International Journal of Research in Advent Technology, Vol.5, No.4, April 2017
E-ISSN: 2321-9637
Available online at www.ijrat.org

A Review of Techniques Used for Sentiment Analysis on Twitter Data

Alind Jain , Shreem Kapoor , Anshu Malhotra(Profesor)
Computer Science Department
Bharti Vidyapeeth College of Engineering
New Delhi, India
Email: alindjain1@gmail.com ,shreemkapoor@gmail.com, anshu.malhotra@bhartividyapeeth.edu

Abstract- "What are people's thoughts?" has been the question on people's mind for a long time and that's what drives us to act and make decisions. And now we are able to analyse what people think with the help of social networks, forums and microblogging services. These sites have made it possible for us to find out the sentiment or opinions of people with whom we have personally no relation to, that is the people we don't know exist. Today almost every site has a forum or a blog where they are able to leave their comments or reviews about a product or services. Twitter which is one of the largest microblogging website, where people blog about almost everything.

Index Terms- n-gram categorisation; Pre-Processing of Data; Bag of Words; Tokenisation; Machine Learning Approach; Lexicon Method; Special Method .

1. INTRODUCTION

Sentiment Analysis also Known as Opinion Mining is on the rise and has gained much attention in direct response to the increasing interest in system which deal with opinion Mining. Many companies have begun to understand that it is a rich mine of information. And the companies employ analysis to get a better and informed decision in context to their sales and their advertisements. In this we strive to tackle the problems of n-gram categorisation also called sentiment polarity categorisation, which is the main problem in sentiment analysis.

Fig. 1: Graphical display of opinion-related

2. A REVIEW ON SENTIMEN ANALYSIS

2.1. *Sentiment Analysis*

Our most fundamental issue that arises is the n-gram categorisation. n-gram categorisation can be defined as finding the sentiment polarity of a word or a set of words, where trail of words are being assorted into phrases called n-grams. N-grams method can decrease bias, but may increase statistical sparseness. Like a unigram can be defined as a single word and bigram can be defined as a set of two words and so on. Sentiment polarity would be, if received a text, we have to categorize that piece of text into one of the sentiment polarity and that can be either positive, negative or neutral [4] as shown in Figure 1.

N-gram classification can be basically defined as having mainly three main levels of complexity classification, namely the document level, the sentence level and lastly the entity level or unigram level.
As the name suggests the document level take into consideration of the document as a whole, the sentence level gives the polarity to a single sentence, and the unigram level looks at the polarity of the words.
N-gram categorisation has been proven to show an improvement in Text classification, but there is a unique solution for which sizes of N-gram for employment.

The simplest method for unigram categorisation would be the bag of words method (BOW). The BOW model, the document gets a categorisation as a vector of words in Euclidian space where each word is

independent from others. This bag of 2 single words is commonly called a stock of unigrams. The independence of unigrams means that the occurrence of one unigram in the text will not become the appearance of any other unigram.

The main two methods of sentiment analysis, which are shown in Figure 2 are machine learning based method and lexicon-based approach, both rely heavily on the bag-of-words. In the machine learning supervised method, the classifiers are using the n-grams as attributions. In the lexicon-based method the unigrams or n- which are found in the lexicon are assigned a sentiment polarity score, the overall polarity score of the document is then computed which is sum of the polarities of the unigrams.

But we have come to realise that the Lexicon Based method and the Machine learning approach both involve the preparation of data and the pre-processing of data.

3. DATA ACQUISITION

3.1.1. Pre-Processing of Data

- A string is cut down into words plus other elements called tokens. The separators for defining individual words are mostly whitespace while other symbols can and are also used, this is called Tokenisation or the creation of Bag of Words (BOW). [11]
- All tokens are converted into Lowercase.[3]
- Replacing words with their stems or roots. The memory and procession of the BOW gets reduced due to this. For example words like decision, decide, decisive and decisively can always be correlated with one word and counted only once. Keeping in mind words with different meaning should not be grouped together and after Stemming the words remain different.[4]
- Removal of connecting Functions like prepositions, articles, conjunction, etc.[4]
- Part of Speech Tagging(POS) refers to giving a tag to each word or unigram of a sentence and defining as to which part of the Grammatical Speech it belong to that is noun, adjective, verb, conjunction, etc. [8]
- When we use a negative word like Not, Don't, No, Can't, etc. then the whole meaning of the sentence changes and this is important as the occurrence of one negation or a negative word can change the whole meaning of the text or the sentence. [10]
- Symbols used in twitter are not very useful to predict the sentiment of the tweet, these include URL's, @-mentions, hash-tags, etc.[13]

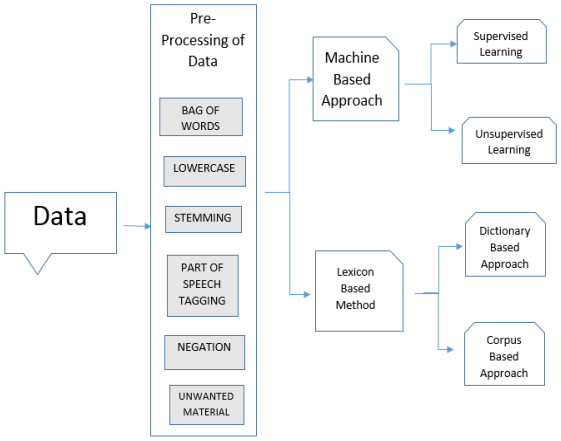

Fig. 2 Sentiment Classification Techniques [5].

4. PROCESSING OF DATA

4.1.1 Machine Learning Based Approach

An algorithm is implemented that makes use of labelled data, snips features that trains a model which has different classes and makes a function which can be used to sorting out new examples [6]. The Labelled data has the primary dataset which has the sentiment allocation or classes pre-defined or supervised data [7] for the training of the model. The goal is to build an algorithm which learns, with the input of the model after the training will give us accurate classification of unlabeled documents.

The researchers classified texts as being mainly negative and positive but review of the Twitter Dataset, they have realised some tweets only state facts and thus, cannot be put into the class of negative or positive. Due to this a new class arose which was called neutral and didn't affect the overall sentiment polarity of a document.

4.1.2 Lexicon Based Approach

The lexicon Based method relies on a dictionary or lexicon of words which have pre-defined polarity [3]. A Dictionary is created for the analysis of the document, after which we assign them different

polarity. And all of this is done after the pre-processing of data.

Sentiment Score Calculation is done where each word from a text is compared against the lexicon. If the word exists in the lexicon than the score of the word I added to the total sentiment of the score of the document.

Lexicon Generation

- Hand-tagging Lexicon- Manually tagging of words as positive or negative. It involved reading several thousand of text documents and selecting the words which are carrying sentiment.[4]
- Constructing a dictionary from labelled data- The data gets tokenised and a BOW is created.[11]
- Bootstrapping- It is extending of the lexicon, in which two adjectives are conjoined with the word "and" and it is known that "and" mostly conjoin words with the semantic introduction. Example: "The weather is cool and awesome."

As the words Cool and Awesome are joined with "And" it would be considered that both words carry the same sentiment. If the word "cool" was in the lexicon than a new word "awesome" would be added in the dictionary. [3]

Every Word is assigned a sentiment polarity [5], from 1 to 0;

1 corresponding to a positive sentiment while 0 being a negative. Where the range [0.6, 0.4] can be considered as neutral keeping in mind some tweets are just facts and hold no polarity.

The polarity Score as defined by the [3] of sentence can be calculated by the following formula:

Polarity Score= 2*Goodness – 1. [3]

Example

The Polarity Score of "Like" would be 0.463*2-1= **-0.094** which is close to zero and tells the word is neutral

The Polarity Score of "benefit" which is an absolute positive word would be 1

1*2-1=**1**

Examples for the Polarity Score Calculator

3. Special Case

The accuracy of the analysis can be increased by collaborating the two approaches, namely the Machine Learning Method and the Lexicon Based Approach. The results were proven to show that it gave positive data as compared to the approaches working independently.

equations should be numbered consecutively, with the number set flush right and enclosed in parentheses. The equation numbers should be consecutive within the contribution

4. CONCLUSION

Opinion Mining or Sentiment Analyses is a field where we study people's emotions(basically happy, angry, sad, neutral, etc.) , their attitude about a certain topic and their Sentiments.

It can be likely that many other applications are not discussed in this review. It is understood that the Sentiment Analysis and Opinion Mining are more likely dependent on the topic at hand. From the above review we have to come to conclude that the special case that involves both the Machine Learning Method and Lexicon Based Approach consistently outperforms the independent method, as they both make up for each other where they lack . We can also give in to conclusion that it would be easier in classification of tweets and if we improve the training dataset and also keep updating it from time to time we can get accurate results.

Acknowledgments

We want to thank Ms Anshu Malhotra, who is a Professor at Bharati Vidyapeeth's College of Engineering for assistance with this paper.

REFERENCES

[1] Hassan Saif, Yulan He, harith Alani "Alleviating Data Sparsitiy for Twitter Sentiment Analysis," http://oro.open.ac.uk/38501/1/hassan-MSM2012.pdf .

[2] K.; Broder, John Dodd, "Twitter Sentiment Analysis", http://trap.ncirl.ie/1868/1/johndodd.pdf .Hassan Saif, Yulan He, harith Alani "Alleviating Data Sparsitiy for Twitter Sentiment Analysis," http://oro.open.ac.uk/38501/1/hassan-MSM2012.pdf .

[3] John Dodd, "Twitter Sentiment Analysis", http://trap.ncirl.ie/1868/1/johndodd.pdf .

[4] O. Kolchyna∗ , T. T. P. Souza, C. Treleaven, T. Aste, "Twitter Sentiment Analysis", https://arxiv.org/pdf/1507.00955v1.pdf .

[5] Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani∗, Veselin Stoyanov, "SemEval-2016 Task 4: Sentiment Analysis in Twitter ", http://alt.qcri.org/semeval2016/task4/data/uploads/semeval2016_task4_report.pdf .
[6] Walaa Medhat, Ahmed Hassan, Hoda Korasgy, "Sentiment Analayis Algorithm and Applications" http://www.sciencedirect.com/science/article/pii/S2090447914000550?np=y .
[7] Bo Pang and Lillain Lee, "Opinion Mining and Sentiment Analysis", http://www.cs.cornell.edu/home/llee/omsa/omsa.pdf .
[8] GebreKristos Gabreselassie Gebremeskel, "Sentiment Analysis and Twitter Posts about News", file:///C:/Users/mickey/Desktop/Research%20Paper/TH-01.pdf
[9] Afroze Ibrahim Buqapuri, "Twitter Sentiment Analysis", https://arxiv.org/ftp/arxiv/papers/1509/1509.04219.pdf .
[10] Jon Tatum, John Travis Sanchez, "Twitter Sentiment Anlysis", https://pdfs.semanticscholar.org/f7ad/fb74f2d4536186069e11c4354fac77efdbec.pdf .
[11] G.Vinodhini, RM.Chandrasekaran, "Sentiment Analysis and Opinion Mining: A Survey", https://www.ijarcsse.com/docs/papers/June2012/Volume_2_issue_6/V2I600263.pdf .
[12] Xing Fang and Justin Zhan, "Sentiment Analysis using product review Data", http://journalofbigdata.springeropen.com/articles/10.1186/s40537-015-0015-2 .
[13] Demo from https://www.lexalytics.com/ .
[14] Hao Wang, Dogan Can, Abe Kazemzadeh, François Bar and Shrikanth Narayanan, "A System for Real-time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle", http://delivery.acm.org/10.1145/2400000/2390490/p115-wang.pdf?ip=43.252.28.119&id=2390490&acc=OPEN&key=4D4702B0C3E38B35%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35%2E6D218144511F3437&CFID=681900521&CFTOKEN=53196714&__acm__=1477250728_039e869cb8a0448bd7be66fdb930dfa2 .
[15] Md. Daiyan , Dr. S. K. Tiwari , Manish Kumar , M. Aftab Alam, "A Literature Review on Opinion Mining and Sentiment Analysis", http://www.ijetae.com/files/Volume5Issue4/IJETAE_0415_46.pdf .
[16] Thelwall, M., Buckley, K., & Paltoglou, G. (in press). Sentiment in Twitter events. Journal of the American Society for Information Science and Technology
[17] Part of speech Tagging , http://www1.cs.columbia.edu/~kathy/NLP/ClassSlides/Class5-POS/mypos.ppt .
[18] Comparing Data-Driven Learning Algorithms for PoS Tagging of Swedish (2001), http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.12.9840 .
[19] 100 happy tweets 100 sad tweets from16.44 on Thursday 14/04/2011, http://www.scribd.com/doc/56012075/100-sample-happy-tweets .
[20] Applications in Social Media: Sentiment Analysis , Neil Glassman ,http://socialtimes.com/sentiment-analysis-socialmediamarketing32_b59924 .
[21] Is it possible to guess a user's mood based on the structure of text?, http://stackoverflow.com/questions/933212/is-it-possible-to-guess-a-users-moodbased-on-the-structure-of-text
[22] Sentiment Tutorial, http://alias-i.com/lingpipe/demos/tutorial/sentiment/read-me.html .
[23] Python 2.7, http://www.python.org/ftp/python/2.7.2/python-2.7.2.msi.