Supervised Approach of Sentimentality Extraction from Bengali Facebook Status

Conference Paper · December 2016

DOI: 10.1109/ICCITECHN.2016.7860228

4 authors, including:

Md Saiful Islam
Shahjalal University of Science and Technology
70 PUBLICATIONS **237** CITATIONS

SEE PROFILE

Md Afjal Hossain
Shahjalal University of Science and Technology
3 PUBLICATIONS **13** CITATIONS

SEE PROFILE

Md Ashiqul Islam
8 PUBLICATIONS **31** CITATIONS

SEE PROFILE

Supervised Approach of Sentimentality Extraction from Bengali Facebook Status

Md. Saiful Islam
Computer Science and Engineering
Shahjalal University of Science & Technology
Sylhet, Bangladesh
saiful-cse@sust.edu

Md. Afjal Hossain
Computer Science and Engineering
Shahjalal University of Science & Technology
Sylhet, Bangladesh
afjal.sm19@gmail.com

Md. Ashiqul Islam
Computer Science and Engineering
Shahjalal University of Science & Technology
Sylhet, Bangladesh
rajib.sust47@gmail.com

Jagoth Jyoti Dey
Computer Science and Engineering
Shahjalal University of Science & Technology
Sylhet, Bangladesh
jagothjyotidey91@gmail.com

***Abstract*— Sentiment is the only things that separate human and machine. To simulate the feelings for machines many researchers have been trying to create method and automated the process to extract opinion of particular news, product or life entity. Sentiment Analysis (SA) is a combination of opinions, emotions and subjectivity of a text. Currently SA is the most demanding task in Natural Language Processing. Social networking site like Facebook are mostly used in expressing the opinions about a particular entity of life. Newspaper published news about a particular event and user expressed their feedback in news comments. Online product feedback is increasing day by day. So reviews and opinions mining play a very important role in understanding people satisfactions. Such opinion mining has potential for knowledge discovery. The main target of SA is to find opinions from text extract sentiments from them and define their polarity, i.e positive or negative. In this domain most of the model was designed for English Language. This paper describes a novel approach using Naïve Bayes classification model for Bengali Language. Here a supervised classification method is used with language rules for detecting sentiment for Bengali Facebook Status.**

Keywords: Antonyms word, Naïve Bayes Rules, N-gram model, Parts of Speech (POS) Tagger, Stemming.

I. INTRODUCTION

Smart phone on everybody's hand, Personal computer on every yard, people like to share information. They often use blogs, forum, e-news and social networking sites like Facebook, twitter to express their views and opinions. Huge amount of content is generated day by day thus mining data and extracting user sentiment is an important task. [1]

Sentiment analysis is the process of using text analytics to mine various sources of data for opinions. Often, sentiment analysis is done on the data that is collected from the Internet and from various social media platforms. Politicians and governments often use sentiment analysis to understand how the people feel about themselves and their policies.

Bengali is the language native to Bangladesh and the Indian state of West Bangla. Bangla is the national language in Bangladesh and second most spoken language in India. With about 250 million native and about 300 million total speakers are worldwide. It is the seventh most spoken language in the world by total number of native speakers and the eleventh most spoken language by total number of speakers [2].

Bangla language structure is very flexible in compare to English. Suppose the general structure of an English language: Subject + Verb + Object. (Example: I love you). Any other ordering of this sentence is incorrect. But in Bengali language আমি তোমায় ভালোবাসি (I love you) any order of this sentence is correct. Bangla is a free order, high morphological language. Therefore, Data collection, generation, anomaly detection, features extraction take high challenges. The primary goal of this paper is to analyze the sentiment of individual Facebook status.

II. RELATED WORK

In Bangla language natural language processing is not rich as English language. In the field of sentiment analysis, small amount of work has been done in Bangla language. Lack of available datasets or dependable API's like SenticNet, SentiWordNet, WordNet-Affect were difficult to continue the sentiment analysis task. In early stage Das and Bandopadhya design and developed a sentiwordnet for Bengali language using English-Bangla dictionary. 35805 words were created by them [3].

K. M. Azharul Hasan Mosiur Rahman, Badiuzzaman [4] design a model by using the WorldNet API to get the senses of each word according to its parts of speech and SentiWordNet API to get the prior valence (i.e. polarity) of each word. They calculate the total positivity, negativity and neutrality of sentence or document with respect to total sense. Model accuracy is 76 %.

Shaika Chowdhury Wasifa Chowdhury [5] design a model by using SentiWordNet, WorldNet API for lexical analysis and implement support vector machine and maximum entropy model to detect the sentiment of micro blog. Besides, they gave preference of emoticons. Model accuracy up to 93% by using unigram and emoticon features.

In Natural Language Processing, English is the most popular language for research. For sentiment analysis there many model designed and developed for extract the sentimentality. Most approaches used in this area are [6]

- Subjective Lexicon
- N-gram modeling
- Supervised Classification Method

We choose supervised classification method to detect the sentimentality of Bengali Facebook status.

In sentiment analysis many machine learning approaches were taken for detect the sentimentality. Naïve based classification, Maximum entropy and support vector machine, N-gram approach along with POS information is used to perform machine learning for determining the polarity in English Language(i.e. positive or negative) [7]

A deep neural network approaches were taken for assign polarity of English text. It proposed an efficient embedding for modeling higher-order (n-gram) phrases that projects the n-grams to low-dimensional latent semantic space, where a classification function can be defined. [8]

We choose Naïve based classification method along with Bi-gram and linguistic analysis for sentiment analysis in Bengali Facebook Status.

III. Our Methodology

A. *Data Domain*

Two most popular and common social networking sites are Facebook and Twitter. Generally, in South Asia Facebook is more popular and common than Twitter. People from South Asia used Facebook to Ads their local product, services. Many followers provide product feedback by comments or status at many Facebook Groups. They provide their comments by their native language as Bengali. Facebook users are increasing day by day and in the meantime Bengali contents are increasing too. As almost 300 million of people in the world are using Bengali and most of them use Facebook So Facebook is the best option to retrieve data for our thesis work. Many works has already done to analysis the sentiment of Twitter data. But sentiment analysis on Facebook Bengali data is exiguous.

B. *Data Collection*

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

TABLE-I. Showing corpus data statistics.

Parameters	Positive	Negative
Status	1000	1000
Unique Words	2176	2234
Adjective	823	768
Valance Shifter words frequency	306	378

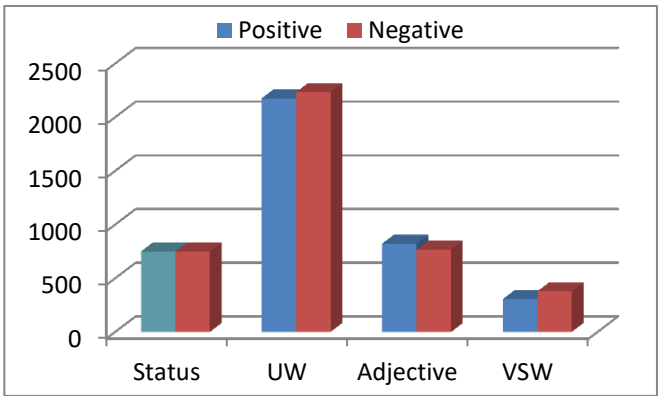

Fig.1. Number of positive, negative status, number of adjective, number of valance shifter (VS) words, Unique Words (UW).

C. *Preprocessing*

Data preprocessing and cleaning step is playing an important role in machine learning. Preprocessing includes remove symbols like hashtags (#), websites URLs etc. Stemming data words is also a part of preprocessing. For words stemming we used nltr open source software [9]

D. *Negation Handling*

In Bengali some words are used to negate the sentence polarity. Those words are '(na)না, (ni)নি,(nei) নাই,(neo) নও,(nay) নয়' etc.

These words are called valance shifting words. Those valance shifting words play an important role in our sentiment detection.

TABLE-II. Showing Bengali sentence and their polarity

Sentences	Polarity
আমি ভাল ফুটবল খেলি(ami valo football kheli)	Positive
আমি ভাল ফুটবল খেলি না(ami valo football kheli na)	Negative
আমি মন্দ ছেলে না(ami mondo chele na)	Positive
আমি মন্দ ছেলে(ami mondo chele)	Negative

In TABLE-II shows one valance shifter word can change the polarity of a sentence (row-1 & row-2, row-3 & row-4). Those valance shifter words occurs in positive status and also negative status and play an important role to change the polarity of sentiment of text. As its play an important role, we have to care about this fact in our thesis works. We normalize that status by the help of linguistic analysis

E. Linguistic analysis

Every language has its own rules for combining words to create sentences. Syntactic analysis attempts to define and describe the rules that speakers use to put words together to create meaningful phrases and sentences. Bengali sentence is divided into three classes a) Simple b) Complex c) Compound.

We normalized our status by considering some rules that are define in Bengali Grammars [10]. For this normalization we build manually a corpus of Synonym- Antonym of Bengali words. Corpus has 1200 unique words. A small portion of this corpus given below:-

TABLE-III. Showing Word and Antonym of this word (a small portion)

Word	Antonym
অকর্মক	সকর্মক
অধিত্যকা	উপত্যকা
সক্ষম	অক্ষম
অনন্ত	সান্ত
উজান	ভাটি
উত্থান	পতন
উন্নয়ন	অবনমন

Following method we apply to normalize a status:-

- Detect simple sentence with one adjective.
- Find "(na) না, (ni)নি, (nei)নাetc " valance shifting word placed at the right side of a simple sentence.
- Remove the valance shifting word.
- Replace the adjective with its antonym [10].

TABLE-IV. Applying Normalization

Before Normalization	After Normalization
আমি ভাল ফুটবল খেলি না। (polarity- Negative)	আমি খারাপ ফুটবল খেলি । (polarity -negative)
আমি মন্দ ছেলে না। (polarity - positive)	আমি ভাল ছেলে । (polarity- positive)

F. Unigram and Bigram

The texts consist of sentences and also sentences consist of words. Human being can easily understand linguistic structures and their meanings, but machines are not enough smart to successful on natural language comprehension yet. So, we try to teach some languages to machines like as an elementary school kid. We used Unigram and Bigram as features in Naïve Bayes classification model.

A unigram is just one single word. But a bigram is a word pairs. The bigrams within a sentence are all possible word pairs formed from neighboring words in the sentence.

IV. PROPOSED ALGORITHM

Bayes' Theorem is a theorem of probability theory originally stated by the Reverend Thomas Bayes. We use Laplace (add-1) Smoothing for Naïve Bayes. Our target to give a document d in a class $c^* = \arg\max_c P(c|d)$. Naive Bayes (NB) classier by first observing that by Bayes' rule

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)} \quad (1)$$

Where P(d) plays no role in selecting c*. To estimate the term P (d|c), Naive Bayes decomposes it by assuming the f_i's are conditionally independent given d's class

$$P_{NB}(c|d) = \frac{P(c)(\prod_{i=1}^{m} P(f_{i|c})^{n_i * d})}{P(d)} \quad (2)$$

Our training method consists of relative-frequency estimation of P (c) and P (f_i|c), using add-one smoothing.

Polarity of Facebook status has been calculated by following method:-

1) Input a set of Positive and Negative status
2) Perform preprocessing at this set of data
3) Stemming every words from this set
4) Detect the simple sentence with one adjective, if valance shifting words occur in this sentence apply the Linguistic analysis method describe above
5) Count Unigram and Bigram of the data words
6) Measures the Prior probability and conditional probability.

7) Apply Laplace smoothing on this data and learn the parameters
8) From query text d choose

$$C_{MAP} = \text{argmax } P(c|d) \quad (3)$$

9) Maximum score from a class between two classes are our desire output class.

V. Result Evaluation

In this article, we used Naïve Bayes Classification model to classify Facebook status in positive or negative which gives us a satisfactory result with f-score 0.72. We collected around 2000 status update from 70 users. K. M. Azharul Hasan Mosiur Rahman, Badiuzzaman [4] design a model by using Naïve Bayes where Bengali sentences were translated to English by using API and used WorldNet, SentiWordNet API. TABLE-I contains the corpus statistics. Fig.1 is shown positive and negative status count for training and testing phase. Classifier accuracy is shown in terms of precision, recall and F-score.

TABLE-V. Result Evaluation

Method	Precision	Recall	F-score
Naïve Bayes with Unigram	0.65	0.56	0.60
Naïve Bayes and Bigram	0.77	0.68	**0.72**

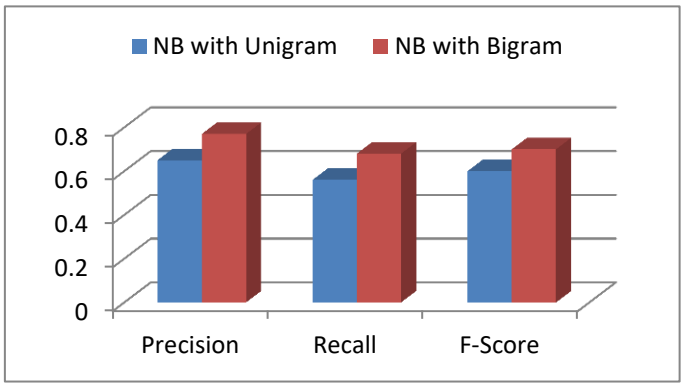

Fig.2 Naïve Bayes (NB) Classification Result Analysis with Unigram & Bigram

VI. Acknowledgment

We wish to express our profound sense of gratitude to our supervisor Assistant Professor Md. Saiful Islam for introducing us to this research topic and providing his valuable guidance and unfailing encouragement throughout the course of the work. Thanks SUST-NLP Research members for helping us to collect data. We collected data from Bengali first Search Engine Pipilika [16]. We are immensely grateful to them for their constant advice and support for successful completion of this work.

VII. CONCLUSION

Sentiment analysis is the most interesting and newly emerged research topic. It will open a new door for the writers, bloggers, and businessman. One can easily know the percentage of product acceptance and make their strategy to improve the product quality.

We used several supervised machine learning methods. It gives us approximately satisfactory accuracy. Our model runs on a small dataset. In future, data corpus can be enhanced and improve processed algorithm to achieved better accuracy. The approach presented here is flexible and suggests promising avenues for further research.

VIII. REFERENCES

[1] Sneha Mulatkar , Sentiment Classification In Hindi, International Journal of Scientific and Technology Research Volume 3, Issue 5, May 2014

[2] wikipedia.org, 'Bengali Language', [Online]. Available: https://en.wikipedia.org/wiki/Bengali_language [Accessed: Retrieved April 21, 2015]

[3] Amitava Das, Sivaji Bandopadaya, SentiWordnet for Bangla, Knowl-edge Sharing Event -4: Task, Volume 2,2010

[4] K M Azharul Hasan, Mir Shahriar Sabuj, Zakia Afrin (2015) Opinion Mining using Naïve Bayes In: IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE) pp. 511-514, IEEE.

[5] Shaika Chowdhury, Wasifa Chowdhury, "Performing sentiment analysis in Bangla microblog posts", ICIEV, 2014, 2014 International Conference on Informatics, Electronics & Vision (ICIEV), 2014 International Conference on Informatics, Electronics & Vision (ICIEV) 2014, pp. 1-6, doi:10.1109/ICIEV.2014.6850712

[6] Amandeep Kaur, Vishal Gupta, A Survey on Sentiment Analysis and Opinion Mining Techniques, Journal of Emerging Technologies in Web Intelligence, Vol. 5, No. 4, 2013

[7] Bo Pang, Lillian Lee, and Shivakumar 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

[8] Farah Benamara, Carmine Cesarano, Antonio Picariello, Diego 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.

[9] nltr.org ‘snltr-software’ [Online]. Available: http://nltr.org/snltr-software/ [Accessed: Retrieve April 3, 2015]

[10] banglaacademy.org, বাক্য রূপান্তর, [Online]. Available: http://www.ebanglalibrary.com/banglagrammar/বাক্য-রূপান্তর [Accessed: Retrieve April 3, 2015]

[11] Peter Turney. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews, pages 417424, 2002.

[12] Sentiment classification based on supervised latent n-gram analysis Bespalov, Dmitriy, et al., 2011

[13] T. Joachims, Making large-Scale SVM Learning Practical. Advances in Kernel Methods-Support Vector Learning, B. Schlkopf and C. Burges and A. Smola (ed.), MIT-Press, 1999

[14] Bespalov, Dmitriy, et al. ”Sentiment classification based on supervised latent n-gram analysis.” Proceedings of the 20th ACM international conference on Information and knowledge management. ACM, 2011.

[15] Kristina Toutanova, Dan Klein, Christopher Manning, and oram Singer. ”FeatureRich art-of-Speech agging with a Cyclic Dependency etwork.” In roceedings of AAC, pp. 252-259

[16] Bengali Search Engine Pipilika, Available at: http://www.pipilika.com/