OPINIONIST-A Lexicon and SVM Based Tool for Analyzing Opinion on Social Media

Bushra Fazal, Tazeen Muzammil, Lubna Siddiqui, Sumayya Abid, Sadia Waheed, Saira Yaseen

Abstract—The world of technology is developing rapidly and people are adopting to social media fast. To this day twitter is one of the most popular social network. Personal opinions regarding "everything" is shared on the social media. This data is very useful for gathering public opinion. Opinionist provides a web based platform for user to fetch tweets on a desired topic and displays the result in the form of graphical notations. This makes it easier for the user to determine whether the public opinion is positive or negative about the particular topic. The Opinionist uses two sentiment analysis techniques namely Lexicon Bayes approach and Support Vector Machine Algorithm. Using the result of the two algorithms and produces an average and displays it along with the results produced by the two algorithm. Use of averaging reduces extreme values giving a more general and realistic value**.**

Index Terms—Stop-words, Opinion Mining, Sentiment analysis, Lexicon, Support Vector Machine

1 INTRODUCTION

THEREare a large number of tweets are present on twitter for a single topic. It is very difficult and tiring work to reach a final decision that weather the public opinion is going in favor or against the topic.

The purpose of Opinionist is to provide a web based application reads the tweets about a user selected topic and classifies them as positive negative or neutral. The results are depicted in graphical format. Opinionist is based on natural language processing and also provides a comparison of multiple opinion mining techniques namely lexicon based approach and Support Vector Machine algorithm. Using averaging any extreme values produced are averaged out giving a more general result.

Lexicon based approach assumes that the contextual sentiment of a text in other words polarity of a text is the sum of contextual sentiment or polarity of each word, sentence or phrase [1]. In Opinionist the lexical analyzer takes in the XML file produced by the-Preprocessor and extracts data tables. It sums the polarity of words in one sentence to get the sentiment orientation of the sentence. Then sums the sentiment orientation of the sentence to give the sentiment score of the tweet. In the end sums the scores of all tweets to give the total polarity. In case of negation the sentence polarity is reversed. Negation is the use of words like not and hardly before any word. For example good is a positive polar word. Using not before good makes the overall polarity negative. Negation in lexicon approach is handled at sentence level.

The lexical analyzer starts with the word table of the data set. It sums up the polarity of all words with the same comment and sentence id. One tweet is treated as one comment after all sentence polarity have been calculated it sums all the sentence polarities with the same comment id. It then sums the comment polarities to produce total positive, negative and neutral polarities. Each polarity is divided by the total polarity then multiplied by hundred to give the percentage polarity [2].

After cleaning noisy data (stop words removal) text is converted into a pattern "positive negative positive positive" by assigning Labels to each word of the sentence(Labels are fetched from the Word Table from xml dataset). The preprocessed text is not suitable because SVM has defined Input and output format Firstly the document is converted into the required format for SVM [3].Then a model is trained for input-output mapping The model is trained using the training set in which numeric values are assigned to patterns like "positive negative =0" whereas zero is representing neutral ,which means that if there are two words 1 is positive 1 is negative ,the overall sentence will be neutral. The training set is shown below:

Table 1: SVM Training Set

Text	Ispositive	Text	IsPositive
Positive	1	Positive positive positive	1
Negative	-1	Positive negative positive	1
Positive positive	1	Positive positive negative	1
Positive negative	0	Negative positive positive	1
Negative positive	0	Negative negativee positive	-1
Negative negative	-1	Neutral	0

- *Bushra Fazal is a lecturer at department of Computer & Software Engineeringof Bahria University Karchi campus,Pakistan PH+92 21-111-111-028 E-mail: luna.siddiqui@bimcs.edu.pk*
- *Tazeen Muzammil is a Senior Assistant Professor r at department of Computer & Software Engineeringof Bahria University Karchi campus,Pakistan PH+92 21-111-111-028 E-mail: luna.siddiqui@bimcs.edu.pk*
- *Lubna Siddiqui is an Assistantt Professor at department of Computer Science of Bahria University Karchi campus,Pakistan PH+92 21-111-111-028 E-mail: luna.siddiqui@bimcs.edu.pk*

A class (positive or negative or neutral) label is assigned to each sentence in the whole document .On the basis of class label comment of sentences ,labels are assigned to comments .After the labelling of comments overall text polarity is calculated in the form of percentages for positive ,negative and neutral text.

2 MEHTODOLOGY

Starting with a provided keyword by user, related tweets are collected from twitter using "LINQ to Twitter" Wrapper for twitter Stream API [4]. Then hyperlinks, citations and extra characters are removed from tweets.

After gathering tweets they are passed through the preprocessing module. The steps involved in preprocessing are POS tagging, Stop words removal, tokenizing, getting scores from AFINN List (contain list of 2477 words and phrases scores from range -5 to +5) [5], negating the score if the previous word is a negating word (not, couldn't, never).

The preprocessing module takes the whole text as input. It first removes stop words from the text. The resultant text is then passed through Stanford NLP POS tagger. The POS tagger marks the parts of speech in the text. The tagged text is then broken into paragraphs then sentences and finally tokens are generated. Here words are considered tokens. The generated tokens are checked against the "AFINN" list to generate polarity scores.

After performing above steps of preprocessing the tweets written to xml file along with their each sentence, sentiment words (adjectives, Verbs and Adverbs), scores and POS tags.

By taking the scores from xml dataset and afore mentioned algorithms are applied to generate sentence, paragraph and document level polarity. The results of the algorithms are averaged. The percentage polarity s calculated for each algorithm as well as for average. Finally the three results are displayed in graphical format on Opinionist.

2.1 Opinionist How It Works

The Opinionist is divided into 3 major modules: website, preprocessing and algorithm implementation. These s modules have multiple sub modules.

The website module takes in the user input and uses twitter API to fetch tweets. It also displays the final result in graphical format. It provides the ability to run sentiment analysis to only registered users and maintains an activity log. Anyone can get registered as a user. The user can logout any time and can view options that are authorized to it.

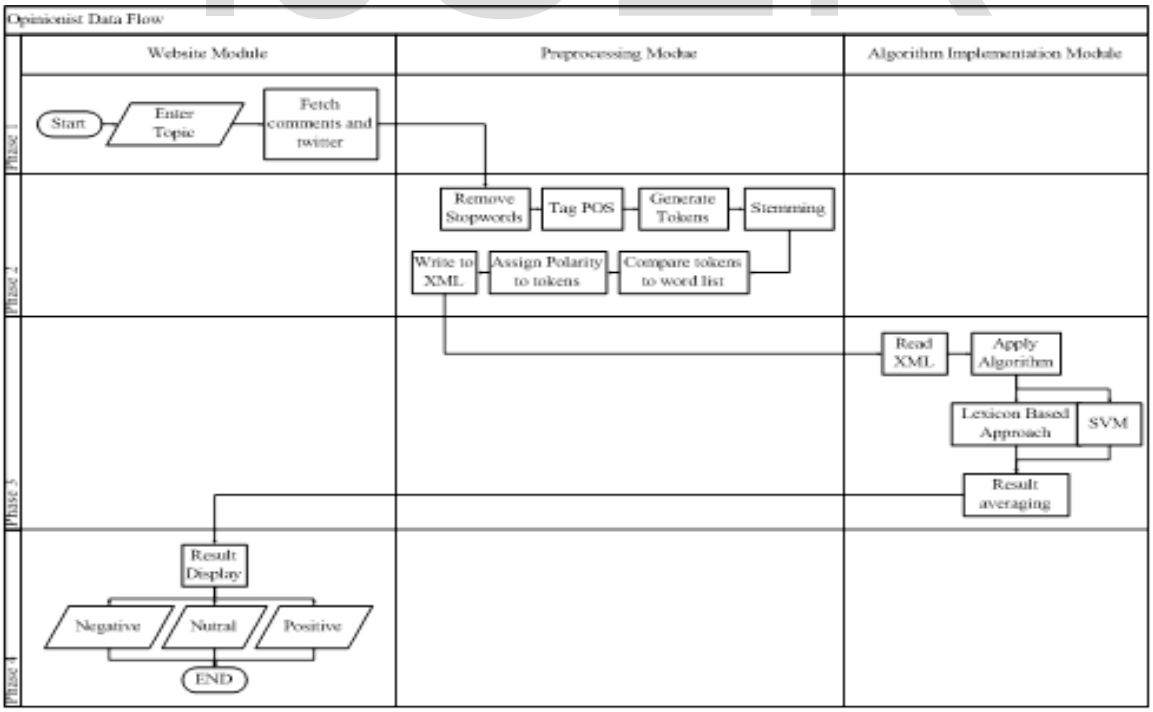

Figure 1: Complete Overview of the Opinionist

The preprocessing module cleans the data retrieved through the API before they can be processed to obtain the results. It consists of submodules: Stop-words remover, POS Tagger and Tokenizer stemmer. The stop words remover removes the frequently occurring words that have no polarity, the POS tagger label the word with the part of speech they belong to and the tokenizer generates tokens passes them through stemmer and then compares the tokens to a predefined list to determine their polarity [6]. The preprocessing module uses XML to save the results.

The algorithm Implementation module follows the factory pattern and has submodules: Lexicon Based Approach, SVM and averaging. The first two are opinion mining algorithms and the latter averages the results generated by the two algorithms and produces final results. This improves the accuracy of the results generated. This module uses the XML generated by the previous module as an input and
forwards the result to the website module where they are displayed to the user as shown in figure 2.

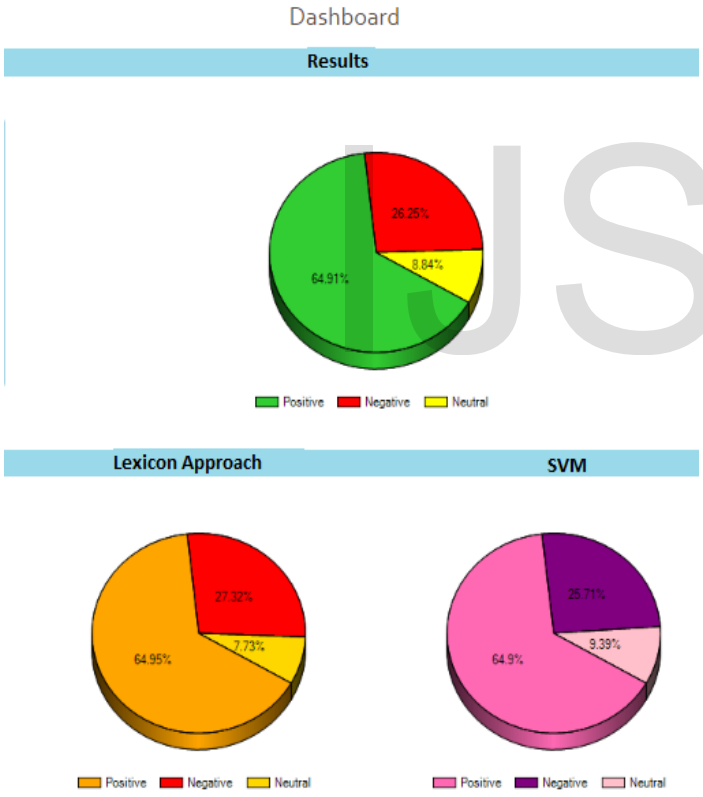

Figure 2: Graphical depiction of results on Opinionist website.

The chart under lexicon approach displays the results produced by the lexical algorithm. The chart under SVM displays the results produced by SVM algorithm. As it can be seen that the values produced by the two algorithms differ. The chart under the results heading displays the average of the two algorithms. The details of the labeling can also be viewed as shown in Table 2.

Table 2: Detailed display of labels assigned to the tweets

C_ Id	Comments	Polarity
0	iPhone sales fall second quarter row	neutral
1	Australian banks demand access NF chip used iPhone complete Apple pay	negative
2	iPhone leather sleeve multiple choice	neutral
3	MicroSoft Pix app uses artificial intelligence auto enhance iPhone photos	neutral
4	Water red heavy duty schock proof hybrid rugged hard case Apple iPhone C	negative
5	My iPhone black thanks Naudi	positive
6	Car.amp motor bike Tom Tom universal high speed dual charger iPod iPhone, smart phone	neutrals
7	My phone dying I micro usb converter.I need cable electricity.Why Android user use iPhone charger.dats sir	negative
8	Suga belly lost mode iPhone.Do use device include iPhone its lost use track phone	negative
9	Shock proof hybrid fashion soft Hard PC Case Cover Apple iPhone SE SG	negative
10	Mobile juice pack air iPhone black charging phone case	neutral
11	CRP disabilities iPhone App Bone. Conduct implant helps. Year. Old hearing Back	positive
12	Wants play real machine like iPhone	positive
13	Funny Pics Depot Android VS iPhone	neutral
14	Apple support HELL SMALLER iPHONE too damn small. Nobody small keyboard amp more. MarkSimoneNY FoxNews LouDobbs CNN	negative
15	eienacarstoiu Experts Apple doesn't sell enough Apples earnings beat impressive	positive
16	lizardiuvr I m happy longer iPhone camera quality pokemon go.I use snapchat filters.Deression CU	neutral
17	Apple iPhone GB gold bell mobility smartphone	neutral
18	Opened park request iPhone EI Camino Del Mar Resubmitted	neutral
19	Awesome users just follow me via find unfollower	neutral
20	Apple announces carekit available April	neutral

3 RESULTS

Both algorithms (SVM and Lexicon are giving good results) .Working on different methods both algorithms provide good results .In our case the accuracy for both algorithms is dependent on the AFINN list (If the word is present in AFINN, it is included, otherwise it is ignored) But if efficiency is concerned Lexicon is best because SVM take few seconds in training model (also it is dependent on Libsvm Api).

4 CONCLUSION

Sentiment analysis is a developing natural language processing technique. There are multiple approaches to sentiment analysis but there are still hindrances in acquiring a full human like opinion or judging an opinion to a high accuracy. To deal with context language models are introduced .Recognition of sarcasm is still not possible and thus for these reasons further enhancement in language processing modules is required.

REFERENCES

[1] V. Y. a. H. E. Prabu Palanisamy, "Serendio: Simple and Practical lexicon based approach to Sentiment Analysis," *Serendio Software Pvt Ltd*, p. 6.

[2] S. N. a. R. J. Richa Sharma, "OPINION MINING OF MOVIE REVIEWS AT DOCUMENT LEVEL," *International Journal on Information Theory (IJIT), Vol.3, No.3,*, p. 9, 2014.

[3] M. V. G. M. V. K. ,. M. K. D. Ms. Gaurangi Patil, "Sentiment Analysis Using Support Vector Machine," in *International Journal of Innovative Research in Computer and Communication Engineering* , 2014.

[4] J. Mayo, "Documentation," 26 april 2015. [Online]. Available: https://linqtotwitter.codeplex.com/documentation.

[5] F. Å. Nielsen, "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs," MARCH 2011.

[6] "Alphabetical list of part-of-speech tags used in the Penn Treebank Project," 2003. [Online]. Available: www.ling.upenn.edu.