives to perform blog sentiment analysis. They considered the fact that opinions are mainly expressed by adjectives and pre-defined lexicons fail to identify domain information. Chesley et al. [9] performed topic and genre independent blog classification, making novel use of linguistic features. Each post from the blog is classified as positive, negative and objective. Turney [40] worked on product reviews. Turney used adjectives and adverbs for performing opinion classification on reviews. He used PMI-IR algorithm to estimate the semantic orientation of the sentiment phrase. He achieved an average accuracy of 74 percent on 410 reviews of different domains collected from Epinion. Hu and Liu [22] performed feature based sentiment analysis. Using Noun-Noun phrases they identified the features of the products and determined the sentiment orientation towards each feature. Bo Pang, Lillian Lee and Vaithyanathan[35] tested various machine learning algorithms on Movie Reviews. They achieved 81 percent accuracy in unigram presence feature set on Naive Bayes classifier.

Chapter 3

General Approach and System Architecture

3.1 General approach of sentiment analysis

there are many work has been done for english language. But a small work on bangla language has been done in sentiment analysis. Some of common and wel known approaches used for sentiment analysis are given below-

3.1.1 Using Subjective lexicon

In this approach, There are a list of word for each word has a weighted value such as positive, negative e.t.c. A list of such words and phrases is called a sentiment lexicon (or opinion lexicon). in this approach, we combine the value of all word in given text and then we get the combined positive,negative and objective score which give the idea of nature of the given text. Below we will discuss a selection of methods to determine the sentiment of a single word.

Using WordNet

WordNet is a lexical database for the English language.[1] It groups English words into sets of synonyms called synsets, provides short definitions and usage examples, and records a number of relations among these synonym sets or their members. WordNet can thus be seen as a combination of dictionary and thesaurus. While it is accessible to human users via a web browser,[2] its primary use is in automatic text analysis and artificial intelligence applications. The database and software tools have been released under a BSD style license and are freely available for download from the WordNet website. Both the lexicographic data

(lexicographer files) and the compiler (called grind) for producing the distributed database are available.(source : wikipidea)

Kamps and Marx developed an automatic method [48] using the lexical database WordNet to determine the emotional content of a word along as good dimensions. In essence, the WordNet database consists of nodes (the words) connected by edges (synonym relations). Kamps and Marx define a distance metric between the words in WordNet, called minimum path-length(MPL). This distance metric counts the number of edges of the shortest path between the two nodes that represent the words. For example, the words good and big have a MPL of 3. The shortest path from the word good to the word big is the sequence good, sound, heavy, big.

Sentiment of Sentences

Above the previous section we see the method of single word. Sentiment of the Sentence has a bag of word. its also called the bag-of-word apporach, has some importance drawbacks. there are several approach in this field.There are several approaches in this field; we mention here briefly Mulder and al.'s article [49], which discusses the successful use of an affective grammar. They note that simply detecting emotion words can tell whether a sentence is positive or negative oriented, but does not explain towards what topic this sentiment is directed. They combined a lexical and grammatical approach:

1. lexical, because they believe that affect is primarily expressed through affect words, and

2. grammatical, because affective meaning is intensified and propagated towards a target through function-words and grammatical constructs.

3.1.2 Using N-Gram Modeling

In this approach, from a given training data we make N-Gram model (uni-gram, bi-gram , tri-gram and combination of same) and perform classification of the test data using the model formed.

3.1.3 Using Machine Learning

In this method, we extract features from the text and learn the model using the training corpus by selecting a set of relevant features. While forming features, information is incorporated at different levels such as syntactic information, lexical information, part of speech

information, negation words such as not, no, none which reverses the polarity, abbreviation, punctuations etc to perform supervised or semi-supervised learning .

3.2 System Architecture

The system Architecture is given Figure 2.1. the system is decomposed into tow main step:

1. Pre-Processing.
2. Main Processing.

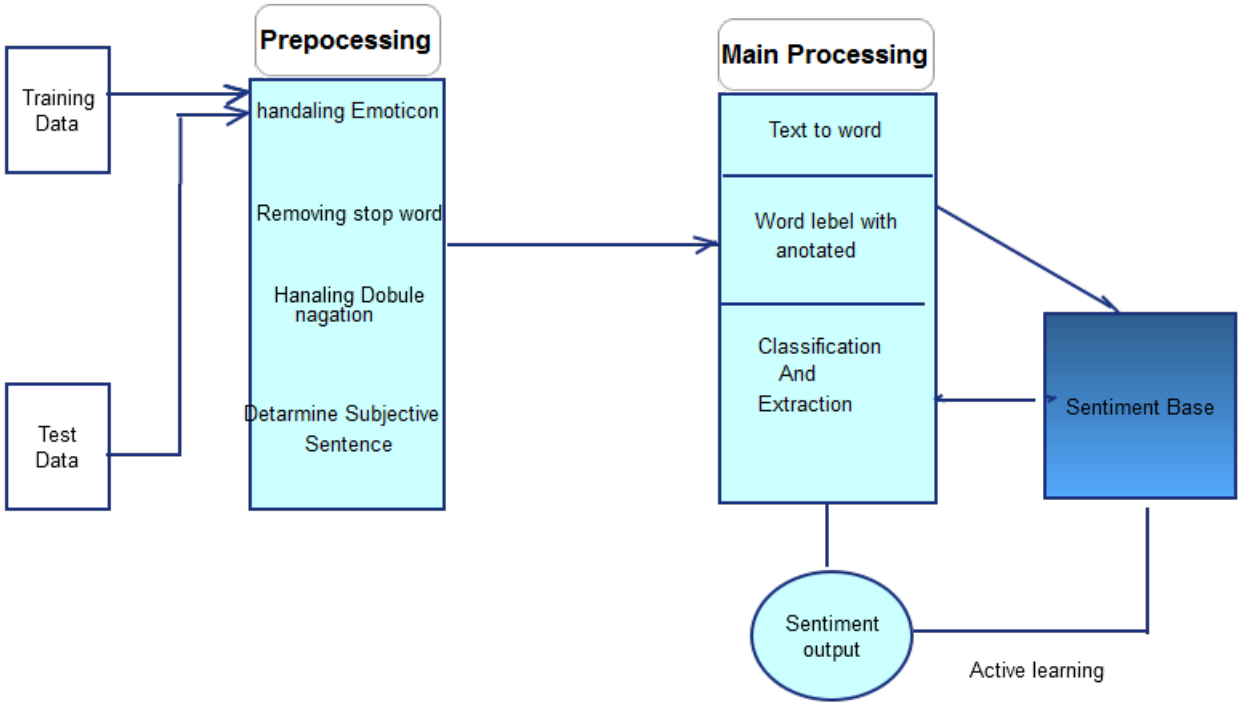

FIGURE 3.1: Architecture of sentiment analysis

Chapter 4

Pre-processing

4.1 Data Collection

Using the Facebook API Access we can collect the facebook status,facebook cooment,like e.t.c. we can also collect the twitter data though the twitter API also. we have collect the 30,000 facebook comment and status. we got this data in excle formate. All the collected data list are given below:

TABLE 4.1: Data Collection

origin	Total number of data
Ekattor Tv Status	92
Emran H Sarkar Status	100
Ekattor Tv Comment	3660
Magistrate Comment	7,000+
Emran H Sarkar Comment	20,000+

4.2 Data Analysis

A basic task in sentiment analysis is classifying the polarity of a given text at the document, sentence, or feature/aspect levelwhether the expressed opinion in a document, a sentence or an entity feature/aspect is positive, negative, or neutral. Advanced, "beyond polarity" sentiment classification looks, for instance, at emotional states such as "angry", "sad", and "happy".

we analysis a partion of data manualy which is discribe avobe. we divide sentiment into six basic catagory. "happy","sad","angry","surprise","disgust","fear". this analysis is given in tabular form in below:

TABLE 4.2: Manual Data Analysis

	Happy	Sad	Angry	Surprise	Disgust	Fear	Total
Ekattor Tv Status	24	14	3	12	7	4	64
Emran H Sarkar Status	13	22	9	17	5	6	72
Ekattor Tv Comment	153	274	226	162	56	81	952
Magistrate Comment	447	178	157	107	81	53	1023
Emran H Sarkar Comment	532	665	845	211	306	191	2750

4.3 Pre-processing

In this system, Pre-processing mainly used to simplyfiy the input data that is become suitable for the main process. we classify the pre-processing into basic five stage .

1.Basic Operation and Cleaning

2.Handaling Emoticon

3.Removing Stop word

4.Handaling Double Nagation.

5.Determining subjective sentence.

In our sentiment analysis system, we have implemented different forms of text pre-processing (Figure 4.1) directly over the raw training or testing data. A detail about the data set is left for the discussion in the next chapter while we limit our discourse in the context of only methodology here.

4.3.1 Basic Operation and Cleaning

This first module manages basic cleaning operations, which consist in removing unimportant or disturbing elements for the next phases of analysis and in the normalization of some misspelled words. In order to provide only significant information, in general a clean tweet should not contain URLs, hashtags (i.e. #happy) or mentions (i.e. @BarackObama). Furthermore, tabs and line breaks should be replaced with a blank and quotation marks with apexes. This is useful in order to obtain a correct elaboration by Weka (i.e. not closing a quotation mark causes a wrong read by the data mining software causing a fatal error in the elaboration). After this step, all the punctuation is removed, except for apexes, because they are part of grammar constructs such as the genitive.

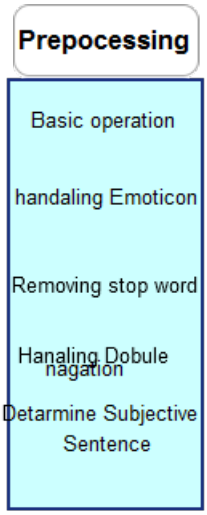

FIGURE 4.1: stage of Pre-processing

4.3.2 Emiticon Handaling

An emoticon is a pictorial representation of a facial expression using punctuation marks, numbers and letters, usually written to express a person's feelings or mood.

In Western countries, emoticons are usually written at a right angle to the direction of the text. Users from Japan popularized a kind of emoticon called kaomoji ; often confused with emoji in the West) that can be understood without tilting one's head to the left. This style arose on ASCII NET of Japan in 1986.

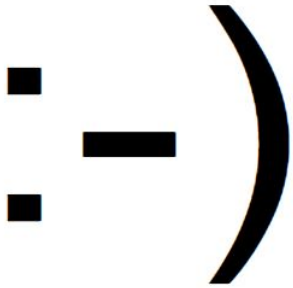

FIGURE 4.2: Smile emiticon (source: wikipidea)

As social media has become widespread, emoticons have played a significant role in communication through technology, and some devices have provided stylized pictures that do not use punctuation. They offer another range of "tone" and feeling through texting that portrays specific emotions through facial gestures while in the midst of text-based cyber communication.

This module reduces the number of emoticons to only two categories: smile positive and smile negative[50], as shown in Table 1

TABLE 4.3: List of substituted Emoticons

smile positive	smile negative
0:-)	>:(
:)	;(
:D	>:)
:*	>:)
:0	:(
:P	:|
:)	>:/

4.3.3 Removing Stop word

there a lot of word which do not convey any significance. if we want to remove this it does not any effect for sentiment analsis.In the sentiment analysis many of the words(abong,tai etc) is called stop word. If we remove this word in pre-processing stage then the efficiency of the system is increaseing.

As there is no single universal list of stop words used by all processing of natural language tools, and indeed not all tools even use such a list, we have maintained a list of such selective stop words and discarded them from the input text.

4.3.4 Handaling Double Nagation

The aim of sentiment analysis is to find out the positive and negative feelings, emotions and opinions written in a text. These sentiments are based on the meaning of words used in text according to different scenarios and situations. There are a variety of ways used to express the same feeling in a written text by using different grammatical rules. These grammatical rules contain negations that are very frequently used in text that completely change the meanings of words. In other words, negation identification and detecting its scope with in a sentence (text) are necessary in finding out the sentiments from a piece of

text. Although negation identification is an important aspect of sentiment analysis, it is yet to be properly addressed. In general, the efforts put into sentiment analysis of sentences having negation terms in them are less efficient with respect to general sentiment analysis. Negation identification is not a simple task and its complexity increase. Amna Asmi and Tanko Ishaya [50] decribe in breif in their paper about the english negation word. Here he said that if we change the nagation word then it can be change the whole meaning . so we need to be carefull about this.

for

4.3.5 Determining subjective sentence

Focuses on determining subjective words and texts that mark the presence of opinions and evaluationsvs. objective words and texts, used to present factual information. Subjective sentence means the sentence which contain sentiment or opinion. there are many sentence which is not effect on sentiment . we don't need to consider this types of sentence for determining Sentiment or opinion.

for example:

if there are total n sentence for test. we can find only m sentence for determining sentiment or opinion. here,

always $m <= n$.

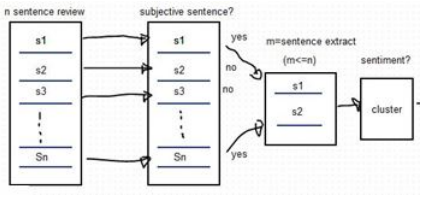

FIGURE 4.3: Detarming Subjective Sentence

Chapter 5

Main Processing

From the chapter 4 we can get m number of sentence. which is also normalized form. in the main processing the are three step:

1.Text to word.

2.Word lebel with annotated value.

3. Classification and extraction;

5.1 Text to word

we find the meaningful and usefull unit such as words,sentence or topics. farther we need to applying Naive Bayes(NB) Classifier, which uses a bag of word appoarch so that we have tokenized the input documents into words.

We have used StringTokenizer class of Java to tokenize our input documents. After this step a document becomes stream of words like below:

TABLE 5.1: documents to word

input text	Tokenized Output
i am a student	i,am , a ,student

5.2 Word lebel with annotated value

with the help of sentiwordnet we decide the annotated value of the word

5.2.1 SentiWordNet

SENTIWORDNET is the result of the automatic annotation of all the synsets of WORDNET according to the notions of positivity, negativity, and neutrality. Each synset s is associated to three numerical scores Pos(s), Neg(s), and Obj(s) which indicate how positive, negative, and objective (i.e., neutral) the terms contained in the synset are. Different senses of the same

term may thus have different opinion-related properties. For example, in SENTIWORDNET 1.0 the synset [estimable(J,3)]1, corresponding to the sense may be computed or estimated of the adjective estimable, has an Obj score of 1.0 (and Pos and Neg scores of 0.0), while the synset [estimable(J,1)] corresponding to the sense deserving of respect or high regard has a Pos score of 0.75, a Neg score of 0.0, and an Obj score of 0.25.

Each of the three scores ranges intheinterval [0.0,1.0], and their sum is 1.0 for each synset. This means that a synset may have nonzero scores for all the three categories, which would indicate that the correspond ingterms have, in the sense indicated by the synset, each of the three opinion related properties to a certain degree.

5.2.2 Creating Frequency Table

From each of the mega documents (positive and negative mega document for two classes) we have created in the last section, we have reformed them into frequency tables (one for each). Each frequency table contains the overall frequencies of each word, making a table (Figure: 5.1) of two columns.

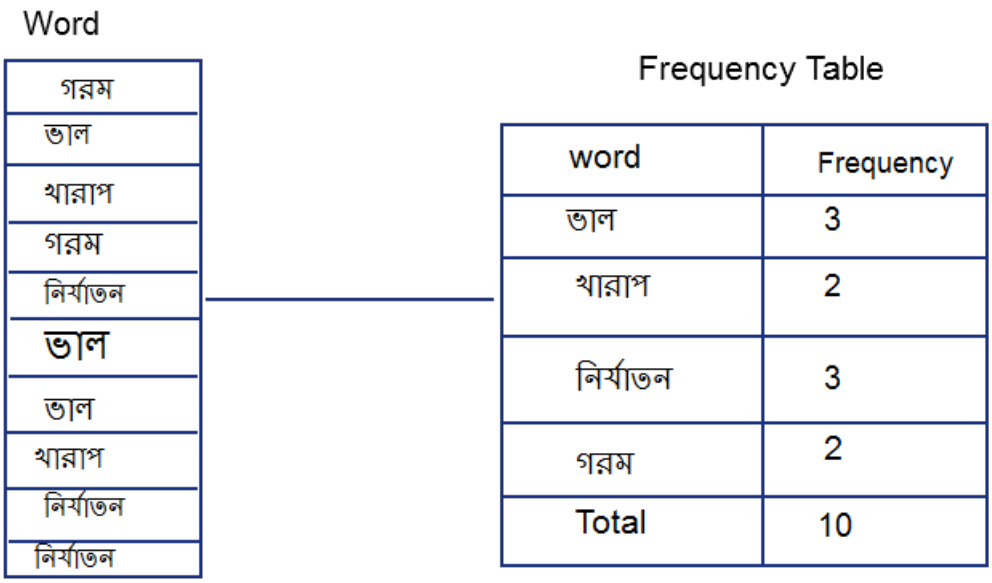

FIGURE 5.1: Frequency table of word

5.3 Classification and extraction

With the bag-of-words model we check which word of the text-document appears in a positive-words-list or a negative-words-list. If the word appears in a positive-words-list the total score of the text is updated with +1 and vice versa. If at the end the total score is positive, the text is classified as positive and if it is negative, the text is classified as negative. Simple enough!

With the Naive Bayes model, we do not take only a small set of positive and negative words into account, but all words the NB Classifier was trained with, i.e. all words presents in the training set. If a word has not appeared in the training set, we have no data available and apply Laplacian smoothing (use 1 instead of the conditional probability of the word). The probability a document belongs to a class C is given by the class probability P(C) multiplied by the products of the conditional probabilities of each word for that class.

$P = P(C) \cdot \prod_i P(d_i|C) = P(C) \cdot \prod_i^n \frac{count(d_i,C)}{\sum_i count(d_i,C)} = P(C) \cdot \prod_i \frac{count(d_i,C)}{V_C}$

Here count(d_i, C) is the number of occurences of word d_i in class C , V_C is the total number of words in class C and n is the number of words in the document we are currently classifying. V_C does not change (unless the training set is expanded), so it can be placed outside of the product:

$P = \frac{P(C)}{V_C^n} \cdot \prod_i^n count(d_i, C)$

5.3.1 Boolean Multinomial Naive Bayes on a test document d

First remove all duplicate words from d

Then compute NB using the Same Equation:

$$c_{NB} = \underset{c_j \in C}{\operatorname{argmax}} P(c_j) \prod_{i \in positions} P(w_i \mid c_j)$$

we are trying to compute the prior probably for particular document being in a class j, we have the count of all the documents, and out of those documents how many of the documents are in class j (in training set).

5.4 Active Learning

In our sentiment analysis system, weve incorporated the provision for Active Learning by means of which a user may give feedback about the output of a given test document by

suggesting which class it really belongs if he/she thinks that the output of the system is wrong. The system will then add that test document as part of its training documents with the corresponding class. This way our system can eventually adopt with immense domain and able to learn new things and could be more accurate with each feedback.

Chapter 6

Experiment and Evaluation

Because of lackness of time, we can not done whole experiment. But our expected result discuss below:

we have about 20,000 facebook status and comment in excel formate. we describe chapter 4 of our facebook data. we have manually done this task(analysis). here about 5,000 Data are manually doing this. we can use this data set for training set.

our expected result are describe though "Recall" and "Accuracy" which are measured for text categorization purpose. lets see the table:

TABLE 6.1: Contingenct Table for Recall & Accuracys Calculation

class	Correct	Not Correct
Positive	t_p	f_p
negative	f_n	t_n

The meaning of the four terminology in the above table (Table: 6.1) are mentioned below:

- t_p = True positive (looking for positive and system also detected it as positive).
- f_p = False positive (was actually negative but system detected it as positive).
- f_n = False negative (was actually positive but system detected it as negative).
- t_n = True negative (looking for negative and system also detected it as negative).

Now we are in a condition to introduce the equation of Recall and Accuracy:

Recall = $\frac{t_p}{t_p+f_n} = \frac{t_n}{t_n+t_p}$

Accuracy = $\frac{t_p+t_n}{(t_p+f_p)+(f_n+t_n)} = \frac{t_p+t_n}{No.of totaltestdocuments}$

The calculation for the recall measure is done separately for positive and negative classes which is clear from the above equation of recall whereas the calcu lation for accuracy is a combined measure which acts as an average performance for both the positive and negative scenarios.

Chapter 7

Future Work And Conclution

Sentiment Analysis has been quite popular and has lead to building of better products, understanding users opinion, executing and managing of business decisions. With rapidly increasing technology, the early approach of word-of-mouth has been shifted towards the mass opinion what the people like and appreciate in majority. People rely and make decisions based on reviews and opinions. The rise in user-generated content for Hindi language across various genres- news, culture, arts, sports etc has open the data to be explored and mined effectively, to provide better services and facilities to the consumers.

we have tried to experiment but we are failing because of lackness of time. next, we do the experiment. here we are tried to extract only positive or negative. The more we can train our system, the better it will perform. The current performance of our system is noticeable campared to many of the baseline methods.

We need to come with a notion of prior sentiment polarity for set of words in the form of a subjective lexicon. Then we can explore and dig in depth regarding the task of sentiment classification for the web text and improve over same. The main issues while working with bangla languages comes- while handling the morphological variations, identifying context, performing word sense disambiguation and handling multiple spellings with each of them in itself is a research problem. Much work needs to be done at this level to address the above challenges that can boost up and help the research in the related areas. As discussed above there is lack of annotated datasets and resources for bangla languages, so it needs considerable focus and time to be given. The basic resources like part of speech tagger, morphological analyzers, named entity recognizers and parsers have also not yet reached the state of art accuracy and needs improvement. Once we have sufficient data to experiment with, various machine learning techniques can be easily used and applied to learn from the text more effectively.

References

[1] D. Das and S. Bandyopadhyay .Emotions on Bengali Blog Texts: Role of Holder and Topic [2] Bing Liu. Sentiment analysis: A multi-faceted problem. IEEE Intelligent Systems, 25(3):7680, 2010.

[1] A. Agarwal, F. Biadsy, and K. R. Mckeown. Contextual phrase-level polarity analysis using lexical

affect scoring and syntactic n-grams, 2009. [2] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau. Sentiment analysis of twitter data. In

Proceedings of the Workshop on Languages in Social Media, LSM 11, pages 3038, Stroudsburg, PA,

USA, 2011. Association for Computational Linguistics. [3] S. Baccianella, A. Esuli, and F. Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for

sentiment analysis and opinion mining. In Proceedings of the Seventh conference on International

Language Resources and Evaluation (LREC10), Valletta, Malta, may 2010. European Language

Resources Association (ELRA). [4] A. Balahur, R. Steinberger, M. A. Kabadjov, V. Zavarella, E. V. der Goot, M. Halkia, B. Pouliquen, and

J. Belyaeva. Sentiment analysis in the news. In LREC. European Language Resources Association, 2010. [5] F. Benamara, C. Cesarano, A. Picariello, D. Reforgiato, and V. S. Subrahmanian. Sentiment analysis:

Adjectives and adverbs are better than adjectives alone. In Proceedings of the International Conference

on Weblogs and Social Media (ICWSM), 2007. Short paper. [6] J. Blitzer, M. Dredze, and F. Pereira. Biographies, bollywood, boomboxes and blenders: Domain

adaptation for sentiment classication. In ACL, 2007. [7] K. Bloom, N. Garg, and S. Argamon. Extracting appraisal expressions. In Proceedings of Human

Language Technologies/North American Association of Computational Linguists, 2007. [8] R. M. Carmen Banea and J. Wiebe. A bootstrapping method for building subjectivity lexicons for

languages with scarce resources. In Proceedings of LREC08, 2008. [9] P. Chesley. Using verbs and adjectives to automatically classify blog sentiment. In Proceedings of

AAAI-CAAW-06, the Spring Symposia on Computational Approaches, 2006. [10] A. Das and S. Bandyopadhyay. SentiWordNet for Bangla. 2010. [11] A. Das and S. Bandyopadhyay.

SentiWordNet for Indian Languages. 2010. 60 [12] D. Das and S. Bandyopadhyay. Labeling emotion in bengali blog corpus a fine grained tagging at

sentence level. In Proceedings of the Eighth Workshop on Asian Language Resouces, pages 4755,

Beijing, China, August 2010. Coling 2010 Organizing Committee. [13] K. Dave, S. Lawrence, and D. M. Pennock. Mining the peanut gallery: Opinion extraction and

semantic classification of product reviews. pages 519528, 2003. [14] D. Davidov, O. Tsur, and A. Rappoport. Enhanced sentiment learning using twitter hashtags and

smileys. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters,

COLING 10, pages 241249, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. [15] N. A. Diakopoulos and D. A. Shamma. Characterizing debate performance via aggregated twitter

sentiment. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 10,

pages 11951198, New York, NY, USA, 2010. ACM. [16] G. Draya, M. Planti, A. Harb, P. Poncelet, M. Roche, and F. Trousset. Opinion mining from blogs. In

International Journal of Computer Information Systems and Industrial Management Applications

(IJCISIM), 2009. [17] A. Esuli and F. Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In

Proceedings of the 5th Conference on Language Resources and Evaluation (LREC06, pages 417422,

2006. [18] A. Go, R. Bhayani, and L. Huang. Twitter sentiment classification using distant supervision. pages

16, 2009. [19] N. Godbole, M. Srinivasaiah, and S. Skiena. Large-scale sentiment analysis for news and blogs. In

Proceedings of the International Conference on Weblogs and Social Media (ICWSM), 2007. [20] V. Hatzivassiloglou and K. R. McKeown. Predicting the semantic orientation of adjectives. In

Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth

Conference of the European Chapter of the Association for Computational Linguistics, ACL 98, pages

174181, Stroudsburg, PA, USA, 1997. Association for Computational Linguistics. [21] B. He, C. Macdonald, J. He, and I. Ounis. An effective statistical approach to blog post opinion

retrieval. In Proceedings of the 17th ACM conference on Information and knowledge management, CIKM

08, 2008. [22] M. Hu and B. Liu. Mining and summarizing customer reviews. In KDD, pages 168177, 2004. [23] T. W. Intelligent and T. Wilson. Annotating opinions in the world press. In In SIGdial-03, pages 1322,

2003. 61 [24] A. Joshi, B. A. R, and P. Bhattacharyya. A fall-back strategy for sentiment analysis in hindi: a case

study, 2010. [25] J. Kamps, M. Marx, R. J. Mokken, and M. D. Rijke. Using wordnet to measure semantic orientation

of adjectives. 2004. [26] A. Karthikeyan. Hindi english wordnet linkage. [27] L.-W. Ku, Y.-T. Liang, and H.-H. Chen. Opinion extraction, summarization and tracking in news and

blog corpora. In AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs, 2006. [28] P. Melville, W. Gryc, and R. D. Lawrence. Sentiment analysis of blogs by combining lexical

knowledge with text classification. In Proceedings of the 15th ACM SIGKDD international conference on

Knowledge discovery and data mining, KDD 09, 2009. [29] G. A. Miller. Wordnet: a lexical database for english. Commun. ACM, 38(11):3941, Nov. 1995. [30] S. min Kim. Determining the sentiment of opinions. In Proceedings of COLING, pages 1367 1373,

2004. [31] S. min Kim and E. Hovy. Identifying and analyzing judgment opinions. In Proceedings of

HLT/NAACL-2006, pages 200207, 2006. [32] D. Narayan, D. Chakrabarti, P. Pande, and P. Bhattacharyya. An experience in building the indo

wordnet - a wordnet for hindi. In First International Conference on Global WordNet, 2002. [33] T. Nasukawa and J. Yi. Sentiment analysis: capturing favorability using natural language processing.

In Proceedings of the 2nd international conference on Knowledge capture, K-CAP 03, pages 7077, New

York, NY, USA, 2003. ACM. [34] B. Ohana and B. Tierney. Sentiment classification of reviews using sentiwordnet. 2009. [35] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment classification using machine learning

techniques. In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing

(EMNLP), pages 7986, 2002. [36] D. Rao and D. Ravichandran. Semi-supervised polarity lexicon induction. In Proceedings of the 12th

Conference of the European Chapter of the Association for Computational Linguistics, EACL 09, pages

675682, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. [37] E. Riloff and J. Wiebe. Learning extraction patterns for subjective expressions. In Proceedings of the

2003 conference on Empirical methods in natural language processing, EMNLP 03, pages 105112,

Stroudsburg, PA, USA, 2003. Association for Computational Linguistics. 62 [38] H. Saggion and A. Funk. Interpreting sentiwordnet for opinion classification. In LREC, 2010. [39] P. J. Stone, D. C. Dunphy, M. S. Smith, and D. M. Ogilvie. The General Inquirer: A Computer

Approach to Content Analysis. MIT Press, Cambridge, MA, 1966. [40] P. Turney. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of

reviews, 2002. [41] J. Wiebe. Learning subjective adjectives from corpora. In Proceedings of the Seventeenth National

Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial

Intelligence, pages 735740. AAAI Press, 2000. [42] J. M. Wiebe, R. F. Bruce, and T. P. OHara. Development and use of a gold-standard data set for

subjectivity classifications. In Proceedings of the 37th annual meeting of the Association for Computational

Linguistics on Computational Linguistics, ACL 99, pages 246253, Stroudsburg, PA, USA, 1999.

Association for Computational Linguistics. [43] T. Wilson. Recognizing contextual polarity in phrase-level sentiment analysis. In Proceedings of

HLT-EMNLP, pages 347354, 2005. [44] H. Yu and V. Hatzivassiloglou. Towards answering opinion questions: separating facts from opinions

and identifying the polarity of opinion sentences. In Proceedings of the 2003 conference on Empirical

methods in natural language processing, EMNLP 03, pages 129136, Stroudsburg, PA, USA, 2003.

Association for Computational Linguistics. [45] C. Zhang, W. Zuo, T. Peng, and F. He. Sentiment classification for chinese reviews using machine

learning methods based on string kernel. In Proceedings of the 2008 Third International Conference on

Convergence and Hybrid Information Technology - Volume 02, pages 909914, Washington, DC, USA,

2008. IEEE Computer Society. [46] W. Zhang and et al. Opinion retrieval from blogs. In Proceedings of the sixteenth ACM conference

on Conference on information and knowledge management (2007), 2007. [47] Huifeng Tang, Songbo Tan, and Xueqi Cheng. A survey on sentiment detection of reviews. Expert

Systems with Applications, 36(7):1076010773, 2009. [48] Jaap Kamps, MJ Marx, Robert J Mokken, and Maarten De Rijke. Using wordnet to measure

semantic orientations of adjectives. 2004. [50]Amna Asmi and Tanko Ishaya. Negation Identification and Calculation in Sentiment Analysis