Web Sentiment Analysis for Scoring Positive or Negative Words using Tweeter Data

A Pappu Rajan
Research Scholar, Department of Computer Science
St.Xavier's College, Palaymkottai
Tamil Nadu , India

S.P.Victor
Research Guide ,Department of Computer Science
St.Xavier's College,Palayamkootai
Tamil Nadu , India

ABSTRACT
As people are free to say their opinions on anything using various social networking sites like Twitter, Facebook, Discussion forums, and blogs. Particularly Microblogging and text messaging have emerged and become dominated tool over the web. Twitter messages (tweets) is often used to share opinions and sentiments about the surrounding world. The availability of social content generated on sites such as Twitter creates new opportunities to study public opinion about the entity. This analysis we took twitter data for sentiment classification. The Sentiment analysis is done on a per-Tweet basis. The words in each Tweet are compared with those in other Tweets that have been previously labeled as "positive", or "negative". After looking at these words, the algorithm then judges whether the text in the Tweet is positive or negative based on the likelihood for each possibility. The overall objective of this paper is to determine the sentiment of the text, whether it is positive or negative, which is extended to strength of polarity also this approach is used to obtain the significant features and to analyzing the overall sentiment for each object by computing the weighted average for all the sentiments in the textual data.

Keywords
Web sentiment analysis, Opinion mining

1. INTRODUCTION
In the decision making process each and every piece of information are very important. After arriving internet world user doesn't bother about other opinions from individuals newspaper, surveys, opinion pools, consultants because web analytics introduce new system called opinion mining, which is find out the opinions and experience of other people over the internet using digital social media network like Facebook , reviews, forums, blogs, Twitter, micro-blogs, etc., Indeed, according to surveys about 6 in 10 (60%) online shoppers say user generated customer product reviews have a significant or good impact on their buying behavior.[1][2] Also Data from the 2011 Social Shopping Study indicates that 50% of consumers spend 75% or more of their total shopping time conducting online product research, with 15% spending 90% or more of their shopping time in this manner. Another surveys by Deloitte Consumer Products Group found that almost two-thirds (62%) of consumers read consumer written product reviews online. In fact, a recent study by Deloitte found that "82% of purchase decisions have been directly influenced by reviews". The objective of this paper is to throw lime light on determine the sentiment of the text, whether it is positive or negative, which is extended to strength of polarity. With the explosion of Web 2.0 platforms such as blogs, discussion forums, peer-to-peer networks, and various other types of social media. Consumers have at their disposal a soapbox of unprecedented reach and power by which to share their brand experiences and opinions, positive or negative, regarding any product or service. As major companies are increasingly coming to realize, these consumer voices can wield enormous influence in shaping the opinions of other consumer and, ultimately, their brand loyalties, their purchase decisions, and their own brand advocacy. Companies can respond to the consumer insights they generate through social media monitoring and analysis by modifying their marketing message, brand positing, product development, and other activities accordingly[3][4]

2. SENTIMENT ANALYSIS / OPINION MINING
Sentiment analysis, also called Opinion mining, is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics and their attributes. In general opinion cannot structure a problem but it can subjective and in case opinion gathered from many people it should be summarized.

The notion of an opinion mining is given by [Jin.2006, Liu, 2010]. They put most impact on their work and said that the basic components of an opinion are:

Opinion holder: it is the person that gives a specific opinion on an object.

Object: it is entity on which an opinion is expressed by user.

Opinion: it is a view, sentiment, or appraisal of an object done by user.

There are two types of opinion: Regular and Comparative. Regular opinion is expressions on some target entities, which can be classified into direct and indirect opinion. Comparative opinion is Comparisons of more than one entity.[6][8]

An *opinion* is a quintuple
(e_j, a_{jk}, so_{ijkl}, h_i, t_l),
Where

e_j, is a target entity / Named Entity Extraction

a_{jk} is an aspect/feature of the entity *ej* / Information Extraction

so_{ijkl} is the sentiment value of the opinion from the opinion holder h_i on feature a_{jk} of entity e_j at time t_l. so_{ijkl} is +ve, -ve, or neu, or more granular ratings. / so_{ijkl} is Sentiment identification.

hi is an opinion holder. / information / Data extraction

tl is the time when the opinion is expressed.

3. REVIEW OF LITERATURE

Koweika et al. [Kow, 2013], presented a paper for sentiment Analysis for Social Media . It covered Social media has become one of the biggest forums to express ones opinion. The journal also says that with the data from Twitter, we could classify whether the tweets are positive or negative. Gender prediction and Age prediction can also be done based on the words and slang the people use in their opinions.

Arti Buche [Art ,2013] , presented a paper was Opinion Mining and Analysis: A Survey. It clearly explained that the Sentiment analysis is a type of Natural Language Processing which is used to track products, brands in the Web. It has thus become a necessity for companies to collect data from various sources such as blogs, review sites, Micro-blogs and there-by determining whether they are viewed positively or negatively using part of speech tagging.

Preslav Nakov [Pre, 2013] has introduced Sentiment Analysis in Twitter. Researcher explained Twitter maintains information about who follows who. Retweets and tags inside of tweets provide discourse information. The opinions and reviews collected from Twitter and SMS are classified to sentiment using contextual phrase-level polarity. The sentiments were classified in three ways positive, negative or objective[9][11]

Jisha Manjaly [Jis, 2013] has proposed a new system in Twitter based Sentiment Analysis for Subject identification. It discussed about Social media such as blogs, Twitter, Facebook are widely used for participatory information sharing and collaboration. The opinions are then classified to positive or negative and neutral which is further divided into the emotional states such as sad, happy and angry

Sindhu C [Sid ,2013] , presented a Survey on Opinion Mining and Sentiment Polarity Classification**.** Sentiment analysis refers to computational techniques for analyzing the opinions that are extracted from various sources like the blog posts, comments on forums, reviews about products, policies or any topic on social networking sites or tweets. The process of selecting the opinionated sentences and ignoring the factual sentence is called Subjectivity Detection which is then preprocessed by tokenizing, stop words filtering and stemming. [21]. The expressed opinion in a sentence is classified into six emotion as: Anger, disgust, fear, happiness, sadness and surprise.

4. TASK OF OPINION MINING

4.1 Motivation and Objective

From the review of literature, a number of approaches are used to identify the significant features of opinion mining and to determine the sentiment of the text, whether it is positive or negative, which is extended to strength of polarity. The aim of this approach is used to obtain the significant features and to analyzing the overall sentiment for each object by computing the weighted average for all the sentiments in the textual data.

4.2 Data Set

Tweets are collected using R tool from following five companies. In this work, tweets about specific company are used as the hash tags (e.g.: #TCS). The companies are chosen in such way there are more people talking about it in twitter. All the companies are listed in NSE. The following are the companies according to verticals:

Table 1.0 Lists of Companies

Company Name	**Type of Industry**
Bharti Airtel Ltd.	Telecommunication - Services
Titan Industries Ltd.	Retail
Bosch Ltd.	Automobiles
Tata Consultancy Services Ltd.	Computers – Software
Colgate Palmolive (India) Ltd.	FMCG

The Sampling technique adopted for this project is Topic-based sampling based on sentence level opinion mining, since we have collected textual data regarding specific hash tags. For one day, about 2290 tweets were analyzed. So, for 50 days we obtained a sample size of 114,500 tweets for the entire project.

4.3 Steps for Sentiment analysis

In opinion mining are various types of sentiment analysis as: word level , feature-level, entity-level, sentence-level, document-level ,. Data set are collected from different Twitter by web crawling, in this step will be explained very clearly below paper. Extracting data from SMN (Twitter) : using Twitter API REST API s are having the following resources : Time lines , tweets search, streaming , direct message, friends and followers ,users ,suggested users ,favorites , lists, saved searchers, place and Geo , trends , spam reports, OAuth, help. These APIs use the pull strategy for data retrieval. To collect information a user must explicitly request it. Streaming APIs provides a continuous stream of public information from Twitter. These APIs use the push strategy for data retrieval. Once a request for information is made, the Streaming APIs provide a continuous stream of updates with no further input from the user. Opinion Retrieval involves retrieving desired information from bag-of-words or Twitter textual data to measure ad hoc information retrieval effectiveness in the standard way, we need a test collection consisting of three things: 1) A document collection. 2) A test suite of information needs, expressible as queries or tags 3) A set of relevance judgments, standard a binary assessment of either relevant or non-relevant for each query-document pair. Sentiment Extraction: Finding or discovering of target entity. It uses various method to extract the sentiment from sentiment document using unsupervised learning, supervised learning and lexicon based approach. Sentiment Classification: Positive / Negative -Score Analysis: To find weather a piece of text is opinionated or not and to find the polarity of the text. This classification may be binary or multiclass classification.

4.5 Building a Lexical Sentiment Analysis for Scoring Positive or Negative word Analysis

To get around the potential; issue of having an unsuitable lexicon, we constructed lexicon automatically for each dataset. Because we are using data collected directly from Twitter, we do not have explicit positive or negative labels.

4.6 Steps for performing lexical sentiment analysis

Step 1. Read opinion data set

Step 2. Clean up the opinion, to remove noise data

Step 3 Split whole opinion based sentence into opinion word

Step 4. Find the number of positive and negative opinion

Step 5. Compare number of positive and negative opinion from multi set

Step 6. Score opinion: Number of positive words – number of negative words

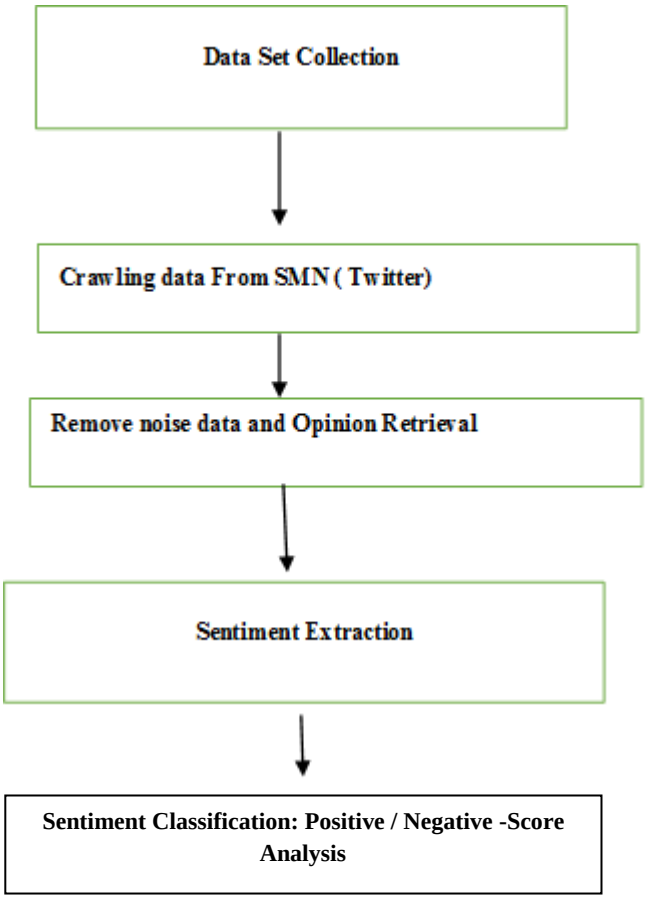

Fig.1.0 Process of Sentiment Analysis

4.7 Bag of Words

The main process of lexical sentiment analysis is to compare the Tweets crawled from Twitter with the bag of words (BOW) containing positive and negative words.[12] [15] Some of the words in BOW are given in the table:

Table 2.0 Few List of Positive and Negative words from the Bag of Words

Positive Words	Negative Word
Amazing	Annoying
Beautiful	Cheating
Happy	Bad
Gorgeous	Impolitely
Ecstatic	Hideous
Fantastic	Accursed
Pleasant	Overblown
Marvelous	Perplexing

The Tweets are compared with the bag of words as above and classified as positive or negative. The strength of polarity (i.e.) very positive, positive, slightly positive is determined by the frequency of positive or negative words repeated in a single tweet

Some of the Tweets and its classification are given below:

Table3.0 Sentiment Analysis Data

#TCS focuses on making its customers STRONGER while designing #JDEdwards solutions - more on its offerings here http://t.co/b66XTqt8CD	Positive
Highly impressed with the way #TCS has digitised Indian Passport procedures. They've officially put an end to agents/bribes/queues/confusion.	Positive
Instantly block access to swipe cards. Don't do like this #TCS	Negative
So happy to be placed in #TCS. Whenever some1 asks wat r u doing i just say tcs n rest all is explained :)	Positive

The chart below represents the sentiment analysis on November 1st, 2013 .

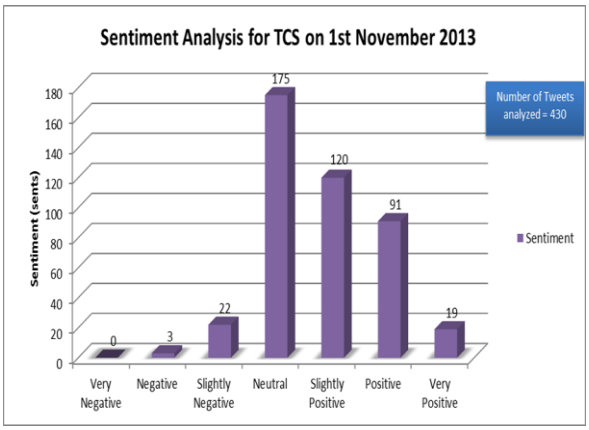

Fig 2.0 Sentiment Analysis Chart

5. SCORING SENTIMENTS

Overall negative = (0*-3) + (3*-2)+(22*-1)=28

Overall positive = (175*0) + (120*1)+(91*2)+(19*3)=359

Overall Sentiment = 359-28=331

The Overall sentiment for TCS is seems to be very positive, since the bar char is skewed towards right (i.e). Positive Side. There are no very negative tweets, very few negative and slightly negative tweets, which shows that buzz about TCS is more positive in nature.

5.2 Overall opinion analysis

The Fig 3.0 chart represents the trends of all the five companies (i.e.) Airtel, Tata Consultancy Services, Titan Industries, Colgate Palmolive and Bosch by using the overall sentiment in this research

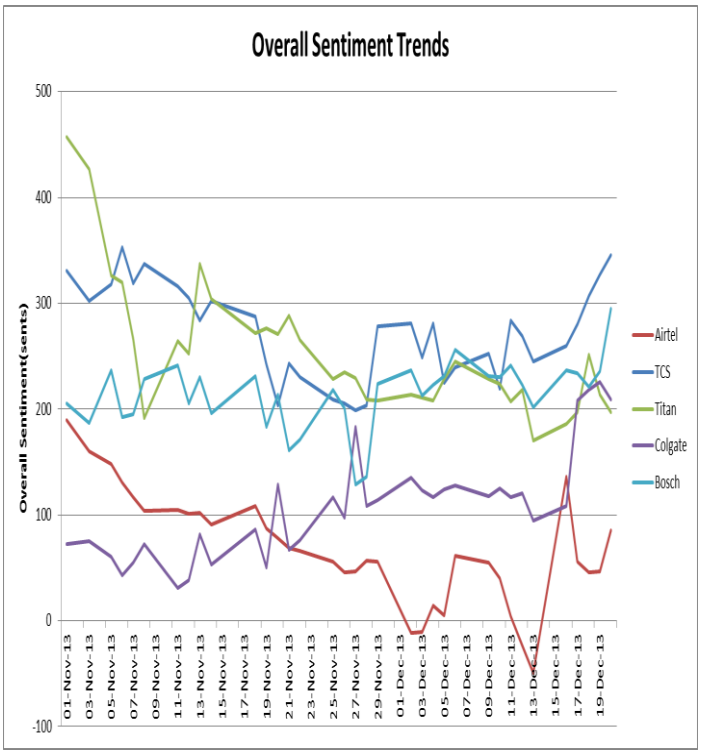

Fig.3 Overall Sentiment Flow

From the above trend of overall sentiment of various companies, where the overall sentiment for Bharti Airtel went down to negative on 13th of December, 2013 and it is due to more tweets relating to bad customer service. Comparatively, overall sentiment for TCS is more positive than all the other companies where the sentiment is nearly above 320. The overall sentiment for Bosch is rather stable and it has increased nearly to 300 on the last day of analysis. On the other hand, Titan industries had a very positive buzz on 1st November, 2013, which declined to around overall sentiment of 200 on 19th December, 2013. While, the overall sentiment for Colgate Palmolive was very low on 1st November, 2013 but at the later stage the sentiment has increased drastically above 205 on 19th December, 2013.

5.3 Overall Sentiment 7 point scale data

The lexical sentiment analysis was performed over five companies. We have analyzed about 2290 tweets per day. So, for 50 days from 1st November, 2013 to 20th December, 2013, a total of 114,500 tweets were analyzed for the entire project. The following table represents the polarity of sentiment in very negative to very positive scale (i.e) a 7 point scale.

Note : A : Very Negative B:Negative C: Slightly Negative D: Neutral E : Slightly Positive F :Positive G : Very Positive H: Overall Sentiment % : Percentage

Table 5 Sample Overall Opinion 7 point data

Date	A	B	C	D	E	F	G	H	%
01-Nov-13	0	3	22	175	120	91	19	331	76.98
03-Nov-13	1	7	25	172	123	85	17	302	70.23
------	--	--	--	--	----	---	---	---	---
20-Dec-13	0	6	10	174	125	102	13	346	80.47

6. CONCLUSIONS

This research introduce the theoretical basic of opinion mining. The proposed approach determines the sentiment of the text, whether it is positive or negative, which is extended to strength of polarity and also which was obtain the significant features and to Analyzing the overall sentiment for each object by computing the weighted average for all the sentiments in the textual data. For further research the Stock prices of above mentioned companies are collected from the official website of National Stock Exchange (NSE) for the same period , so Comparing the overall sentiment of each object with its Stock Prices and Comparing the predicted results of Closing prices using ALM with the values predicted using Artificial Neural Networks. There still remain many areas for further research, such as the design of efficient algorithms for opining mining from the positive and negative sentiment result.

7. REFERENCES

[1] Curator. S (2013, September 23).Social media: a rich source of customer sentiment for you to mine. Whatech Channel.

[2] Koweika A.,Gupta A.,Sondhi K.(2013).Sentiment analysis for social media. International Journal of Advanced Research in Computer Science and Software Engineering

[3] Bissattini C., Christodoulou K.(2013).Web sentiment analysis for revealing public opinions, trends and making good financial decisions. Journal of Advanced Research in Computer Science and Software Engineering

[4] Tulankar S.,Athale R.,Bhujbal S (2013). Sentiment analysis of equities using data mining techniques and visualizing the trends. International Journal of Computer Science Issues..

[5] Qiu M.,Yang L., Jiang J. (2013). Mining user relations from online discussions using sentiment analysis and probabilistic matrix factorization. Proceedings of NAACL-HLT, Atlanta, Georgia.

[6] Buche A., Chandak M.B., Zadgaonkar A.(2013).Opinion mining and analysis: a survey. International Journal on Natural Language Computing(IJNLC).39-48.

[7] Cataldi M., Ballatore A., Ilaria T. (2013.Good location, terrible food: Detecting feature sentiment in user-generated reviews. International Journal of Social Network Analysis and Mining (SNAM).

[8] Manjaly J.S. (2013).Twitter based sentiment analysis for subject identification. International Journal of Advanced Research in Computer and Communication Engineering

[9] Zuell B., Preradovic N.M., (2013).Methods and usage of sentiment analysis in the context of the TV industry. International Journal of Recent Advances in Information Science..

[10] Jagtap V.S., Pawar K. (2013).Analysis of different approaches to sentence-level sentiment classification. International Journal of Scientific Engineering and Technology. 164-170. Mullen T. and Malouf R.,Taking sides: User classification for informal online political discourse. Internet Research, 2008.

[11] Nakov P.,Kozareva Z., Ritter A.(2013),"Sentiment analysis in Twitter. Second Joint Conference on Lexical and Computational Semantics". Atlanta, Georgia.

[12] Younggue B.,Hongchul L.(2012),"Sentiment analysis of Twitter audience: Measuring the positive or negative influence", Journal of the American Society for Information Science and Technology.

[13] Savage N. (2011, March)," Twitter as medium and message", Communications of the ACM *Society*, Issue: 3, Vol.54. pp: 18-20.

[14] ZHU Jian, XU Chen, and WANG Han-shi, 2010. Sentiment classification using the theory of ANNs, The Journal of China Universities of Posts and Telecommunications, 17(Suppl.): pp. 58–62.

[15] X. Ding, B. Liu, and P. S. Yu, 2008. A holistic lexicon-based approach to opinion mining, Proceedings of the Conference on Web Search and Web Data Mining (WSDM).

[16] A.Pappu Rajan ,(2013)," A Study on Security Threat Awareness among Students Using Social Networking Sites, by Applying Data Mining Techniques", International Journal Of Research In Commerce, IT & Management, Vol. No. 3 , Issue No. 09 : ISSN 2231-5756

[17] Kwak, H., Lee, C., Park, H. &Moon, S. What is Twitter, a social network or a news media? Proceedings of the 19th international conference on World Wide Web, 591–600 (2010).

[18] Changhua Yang, Kevin Hsin-Yih Lin, and Hsin-Hsi Chen.2007. Emotion classification using web blog corpora.In WI '07: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, pages 275–278,Washington, DC, USA. IEEE Computer Society.

[19] B. Liu. Web Data Mining: Exploring Hyperlinks, Contents and Usage Data. Springer, 2007

[20] Dongcheol,K.and Soon---Ho,K.(2012)Investor Sentiment from Internet Message Postings and Predictability of Stock Returns. Working paper, Korea University Business School, Seoul 136---701, pp. 2---53.

[21] Kumar,A.and Teeja, M.S. (2012) Sentiment Analysis on Twitter. IJCSI International Journal of Computer Science Issues, Vol.9, Issue 4, No 3, pp. 372---373.