2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 February, 2019

Bengali VADER: A Sentiment Analysis Approach Using Modified VADER

Al Amin, Imran Hossain, Aysha Akther* and Kazi Masudul Alam
DGTED Lab, Computer Science and Engineering Discipline
Khulna University
Khulna-9208, Bangladesh
{alamin1512, imran1503, aysha, kazi}@cseku.ac.bd

***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 English 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 requirement of Bengali to English translation using tools such as Google Translator, MyMemory Translator, etc. Our experiments demonstrate that the modified Bengali VADER significantly improves the sentiment analysis result of Bengali text over the current model.**

***Index Terms*—Sentiment analysis, Bengali VADER, Bengali Lexicon, Bengali to English Translation.**

I. Introduction

Sentiment is a term related to sensitiveness or emotional feelings. Sentiment is a genuine and refined sensibility, which is affected by feeling as opposed to reason or reality [1]. Whereas, analysis is the detailed examination that is required to understand the nature or to determine the essential features of something complex through analytical study [2]. According to [3], sentiment analysis determines the tendency of individual's opinions through information mining by using natural language processing (NLP), computational linguistics and textual analysis on emotional data gathered from the Internet, online social media and similar sources. On the other hand, Oxford dictionary describes sentiment analysis as the process that 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 [4].

Business organizations, governments, brand management companies, and data scientists are demonstrating increasingly higher interest about sentiment analysis and its various applications. Data analytics can provide significant insight about the users perception about any topic that can: improve business strategies, improve brand impression, improve campaign management, modify marketing strategies, generate business leads, test business KPIs, etc. Related opinion data is available in the comments of news articles, comment/status section of social medias (i.e. Facebook, Twitter, Instagram), blogs, review of products, etc. We can easily conclude that business organizations can improve their bottom line using sentiment analysis in different data segments. Similarly, it is important for the governments to understand the sentiments of the citizens to address important public opinions as well as update political agendas to avoid unwanted circumstances. In all these cases, sentiment analysis can be an important game changing tool.

Multiple international languages such as English, French are exploring the sentiment analysis field for a while. Though Bengali is one of the most used languages in the world, it is well behind in the race. As Bengali is used everyday by more than 250 million people of the world, primary language in Bangladesh and secondary language in India [5] [6] [7], it has huge potential for the business as well as governments. 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, lower than Chinese and Hindi [8].

As a result, researchers should focus more on Bengali text analysis. Since there are very few works on sentiment analysis of Bengali text, in this paper we have put some light in that direction. Mobile adoption as well as Internet growth is very high in Bangladesh, which is contributing rich sentimental data in the online world. From this perspective, we have modified a popular English sentiment classification tool VADER [9] and upgraded it to support Bengali sentiment analysis without using intermediate Bengali to English translation.

Rest of the paper is organized as Section II describes state of the art works in the field of Bengali text analysis. Section II-C briefly discusses about the methodology of VADER operation in English text. In Section III, we discuss our approach to build the Bengali polarity lexicon from VADER English lexicon and the operation methodology of the proposed Bengali VADER, which is followed by Results and Discussion in Section IV. Finally, we conclude the paper in Section V with possible

*Corresponding Author

978-1-5386-9111-3/19/$31.00 ©2019 IEEE

future works.

II. Literature Review

A. Lexicon

A lexicon is the vocabulary of a man, dialect, or a department of expertise that stocks the lexemes in that linguistics. Polarity lexicons are that which have a listing of words with initial level of polarities. It is one of the main resources for analyzing the sentiments and opinion expressed in texts in an computerized way. There are primarily three ways to build polarity lexicons: interpreting existing lexicons from different languages, extracting polarity lexicons from corpora [10], and annotating sentiments Lexical Knowledge Base [11] [12] [13] [14].

For major languages there are well known manually constructed lexicons such as General Inquirer [15], OpinionFinder [16], SO-CAL [17], etc. In [18] and [19] authors analyzed the methodology of translating English resources to Romanian and Spanish respectively. There are several English polarity lexicons available online such as SentiWordNet[1], VADER but hardly any Bengali polarity lexicon.

B. English Language

English language is rich for notable works on sentiment analysis such as VADER. In this model, researchers combined qualitative and quantitative methods to produce a gold-standard sentiment lexicon, which was later empirically validated against especially receptive microblog type contexts. VADER combines important lexical features obtained from five generalized rules that embody grammatical and syntactical conventions of human speech. It also retains the advantages of conventional sentiment lexicons such as LIWC [20] [21].

In [14] authors explored three strategies to build polarity lexicons: interpreting existing lexicons from other languages, explaining sentiments from lexical knowledge bases and extracting polarity lexicons from corpora. All of the models require different degrees of human effort. Here, Spanish lexicon [22] has been translated by means of the Elhuyar Spanish-Basque[2] dictionary, where every Spanish word incorporates five interpretations.

C. VADER

VADER combines qualitative and quantitative methods to produce a gold-standard sentiment lexicon. A good number of sentiment analysis models rely significantly on basic sentiment (or opinion) lexicons. A *sentiment lexicon* is a listing of lexical capabilities (e.g. words) which can be commonly categorized with their semantic orientation as either positive or negative [23]. After producing and validating lexicon, they have used the architecture shown in Figure 1 to evaluate the sentiment of texts. The architecture mainly follows the steps described below to calculate the polarity of the sentences:

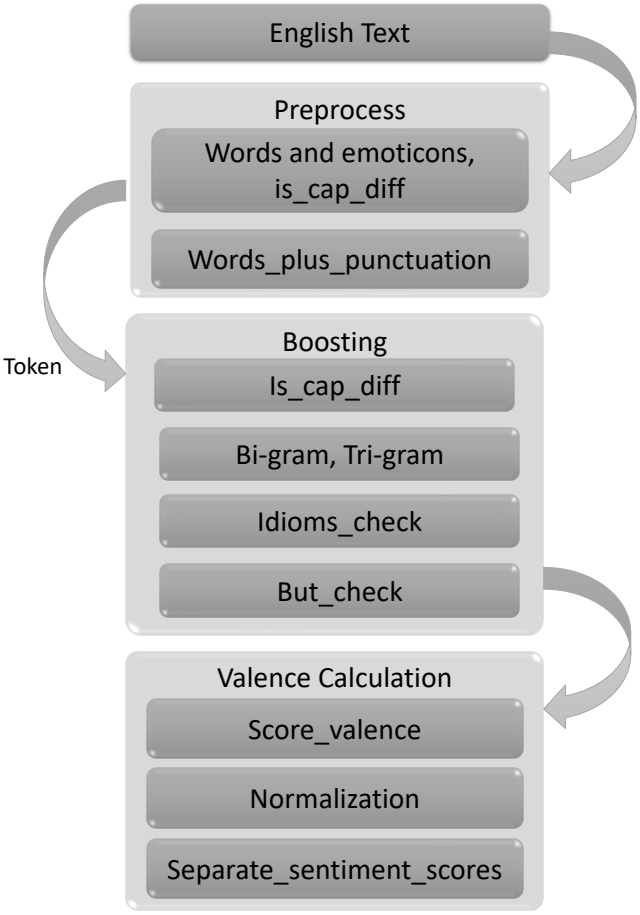

Figure 1. Implementation architecture of English VADER[3]

1) Preprocessing: In preprocessing step, input English text is tokenized. At first, words, emoticons and all capitalized words are tokenized. Later, words and punctuation marks are tokenized. Some punctuation marks affect the valence of words, which is kept with the word.

2) Boosting: Once tokenization is complete all the tokens are checked for valence boosting purpose. VADER uses bi-gram and tri-gram model for boosting. If a boosting word such as "extremely, very, great" are found then the valence of the word is boosted. Then all capital words are considered and the valence is further boosted. The text is also checked for idioms and phrases, if found then the valence is boosted again. Later, it is looks for 'but' word, if found then the sentence is divided into two parts and valence is calculated on two different parts. Next, overall valence is calculated for such a sentence.

3) Valence Calculation: In this step, valence of a sentence is measured, which is between -4 to +4. This calculated value is further normalized to range between -1 to +1. In this way, every sentence is assigned with their respective polarity.

[1] http://sentiwordnet.isti.cnr.it/
[2] http://hiztegiak.elhuyar.eus
[3] VADER Sentiment, https://github.com/cjhutto/vaderSentiment

D. Bengali Language

Though Bengali language sentiment analysis is far behind than English language sentiment analysis, Bengali polarity lexicon based works can learn from popular works such as SentiWordNet [24], VADER [9] etc. In [25], authors classify sentiment from sentence or an entire document as positive, negative or neutral. Firstly, they have used the WordNet to get the senses of words according to their respective parts of speech and later applied SentiWordNet to get their prior valence or polarity.

The authors of [26] have translated the SentiWordNet in Bengali and used its polarity to calculate the valence of Bengali texts. They have developed training corpus using semi-supervised bootstrapping method; used Support Vector Machine (SVM) and Maximum Entropy (MaxEnt) for classification; and later combined various set of features to compare the performance of these two machine learning models. They also built a Twitter-specific Bengali sentiment lexicon for procedure-based classification using binary features.

Das et. al. [27] constructed an opinion outline system based on Bengali news corpus. The system identifies sentiment information from each document, which is further aggregated using topic-sentiment model and later summarized. Topic-sentiment model uses theme clustering (k-means), which further uses document level theme relation graph to achieve discourse level theme identification and topic sentiment aggregation. Finally standard page rank algorithms are utilized for information retrieval purpose.

III. Proposed Bengali VADER

We have developed a model to identify sentiments from Bengali text. Figure 2 represents our developed system that includes preprocessing, boosting and Bengali valence creation. This paper focuses on two interrelated things: 1) development and validation of Bengali polarity lexicon and 2) extraction of sentiment intensity from Bengali text using our methods. These processes are described accordingly in the following sections.

A. Dictionary

We have created two dictionaries of negation words and booster words[4]. They are used for negation and boosting of a sentence respectively.

1) Negation List: We have created a list of words that are common in Bengali to represent negativity of the texts. If any of these words are present in the text, then the polarity of the text is reversed to either positive or to negative. Some example words are:

(1) 'না', 'নি', 'নয়', 'নাই', 'নেই'

2) Booster Dictionary: Booster dictionary contains Bengali words that are used to boost valence of any text. If any one of them are present in the text then polarity will be increased in our model. The dictionary contains words such as:

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

[4]https://github.com/mkazi078/bengalisentiment

B. Bengali Lexicon

We have constructed a lexicon translating VADER [9] lexicon using a bilingual dictionary and given their corresponding polarity. The valences of the words are given in a range of -4 to +4, where -4 represents most negativity, +4 represents most positivity, and 0 represents neutral. There are more than 3000 words in our compiled lexicon, which is continuously growing.

C. Preprocessing

In the preprocessing stage, we have removed punctuation as well as unnecessary stop words and conducted stemming on the words. The steps are:

1) Punctuation removal: The text is tokenized i.e. split into words and punctuation marks are removed from the texts. For example, the following text

(3) সে খুব বেশি ভয় পায়নি ।

will appear as

(4) 'সে', 'খুব', 'বেশি', 'ভয়', 'পায়নি' '।'

After punctuation removal will be presented as:

(5) 'সে', 'খুব', 'বেশি', 'ভয়', 'পায়নি'

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' appears as:

(6) 'খুব', 'বেশি', 'ভয়', 'পায়নি'

3) Stemming: Natural language processing uses stemming process to reduce derived words into their original form or stem. In our proposed Bengali VADER model, we stemmed a Bengali word to its root form so that it can be easily compared with the lexicon. We first verify if the word needs to be stemmed. If the word has 'ে', 'ো', 'ি', 'ী', 'ো' etc. at the end, then the word is stemmed. Then it is analyzed whether the word has 'ের', 'টা', 'টি' etc. at the end, if found they are removed. In the third step, we check for 'না', 'নি' at the end of the word and 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'

D. 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 a booster word following three processes have been used:

1) Bigram: We have selected bigram or digram tokens i.e. two adjacent tokens in a sequence from a sentence. This pair of two adjacent tokens is again used to boost word valence. Example of bigram:

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

2) Trigram: Trigram is a special case of the n-gram where n=3. In this case, a pair of three consecutive tokens is again used to boost a word valence. 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, and for trigram, it is multiplied by 0.75 [9].

3) Negation: Usually, overall sentiment of a sentence is affected by negation words. But, use of negation words is different in Bengali from English. In case of English, it is used in the middle of the sentences, whereas 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 i.e. existing valence is reversed.

E. Valence Calculation

After completing all the previous steps, the final step is to calculate the valence. This model guides in finding the correct valence for any text. At any time, the system outputs sentiment scores in three classes: positive, negative and neutral. It further requires normalization.

1) Normalization: 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 an useful metric to get a single unidimensional measurement about sentiment. It is actually called "normalized weighted composite score". To normalize the score, we use the equation:

$$Normalizedscore = \frac{score}{\sqrt{(score * score) + alpha}} \quad (1)$$

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

2) Separation of sentences: If the valence of the text is less than 0 and upto -1 then it expresses negativity. If it is 0 then it represents neutral and if the valence is greater than 0 upto +1 then the text expresses positivity. Every sentence is marked as positive, negative and neutral according to their calculated polarity.

IV. Results and Discussion

After construction of polarity lexicon and applying methods accordingly which has been described in the previous section, 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 between 0 to +1, where +1 represents the highest intensity of positivity. On the contrary, if the polarity is between 0 to -1 then it represents the negativity of a sentence, where -1 is the highest intensity of negative valence. And if the polarity is 0, then the sentence is considered neutral in our Bengali VADER model.

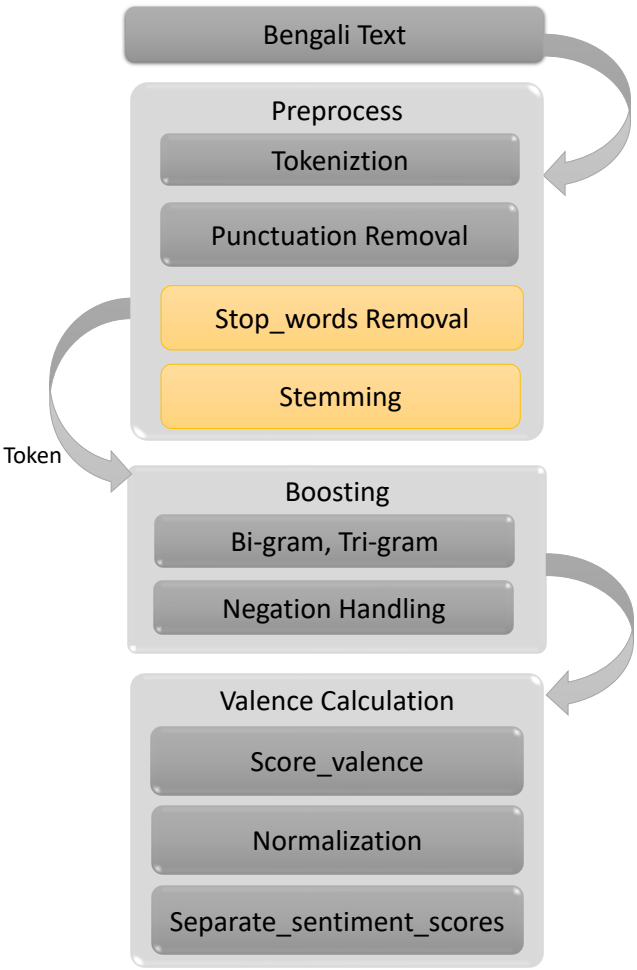

Figure 2. System Architecture of proposed Bengali VADER

We have evaluated the sentiment for a sentence using VADER, where VADER first translates the given Bengali text to English and gives the polarity using its English polarity lexicon. To translate the Bengali into English VADER uses MyMemory translator. Then we have used Google translator and Python 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 I.

From Table I we can see that for the Bengali sentence 'সে পরিশ্রমী না' VADER gives positive score using all the three of its translators 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 by using its translator, whereas our proposed Bengali VADER correctly determines it as positive.

Table I
COMPARISON BETWEEN VADER AND OUR PROPOSED BENGALI VADER

Sentences	VADER			Our Proposed System
	Python Translator	Google Translator	MyMemory Translator	Bengali VADER
আমি আমার মত ভাল আছি	0.5106	0.5106	0.5106	0.6597
ব্যর্থতা সাফল্যের চাবি ।	0.1027	0.1027	0.1027	0.1027
নিজের উপর আত্মবিশ্বাস থাকা ভাল ।	0.4404	0.4404	0.4404	0.7096
তিনি একজন আদর্শ শিক্ষক	0.5267	0.5267	0.5267	0.5267
সে ভূত ভয় পায় ।	-0.4404	-0.4404	-0.4404	-0.7003
আমার সৌভাগ্য হয়নি ।	-0.3412	-0.3412	-0.3412	-0.3818
সে পরিশ্রমী না	**0.0762**	**0.0762**	**0.0762**	**-0.4767**
কি দারুণ খবর ।	0.6249	0.6249	0.6249	0.4767
ফাঁদে পা দিবেন না ।	**-0.3182**	**-0.3182**	**-0.3182**	**0.3182**
সে লেখাপড়ায় মনযোগী নয়	-0.2584	**0.0**	-0.2584	**-0.3818**
বুদ্ধিমত্তার সাথে লেগে থাকলে আপনিও সফল হবেন ।	0.7783	0.7783	0.7783	0.7003
আমি আমার মা কে অনেক ভালবাসি ।	0.6369	0.6369	0.6369	0.6697
বিদেশি সাহায্যের গতি বাড়লে আর রাজস্ব আদায়ের প্রবৃদ্ধি বাড়লে ঋণের পরিমাণ কমে আসবে ।	**-0.0516**	**-0.0516**	**-0.3612**	**0.4215**
সে খুব বেশি ভয় পায়নি ।	0.0	0.0	0.0	0.5034
চকচক করলেই সোনা হয় না	0.0	0.0	0.0	0.0

Another Bengali sentence 'সে লেখাপড়ায় মনযোগী নয়' is a negative sentence and is correctly detected by VADER using MyMemory or Python translator but Google translator fails in this case. So, we can conclude that intermediate translator can not be an efficient way to translate Bengali sentences into English and measure their sentiment polarity as the result largely depend on the translator and how they are translated. From Table I 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. At the same time, since our process avoids intermediate Bengali to English translation, it performs faster than other processes.

V. CONCLUSION AND FUTURE WORKS

In this paper, we present the development of Bengali polarity lexicon from existing English lexicon of VADER. We have modified the existing VADER architecture to accommodate Bengali sentiment polarity detection. In our methodology, we use stemming, enlist Bengali boosting words, apply bigram-trigram 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 readily available translation based model. From our experimental analysis, we demonstrate that our proposed Bengali VADER performs efficiently than the other systems. Since, our system avoids intermediate Bengali to English translation it also performs faster. Bengali polarity detection system is useful to understand users perception about any product, observation about political scenarios, etc. By improving the Bengali lexicon, bigram-trigram list we can produce much better results. In future, we will apply machine learning approaches so that it can give better performance in some sentences which consists of words that have both a positive and negative meaning.

REFERENCES

[1] T. Dictionary. Sentiment. Accessed 2018-06-12. [Online]. Available: http://www.dictionary.com/browse/sentiment

[2] M. Webster. Analysis. Accessed 2018-06-12. [Online]. Available: https://www.merriam-webster.com/dictionary/analysis

[3] Technopedia. Sentiment Analysis. Accessed 2018-06-12. [Online]. Available: https://www.techopedia.com/definition/29695/sentiment-analysis

[4] O. Dictionary. Sentiment Analysis. Accessed 2018-06-12. [Online]. Available: https://en.oxforddictionaries.com/definition/sentiment_analysis

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

[6] 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.

[7] 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.

[8] 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

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

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

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

[16] 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.

[17] 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.

[18] 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.

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

[20] 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.

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

[22] 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.

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

[24] 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.

[25] 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.

[26] 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.

[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.