Volume : 4 | Issue : 10 | Special Issue Oct- 2015 • ISSN No 2277 - 8179

A Lexical Approach for Tweets Sentiment Classification

Computer Science

KEYWORDS: Sentiment Analysis (SA), Big Data, Negation, Opinion Mining, Twitter dataset, evaluation

Sneha Vishwakarma Manav Rachna International University, Faridabad

Suresh Kumar Manav Rachna International University, Faridabad

ABSTRACT

In the past few years, there has been a huge growth in the use of microblogging platforms such as Twitter, Facebook, LinkedIn, Instagram etc. Millions and billions of people share their view on different aspects on them every day which generates many yottabytes of data, the data we now call as BigData. This paper focus on Twitter, the most popular microblogging platform, it lets you connect with people, express yourself, and discover more about all the things you like. Here a new approach for sentiment analysis in Twitter is proposed. Sentiment Analysis is considered as a big data task with increase in social media on the web. Conventional approaches are not that efficient to handle the vast amount of sentiment data. A typical tweet contains word variations, negations, phrases, hashtags, emoticons etc. Main focus of the research was to find such a method that can perform sentiment analysis on big data sets more effectively. The method proposed classifies the tweets into positive, negative or neutral in a fast and accurate manner. Here, the impact of the negations and blind negations are investigated for sentiment analysis.

1.INTRODUCTION

In the past few years, social networks have increased their popularity to become the limelight among the internat users. An average internet user spends more time on social networks such as Facebook, Twitter, Instagram etc. Millions of Internet users use microblogging to talk about their daily activities and to seek or share information. These published information might also include real-time opinions and feelings on certain topics, for example likes or dislikes, positive or negative, love or hate etc.

There are different microblogging platforms among which Twitter has become the most prevalent platform. Since twitter's inception in 2006 it has grown at an unbelievable rate. Tweets can be used to complete tasks like sentiment analysis.

Sentiment Analysis or Opinion Mining analyses people's sentiments, opinions, attitudes, and emotions from available data. In natural language processing this is one of the most active research topics and is also widely studied in data mining, Web mining, and text mining. In fact, this research has spread outside of computer science to management sciences and social sciences due to its importance to business and society as a whole. The growing importance of sentiment analysis coincides with the growth of social media such as reviews, forum discussions, blogs, micro-blogs, Twitter, and social networks.

Companies such as Twitratr (twitrratr.com), tweetfeel (www.tweetfeel.com), Sentiment140 (www.sentiment140.com) and Social Mention (www.socialmention.com) are just a few who advertise Twitter Sentiment Analysis as one of their services.

There are also many names and slightly different tasks, e.g. sentiment analysis, opinion mining, opinion extraction, sentiment mining, subjectivity analysis, affect analysis, emotion analysis, review analysis etc.

Sentiment Analysis applications are being applied in almost every business and social domain because opinions are central to almost all human activities. Our beliefs and perceptions of reality, and the choices we make, are largely conditioned on how others see and evaluate the world. For this reason when we need to make a decision we often seek out the opinions of others. This is true not only for individuals but also for organizations.

Sentiment Analysis is an application of Natural Language Processing, text analysis and computational linguistics for identifying, evaluating, and extracting subjective information in source materials [5]. In other words Sentiment Analysis implies for extracting opinions, emotions and sentiments in text.

Sentiment analysis provides companies with a means to estimate the extent of product acceptance and to determine strategies to improve product quality. It also facilitates policy makers or politicians to analyse public sentiments with respect to policies, public services or political issues [6].

Sentiment Classification technique can be roughly divided into machine learning approach, lexicon based approach and hybrid approach. The Machine Learning Approach (ML), shown in Figure 1, applies the famous ML algorithms and uses linguistic features. The Lexicon-based Approach, shown in Figure 2, relies on a sentiment lexicon, a collection of known and precompiled sentiment terms. The Hybrid Approach combines both the approaches and is very common with sentiment lexicons playing a key role in the majority of methods[15].

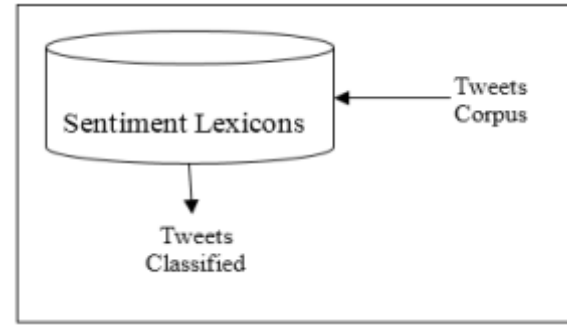

Figure : Lexicon based approach

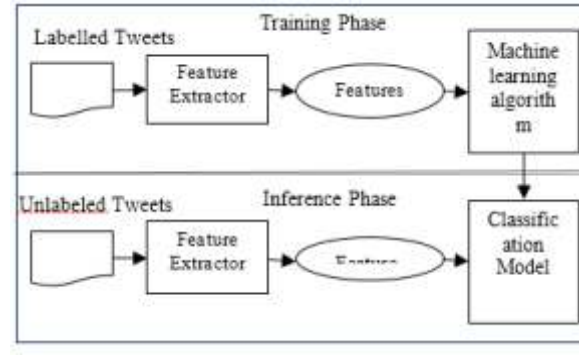

Figure : Machine learning approach

In this research a lexicon based approach is used to perform sentiment analysis.

Machine Learning techniques are not used because although they are more accurate than the lexicon based approaches, they take too much time performing sentiment analysis as they have to be trained first and hence are not efficient in handling big sentiment data.

Liu[1] defined a sentiment as a quintuple "<oj, fjk, soijkl, hi, tl >" where, oj is a target object, fjk is a feature of the object oj, soijkl is the sentiment value of the opinion of the opinion holder hi on feature fjk of object oj at time tl ,soijkl is positive, negative or neutral, or a more granular rating, hi is an opinion holder, tl is the time when the opinion is expressed."

In the past decade, sentiment analysis has become a hot research field. For instance, IBMSPSS [2] provides quantitative sentiment summaries of survey data to assist businesses in understanding consumer attitudes. LexisNexis [3] complies consumer confidence and brand perception summaries using news media, while OpSec [4] also mines user-generated data (social media).

Wall Street has also started to use SA in their trading algorithms with companies like OpFine[9] providing up-to-date sentiment tracking of financial news.

Sentiment Analysis is generally carried out in three steps. First, the subject towards which the sentiment towards which the sentiment is directed is found then, the polarity of the sentiment is calculated and finally the degree of the polarity is assigned with the help of a sentiment score which denotes the intensity of the sentiment.

There are certain limitations while doing Twitter analysis. Firstly, while getting status of user timeline the method can only return a fixed maximum number of tweets which is limited by the twitter API.

Secondly, while requesting tweets for a particular keyword, it sometime happens that the numbers of tweets retrieved are less than the number of requested tweets.

Thirdly, while requesting tweets for a particular keyword, the older tweets cannot be retrieved.

2. RELATED WORK

Sentiment Analysis is well known research topic from many years. The main approach for twitter data analysis focuses on classification of each tweet into positive or negative.
The approaches used so far can be classified into Machine Learning Approach and Lexicon based Approach.

2.1.Machine Learning Approach

Supervised methods are the machine learning approach based on training classifiers, such as Naïve Bayes(NB), Maximum Entropy(MaxEnt), Semi-Supervised Classifier, and Support Vector Machine(SVM).

They work by training algorithms with training data sets before applying it to the actual data set.

Training data are difficult to obtain because of the continuously changing and evolving Twitter data. Aiming to overcome this limitation approaches to automatically generate training data are used but they aren't considered very accurate [8].

Rudy [6] used SVM in its approach. The sentiment classification was done using discriminative classifier. This approach is based on structural risk minimization in which support vectors are used to classify the training sets into different classes based on predefined criteria. Multiclass SVM can also be used for text classification [10].

Rui[7] made a comparative study of the effectiveness of ensemble techniques for sentiment classification. The ensemble framework is applied to sentiment classification tasks, with the aim of efficiently integrating different feature sets and classification algorithms.

Ziqiong Zhang[9] used standard machine learning techniques naive Bayes and SVM into the domain Cantonese-written restaurant reviews to automatically classify user reviews as positive or negative.

2.2.Lexicon Based Approach

In lexicon based method instead of using training data pre built dictionaries of words with associated sentiment orientations is used such as, SentiWordNet.

This method works on an assumption that the collective polarity of tweets is the sum of individual tweets polarity.

Sentiment Analysis for twitter is more challenging because of the problems like, use of short length status message, informal words, negations, emoticons etc.

Presence of Negation words reverses the polarity of a sentence. Taboada[19] performed sentiment analysis while handling negation and intensifying words.

Opinion words are employed in many sentiment classification tasks. Positive Opinion words are used to express some desired states, while negative opinion words are used to express some undesired state.

There are also opinion phrases and idioms which together are called opinion lexicon. There are three main approaches in order to compile or collect the opinion word list[15]. Manual approach being the first, is very time consuming and it is not considered alone. It is usually combined with one of the other two methods. The two automated approaches are Dictionary-based approach and Corpus-based approach.tweet thus classifying into positive, very positive, negative, very negative, and neutral.

Flow Chart of Proposed System:

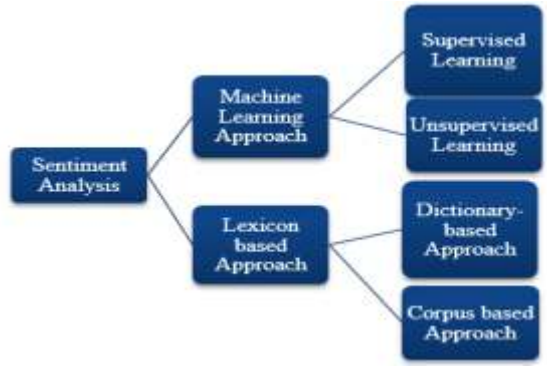

Figure : Sentiment classification techniques

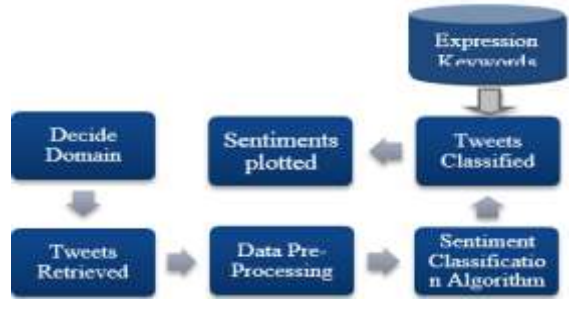

Figure : Flow chart of proposed system

3. PROPOSED APPROACH

The proposed approach in this research focuses on Lexicon based technique. A dictionary of sentiment bearing words with their polarities was used to classify the tweets into positive or negative or neutral sentiment.

3.1.Twitter hashtag(#)

Hashtag is a word which is prefixed by hash sign(#). Use of this symbol indicates that we are directly pointing to a subject hence reducing the need for using complete name.

3.2.Investigating Negation and Blind Negation

Presence of Negation word in a tweet reverses the polarity of a tweet. For example, "This Mobile is not good." Here the word "good" has positive polarity but keyword "not" reverse the polarity to be negative.

If Blind Negation words are present then the tweet is classified into negative polarity. For example, "The performance of couples in Nach Baliye needs to be better."

Default sentiment words are the one with clear positive and negative polarity. For example, "accurate", "bravo", "better", "dynamic" etc refer to words with positive polarity. And words like, "aggressive", "bad", "angry", "cheat" etc refer to negative polarity.

3.3.Evaluation of Sentiment

Sentiment is evaluated on the basis of SentiPoint final score. If the score is equal to 1 then the sentiment of the tweet is positive, if the score is equal to -1 then the sentiment of the tweet so processed is negative or else if SentiPoint remains 0 then neutral i.e. the tweet expresses no sentiment.

Every tweet is processed to give the SentiPoint and be classified into any of the three classes i.e. positive, negative or neutral.

SentiPoint is calculated by using proposed algorithm. The existing approach aggregates the senti score of each

Expression Keywords- This includes the list of positive keywords, negative keywords, negation keywords, and blind negation keywords.

Decide Domain- Decide the domain you would like to process for sentiment analysis. The domain i choose is #AAP i.e. tweets related to Aam Aadmi Party.

Tweets Retrieved- Tweets are retrieved from the micro blogging site Twitter for which a handshaking protocol for authenticity is required.

Data Pre-Processing- Pre-processing is necessary because the tweets are in unstructured form they consists of punctuations, stop words, URLs etc. which needs to be removed significantly.

Sentiment Classification Algorithm- Here the corpus so formed is processed to retrieve the sentiment.

Tweets Classified- The tweets are classified into positive, negative, and neutral sentiment.

Sentiments Plotted- The sentiments so classified are plotted into histogram which shows the difference in the existing approach and the proposed approach.

Algorithm:

Input Data: Pre-processed Twitter Data

a) Removed Punctuations
b) Transformed to Lower case
c) Removed Hyperlink
d) Removed Numbers

Expected Result: Positive, Negative or Neutral

Requirements:

a) List of Positive Senti Words
a) List of Negative Senti Words
b) List of Negation Words
c) List of Blind Negation Words

Algorithm Proposed:

1. SentiPoint = 0
2. If BlindNegation then
3. SentiPoint = -1
4. Else
5. If PositiveWord then
6. SentiPoint = 1
7. Else If NegativeWord then
8. SentiPoint = -1
9. If Negation then
10. SentiPoint *= -1
11. If SentiPoint > 0 then
12. print "Positive"
13. else If SentiPoint < 0 then
14. print "Negative"
15. else print "Neutral"

Lexicon Corpus Details

Number of Positive Words	2006
Number of Negative Words	4783
Number of Negation Words	23
Number of Blind Negation Words	4

1.RESULT AND EVALUATION

The algorithm proposed when implemented gave a better result which as discussed before was my objective. More effective and efficient result was found.

Note: The dataset consisted of 2000 tweets of Aam Admi Party on 25-Apr-2015. It was that period when many went against this party after a farmer committed suicide at Jantar Mantar three days before the tweets were extracted.

The algorithm was implemented over R Studio.

Table 1: Sentiment Count using Existed Method

Sentiment Type (SentiPoint)	Count Value
Positive (>0)	566
Negative (<0)	842
Neutral (0)	592

Same is plotted in the histogram Figure 5. Aggregate of positive and negative sentiment is given.

Table 2 : Sentiment Count using Proposed Method

Sentiment Type (SentiPoint)	Count Value
Positive (1)	645
Negative (-1)	916
Neutral (0)	439

Same is plotted in the histogram Figure 6. We can clearly see here the number of positive and negative tweets has increased and neutral tweets are reduced simultaneously. Which clearly shows that ambiguities present in existing approach is improved at certain level.

Table 3 : Performance of Proposed Method (in seconds)

User Time	System Time	Time Elapsed
1.97	0.05	2.03

Table 4 : Improvisation in existing approach

Improvisation in Positive Tweets	79
Improvisation in Negative Tweets	74
Tweets properly classified	(79+74) = 153

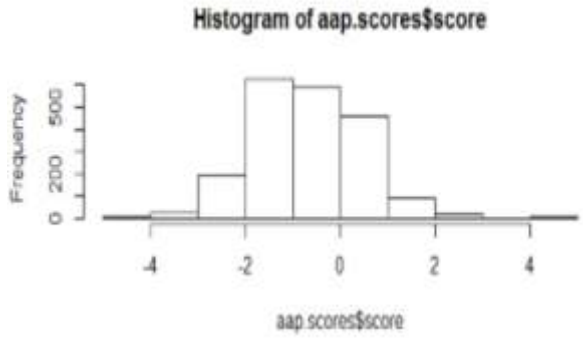

Figure : Histogram created from existing approach

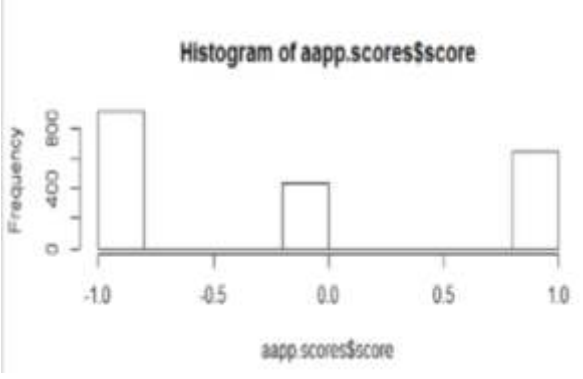

Figure : Histogram from proposed approach

5. CONCLUSION AND FUTURE WORK

The basic task of opinion mining is polarity classification. Polarity classification occurs when a piece of text stating an opinion on a single issue is classified as one of two opposing sentiments.

For classification of such opinions lexicon based method is simple, viable and practical approach where no training sets are required. Investigation of Negation Words and Blind Negation words helped to give more effective result which can be further improvised when factors like emoticons, idioms, phrases etc. will be included.

Considering the stats above I can say that the approach proposed worked really well in terms of both accuracy and speed.

6. ACKNOWLEDGMENT

I am really thankful to my research guide Dr. Suresh Kumar for providing me with necessary support and guidance. I would also like to express my gratitude to various reviewers who will go through my paper.

REFERENCE

[1]http://en.wikipedia.org/wiki/Sentiment_analysis|[2]http://www-01.ibm.com/software/analytics/spss/|[3]http://www.lexisnexis.com/risk/about/data.aspx|[4]http://opsecsecurity.com/ | [5]Liu. Sentiment Analysis and Subjectivity. Handbook of Natural Language Processing, Second Edition, (editors: N. Indurkhya and F. J. Damerau), 2010.|[6]Rudy Prabowo, Mike thelwall, "Sentiment Analysis: A combined approach." Journal of informetrics 3(2009) 143-157.|[7]Rui Xia, Chengquing Zong, Shoushan Li, "Ensemble of feature sets and classification algorithms for sentiment classification", ELSEVIER, Information Sciences 181 (2011) 1138-1152.|[8]Songbo Tan, Jin Zhang, "An empirical study of sentiment analysis for chinese documents", ELSEVIER, Expert Systems with Applications 34 (2008) 2622-2629.|[9]Zinqiong Zhang, Qiang Ye, Zili Zhang, Yijun Li, "Sentiment Classification of Internet restaurant reviews written in Cantonese", Expert Systems with Applications Vol 50 Issue 4 (2011) 743-754.|[10Kaiquan Xu, Stephen Shaoyi Liao, Jiexun Li, Yuxia Song, "Mining comparative opinions from customers reviews for Competitive Intelligence", ELSEVIER, Decision Support Systems 50 (2011) 743-754.|[11]Raisa Varghese, Jayasree M, "Survey on Sentiment Analysis and Opinion Mining", International Journal of Research in Engineering and Technology, Vol. 02 Issue: 11, Nov(2013), 2321-7308.|[12]Hassan Saif, Yulan He, Miriam Fernandez, Harith Alani, "Contextual semantics for sentiment analysis of Twitter", ELSEVIER, Information Processing and Management xxx (2015) xxx-xxx.|[13]Alvaro Ortigosa, Jose M. Martin, Rosa M. Carro, "Sentiment analysis in Facebook and its application to e-learning", ELSEVIER, Computers in Human Behavior 31 (2014) 527-541.|[14]Erik Cambria, Bjorn Schuller, Yunquing Xia, Catherine Havasi, "New Avenues in Opinion Mining and Sentiment Analysis", IEEE, Knowledge-Based Approaches to Concept-Level Sentiment Analysis 13 (2013) 1541-1672.|[15]Walaa Medhat, Ahmed Hassan, Hoda Korashy, "Sentiment analysis algorithms and applications: A survey", ELSEVIER, Ain Shams Engineering Journal 5 (2014) 1093-1113.|[16]Emma Haddi, Xiaohui Liu, Yong Shi, "The Role of Text Pre-processing in Sentiment Analysis", ELSEVIER, Information Technology and Quantitative Management, Procedia Computer Science 17 (2013) 26-32.|[17]Dr. Goutam Chakraborty, Murali Krishna Pagolu, "Analysis of Unstructured Data: Applications of Text Analytics and Sentiment Mining", SAS (2013) 1288-2014.|[18]Alexander Pak, Patrick Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining".|[19]Penchalaiah. C, Murali. G, Suresh Babu. A, "Effective Sentiment Analysis on Twitter Data using: Apache Flume and Hive", International Journal of Innovative Science, Engineering & Technology, Vol 1 Issue 8, October (2014) 2348-7968.