Bengali Sentiment Analysis Using Modified English VADER

Imran Hossain

Al Amin

Computer Science and Engineering Discipline
Khulna University
Khulna-9208, Bangladesh
January, 2019

Bengali Sentiment Analysis Using Modified English VADER

Imran Hossain
Roll No: 150203

Al Amin
Roll No: 150212

Computer Science and Engineering Discipline
Khulna University
Khulna-9208, Bangladesh

January, 2019

Bengali Sentiment Analysis Using Modified VADER

A thesis submitted in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science and Engineering (CSE).

Dr. Kazi Masudul Alam Thesis Supervisor

Associate Professor

Computer Science and Engineering Discipline

Dr. Abu Shamim Mohammad Arif Second Examiner

Professor

Computer Science and Engineering Discipline

Dr. Md. Anisur Rahman Head of the discipline

Professor

Computer Science and Engineering Discipline

Acknowledgement

First of all, we thank the Almighty Allah, the most gracious, the most merciful, who has given us the ability to complete this study within the stipulated time.

Secondly, we would like to express our heartfelt thanks and gratitude to our honourable supervisor **Dr. Kazi Masudul Alam**, associate professor, Computer Science and Engineering Discipline, Khulna University, Khulna, for encouraging us to think and do something out of box. Without his continuous cooperation, suggestion, observation and inspiration, we couldn't reach our goal. He always provided us with his precious time, valuable comments, suggestions and support throughout this thesis.

We want to express our deepest gratitude to our honourable Head **Dr. Md. Anisur Rahman**, professor, Computer Science and Engineering Discipline, Khulna University, Khulna, for continuous inspiration and valuable advice.

Moreover, we would like to express our gratitude to all our respected teachers of Computer Science and Engineering Discipline for their goodwill support and advice.

We also thank our friends and seniors for their heartiest cooperation and encouragement that were very much helpful to us. Finally, we are very grateful to our family who has supported in our every step of life.

Abstract

Sentiment analysis is an essential field of natural language processing (NLP) that classifies the opinion expressed in a text according to its polarity (e.g., positive, negative or neutral). Bengali NLP research is lagging behind English NLP, where there are very few works on Bengali sentiment analysis. In this paper, we approach this issue by modifying a popular tool VADER to support Bengali sentiment polarity identification. We have compiled a Bengali polarity lexicon from the English polarity lexicon of VADER. Furthermore, we have modified the functionalities of English VADER, so that it can directly classify Bengali text sentiments without the previous requirement of Bengali to English translation using tools such as Google Translator, MyMemory Translator, etc. Our experiments show that the modified Bengali VADER significantly improves the sentiment analysis result of Bengali text.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

Sentiment is a term related to sensitiveness or emotional feelings. Sentiment is a genuine and refined sensibility, an inclination to be affected by feeling as opposed to reason or reality[1]. The detailed examination that is done in order to understand the nature or to determine the essential features of anything complex through study is called analysis[2]. The sort of information mining that measures and determines the tendency of individuals' opinions through natural language processing (NLP), computational linguistics and text analysis is known as sentiment analysis; which are utilized to extract and analyze emotional data from the Internet - generally social media and similar sources[3]. According to the Oxford dictionary, sentiment analysis is the process which identifies and categorizes opinions expressed in a piece of text, in order to determine whether the user's view is positive, negative, or neutral towards a specific theme, item or so forth.

Interest from various brands, companies, organiztions and researchers has increased in sentiment analysis and its application in recent years to business analytics and other purposes. In business the application of sentiment analysis cannot be underestimated. For the whole brand revival sentiment analysis can play a major role. Analyzing the contents of social media one can get an excellent source of information and which can provide insights that can: determine marketing strategy, improve campaign success, improve product messaging, improve customer service, test business KPIs, generate leads. In a nutshell, it can be said that one's bottom line can be improved using social media analysis. In political view it is also important to know the public opinion about

[1] http://www.dictionary.com/browse/sentiment
[2] https://www.merriam-webster.com/dictionary/analysis
[3] https://www.techopedia.com/definition/29695/sentiment-analysis

any law or policies of the government, analyzing the sentiment of the people the government can take decision accordingly to avoid unwanted circumstances.

Bengali is one of the most used languages in the world. Almost 250 million population of the world speak in Bengali. It is the primary language in Bangladesh and secondary language in India [2] [3] [4]. In 2015, the British Council published a report on the most important languages for the future. They considered several different factors, one of them is languages spoken in the fastest-growing emerging economies by 2050. Out of these emerging economies, Bengali is expected to be the third most commonly spoken language, below Chinese and Hindi [5]. So even if it is not a top-priority language for business right now, that could change soon.

However, there has not been done many notable works on sentiment analysis on Bengali compared to English. As technology is spreading everywhere the amount of Bengali text is increasing on the internet, so it is essential to have an efficient tool to analyze the sentiment from Bengali text. In this literature we have discussed Bengali polarity lexicon in the section 3.1 that we have created and then have briefly discussed VADER [1] in the section 2.2.1. In the section 3 we have discussed how we have modified the VADER to analyze sentiment from Bengali language. We have exhibited a study on the three different techniques to build models for text classification. The first two techniques are unigram lexicon-based approach and N-gram lexicon based (i.e. Bi-gram, Tri-gram) approach and Third technique is TF-IDF model which is common in text classification. The performance of these techniques on several different machine learning algorithms is also used.

1.2 Motivation

Though Bengali is one of the major languages of the world there has been done few works on Bengali sentiment analysis. Bengali language processing is now the demand of time. Because of few works on Bengali it is still a resource constrained language.There is very few well known Bengali polarity lexicon. There has been some work done on Bengali but most of them is domain based. We wanted to make a contribution in Bengali language processing, specially in sentiment analysis. So to enrich the Bengali language processing and developing its modern tool is the main motivation of our work.

1.3 Objectives

The objective of our work is to create a Bengali polarity lexicon and modify the VADER system for Bengali sentiment analysis. Moreover our work covers the following objectives:

- Develop the tool for Bengali language processing.
- Develop an efficient system for Bengali sentiment analysis.
- Enrich the Bengali lexical resources.
- Construct Bengali polarity lexicon that would be helpful for sentiment scoring.

1.4 Research Publication

"Bengali VADER: A Sentiment Analysis Approach Using Modified VADER." In the 2nd International Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 February, 2019.

1.5 Organization of the Thesis

This thesis has been organized into five chapters. Each chapters gives distinct concept.

Chapter 1 (Introduction): Introduction of our research area has been explored in this chapter.

Chapter 2 (Background and Related Works): This chpater presents the basic concepts and background of sentiment analysis and lexicon. English VADER is also discussed in this chapter.

Chapter 3 (Proposed Bengali VADER): A detail discussion of our proposed methodology of Bengali VADER and created lexicon is presented in this chapter.

Chapter 4 (Results and Discussion): Exprimental results and evaluation of reults are shown.

Chapter 5 (Conclusion and Future Works): Summarization of our research work. Some limitaions and future plan of our research is also included.

Chapter 2

Background and Related Works

In this chapter, the related concepts on sentiment analysis, its application and background of the work is presented. More importantly the detail about VADER is discussed.

2.1 Backgound

Sentiment analysis, a filed of natural language processing (NLP) bears a great importance in todays world. VADER which is the short form of 'Valence Aware Dictionary for sEntiment Reasoning' a popular system for sentiment analysis in English language. SentiWordNet is another popular tool for sentiment analysis, which is a lexical resource. SentiWordNet assaigns positivity, negativity and objectivity score for each word.

Sentiment analysis can be done mainly in three approaches [6]: (i) Machine learning approach, (ii) Lexicon based approach and (iii) Hybrid approach. All the three methods have some advantages and disadvantages.

2.1.1 Machine learning approach

Machine learning approach uses various classification techniques such as: Naive Bayes Classification, Support Vector Machine etc. to classify text. This approach need two datasets, one for training and another for testing. To apply machine learning approach a model TF-IDF is constructed at first, then the techniques are used for classification.

2.1.1.1 TF-IDF

TF-IDF or most commonly known as TFIDF is the short form of 'term frequency-inverse document frequency'. The term is used to reflect how important a word in a collection of documents [7]. It is used to get the importance or weight of a word. The weight is calculated using the following equation:

$$weight = tf * log(\frac{N}{df}) \tag{2.1}$$

where tf is term frequency, N is total number of documents and df is the document frequency.

2.1.1.2 Cross-validation

Cross-validation also known as rotation estimation is one of the popular model validation or prediction technique which is used to know how well a system will work in practice [8]. We used k-fold cross validation to test our system. The method randomly partition the data into k equal size subsamples, then randomly separate one subsample set for testing and the remaining $k-1$ are used for training. Used $k = 10$ for our system testing.

2.1.1.3 Naive Bayes

Based on Bayes' theorem with the freedom of independence assumptions between features naive Bayes is the family of simple probabilistic family [9]. The equation for the classifier:

$$p(C_k|x) = \frac{p(C_k)p(x|C_k)}{p(x)} \tag{2.2}$$

2.1.1.4 Support Vector Machine

Support vector machine also commonly known as it's short form SVM is a supervised learning model which analyzes the data using it's traing model [10]. We used it's linear classifier to train and test our data.

The advantage of machine learning approach is its ability to quick adaptation and creation of trained models for specific reasons and contexts. But to apply model every data need to be modeled.

2.1.2 Lexicon based approach

In lexicon based approach a polarity lexicon is used and match with the words to determine polarityl. A lexicon, word-store, wordbook or word-stock is the vocabulary of a man, dialect, or department

of expertise. A lexicon is a language's stock of lexemes in linguistics. Polarity lexicons are that which have a listing of words with earlier polarities. It is one of the main resources for analyzing the sentiments and opinion expressed in texts in an computerized way.

2.1.2.1 Classification of Lexicon

There are primarily three ways to build polarity lexicon: (i) interpreting existing lexicons from different languages, (ii)extracting polarity lexicons from corpora [11], and (iii) annotating sentiments Lexical Knowledge Bases [12] [13] [14] [15]. The first technique is fast but my be error porne if proper good bilingual dictionary is not used and proper care is not taken about the polarity score. Second technique can be relatively more accurate. The third one is a time consuming and difficult task.

2.1.2.2 Lexicon in other Languages

For major languages there are well known manually constructed lexicons, such as, General Inquirer [16], OpinionFinder [17], SO-CAL [18]. Literature [19] and [20] analyze the approach of translating English resources into Romanian and Spanish, respectively.

2.1.3 Hybrid approach

Hybrid approach is the combination of both machine learning and lexicon based approaches [6]. It combines machine learning and lexcon based approach using semantics rules. It has the potential to improve the preformance of sentiment classification. The main advatage of this approach is that it can detect and measure the sentiment at concept level.

2.1.4 Dataset

Dataset or collection of data is one of the major elements in natural language processing. Though the method we have used to analyze sentiment is lexicon based it is important to use a good dataset to test the benchmark of the system. As Bengali is resource constrained language it is very difficult to get a good dataset. We used the following three dataset in our system.

2.2 Related Works

There are a few works of calculating sentiment from Bengali using Bengali polarity lexicon, but there are some works in English as SentiWordNet [21], VADER etc. In Bengali there are some other methodologies have been used to detect sentiment as in the literature [22] to compute the total positivity, negativity or neutrality of the sentence or document firstly they used the WordNet to get the senses of words according to their respective parts of speech and then used SentiWordNet to get their prior valence or polarity.

In the literature [23] they have translated the SentiWordNet in Bangla and used its polarity to calculate the valence of the Bangla text. They have used a semi-supervised bootstrapping approach for the development of the training corpus and for classification used Support Vector Machines (SVM) and Maximum Entropy (MaxEnt) and then they combined various set of features to analyse the performance of these two machine learning algorithms comparatively. For the procedure-based classifier they also buit a Twitter-specific Bengali sentiment lexicon and as a binary feature the classifiers used.

In English language, there are notable works on sentiment analysis. VADER (for Valence Aware Dictionary for Sentiment Reasoning) is one of them. Firstly they combined qualitative and quatitative methods to produce a gold-standard sentiment lexicon, and then empirically validated the lexicon which is especially receptive to microblog-like contexts. Next, considering the five generalizable rules that embody grammatical and syntactical conventions that humnas use when expressing or emphasizing sentiment intensity, they combined these lexical features. VADER retains the advantages of conventional sentiment lexicons like LIWC [24] [25].

In the literature [15] they have explored three strategies to build polarity lexicons: interpreting existing lexicons from other languages, explaining sentiments Lexical Knowledge Bases and extracting polarity lexicons from corpora. Different degrees of human effort is required for each of these method. Spanish lexicon $ElhPolar_{es}$ [26] has been translated by means of the Elhuyar Spanish-Basque dictionary[1]. For every Spanish word in the lexicon, the initial five interpretations are incorporated into the translated lexicon Lex_{pr}.

In the literature [27] the construction of an opinion outline system have been described that works on Bengali news corpus. At first the sentiment information in each document is indentified by the system, and aggregated and summarized by it. To identify and aggregate the sentiment the system followed a topic-sentiment model. Theme clustering (k-means) and document level

[1] http://hiztegiak.elhuyar.eus

theme relation graph represntation was used to achieve discourse level theme identification and topic sentiment aggregation which is the designed topic-sentiment model. Finally standard page rank algorithms which is utilized in Information Retrieval (IR) for candidate precis sentence selection used the document level theme relation graph. Bengali being a asset-restrained language on paper [27] they described the construction of annotated best quality corpus; they also discussed about the acquisiton of linguistics geat for lexico-syntactic and discourse stage features extraction.

2.2.1 VADER

VADER is the short form of 'Valence Aware Dictionary for sEntiment Reasoning' [1]. They have combined qualitative and quantitative methods to produce, and then empirically validate a *gold-standard* sentiment lexicon. A tremendous number of sentiment analysis processes rely significantly on a basic sentiment (or opinion) lexicon. A *sentiment lexicon* is a listing of lexical capabilities (e.g., words) which can be commonly categorised in keeping with their semantic orientation as either positive or negative [28]. After producing and validating lexicon, they have used the architecture shown in Fig. 2.1 to evaluate the sentiment of texts. The architecture mainly follows the steps described below to calculate the polarity of the sentences:

2.2.1.1 Preprocess

In this step after taking the text input, it tokenizes the text. Tokenization is done in the following two styles:

- **Words and Emoticons:** In this step, the text is tokenized according to words, emoticons and all-caps words.
- **Words Plus Punctuation:** Tokenization is done of words and punctuation marks. Some punctuation marks affect the valence of words, that is why it is kept with the word.

2.2.2 Boosting

After tokenization, the tokens are checked for boosting words. To check the boosting words, they have used bi-gram and tri-gram. If a boosting word, e.g. - "extremely, very, great, etc." is found then the valence of the word is boosted. Then it is checked if the word is all caps. If the word is all caps, then it is boosted. The text is also checked for idioms and phrases, if found then it is boosted.

Then it is checked for ‘but’ word, if found then the sentence is divided into two sections and valence is calculated as two different sentences and then is calculated the overall valence of the sentence.

2.2.3 Valence Calculation

In this step, the valence of the sentence is calculated. The calculated valence is between -4 to +4 then it is normalized to range between -1 to +1. Every sentence is given their respective polarity.

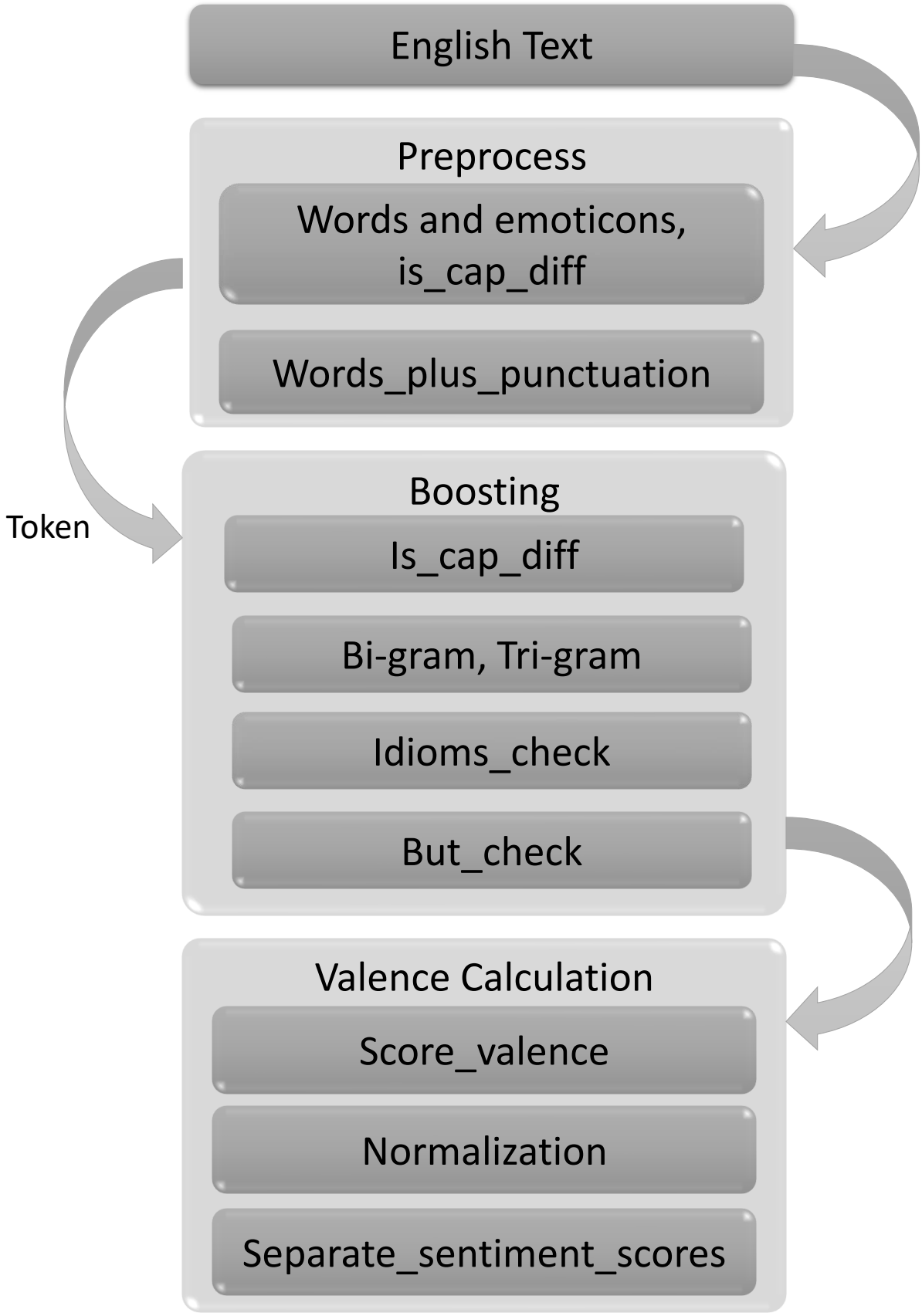

Figure 2.1: Implementation architecture of English VADER [1]

Chapter 3

Proposed Bengali VADER

We have created our methodology to extract expressed sentiment from texts, and this includes pre-processes, tokenization, negation handling, etc. Fig. 3.1 shows an architecture of the research process and summarizes the methods used in this study. This paper focuses on two interrelated things: (1) the development and validation of a Bengali polarity lexicon, which is described on section 3.1 and (2) extracting the intensity of sentiment expressed in Bengali text using our methods. The processes are described accordingly in the below sections.

3.1 Improved Lexicon

There are several English polarity lexicons available online such as SentiWordNet[1], VADER, etc. but we could not locate any Bengali polarity lexicon. We have used two methods to construct bengali polarity lexicon: (1) Translation Approach and (2) Manual approach.

3.1.1 Translation Approach

As we could not find any polarity lexicon in Bengali, we constructed a lexicon translating VADER [1] lexicon using a bilingual dictionary and given their corresponding polarity. The valence of the words are given in a range of -4 to +4, where -4 represents most negativity and +4 represents most positivity and 0 represents as neutral. This lexicon consists of almost 3000 polarity words.

[1] http://sentiwordnet.isti.cnr.it/

3.1.2 Manual Approach

To enrich our translation based polarity lexicon we have manually created another 3500+ polarity words. Firstly we collected the Bengali words from [29]. Then we built an website[2] to know people's opinion about the words giving their polarity. Then the given polarity for each word was averaged and rechecked.

3.1.2.1 Bi-gram and Tri-gram lexicon

For bi-gram lexicon there is a pair of two words and tri-gram lexicon is constructed as a pair of three words. The words are concatanated using _. We manually gave polarity to words in this lexicon. Example of bi-gram is 'স্নায়বিক_ দুর্বলাবস্থা' and tri-gram is 'হামাগুড়ি_দিয়া_আরোহণ'

3.2 Dictionary

We have created two dictionaries of negation words and booster words[3]. They are used for negation and boosting of a sentence respectively.

3.2.1 Negation List

We created a list of words that are used to check the negativity of the texts. If the words of negation list exist in the text, then the polarity of the text is reversed to positive or negative. The list contains words such as:

(1) 'না', 'নি', 'নয়', 'নাই', 'নেই'

3.2.2 Booster Dictionary

Booster dictionary contains the words that are used to boost the valence of the text if it contains any of the boosting words in it. The dictionary contains words such as:

(2) 'বেশি', 'খুব', 'অনেক', 'অতি', 'অতিশয়', 'বহুত', 'অধিক', 'অধিকতর'

[2]http://imran03.pythonanywhere.com/
[3]https://github.com/Imran-cse/Bengali-Sentiment-Analysis

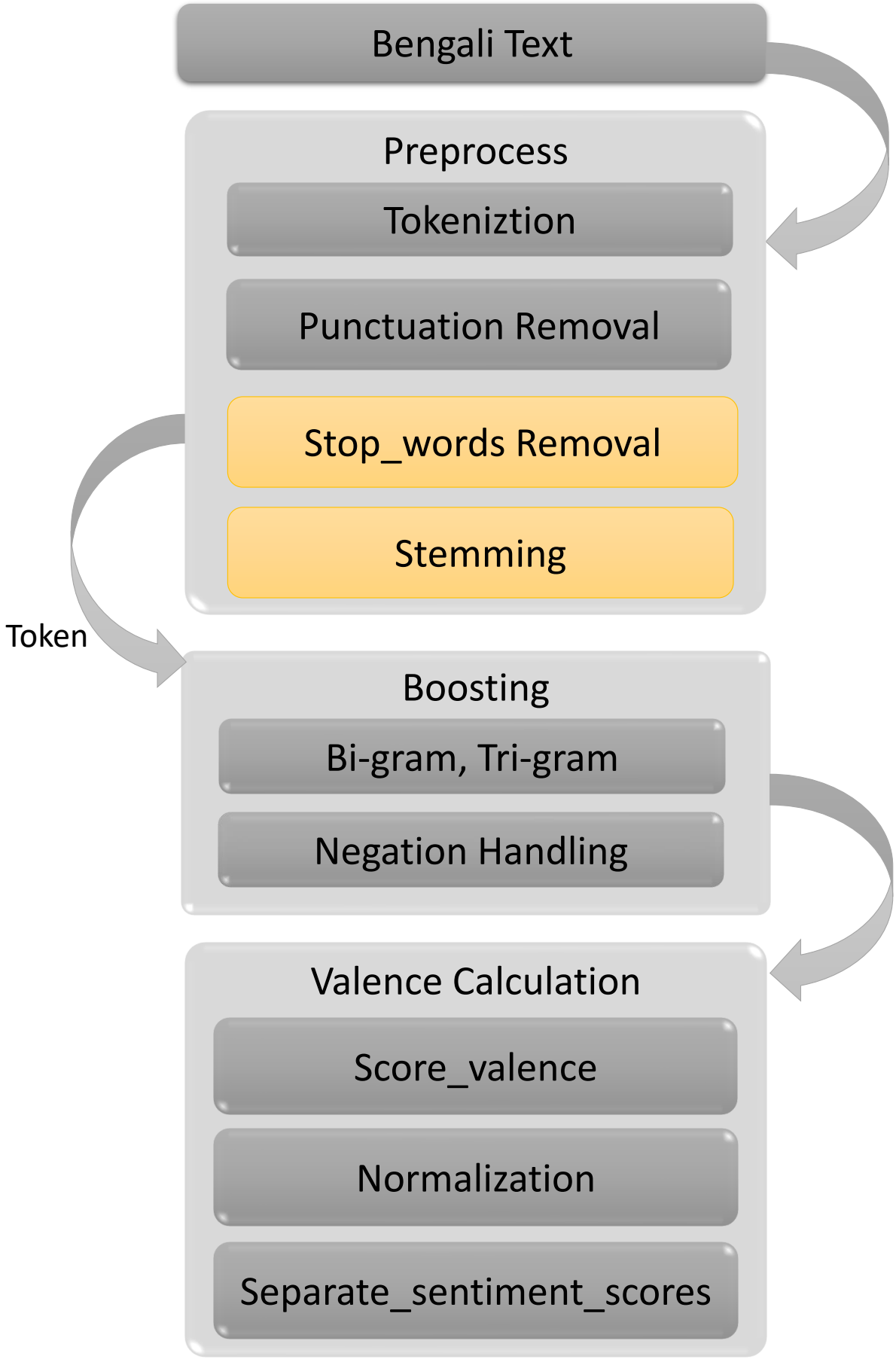

Figure 3.1: System Architecture of Bengali VADER

3.3 Preprocess

Preprocessing is a consequential step in the data mining process. These includes tokenization and punctuation removal, stop words removal, stemming, and N-gram (i.e. Bi-gram, Tri-gram) word generator. Before starting to process the text. We have done preprocessing in the following steps:

3.3.1 Tokenization and Punctuation removal

The text is split into words or tokenized in this step. The next step is to remove punctuation marks from the texts; this removes all the unnecessary punctuation marks from the list of tokens. Original text:

(3) সে খুব বেশি ভয় পায়নি ।

After tokenization the sentence looks like this:

(4) ‘সে’, ‘খুব’, ‘বেশি’, ‘ভয়’, ‘পায়নি’ ‘।’

Then removing punctuation marks the token list looks like:

(5) ‘সে’, ‘খুব’, ‘বেশি’, ‘ভয়’, ‘পায়নি’

3.3.2 Bengali stop-words removal

Stop words are those which are ignored during sentiment analysis. We have created a list of Bengali stop words. The list contains words such as ‘সে’, ‘আমি’, ‘এবং’ etc. Whenever these words are found in the token list they are removed from the token list. After removing the stop-words the token list of ‘সে খুব বেশি ভয় পায়নি’ looks like this:

(6) ‘খুব’, ‘বেশি’, ‘ভয়’, ‘পায়নি’

3.3.3 Stemming

Stemming is the process of reducing inflected (or something derived) words to their word stem, base or root form - generally a written word form. Stemming is important in natural language processing (NLP). We stemmed a word to give its root form so that it can be easily compared with the lexicon. We first checked if the word needs to be stemmed, if the word has ‘ে’, ‘া’, ‘ি’, ‘ ী’, ‘ো’ etc. at the end, then the word is stemmed. Then word is checked if it has ‘ের’, ‘টা’, ‘টি’ etc. at the end then

these are removed. In the third step we check for ‘না’, ‘নি’ at the end of the word if found the word is split there.

Example of stemming ‘ে’:

(7) ‘ফাঁদে’, ‘পা’, ‘দিবেন’, ‘না’

(8) ‘ফাঁদ’, ‘পা’, ‘দিবেন’, ‘না’

Example of stemming ‘এর’:

(9) ‘বিপদের’, ‘সময়’, ‘না’

(10) ‘বিপদ’, ‘সময়’, ‘না’

Example of stemming ‘নি’:

(11) ‘খুব’, ‘বেশি’, ‘ভয়’, ‘পায়নি’

(12) ‘খুব’, ‘বেশি’, ‘ভয়’, ‘পায়’, ‘নি’

3.3.4 N-gram word generation

After processing in the previous steps in this step the data is prepared for giving its polarity by n-gram. That word are made as bi-gram or tri-gram joining them with the help of _, so that it can easily be matched against our constructed bi-gram and tri-gram lexicon. Example of n-gram: ‘খুব_বেশি’, ‘বেশি_ভয়’, ‘ভয়_পায়’, ‘পায়_নি’

3.4 Boosting word valence

In this step, we search if the text has any booster word. If the text contains any booster word included in the booster dictionary, then it intensifies its valence according to the position of the booster word in the sentence. To identify the position of booster words three processes have been used:

3.4.1 Bigram

A bigram or digram is a sequence of two adjacent tokens. It creates the pair of two adjacent tokens. Bigram is done to check for booster word. Example of bigram:

(13) ‘খুব বেশি’, ‘বেশি ভয়’, ‘ভয় পায়’, ‘পায় নি’

3.4.2 Trigram

Trigram is a special case of the n-gram where n=3, that means the pair of three consecutive tokens, and we did this after bigram to check if the booster words exist there. Example of trigram:

(14) 'খুব বেশি ভয়', 'বেশি ভয় পায়', 'ভয় পায় নি'

For any word, if the boosting word is found in the booster dictionary, then for bigram the valence of the token is multiplied by 0.9 [1], and for trigram, it is multiplied by 0.75 [1].

3.4.3 Negation

Though negation words do not express any sentiment, they affect the overall sentiment of a sentence. Use of negation word is different in Bengali from English, while in English it is used in the middle of the sentences in Bengali it is used normally at the end of the sentences. At this step, negation word is searched according to the negation list constructed before. If a negation word is found, then the valence of the sentence is multiplied by -1, that means the valence is reversed.

3.5 Valence Calculation

After processing the text in all the previous steps, this is the final step in calculating the valence. As our goal was to calculate the correct valence for text this one is important. The system outputs sentiment scores to 3 classes of sentiments:

- Positive
- Negative
- Neutral

The score is compound score which is a normalized score. Normalization is described on section 3.5.1 The valance of the Bengali sentence 'সে খুব বেশি ভয় পায়নি ।' is -1.1284 before normalization.

3.5.1 Normalization

The compound score is computed by summing the valence of each word in the lexicon, adjusted with rules, and then normalized to be generally between -1 (extreme negative) and +1 (extreme positive).

This is the most useful metric to get a single unidimensional measure of sentiment. It is actually called “normalized weighted composite score”. To normalize the score, we use the equation:

$$Normalizedscore = \frac{score}{\sqrt{(score * score) + alpha}} \tag{3.1}$$

where $alpha = 15$ is approximated max expected value and $score$ is the calculated score to be normalized.

The normalization score of the sentence ‘সে খুব বেশি ভয় পায়নি ।’ is -0.2797

3.5.2 Separation of sentences

If the valence of the text is less than 0 to -1 then it expresses negativity, if it is 0 then neutral and if the valence is greater than 0 to +1 the text express positivity. Every sentence is marked as positive, negative and neutral according to their calculated polarity. As the score for the sentence ‘সে খুব বেশি ভয় পায়নি ।’ is -0.2797 is less than 0 so the sentence is showed as a negative sentence.

Chapter 4

Results and Discussion

In this chapter we briefly discuss about our result in sentence level usnig the unigram lexicon along with the result of VADER. Then we discuss the result of using bi-gram and tri-gram lexicon. We have utilized three different datasets and we have chosen the movie review dataset as the main one for the experiments. Another two datasets are Bangla Tweets and Restaurant review dataset. For training and testing motives, we have split the dataset into two parts using k-fold cross validation (i.e. 10-fold) with a split ratio of 0.9 (90% data for training and 10% data for testing).

4.1 Discussion of VADER and Bengali VADER

After construction of polarity lexicon and applying methods accordingly which has been described in section 3, we get the result of a text if it is positive, negative or neutral depending on the generated polarity of that sentence. A sentence is called positive if its polarity is greater than 0 to +1, where +1 represents the highest intensity of positivity. On the contrary, the polarity of less than 0 to -1 represents the negativity of sentences, where -1 is the highest intensity of negative valence. And if the polarity is 0, then the sentence is neutral.

We have evaluated the sentiment for a sentence using VADER, where VADER first translate the given Bengali text to English and give the polarity using its English polarity lexicon. We have used Google[1] translator to translate the text to English for VADER and analyzed the polarity of the sentences. Those results, as well as polarity analyzed by our system using Bengali polarity lexicon, are shown in Table: 4.1.

From Table: 4.1 we can see that for the Bengali sentence 'সে পরিশ্রমী না' VADER gives positive

[1]https://translate.google.com.bd/

Table 4.1: Comparison Between VADER and Our Proposed System

Sentences	VADER	Our Proposed System
	Google Translator	Unigram+N-gram lexicon
আমি আমার মত ভাল আছি	0.5106	0.4404
ব্যর্থতা সাফল্যের চাবি ।	0.1027	0.1027
নিজের উপর আত্মবিশ্বাস থাকা ভাল ।	0.4404	0.7096
তিনি একজন আদর্শ শিক্ষক	0.5267	0.5267
সে ভূত ভয় পায় ।	-0.4404	-0.7003
আমার সৌভাগ্য হয়নি ।	-0.3412	-0.3818
সে পরিশ্রমী না	**0.0762**	**-0.4767**
কি দারুণ খবর ।	0.6249	0.4767
ফাঁদে পা দিবেন না ।	**-0.3182**	**0.3182**
সে লেখাপড়ায় মনযোগী নয়	**0.0**	**-0.3818**
বুদ্ধিমত্তার সাথে লেগে থাকলে আপনিও সফল হবেন ।	0.7783	0.7003
আমি আমার মা কে অনেক ভালবাসি ।	0.6369	0.6697
বিদেশি সাহায্যের গতি বাড়লে আর রাজস্ব আদায়ের প্রবৃদ্ধি বাড়লে ঋণের পরিমাণ কমে আসবে ।	-0.0516	0.4215
সে খুব বেশি ভয় পায়নি ।	**0.0**	**-0.2797**
চকচক করলেই সোনা হয় না	0.0	0.0
সুন্দরবন আমাদের অহংকার ।	0.34	0.4939
সন্ত্রাসীরা বিস্ফোরক দ্রব্য ব্যবহার করে অরাজকতা সৃষ্টি করে	-0.4588	-0.4215

score using its translator though in Bengali it is a negative sentence. Whereas our proposed Bengali VADER gives the correct polarity of the sentence and determines it as a negative sentence. For the Bengali sentence 'বিদেশি সাহায্যের গতি বাড়লে আর রাজস্ব আদায়ের প্রবৃদ্ধি বাড়লে ঋণের পরিমাণ কমে আসবে ।' VADER detects the sentence as negative using its translator, but our Bengali VADER correctly determines it as positive. The Bengali sentence 'সে লেখাপড়ায় মনযোগী নয়' is a negative sentence and is correctly detected by our system but VADER wrongly. Moreover, the sentence সুন্দরবন আমাদের অহংকার is used as a positive sentence in and our system detects it correctly when two lexicon is used in together. Another sentence shown in table সন্ত্রাসীরা বিস্ফোরক দ্রব্য ব্যবহার করে অরাজকতা সৃষ্টি করে is correctly detected by the system when bi-gram and tri-gram lexicon have been used.

So, it cannot be a good and efficient way to use the translator to translate Bengali sentences into English and determine their sentiments because the results largely depend on the translator and how they are translated. From Table: 4.1 we can see that in most of the cases our proposed system gives better results than VADER's method to analyze sentiment from Bengali language. Not only that but also the time it needs to translate a sentence from Bengali to English can be huge for a significant amount of data. As our Bengali VADER does not need use any translation system, it can give result

in a very low time, and its result is much dependable.

4.2 Evaluation of results

To test the effectiveness of our constructed lexicon and the system we tested our system using machine learning methods like cross-validation, naive bayes, support vector machine (SVM). To analyze the result using these methods we at first found TF-IDF and predicted the value using cross-validation, NB and SVM. To measure the preformance we ue use four measurement accuray, precision, recall and f1-measure. To measure first we calculate true positive (TP), true negative (TN), false negative (FN), and false positive (FP), these are calculated using confusion matrix shown in table 4.2

Table 4.2: Confusion matrix

	Predicted Positives	**Predicted Negatives**
Actual Positive instances	# of True Positive (TP) instances	# False Negative (FN) instances
Actual Negative instances	# of False Positive (FP) instances	# of True Negative (TN) instances

- **Accuray:** It express the number of data that has been correctly classified in total number of data. The equation is expressed as:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.1}$$

- **Precision:** It expresses the portion of true positive data in all instances of predicted positive data. The equation is written as:

$$precision = \frac{TP}{TP + FP} \tag{4.2}$$

- **Recall:** It expresses the portion of true positive data against all actual positive instances. The equation is:

$$recall = \frac{TP}{TP + FN} \tag{4.3}$$

- **F1-measure:** The harmonic mean of precision and recall. The equaiton is,

$$f1 - measure = \frac{2 * precision * recall}{precision + recall} \tag{4.4}$$

As Bengali is resource constrained language it is very difficult to get a good dataset. We used the following three dataset in our system.

4.2.1 Movie Reviews dataset

We collected the english movie review dataset from github [30]. Then translated the dataset using Google translator. The dataset comprises of one thousand and one thousand negative text files. In the movie review dataset, we have used 1800 review text files to train the classifier and 200 review text files to test the model performance. After testing the accuracy of the model on the test set, we have used 10-fold cross validation technique to find out the average accuracy. We have found out that precision, recall, f-measure of the classification using 10-fold cross validation.

Since our proposed system is unsupervised. So, find the polarity of each review using polarity lexicon. We have found out the TP, TN, FP, FN using the polarity and label of review. Then we have found out accuracy, precision, recall, f-measure using the following equations respectively (4.1), (4.2), (4.3), (4.4). The results of these algorithms and our proposed system are shown in below table 4.2.1.

Table 4.3: Results of Movie review dataset

	Proposed	Naive Bayes	SVM
Accuracy	0.7015	0.7364	0.7608
Precision	0.6557	0.7568	0.7658
Recall	0.8520	0.697	0.754
F1-score	0.7407	0.7251	0.7588

4.2.2 Sports dataset

We collected the sports news dataset from github [31]. The dataset comprises of 2600 and 2600 negative text files. In the movie review dataset, we have used 4680 review text files to train the classifier and 520 review text files to test the model performance. After testing the accuracy of the model on the test set, we have used 10-fold cross validation technique to find out the average accuracy. We have found out that precision, recall, f-measure of the classification using 10-fold cross validation.

Since our proposed system is unsupervised. So, find the polarity of each review using polarity lexicon. We have found out the TP, TN, FP, FN using the polarity and label of review. Then we have found out accuracy, precision, recall, f-measure using the following equations respectively (4.1),

(4.2), (4.3), (4.4). The results of these algorithms and our proposed system are shown in below table 4.2.2.

Table 4.4: Sports dataset

	Proposed	Naive Bayes	SVM
Accuracy	0.7615	0.7207	0.7088
Precision	0.7211	0.6911	0.6927
Recall	0.852	0.8128	0.7647
F1-score	0.7811	0.7453	0.7251

4.2.3 Twitter dataset

To check the effectiveness of the system we used variety of dataset [32]. We used twitter dataset. The dataset contains almost 22 thousand tweets in text format. In the Bangla twitter dataset, we have used 19800 reviews to train the classifier and 2200 reviews to test the model performance. After testing the accuracy of the model on the test set, we have used 10-fold cross validation technique to find out the average accuracy. We have found out that accuracy, precision, recall, f-measure of the classification using 10-fold cross validation.

Since our proposed system is unsupervised. So, find the polarity of each review using polarity lexicon. We have found out the TP, TN, FP, FN using the polarity and label of review. Then we have found out accuracy, precision, recall, f-measure using the following equations respectively (4.1), (4.2), (4.3), (4.4). The results of these algorithms and our proposed system are shown in below table 4.2.3

Table 4.5: Results of Tweeter dataset

	Proposed	Naive Bayes	SVM
Accuracy	0.6454	0.6394	0.6456
Precision	0.6195	0.6322	0.6430
Recall	0.758	0.6674	0.6555
F1-score	0.6817	0.6492	0.6491

Chapter 5

Conclusion and Future Works

5.1 Conclusion and Future works

We report the development of Bengali polarity lexicon using existing VADER lexicon of English and systematic evaluation of the sentiment of Bengali sentences using our proposed methodology which is the modification of VADER. We use stemming and list Bengali boosting words to combine with the system so that it can give a better performance. The results are encouraging as the system gives better analytical results than using translated medium, not only that it's time efficiency is much better as it does not need to translate the original text. In this paper, we have directed experiments on three datasets movie review dataset, Bangla twitter dataset, and restaurant review dataset. After playacting sentiment analysis on these datasets using our constructed N-gram (i.e. Bi-gram, Tri-gram) lexicon-based model and TF-IDF model we found out the accuracy, precision, recall, f-measure as shown in fig (1). The accuracy percentages for movie review dataset using our constructed model and TF-IDF model came as 70.15% and 73.64% respectively. But f-measure came as 0.7407 and 0.7251 respectively. In the Twitter dataset, after playacting models, accuracy came as 64.56% and 63.91% respectively. And f-measure came as 0.68 and 0.6492 respectively.

In future we will apply Hybrid approach so that it can give better performance on some sentences which consists of words that has both positive and negative meaning (i.e. ambiguity). We also have a plan to enrich our lexicon to give better performance.

Bibliography

[1] C. H. E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Eighth International Conference on Weblogs and Social Media (ICWSM-14). Available at (20/04/16) http://comp. social. gatech. edu/papers/icwsm14. vader. hutto. pdf*, 2014.

[2] K. Hasan, A. Mondal, A. Saha *et al.*, "Recognizing bangla grammar using predictive parser," *arXiv preprint arXiv:1201.2010*, 2012.

[3] M. A. Islam, K. A. Hasan, and M. M. Rahman, "Basic hpsg structure for bangla grammar," in *Computer and Information Technology (ICCIT), 2012 15th International Conference on*. IEEE, 2012, pp. 185–189.

[4] K. A. Hasan, A. Mondal, and A. Saha, "A context free grammar and its predictive parser for bangla grammar recognition," in *Computer and Information Technology (ICCIT), 2010 13th International Conference on*. IEEE, 2010, pp. 87–91.

[5] T. W. Post. The future of language. Accessed 2018-07-19. [Online]. Available: https://www.washingtonpost.com/news/worldviews/wp/2015/09/24/the-future-of-language/?utm_term=.5158d11a583a

[6] D. Alessia, F. Ferri, P. Grifoni, and T. Guzzo, "Approaches, tools and applications for sentiment analysis implementation," *International Journal of Computer Applications*, vol. 125, no. 3, 2015.

[7] "Tf–idf," Oct 2018. [Online]. Available: https://en.wikipedia.org/wiki/Tf–idf

[8] "Cross-validation (statistics)," Dec 2018. [Online]. Available: https://en.wikipedia.org/wiki/Cross-validation_(statistics)

[9] "Naive bayes classifier," Nov 2018. [Online]. Available: https://en.wikipedia.org/wiki/Naive_Bayes_classifier

[10] "Support vector machine," Dec 2018. [Online]. Available: https://en.wikipedia.org/wiki/Support_vector_machine

[11] 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*. Association for Computational Linguistics, 1997, pp. 174–181.

[12] S.-M. Kim and E. Hovy, "Determining the sentiment of opinions," in *Proceedings of the 20th international conference on Computational Linguistics*. Association for Computational Linguistics, 2004, p. 1367.

[13] J. Kamps, M. Marx, R. J. Mokken, M. De Rijke *et al.*, "Using wordnet to measure semantic orientations of adjectives." in *LREC*, vol. 4. Citeseer, 2004, pp. 1115–1118.

[14] H. Liu and P. Singh, "Conceptnet—a practical commonsense reasoning tool-kit," *BT technology journal*, vol. 22, no. 4, pp. 211–226, 2004.

[15] I. San Vicente and X. Saralegi, "Polarity lexicon building: to what extent is the manual effort worth?" in *LREC*, 2016.

[16] P. J. Stone, D. C. Dunphy, and M. S. Smith, "The general inquirer: A computer approach to content analysis." 1966.

[17] T. Wilson, P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan, "Opinionfinder: A system for subjectivity analysis," in *Proceedings of hlt/emnlp on interactive demonstrations*. Association for Computational Linguistics, 2005, pp. 34–35.

[18] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," *Computational linguistics*, vol. 37, no. 2, pp. 267–307, 2011.

[19] R. Mihalcea, C. Banea, and J. Wiebe, "Learning multilingual subjective language via cross-lingual projections," in *Proceedings of the 45th annual meeting of the association of computational linguistics*, 2007, pp. 976–983.

[20] V. Perez-Rosas, C. Banea, and R. Mihalcea, "Learning sentiment lexicons in spanish." in *LREC*, vol. 12, 2012, p. 73.

[21] S. Baccianella, A. Esuli, and F. Sebastiani, "Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining." in *LREC*, vol. 10, no. 2010, 2010, pp. 2200–2204.

[22] K. A. Hasan, M. Rahman *et al.*, "Sentiment detection from bangla text using contextual valency analysis," in *Computer and Information Technology (ICCIT), 2014 17th International Conference on*. IEEE, 2014, pp. 292–295.

[23] S. Chowdhury and W. Chowdhury, "Performing sentiment analysis in bangla microblog posts," in *Informatics, Electronics & Vision (ICIEV), 2014 International Conference on*. IEEE, 2014, pp. 1–6.

[24] J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic inquiry and word count: Liwc 2001," *Mahway: Lawrence Erlbaum Associates*, vol. 71, no. 2001, p. 2001, 2001.

[25] J. W. Pennebaker, R. J. Booth, and M. E. Francis, "Linguistic inquiry and word count: Liwc [computer software]," *Austin, TX: liwc. net*, 2007.

[26] X. Saralegi, I. San Vicente, and I. Ugarteburu, "Cross-lingual projections vs. corpora extracted subjectivity lexicons for less-resourced languages," in *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, 2013, pp. 96–108.

[27] A. Das and S. Bandyopadhyay, "Topic-based bengali opinion summarization," in *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*. Association for Computational Linguistics, 2010, pp. 232–240.

[28] B. Liu, "Sentiment analysis and subjectivity." *Handbook of natural language processing*, vol. 2, pp. 627–666, 2010.

[29] A. Das and S. Bandyopadhyay, "Sentiwordnet for bangla," *Knowledge Sharing Event-4: Task*, vol. 2, 2010.

[30] F. Jeremiah, "jerofad/sentimentanalysis," Feb 2018. [Online]. Available: https://github.com/jerofad/SentimentAnalysis

[31] S. A. Taher, K. Afsana, and K. A. Hasan, “Shayokh144/bangla dataset for opinion mining,” Jun 2017. [Online]. Available: https://github.com/Shayokh144/Bangla_Dataset_for_Opinion_Mining

[32] “For academics.” [Online]. Available: http://help.sentiment140.com/for-students?fbclid=IwAR3xB7BGXhlH6kWq76i02qtp4g2iQxFT7mtBux5AgSmGA4GTJYnL1AOQaps

Appendix A

Github source

Github source link: https://github.com/mkazi078/bengalisentiment