LEXICON-BASED RULE EXTRACTION FOR SENTIMENT ANALYSIS IN BIG DATA ENVIRONMENTS: A SCALABLE APPROACH FOR SOCIAL MEDIA AND TEXT MINING APPLICATIONS

DR. SEAN O'CONNOR[1]*
[1]TRINITY COLLEGE DUBLIN, DEPARTMENT OF COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE, DUBLIN, IRELAND

ABSTRACT

In today technology world, big data is generating from many sources like government, banks, social media, science, medical field and many others. This data is very much complex in terms of their properties and it requires some new tools and technologies to analyze. Sentiment analysis is the field of big data analysis which discovers the writer's feeling and behaviour from his data. Behaviour of user can be analyzed as positive, negative or neutral. Lots of techniques can be used to do so. If beyond these polarities something can be done it will be very interesting. The main objective of this paper is to draw the next line in sentiment analysis under big data environment. That line is to fetch information like what groups of users want to say. That will be extracted on the basis of their strength. Here strength means the frequency of that information or rule. This research paper proposed a new approach which is an extension to lexicon method has been introduced to find the rule extraction which can help in the area of sentiment analytics under the big data environment via Hadoop like approaches.

Keywords: *Hadoop, Sentiment Analysis, Unstructured Data, Big Data, Rule Extraction.*

I. INTRODUCTION

Sentiment analytics is used to predict the polarity of sentences. It means predicting the nature of a person whether it is positive or negative. It is the sub task of text analytics. As social media is producing lots of unstructured data which can be very useful for some company because it may be customers‟ views on their products and from this data companies can make lot of money after analysing it in a proper way. Sentiment analysis is one of the best ways to predict that whether user thinks positive or negative on their products, services and lot more things. Facebook and Google are doing lots of analysis just to predict the user's behaviour and feeling which helps them in advertisement targeting. Sentiment analysis is also important in application development. It means developing the application which involves users‟ interest and this can be figure out via sentiment analysis. Primarily Companies may analyse sentiment about:

a) Product
b) Services
c) Reputation
d) Competitors

Sentiment Analysis Methods

There are lots of methods which are used to compute the sentiment analysis. They all are used to classify the text and then finds the polarity of text that whether it is positive or negative or neutral. From last some years a lot of methods appear in the field of sentiment or opinion analysis. These methods can be classified into two categories:

a) Machine Learning Methods
b) Lexicon Based Methods

a) Machine Learning Methods

Machine learning methods works on training the algorithm with training data set. After this training, algorithm is applied on the actual data set. For supervised and unsupervised learning machine learning algorithm needs to be trained first. Purpose of Training the algorithm is to achieve the new goal and that is dealing with unknown data. After training the algorithm it can also deal the unknown data. Machine learning methods are specially designed for processing the human language but because of its complexity these methods need some training then only they become able to work. Machine learning methods are more accurate than lexicon based methods.

Naïve Bayes is one of the most famous methods to classify the text. It is very simple and efficient approach in Natural language processing. It basically works on the probability. It categorizes the text on the basis of probability.

Primarily Prior probability is calculated in Naïve bayes method but it works on each word in the sentence independently. There are three phases in naïve bayes. One is learning, second is testing and third is mapping. Primarily algorithm needs training on pre classified data. So that it can also handle the unknown data based on the training. This thing is done in the mapping phase.

Support vector machine (SVM) is another machine learning method used to classify the text which helps in calculation of sentiment analysis. It is very helpful in recognizing the patterns. Use of support vector machine in sentiment analysis is to classify text so that polarities can be assigned to each word which is very essential in the sentiment analysis. There are also other lot of machine learning methods which can be used to classify the text and it can be further used for sentiment analysis.

b). Lexicon Based Methods

Lexicon based methods are the second approach to find out the sentiment analysis or polarity of a document. This methods works on the polarity of whole document or whole sentence instead of individual words which basically sum of polarity of words. This computes the whole polarity of sentence or whole document. Lexicon based methods are completely different from machine learning methods. In lexicon methods algorithm needs no training. In these approaches widely a dictionary has been used with polarity associated with each word. This approach is very much faster in huge amount of data as compare to machine learning methods because they do not need any type of training. Although machine learning methods are more accurate than lexicon but they are not time efficient or fast. Negation and blind negation are the other aspects which can reverse the polarity of whole sentence. For example "Movie was not good". Here good word infers the positive sentiment but nature of whole sentence is negative because of word "not". So this thing has to be computed and it is computed under lexicon based methods.

II. BACKGROUND STUDY

The system proposed in this paper extends the existing sentiment analysis. A hypothesis is formulated that if Lexicon based method is extended to some level that it gives positive, negative sentiments and extract some rules based on the frequency. Lexicon based method is although not very much accurate as compare to machine learning methods but it is much more time efficient. Time efficiency is necessity because whole study is performed on big data or more precisely on unstructured data.

So the main idea in this paper is to extract some interesting rules by extending the lexicon based method which is based on usage of dictionary. This can be very helpful in predicting that what a particular cluster want to say or infer. Here cluster belongs to group of people who express their sentiments.

Main objectives of this paper are:
a) Sentiment Analysis
b) Rule Extraction
c) Frequency of Rule

III. PROPOSED METHODOLOGY

Tools which are used for the data analysis part and where whole experiment is performed is Hadoop. It is a framework which allows processing of huge data. It is the most frequent tool used in big data analysis. Hive is Query language which also runs on the top of hadoop. Purposed work is implemented in Hive over the hadoop. This hadoop is the sandboxed version by hortonwork. Version of hadoop which is used in experiment is 2.1. For the visualization purpose Integration of Microsoft Excel 2013 with hadoop shows the numeric data into graphical form for better and easy understanding. Here is the algorithm which tells the whole working of proposed work in this paper.

Algorithm

(sentiment_count , sentiment_of_tweet)
// Finding the polarities of tweets as positive negative or neutral
Start
sentiment_count = = 0;
for each tweet ti

```
{
for each word wj which is present in
Dictionary
{
if ( Polarity[wj] = = "Positive" )
{
sentiment_count = sentiment_count + 1;
}
if ( Polarity[wj] = = "Negative" )
{
sentiment_count = sentiment_count + 1;
}
}
if sentiment_count of tweet ti > 0
{
sentiment_of_tweet = "Positive";
}
if sentiment_count of tweet ti < 0
{
sentiment_of_tweet = "Negative";
}
if sentiment_count of tweet ti = = 0
{
sentiment_of_tweet = "Neutral";
}
// Break the tweet into words
// Extract the rules and count the frequency of rules
Sentences(ti);
ngrams (sentences (ti)); }
Stop
```

In proposed methodology a lexicon based approach has been used which is a dictionary based approach. Sentiment analysis has been performed on the data set which is taken from twitter. Polarity assignment is very important task in the approach. So for calculating the polarity score has been assigned to each and every word which breaks up from the each tweet. After sentiment analysis rule extraction has been performed on the data set used in the experiment in hive on the top of hadoop framework.

IV. RESULT & DISCUSSION

In this section, the proposed system results have been discussed. Fig. 2 shows the sentiment analysis on the dataset that is taken for experiment. Rules have been extracted from which shows the sentiment of a particular group. Frequency shows the efficiency of that particular rule.

There are four rules which are extracted from the dataset that is taken for the experiment. First rule is suggesting that 72 people are saying that the acting in movie is good. Now this rule implies that it is a positive tweet and 72 people are praising the movie for the acting part.
In another rule 14 people are saying that movie was terrible. It implies that movie was not good and it is a negative rule. So these rules are depends upon the frequency. If frequency of rule will more than that rule will become more efficient. In other part of experiment tweets has been categorized as positive, negative and neutral. As results show the numeric figure of calculated results on data.

V. CONCLUSION

Big data is present everywhere today and it will grow in the future too. Unstructured data is the major component of big data. It is very much complicated because of its structure. It has no structure or model with which this data can be handled. But in this data there is lot of meaning and hidden benefit. This can only come out if it is treated

properly and analysed with an efficient way. Sentiment analysis is the good area which treats this data as a challenge and figure out something useful for the business intelligence. A new approach which is an extension to lexicon method has been introduced to find the rule extraction which can help in the area of sentiment analytics under the big data environment via Hadoop like approaches.

REFERENCE

1. *Olmezogullari, E., Istanbul, Turkey, Ari, I. "Online Association Rule Mining over Fast Data". In IEEE International Congress on Big Data, pp. 110-117, 2013.*
2. *Chetan Kaushik, Atul Mishra "A Scalable, lexicon based Technique for Sentiment Analysis". In International Journal in Foundations of Computer Science & Technology (IJFCST), Vol.4, No.5, September 2014*
3. *A. Nisha Jebaseeli, E. Kirubakaran "A Survey on Sentiment Analysis of (Product) Reviews". In International Journal of Computer Applications (0975 – 888) Volume 47– No.11, June 2012*
4. *Zhu Nanli, Zou Ping, Li Weiguo, Cheng Meng in "Sentiment analysis: A literature review". In IEEE, Management of Technology (ISMOT), International Symposium, pp 572-576, 2012*
5. *Sunil B. Mane, Yashwant Sawant, Saif Kazi, Vaibhav Shinde in "Real Time Sentiment Analysis of Twitter Data Using Hadoop" In International Journal of Computer Science and Information Technologies, Vol. 5 (3), 3098 - 3100, 2014*
6. *Vasu Jain "Prediction of Movie Success using Sentiment Analysis of Tweets". In The International Journal of Soft Computing and Software Engineering [JSCSE], Vol. 3, No. 3, Special Issue: The Proceeding of International Conference on Soft Computing and Software Engineering [SCSE"13], 2013*
7. *Walaa Medhat, Ahmed Hassan, Hoda Korashy "Sentiment analysis algorithms and applications: A survey" Ain Shams Engineering Journal vol 5, issue 4 pp 1093–1113, 2014.*