International Journal of Advanced Research in IT and Engineering

ISSN: 2278-6244
Impact Factor: 6.111

Sentiment Analysis : A Survey

Sangeetha Suresh Harikantra, Dept. of Computer Science and Engineering , NMAM Institute of Technology, Nitte, Udupi district, India

Roshan Fernandes, Dept. of Computer science and Engineering, NMAM Institute of Technology, Nitte, Udupi district, India

Abstract— Recently many networking sites are emerging and also social networking sites have become a part of life. In some of the social networking sites such as Facebook and Twitter, users express their emotion, share some information with friends and also express their views about the product. We have focused on the sentiment analysis of users' views on product. The trend is to identify the opinion of different individuals around the world using Twitter micro blog and provide positive or negative opinion about the product. There will be millions of reviews on single product and it would be impossible for the customer or the organization to read each reviews and judge the quality of the product .Opinion mining tools helps both the customer as well as company; customer can judge the product and company can improve their product by getting feedback. We have done literature survey on various sentiment analysis tools and their behavior with other works done on the subject.

Keywords— Opinion Mining, sentiment analysis, Twitter, product (key words)

I. INTRODUCTION

Social networking site like Facebook and Twitter are used widely throughout the world. Micro blogs provides opportunity to the user to express and share the information. Millions of users share opinion or feedback on different aspect of life every day [1]. The micro blogging websites are rich in data and can consider as source for opinion mining. Twitter is one of the most popular micro blog, in which we can discover what people really think. Many researchers are interested in this area, due do huge source of data obtained in this field. The trend is to identify the opinion of variety of individuals around the globe using micro blog.

Social networking sites has provided platform for the customer to comment on products. Micro blogs connect customer and business organization. The information provided by the customer will help the organization to keep track of their product and services. Also provides the feedback to organization which will lead for the improvement of product and services. Micro blogs also provide the publicity of product with success or failure of company. In the customer point of view, Micro blogs helps in making decision on product. Around 87% of internet users make decision to purchase a product, influenced by micro blogs [2].

Twitter is a micro blog which contains tweets. Millions of users use Twitter to express their views. Each day we find millions of tweets on a product. Manually reading all the tweets and making decision is impractical. Tools have been developed for decision making, which provide positive, negative and neutral opinion on the product. Sentiment is person's positive, negative or neutral feeling. Sentiment analysis on tweet will involves collection of data, extraction, classification, understanding and providing the opinion that are expressed in various tweets.

Twitter data is huge data and may be unstructured or structured form, so it may be difficult to handle data. Sentiment analysis deals with Natural Language Processing(NLP) and also text classification. Text classification using machine learning approaches are Naive Bayes (NB), Maximum Entropy (ME), and Support Vector Machines (SVM) [3]. As we know 140 characters is limitation of tweet, long messages have become in short forms (slang language) [4]. Emoticons are pictorial representation of human facial expressions to express their feeling are also used in tweets. Some tweets may be sarcastic and difficult to understand. Parts of speech tagging (POST) cannot identify such particular piece of sentence.

Sentiment analysis objective is to provide positive, negative or neutral opinion about a product in this paper. Sentiment analysis includes collection of raw Twitter data, removing the irrelevant data, parsing the data, extracting the required tweets. Using appropriate classifier and POS tagger and get the feature words. Based on these feature and some combination of words it will classify into three categories positive, negative and neutral and then find out the overall percentage of positive, negative and neural tweets.

In the customer point of view opinion mining on product helps in collecting the opinion of the product. Since Twitter has users of different region, different field, tweets collected will be varieties. Users of different regions may have different opinion on the product, in the same way with users in different field may give different opinion. So varieties of tweets can be analysed and tweets of users with different mentality can also be analysed. So, various types of tweets can give better result in collecting the opinion. Opinion mining tools helps the customer to give positive, negative or neutral opinion about the product. The result shows the percentage of positive number of tweets,

negative number of tweets and neutral number of tweets collected from the database of Twitter.

Users of the product will share their experience and views on Twitter. Millions of users tweet and millions of tweets on particular product can be retrieved. So, in business point of view opinion mining tools can help in advertising their product. If opinion mining tools provides good opinion, then it helps in progress of company else it may affect negatively on business.

Another application of opinion mining tools is, organization can analyse the product review without any survey. Manual survey on product will be very expensive and also requires man power; the opinion mining tools will conduct this survey online and at very low cost. Customers tweet their opinion about their product and also the problems faced by them. Organization can take these tweets as feedback and improve their product. Hence tool will applicable in improving the product and also improve their productivity.

II. SENTIMENT ANALYSIS

Tweets collected from Twitter must be classified according to the polarity of feature words. There are several different opinion mining tools based on different domain. General steps involved are data collection, data pre-processing, data extraction and result representation.

Data collection: Data collection involves in collection of tweets. Twitter data are private and public. Only public tweets can be accessed and private tweets cannot be accessed. Millions of tweets will be tweeted per day, so huge data must be managed.

Data Pre-processing: In this process only the relevant data are collected and all irrelevant data are removed. Irrelevant data removal involves like replacing of emoticons, uppercase/lowercase identification, URL extraction, removal of stop words, identification of punctuations [8] and slang language replacement.

Data extraction: Here tweets with opinion are helpful; opinion from the tweets is filtered out. We can use any classifier technique for this like Naive Bayes [9]. Then classify tweets according to the polarity as positive, negative or neutral. There can be more than one classifier in this process.

Presenting result: Positive, negative or neutral opinion obtained is then represented in pictorial representation.

There are many problems faced during sentiment classification. In Opinion mining process there must be sufficient data to analyze the opinion about a product, political issues, person or any other aspect based mining. In the data collection process there are many problems faced such as all the data in the tweet are not public. Private data as well as public data are present. Only the public tweets can be accessed and only the owner account can retrieve the private tweets. There are many methods to retrieve the tweets from Twitter. Twitter is social network which does not have the anomalies in order of tweets [7]. Millions of tweets are generated per day; managing this large number of tweets is one of the big tasks. All unwanted data must be removed in the pre-processing steps include slang language, emoticons and other abbreviations. As 140 characters is maximum limit of the tweet, slang language is often used; several different methods to overcome this problem are discussed below. Data on Twitter may be structured or unstructured and appropriate classifier must be used according to the data retrieved.

III. LITERATURE REVIEW

In this survey data sets are taken for the literature review process. Thelwall, Buckley, Paltoglou, Cai, & Kappas, explains that, SentiStrength is a sentiment analysis stand alone tool for short and even informal language also lightweight. It has human level accuracy for short English text. SentiStrength for social media have extended with mood setting and lexicon extension for specific domain like product based tweets. Mood setting includes like repeated letters, emoticons, and exclamation mark. SentiStrength accuracy level is 60.6% for positive emotion and 72.8% for negative emotions [6]. According to Renata Lopes, Demostenes Zegarra Rodriguez and Graca Bressan, SentiMeter-BR is a tool which analyses customer sentiment of Brazil based on particular domain. SentiMeter-BR provided better performance than the SentiStrength tool. Twitter helped them to build the dictionary of this tool. Same as SentiStrength it uses numerical representation to indicate the polarity. Negative indicates from -1 to-5, positive from +1 to +5, emoticons from -1 to+1, slang and strong words from -5 to +5. Separate files are maintained for slang, negative adjective, positive adjective, negative word, and positive word [5].

According to Patricia L V Ribeiro, Li Weigang and Tiancheng Li the opinion mining tool for Twitter has expanded hashtag algorithm and spam detection algorithm. Hashtag algorithm help in collecting many tweets and good result spam detection algorithm uses to analyse spam tweets. Sentiment analysis algorithm uses the supervised learning without the training sets; it only requires lexicon and a set of unlabeled tweets [10]. Alec Go, Richa Bhayani, Lei Huang from Stanford University using Distant Supervision classified the tweets in Sentiment140 tool. Even with emoticons Sentiment140 provides us accuracy of 80% using Naive Bayes, Maximum Entropy, and SVM. In Distance Supervision classification Twittratr is used as baseline [11]. Twittratr is a website which also performs sentiment analysis. Twittratr is Twitter opinion tracker which can categorize the positive and negative tweets depending on the brand, person, topic or product according to Akshat Bakliwal, Piyush Arora, Senthil Madhappan [13]. Diego Terrana, Agnese Augello, Giovanni Pilato talks about the Automatic Unsupervised Polarity Detection [AUPD] in Twitter data, where the classification is made without predefined polarity [14]. Po-Wei Liang and Bi-Ru Dai explains that domain-specific training data to build a generic classification and has also compared the result of opinion miner and unigram in extracting the feature from tweets. Finally concluded that opinion miner gives better result than the unigram. Opinion

miner was created using machine learning techniques and domain specific training for better accuracy [15].

Xujuan Zhou, Xiaohui Tao, Jianming Yong, Zhenyu Yang have proposed opinion mining tool based on lexicon for sentiment analysis named, Tweets Sentiment Analysis Model (TSAM) is proposed based on politicians' sentiment semantic[16]. Seyed-Ali Bahrainian and Andreas Dengel discusses about the aspect based sentiment. They have compared the sentiment polarity detection and sentiment summarization also hybrid sentiment polarity to gain better performance. Accuracy on positive class negative class and overall accuracy is analysed using unsupervised polarity detection, unsupervised ranking detection and hybrid using SVM as baseline, SVM which is trained on unigram feature set [17]. According to Eun Hee Ko, Diego Klabjan explains that cosine similarity, K-Means and Latent Dirichlet Allocation[LDA] is been used and compared. LDA is collection of data which is discrete and three layer hierarchical Bayesian model. Using textual data sets, the data sets with same values have high correlation value and the data sets different values will have low correlation values [18]. T. K. Das, D. P. Acharjya and M. R. Patra have developed system by connecting to Alchemy API by REST call method. Unstructured data analysis is considered, tweets are collected using this API which pulls out the tweets and provides filtering of tweets [19].

Farhan Hassan Khan, Usman Qamar and M.Younus Javed have used emoticons analysis, SentiWordNet analysis and Bags of words and compared with other techniques. Focus was also on the pre-processing of data, and gain better performance[20]. Lu Lin, Jianxin Li, Richong Zhang, Weiren Yu and Chenggen Sun have implemented an opining mining tool called Weibo, which collects the retweets that is tree-structure and retweets comments. It uses association rule mining algorithm and constructed Semantic Orientation from Pointwise Mutual Information and also lexicon based algorithm is built for sentiment analysis. They have compared the result on sentiment orientated classification that is SVM based methods and lexicon based methods and found that lexicon based methods are more efficient than other methods [21].

IV. CONCLUSION

This paper has survey of many opinion mining tools and also different types of steps involved in classification of tweets. Some give higher performance and some provide good accuracy. Some tools used supervised data and some unsupervised data. The tool can either use supervised or unsupervised, the accuracy of sentiment tool must be maintained. It is better to use structured data and use only minimum number of neutral tweets for predicting the opinion. During data collection, the tool can use any efficient database for better accuracy. Pre-processing steps must include slang language and emoticons processing because 140 characters is the limitation and also emoticons are used to express feeling easily. The tool can use any classifier/clustering depending on data and pre-processing steps.

REFERENCES

[1] M.Rambocas and J.Gama, "Marketing Research:The Role of Sentiment Analysis " The 5th SNA-KDD Workshop'11. University of Porto, 2013.

[2] Aliza Sarlan, Chayanit Nadam, Shuib Basri, "Twitter Sentiment Analysis", International Conference on Information Technology and Multimedia (ICIMU), November 18 – 20, Putrajaya, Malaysia, 2014.

[3] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques",In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 79–86, 2002.

[4] Monika Arora, Vineet Kansal, "A Framework for Informal Language: Opinion Mining". International Conference on Computing, Communication and Automation (ICCCA2015) IEEE, 2015.

[5] Renata Lopes Rosa, Demostenes Zegarra Rodriguez, Graca Bressan, "SentiMeter-Br: a Social Web Analysis Tool to Discover Consumers' Sentiment", IEEE 14th International Conference on Mobile Data Management. University of Sao Paulo, SP – Brazil, 2013.

[6] 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.

[7] Zhaoxia WANG, Victor Joo Chuan TONG, Xin XIN, Hoong Chor CHIN, "Anomaly Detection through Enhanced Sentiment Analysis on Social Media Data", IEEE 6th International Conference on Cloud Computing Technology and Science, 2014.

[8] V.Pandey and C.V.K Iyer, "Sentimental analysis of Microblogs" unpublished.

[9] 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.

[10] Patricia L V Ribeiro ,Li Weigang and Tiancheng Li "A Unified Approach for Domain-Specific Tweet Sentiment Analysis", FUSION, 2015

[11] Alec Go, Richa Bhayani and Lei Huang, "Twitter Sentiment Classification using Distant Supervision",unpublished.

[12] Amiya Kumar Tripathy, Revathy Sundararajan, Chinmay Deshpande, Pankaj Mishra, Neha Natarajan, "Opinion Mining from User Reviews", International Conference on Technologies for Sustainable Development (ICTSD-2015), Feb. 04 – 06, 2015, Mumbai, India, 2015

[13] 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.

[14] Diego Terrana, Agnese Augello, Giovanni Pilato, "Automatic Unsupervised Polarity Detection on a Twitter Data Stream", IEEE International Conference on Semantic Computing, 2014.

[15] Po-Wei Liang and Bi-Ru Dai, " Opinion Mining on Social Media Data", IEEE 14th International Conference on Mobile Data Management,2013.

[16] 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.

[17] Seyed-Ali Bahrainian, Andreas Denge, "Sentiment Analysis and Summarization of Twitter Data "IEEE 16th International Conference on Computational Science and Engineering, 2013.

[18] Eun Hee Ko and Diego Klabjan, "Semantic Properties of Customer Sentiment in Tweets", 28th International Conference on Advanced Information Networking and Applications Workshops,2014.

[19] T. K. Das, D. P. Acharjya and M. R. Patra, "Opinion Mining about a Product by Analyzing Public Tweets in Twitter", International Conference on Computer Communication and Informatics (*ICCCI* -2014), Jan. 03 – 05, 2014, Coimbatore, INDIA, 2014.

[20] Farhan Hassan Khan, Usman Qamar and M.Younus Javed ," SentiView: A Visual Sentiment Analysis Framework", International Conference on Information Society, 2014.

[21] Lu Lin, Jianxin Li, Richong Zhang, Weiren Yu and Chenggen Sun, "Opinion Miining and Sentiment Analysis in Social Network:ARetweeting Structure-aware Approach", IEEE/ACM 7th International Conference on Utility and Cloud Computing,2014.