ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 9, May 2016

Opinion Mining on Twitter Data

Sangeetha Suresh Harikantra [1], Roshan Fernandes [2]

PG Student, Department of Computer Science, NMAMIT College, Nitte, Udupi District, Karnakata, India[1]

Associate Professor, Department of Computer Science, NMAMIT College, Nitte, Udupi District, Karnakata, India [2]

ABSTRACT: In today's life most of the individuals are attracted toward social networking site. Many social networking microblog are emerging like Facebook and Twitter, where users express their experiences and emotions. We find millions of tweet per day and considering this as advantage, we focus on finding opinion about product. Opinion mining provides polarity of the tweets collected and provides overall result of product. It can either give positive, negative or neutral opinion about the product. Many researchers are attracted towards this field, to provide effective decision on product, political issues or person. Our system uses machine learning technique and automatically provides the opinion on particular product.

KEYWORDS: Opinion Mining, sentiment analysis, Twitter, product.

I. INTRODUCTION

In recent years all most every individual use any one of the social networking microblog. Social networking site like Facebook and Twitter has facility to share their experiences and emotion with their friends, relatives and colleges. There are millions of users using these site and millions of tweets will be generated per day [1]. Social networking sites are resource of information where issues, information or feedbacks are shared. Our focus is to mine the required information from this resource. Opinion mining tools helps in find the emotions expressed in the tweets and helps in making decision. Product reviews has become one of most important topic. Customers express their shopping experiences and product feature in social networking sites. Most of the user tweet based on product so millions of tweets can be collected on particular product. If any individual wishes to purchase a product, the sentiment analysis tools will be helpful in providing appropriate decision on product based on the tweets.

We have focused on Twitter, the most popular networking site. Twitter has millions of users of different region. Huge number of tweets can be collected during the data-collection process and also twitter is unambiguous than compare to other networking site. Twitter networking site has user all over the world, so customers of different region and different habituality tweet. Varity of tweets with different experience can help sentiment analysis tool to provide the accurate result. Our focus is to provide the accurate result on product reviews. It has been observed that nearly 87% of customers make decision based on internet reviews [2].Sentiment analysis tools provides assistance to customer with no time and less effort. It is difficult for customer to search and read every tweet and then make decision on particular product. Sentiment analysis will analyse each and every tweet and provide the overall result on product. It analysis provides positive, negative or neutral opinion based on the tweets obtained.

Sentiment analysis is useful in both user point of view as well as business point of view. Suppose any individual wishes to purchase a product, opinion tool will help the customer to provide the percentage of positive and negative reviews. In business, the system will give the feedback to the organisation to improve the product and help in improvement of product quality. If the opinion given by the system is good, then the advertisement of product is done and help in productivity of the product. If opinion is bad, it may affect reversely on the productivity and reduce the business.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 9, May 2016

II. RELATED WORKS

There are previously many tools developed in this area. Some of the tools have been analysed as below. Authors Thelwall et al. explained on the strength detection on informal text and explained about the tool called SentiStrength. SentiStrength scales from 1 to 5 for both positive (+1 weak positive to +5 extreme positive) and negative (-1 weak negative to -5 extreme negative) sentiment. It provided accuracy level of 60.6% for positive emotion and 72.8% for negative emotions. It had some limitation like, phrase identification was for only few frequent examples and ambiguous words caused problems to determine polarity. Several experiments were conducted like SentiStrength was compared with standard machine learning classification algorithms in Weka, using frequencies of word list. Also searched for the alternative smaller feature sets to provide better results, feature reduction. Also variation of SentiStrength with other algorithm was observed [3].

Author Linus Philip Lawrence compares two sentiment tools, Semantria and Social Mention. Semantria is owned by company lexalytica, provides text analysis via AI and Excel plugin. Semantria is document level sentiment analysis tool, where any information collected from any database can be given and author has considered the Twitter data. Social Mention can collect information from more than 80 social networking sites. Analysis was done based on 12 different car models and metrics. Semantria uses the algorithm which identifies sentiment phrases and then scale sentiment values from +2.0 to -2.0 to determine polarity [4].

According to Akshat Bakliwal et al. the experiment was conducted with corpus and determined the sentiment score using sentiment bearing phrases. Using Stanford Dataset 87% was achieved; using Mejaj Dataset they achieved 88% and using supervised machine learning approach with Stanford Dataset they achieved 88% accuracy [5]. Patricia L V Ribeiro et al. have used many algorithm for better accuracy, to collect sufficient amount of tweets they used Hashtag Expansion algorithm, to remove noisy tweet they have Spam detection algorithm. Also Graph builder algorithm to create graph with meaningful node relationship. Graph Propagation algorithm and Lexicon Building algorithm are constructed which also includes negation and elongation [6].

Tiara et al. have built a system that gives opinion about the television program, with lexicon based method that uses the dictionary to score each and every word and entity also handles the negation in each tweet [7]. According to the author Xujuan Zhou et al. speak about the providing the opinion on particular event and named the tool TSAM(Tweets Sentiment Analysis Model). They have concentrated on opinion bearing words than using the bag of words method; also they have used the POS tagging and stemming method. Stemming attributes will help to indicate lookup must be performed lemmas or token. Prior polarity is calculated and intensive lexicon with most expensive words is used, valence shifter lexicon entities shift the polarity of an existing opinion towards negative or positive including negation words [8].

Author Seyed-Ali Bahrainian and Andreas Denge, in the feature selection process has used the slang dictionary. SentiStrength lexicon to tag sentiment-bearing words also tags the 'NEGATE' for negation words, and classified into intensifiers, diminishers and negations. It includes sixteen features which indicated the target and its neighbouring words sentiment. Using SVM classifier as baseline conducted comparison with novel hybrid method against the unsupervised method against defined baseline, showing overall accuracy of their hybrid method was efficient [9]. The authors Aliza Sarlan, use both lexicon based method to find the polarity of each tweet and machine learning based approach. They had used part of speech tagger to each tweet post, collected all adjective, made set of top adjectives, then moved the tweet to experimental set including negation words, presence, absence or frequency of each word. Alchemy API and python was used for building the system [10].

III. METHODOLOGY

This section explains about the process carried out in our opinion mining tools. Opinion mining basically follows some steps that are data collection, preprocessing, data extraction and providing the result. Our system determines the

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 9, May 2016

opinion on a product and we have to consider any one product name (like MicroMax mobiles). Design of the system is shown below in figure1.

Data collection: Data collection is a process where the Twitter data is collected. Twitter has huge data so we must be able handle the data. So, in our method we collect tweets based on particular entity. With the help of the Twitter API we search the tweets of product and store it in the JSON format, so that we may easy handle the data which is collected. To access the tweet an account must be created and customer secret key must be obtained. The customer key can help in retrieving the tweets but only the public tweets of product are retrieved. The tweets obtained are in JSON format, so JSON Simple library is used to parse the file. All the tweets with the language English is collected.

Data preprocessing: Data preprocessing is most important process due to lot of irrelevant data present in the data which is collected. Preprocessing involves in replacing of emoticons, handling URL, hashtags, white space, identifying punctuations and lowercase conversion.

Replacing emoticons: Users usually use emoticons to express their emotions. And emoticon express more than words, so all these emoticons must be replaced with appropriate word [13].

URL and Hashtags: Most of the users use URL and hashtags to express more about the topic and some even share information. This is due to the limitation of 140 characters in tweet. But the URL and hashtags may contain more related information about the product. So our system must be able use the links.

Lower case: User tweet may contain upper case and lower case, some words which are not properly used (words like TwITTer DaTa) and it is difficult to find out the meaning. In such cases the all the words have to be converted to lower case to handle the tweets.

Identifying punctuations: Punctuations and white space must be identified and removed to avoid redundant features.

Data Extraction: There are different opinion mining techniques, like document level, sentence level and phrases level. Our system is based on the sentence level. The feature extraction involves following steps.

Tokenisation: Once the data is preprocessed our system will split the sentence into tokens or words. Tokenisation helps reducing complexity in future process.

POS tagger: Next process is to tag the words parts of speech to ever word using NLP. POS tagging helps in linguistic categorising the words in sentence. Our method uses the Stanford POS tagger to tag the word. POS tagger breaks the sentence grammatically and identifies the noun, pronoun, adjective, adverb, verbs etc.

Stopwords: There are some of the commonly used words like “is, a, the” which does not carry any meaning for sentiment analysis are stopwords. The tagged sentence will contain stopwords data and all the stopwords must be removed.

In feature selection process we make our system more efficient and reduce the size of the data. The system identifies the required feature for later use in classification process. Adjectives are most expensive feature to determining the sentiment. Adverb doesn’t have any polarity but in combination with adjective will provide the most important feature for sentiment. So, adverb related to adjectives must also be considered. Some words such as no, not, never changes the polarity of the tweet, these negation words must be identified and invert the polarity.

SentiWordNet: Corpus Statistics or SentiWordNet can be used to identify the weight of features. Our system uses SentiWordNet dictionary, which is lexicon resource for most of the opinion mining tools. SentiwordNet mainly has subjective and objective values of synset, where subjective has positive and negative values. We have determined the score for each tweet such that the value lies between -1 to +1 including negation features. These score will indicate

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 9, May 2016

whether the tweet has positive or negative polarity. Based on score tweets are categorized into strong negative, negative, weak negative, neutral, weak positive, positive and strong positive [14].

Classification and Evaluation: In classification, we aim to classify the data according to predefined classes. Any classifier algorithm can be used for the classification like nearest neighbour, naive Bayes, maximum entropy and Support vector Machine (SVM) [11].

SVM classifier is used in the system as it provides better accuracy compared to others and SVM are universal learners [12]. SVM has different form linear and non-linear and is supervised learning. SVM finds the hyperplane that separates the positive and the negative class. Then the evaluation is done based on the classification and presented in pie chart form. Pie chart represents the percentage of the positive, negative and neutral tweet.

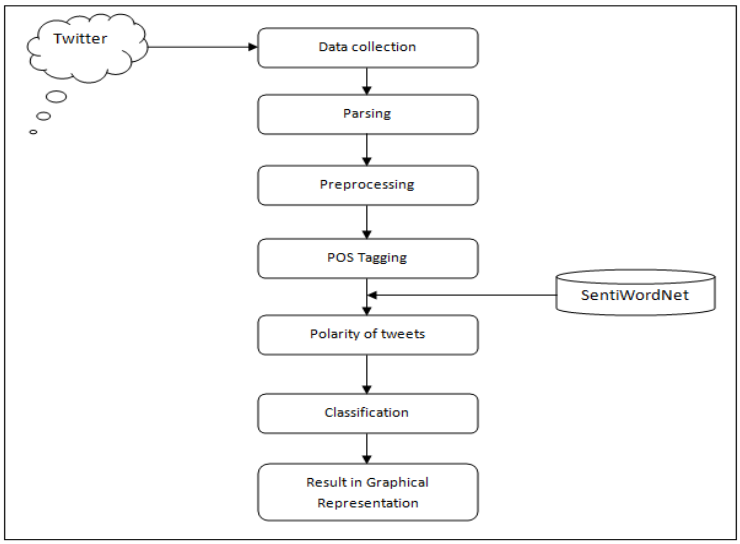

Figure 1: System Design

IV. RESULTS AND DISCUSSION

To retrieve Twitter data developer has to agree the conditions of Twitter developing platform. The tweets or data collected is in the JSON format, which is easy to manage and also language independent. Any programming language can be used for development, our system uses JAVA. Tweets from the JSON file are tagged using Stanford POS tagger. Using Lexicon based approach; values are determined for each tweet and obtained the polarity of tweets. The system then calculates the percentage of positive and negative tweets.

V. CONCLUSIONS AND FUTURE WORK

Twitter data is collected and analysed to develop a system which make good use of resource in field of marketing. The system determines the percentage of positive and negative opinions of customers on a particular product. The result in form of chart will help both the organisation as well as customer, to know about the product. For better results, the system can include database for unstructured data. It must be able to proved more accuracy for different types of sentence, including slang language, comparative and sarcastic sentence.

REFERENCES

[1] Rambocas M, Gama J, "Marketing research: The role of sentiment analysis." Universidade do Porto, Faculdade de Economia do Porto; 2013.
[2] A.K.Jose, N.Bhatia, and S.Krishna, "Twitter Sentiment Analysis", National Institute of Technology Calicut, 2010.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 9, May 2016

[3] Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, "A Sentiment strength detection in short text", Journal of the American Society for Information Science and Technology, 61(12), 2544–2558, 2010
[4] Linus Philip Lawrence "Reliability of Sentiment Mining Tools : A Comparison of Semantria and Social Mention", University of Twente Enschede The Netherlands, 2014.
[5] Akshat Bakliwal, Piyush Arora, Senthil Madhappan Nikhil Kapre, Mukesh Singh and Vasudeva Varma, "Mining Sentiments from Tweets", Association for Computational Linguistics, Proceedings of the 3rd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, pages 11–18, Jeju, Republic of Korea, 12 July 2012 .
[6] Patricia L V Ribeiro, Li Weigang and Tiancheng Li, "A Unified Approach for Domain-Specific Tweet Sentiment Analysis", 18th International Conference on Information Fusion Washington, DC - July 6-9, FUSION, 2015.
[7] Tiara, Mira Kania Sabariah, Veronikha Effendy, "Sentiment Analysis on Twitter Using the Combination of Lexicon-Based and Support Vector Machine for Assessing the Performance of a Television Program", 3rd International Conference on Information and Communication Technology (ICoICT), 2015.
[8] Xujuan Zhou, Xiaohui Tao, Jianming Yong, Zhenyu Yang , "Sentiment Analysis on Tweets for Social Events", Proceedings of the IEEE 17th International Conference on Computer Supported Cooperative Work in Design, 2013.
[9] Seyed-Ali Bahrainian and Andreas Denge, "Sentiment Analysis and Summarization of Twitter Data "IEEE 16th International Conference on Computational Science and Engineering, 2013.
[10] Aliza Sarlan, Chayanit Nadam, Shuib Basri, "Twitter Sentiment Analysis", International Conference on Information Technology and Multimedia (ICIMU), Putrajaya, Malaysia, November 18 – 20, 2014.
[11] Dumais, Susan, et al. "Inductive learning algorithms and representations for text categorization". Proceedings of the seventh international conference on Information and knowledge management. ACM, 1998.
[12] Thorsten Joachims " Text categorization with support vector machines: learningwith many relevant features", Proc. of ECML-98,
10th European Conference on Machine Learning, Springer Verlag, Heidelberg, DE,pp. 137-142,1998.
[13] Diego Terrana, Agnese Augello, Giovanni Pilato, "Automatic Unsupervised Polarity Detection on a Twitter Data Stream", IEEE International Conference on Semantic Computing, 2014.
[14] Eun Hee Ko, Diego Klabjan, "Semantic Properties of Customer Sentiment in Tweets", 28th International Conference on Advanced Information Networking and Applications Workshops, 2014.