Available at: www.researchpublications.org

Review on Sentiment Analysis of Twitter Data

Ms. Swapna R. Kharche[1], Prof. Lokesh Bijole[2]

[1]Department of Computer science & Engg, Padmashri V B Kolte College of Engineering, Malkapur (M.S.) India

[1]swapnakharche@gmail.com

[2]Assistant Professor, Department of CSE, Padmashri V B Kolte College of Engineering, Malkapur (M.S.) India

[2]lokeshmits5588@gmail.com

***Abstract*—Millions of users share their opinions on different aspects of life everyday. Sentiment Analysis is a process of identifying opinions in large unstructured/structured data and analyzing polarity of those opinions. Microblogging website such as twitter is rich source for sentiment analysis. Sentiment analysis over Twitter offers organizations a fast and effective way to monitor the publics' feelings towards their brand, business, directors, etc. In this paper we focused on different approaches used for sentiment analysis of twitter data.**

***Keywords*—Twitter, sentiment analysis, social media, machine learning, hybrid approach**

I. INTRODUCTION

Others opinions can be important when it's time to choose from multiple options or make a decision. When those choices involve valuable resources people consider their peers' past experiences. Now social media provides new tools to efficiently share ideas with everyone connected to the World Wide Web. Opinion mining is process of extracting information based on a people's opinions from data available on websites. Whereas sentiment analysis focus on polarity detection(positive, negative or neutral). Microblogging site Twitter contains large number of short length messages. As more and more users post about products and services they use, social sites like TWITTER, FACEBOOK become valuable sources of people's opinions and sentiments. The dataset collected from all these sites can be effectively and efficiently used for marketing, social networking. For example, manufacturing companies may be interested in the following questions:

- What do people think about our product (service, company etc.)?
- How positive (or negative) are people about our product?
- What would people prefer our product to be like?

Political parties may be interested to know if people support their program or not. Social organizations may ask people's opinion on current debates. The need to collect opinions from social networking sites and draw conclusions that what people like/dislike, has been the most important aspect in today's scenario. In our paper, we study that how twitter would use for sentiment analysis purposes which not only shown people's opinion or point of view towards any matter but also provide their needs, demands from the current scenario.

The objective of this paper is to discuss concept of sentiment analysis of twitter tweet and comparative study of its various techniques.

II. LITERATURE REVIEW

There are various text mining approaches used to mine the twitter data.

T. K. Das, D. P. Acharjya, M. R. Patra [1] developed a system that processes the tweets by pulling data from tweeter posts, preprocessing it and connecting to Alchemy API. Alchemy API is a web service that analyzes the unstructured contents (news, articles, blogs, posts etc.). The three way classification is done by analyzing the collected data. The high end users generate the report in the form of cumulative graphs, pie charts and tables. It can help the management to improve the quality of their product.

Efthymios Kouloumpis, Theresa Wilson, Johanna Moore [2] used three different corpora of Twitter messages in experiments-hashtagged data set (HASH), emoticon data set (EMOT), a manually annotated data set produced by the iSieve Corporation (ISIEVE). The goal for this experiment is two-fold. First, it evaluates whether training data with labels derived from hashtags and emoticons is useful for training sentiment classifiers for Twitter. Second, it evaluates the effectiveness of the features from section for sentiment analysis in Twitter data. This experiment on twitter sentiment analysis showed that when microblogging features are included, the benefit of emoticon training data is lessened.

Pak and Paroubek [3] used a dataset formed of collected messages from Twitter. This paper shows how to automatically collect a corpus for sentiment analysis and opinion mining

purposes. Twitter contains a very large number of very short messages created by the users of this microblogging platform. The contents of the messages vary from personal thoughts to public statements. This paper presents a method for an automatic collection of a corpus that can be used to train a sentiment classifier. This classifier is able to determine positive, negative and neutral sentiments of documents. The classifier is based on the multinomial Naïve Bayes classifier that uses N-gram and POS-tags as features.

A. Shrivatava, S. Mayor and B. Pant [4] developed method which is efficient and time saving to classify millions of tweets posted on twitter. This methodology helps to establish the DOMAIN DICTIONARY that contains the feature terms of individual classified files. They designed Twitter TWEETS PULLER which can pulls 1000 tweets at a time when it is connected to the server site and CLASSIFIER TOOL that classifies features of Twitter Tweets separately.

Shulong Tan, Yang Li, Huan Sun [5] interpreted public sentiment variations on twitter. They proposed two Latent Dirichlet Allocation (LDA) based models, Foreground and Background LDA (FB-LDA) and Reason Candidate and Background LDA (RCB-LDA). The FB-LDA model can filter out background topics and then extract foreground topics to reveal possible reasons. The RCB-LDA model can rank a set of reason candidates expressed in natural language to provide sentence-level reasons. The proposed model helps to discover special topics or aspects in one text collection in comparison with another background text collection.

Anton Barhan, Andrey Shakhomirov [6] investigated and developed a method for automatic sentiment analysis of Twitter messages. For developing this method, they reviewed the existing automatic sentiment analysis methods and studied the text features of social media messages in the context of developing methods for their sentiment analysis.

III. DIFFERENT APPROACHES FOR SENTIMENT ANALYSIS OF TWITTER DATA

There are two main techniques for sentiment analysis: machine learning based and lexicon based. New research studies have used combination of these two methods for better performance.

A. Machine learning based approach

The machine learning (ML) approach used for sentiment analysis mostly belongs to supervised classification in general and text classification techniques in particular. Thus, it is called "supervised learning". In a machine learning based techniques, two sets of documents are needed: training and a test set. A training set is used by an automatic classifier to learn the differentiating characteristics of documents, and a test set is used to check the performance of the automatic classifier. A number of machine learning techniques have used to classify the reviews. Machine learning techniques like Naive Bayes (NB), maximum entropy (ME), and support vector machines (SVM) have achieved great success in sentiment analysis.

Naive Bayes is a simple but effective classification algorithm. Naïve Bayes (NB) algorithm used to classify textual data. NB can perform better on several cases and additionally it has several advantages such as lower complexity and simpler training procedure. However, NB greatly suffers from sparsity when applied to the particularly high dimensional data as in text classification. This is especially the case when the training data consist of very short documents such as tweets and when the training set size is limited because of the cost of manual labeling processes. In order to avoid zero probability problem smoothing methods are used. Murat C. Ganiz, Dilara Torunoğlu [8] used Naïve Bayes (NB) algorithm to classify large amount of textual data. They proposed Wikipedia Semantic smoothing approach to avoid the sparsity problem. Using WEX, they have extended Twitter Sentiment 140 dataset Wikipedia article titles, categories and redirects. The semantic smoothing approach extracts important concepts called topic signatures from the training documents and calculates term probabilities by statistically mapping terms to topic signatures using Expectation Maximization (EM) algorithm.

The support vector machine (SVM) is a statistical classification method proposed by Vapnik. The support vector machine has performed effectively for classification in the literature. SVM ac be use more effectively in combination with SentiWordNet for sentiment classification [11]. SenticWordNet 3.0 is also a publicly available lexical resource which is explicitly devised for supporting sentiment classification and opinion mining applications.

V. B. Raut, Prof. D.D. Londhe [7] presented different approaches, methods and techniques used in process of opinion mining and summarization, and comparative study of these different methods. This study shows that machine learning approach works well for sentiment analysis of data in particular domain such as movie, product, hotel etc.

B. Lexicon based approach

The lexicon based techniques to Sentiment analysis is unsupervised learning as it does not require prior training in order to classify the data. In this approach, classification is done by comparing the features of a given text against sentiment lexicons whose sentiment values are determined prior to their use. Sentiment lexicon contains lists of words and expressions used to express people's subjective feelings and opinions. For example, start with positive and negative word lexicons, analyze the document for which sentiment need to find. Then if the document has more positive word lexicons, it is positive, otherwise it is negative. Antonio Moreno-Ortiz, Chantal Pérez Hernández [9] presented lexicon-Based approaches to Sentiment

Analysis (SA) using sentitext. Sentitext is a web-based, client-server application written in C++ (main code) and Python (server). They perform a test to check whether such lexically-motivated systems can cope with extremely short texts, as generated on social networking sites, such as Twitter. They conclude that differentiating between neutral and no polarity may not be the best decision and it is very difficult to obtain good results in these two categories. Lexicon based approach is suitable for short text in micro-blogs, tweets, and comments data on web [7]. A. Khan et al. [16] proposed rule based domain independent method of sentiment classification at the sentence level. They first classify sentences into objective and subjective and check their semantic scores using the SentiWordNet. The final weight of each individual sentence is calculated after considering the whole sentence structure, contextual information and word sense disambiguation. Their method achieves an accuracy of 86.6% at the sentence level.

C. Hybrid approach

Few research techniques having combination of both the machine learning and the lexicon based approaches used to improve sentiment classification performance. A. Mudinas [14] developed pSenti – a concept-level sentiment analysis system that seamlessly integrates into opinion mining lexicon-based and learning-based approaches. The hybrid approach is important as it gain both stability as well as readability from a carefully designed lexicon, and the high accuracy from a powerful supervised learning algorithm. The hybrid approach pSenti achieved 82.30% accuracy.

Farhan Hassan Khan, Usman Qamar [10] presented a new algorithm for twitter feeds classification based on a hybrid approach. They compare their work with other techniques to prove the effectiveness of the proposed hybrid approach. It resolves the data sparsity issue using domain independent techniques. They achieved an average accuracy of 85.7%.

Zhang et al. [12] employ an augmented lexicon-based method for entity level sentiment analysis. First extract some additional opinionated indicators (e.g. words and tokens) through the Chi-square test on the results of the lexicon-based method. With the help of the new opinionated indicators, additional opinionated tweets can be identified. Afterwards, a sentiment classifier is trained to assign sentiment polarities for entities in the newly identified tweets. The training data for the classifier is the result of the lexicon-based method. They achieved accuracy of 85.4%.

IV. COMPARATIVE STUDY

Most of the researchers reported that machine learning has high accuracy than other techniques. The main limitation of supervised learning is that it generally requires large training data that are very expensive. The lexicon based approach is best for short text which gives better sentiment analysis performance. The main advantage of hybrid approach using a lexicon/learning combination is to attain high accuracy from a powerful machine learning algorithm and stability from lexicon based approach.

There are different approaches used for sentiment analysis of twitter data. Table 1 presents summary of accuracy gain by sentiment analysis of twitter messages using different techniques.

TABLE 1
Accuracy gain by sentiment analysis of twitter messages using different techniques

Paper	Technique	Accuracy
A. Shrivatava, S. Mayor and B. Pant [4]	SVM	70.5%
Murat C. Ganiz, Dilara Torunoğlu [8]	Naïve Bayes (NB)	Highly accurate
Farhan Hassan Khan, Usman Qamar [10]	ML and lexicon	85.7%
Alec Go et al. [13]	Naïve Bayes, Maximum entropy, and Support vector machines.	Approx. 80 % - 82 %
Zhang et al. [12]	ML and lexicon	85.4%
A. Mudinas, D. Zhang, M. Levene[14]	ML and lexicon	82.3%
Sunil B. Mane et al[15]	Naïve Bayes (NB)	72.27%
A. Khan, B. Baharudin [16]	Lexicon	86%

V. FUTURE ENHANCEMENT

Future research includes the development of real time sentiment analysis tool in order to compare the performance with the application like Tweet Feel, Twendz, and Sentiment140 and the use of supervised learning algorithms to further increase the accuracy.

There are lots of research work and more work remaining in this field to elaborate and there are work also present which is not get solved yet such as use of negation with maximum accuracy, give detail about items, use of multiple languages at a time.

VI. CONCLUSION

Due to web and social network, large amount of data are generated on Internet every day. This web data can be mined and useful knowledge information can be analyzed through sentiment analysis process. This paper discussed various techniques, approaches of sentiment analysis of public from twitter social site and comparative analysis of accuracy gain by these approaches. These techniques should be combined to overcome their individual drawbacks and enhance the sentiment analysis performance.

References

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

[2] Efthymios Kouloumpis, Theresa Wilson, Johanna Moore, "Twitter Sentiment Analysis: The Good the Bad and the OMG!" Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, University of Edinburgh, 2011

[3] Twitter as a Corpus for Sentiment Analysis and Opinion Mining Alexander Pak, Patrick Paroubek.

[4] A.Shrivatava, S. Mayor and B. Pant, "Opinion Mining of Real Twitter Tweets," International Journal of Computer Applications, Volume 100– No.19, August 2014.

[5] Shulong Tan, Yang Li, Huan Sun, Ziyu Guan, "Interpreting the Public Sentiment Variations on Twitter," IEEE Transactions on Knowledge and Data Engineering, Vol. 6, No. 1, September 2012.

[6] Anton Barhan, Andrey Shakhomirov, "Methods for Sentiment Analysis of Twitter Messages," Proc.12th Conference of Frucst Association(2012).

[7] Vijay B. Raut et al, "Survey on Opinion Mining and Summarization of User Reviews on Web," (IJCSIT) International Journal of Computer Science and Information Technologies, Vol.5 (2), 2014, pp.1026- 1030.

[8] Wikipedia Based Semantic Smoothing For Twitter Sentiment Classification by Dilara Totunoglu, Gurkan Telseren, Ozgun Sagturk & Murat C.Ganiz IEEE (2013).

[9] Antonio Moreno-Ortiz, Chantal Pere Hernandez,"Lexicon-Based Sentiment Analysis of Twitter Messages in Spanish," ISSN 1135-5948, pp.93-100, 2013.

[10] TOM:Twitter opinion mining framework using Hybrid Classification scheme, Decision Support Systems by Farhan Hassan Khan (2014).

[11] Chihli Hung, Hao-Kai Lin," Using Objective Words in SentiWordNet to Improve Sentiment Classification for Word Of Mouth", IEEE 2013

[12] L. Zhang, R. Ghosh, M. Dekhil, M. Hsu, and B. Liu, "Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis", Technical report, HP Laboratories, 2011

[13] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," CS224N Project Rep., Stanford: 1–12, 2009.

[14] A. Mudinas, D. Zhang, M. Levene, "Combining lexicon and learning based approaches for conceptlevel sentiment analysis", Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining, ACM, New York, NY, USA, Article 5, pp. 1-8, 2012.

[15] Sunil B. Mane et al, "Real Time Sentiment Analysis of Twitter Data Using Hadoop"(IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, 3098 - 3100.

[16] A. Khan, B. Baharudin, K. Khan; "Sentiment Classification from Online Customer Reviews Using Lexical Contextual Sentence Structure" ICSECS 2011: 2nd International Conference on Software Engineering and Computer Systems, Springer, pp. 317-331, 2011s.