ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE)

International Conference on Recent Trends in Computing and Communication (ICRTCC 2015)

Cochin College of Engineering & Technology

Vol. 4, Special Issue 1, June 2015

Impact and Application of Sentiment Analysis using Twitter: A Survey

Hema Krishnan[1], M. Sudheep Elayidom[2], T. Santhanakrishnan[3]

Research Scholar, School of Engineering, CUSAT, Cochin, India [1]

Associate Professor, School of Engineering, CUSAT, Cochin, India [2]

Scientist E, NPOL, Ministry of Defence, Cochin, India [3]

Abstract: Sentiment analysis is a very relevant technique nowadays for social network analysis. Sentiment analysis or opinion mining is the process of automatically extracting knowledge from sentiments or opinions of others about some topic or problem. We can identify opinions in a large unstructured/structured data and analyze polarity of opinions. Twitter is a large and rapidly growing micro blogging social networking website where people express their opinions in short and simple manner of expressions. Tweets can be analyzed to perform sentiment analysis on various entities (products, people etc :). This has become a rapid and effective way of getting public opinion for business networking or social studies. In this paper we focus on different approaches used for sentiment analysis of twitter data.

Keywords: Twitter; social media; sentiment analysis; machine learning; hybrid approach.

I. INTRODUCTION

Sentiment analysis is the field of study that analyses the mood of public about a particular product or topic. It is also called opinion mining which is a technique for detecting and extracting subjective information in text documents. Opinion mining is a large problem space which is used to extract information based on people's opinions from data available on websites. Web contains a large collection of product reviews by customers since it is easy to publish online in recent years.

Micro blogging has become a very popular communication tool among Internet users today. Popular websites such as Twitter, Facebook, and Tumbler etc: are providing services for micro blogging. People share their opinions about products and services they use in these sites. Therefore these micro blogging sites become valuable sources of people's opinions and sentiments. 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 life.

Twitter is a micro blogging site which contains a large number of short length messages often called tweets. The content of the messages vary from personal thoughts to public statements. Dataset collected from twitter can be effectively used in opinion mining and sentiment analysis tasks. Twitter serves as a corpus [1] for opinion mining due to following reasons.

- Collected corpus from twitter can be arbitrarily large since it contains an enormous number of text posts. Also it grows everyday.
- It is possible to collect text posts of users from different social and interest groups.
- We can collect data in different languages

In this paper we focus on twitter data and impact and application of opinion mining and sentiment analysis using twitter data.

II. RELATED WORKS

There are various approaches for mining twitter data. A system that processes the tweets by pulling data from tweeter posts was developed by T.K Das, D.P Acharjya, and M.R Pathra [2]. Data collected from twitter were pre-processed and connected to Alchemy API. Unstructured contents (news, articles, blogs, posts etc :) can be analysed by the web service Alchemy API. The collected data is analysed and high end users can generate reports in the form of cumulative graphs, pie charts and tables. Management can improve the quality of their product by this method.

Pak and Paroubek [1] used a data set formed of collected messages from twitter. In this paper Twitter is considered as a corpus for sentiment analysis and opinion mining purposes. This paper uses a method for automatically collecting twitter data which can be used to train a sentiment classifier. Positive, negative and neutral sentiments of documents can be determined by the classifier. The classifier is based on the multinominal Naïve Bayes classifier that uses N-gram and POS-tags as features.

An efficient and time saving method to classify millions of tweets posted on Twitter was developed by A.Shrivatava, S.mayor and B.Pant [3]. DOMAIN DICTIONARY which contains the feature terms of individual classified files can be established by this methodology. Twitter TWEETS PULLER which can pull 1000 tweets at a time is designed by them. A CLASSIFIER TOOL which classifies features of Twitter Tweets was also developed by them.

A method for automatic sentiment analysis of Twitter was developed by Anton Barhon, Andrey Shakhomirov [4]. The method was developed by reviewing the existing automatic sentiment analysis methods and studying the text features of social media messages in the context of developing methods for their sentiment analysis.

DOI 10.17148/IJARCCE

III. DIFFERENT METHODS FOR TWITTER SENTIMENT ANALYSIS

A. Levels of sentiment analysis

Sentiment analysis can be done at three levels of granularity namely, document level, sentence level and aspect level.

Document level sentiment analysis [5] main task is to identify the polarities of user documents. It aims to automate the task of classifying user documents which is given on a single entity or aspect. In this overall sentiment of documents can be determined by the polarities of different sentiment words used in documents. This classification does not work with forum and blog posting because in such a posting the author may express opinions as multiple products and compare using comparative sentences.

We can classify a sentence as subjective or objective and this task is called subjectivity classification. Sentence level sentiment classification [5] is classifying these resulting subjective sentences into those expressing positive or negative opinions. In this analysis polarity of each sentence is calculated.

In a typical opinionated document, although the general sentiment on the entity may be either positive or negative, the author writes both positive and negative aspects of the entity. Such information is not provided by Document and sentence sentiment classification. These details are obtained only at the aspect level. Here it is assumed that a document contains opinion on several entities and their aspects.

B. Sentiment analysis techniques

There are mainly two methods for Sentiment analysis or opinion mining: Machine learning based and lexicon based. For better performance new research studies use a combination of these two methods which is known as hybrid approach.

Machine learning based approach

Machine learning (ML) approach [14] applies the famous ML algorithms and uses linguistic features. The ML approach used for sentiment classification mostly belongs to supervised classification in general and text classification techniques in particular. Text classification techniques can be roughly divided into supervised and unsupervised learning methods.

Large number of labelled training documents is used by supervised methods. Two sets of documents are needed in machine learning techniques: 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. Reviews can be classified by a number of machine learning techniques. Machine learning techniques like Naïve Bayes (NB), maximum entropy (ME), and support vector machines (SVM) have achieved a great success in sentiment analysis.

Naïve Bayes is a simple but effective classification 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 arises in the case when the training data consists of very short documents such as tweets and the training set size is limited because of the cost of manual labelling processes. Murat C. Ganiz, Dilara Torunoglu [6] classified large amount of textual data using Naïve Bayes algorithm. The sparsity problem was avoided by proposing a smoothing approach. Twitter sentiment 140 data set Wikipedia article titles, categories and redirects were extended using WEX.

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 can be used more effectively in combination with SentiWordNet for sentiment classification [7]. Sentiment classification and opinion mining applications can be explicitly devised by SentiWordNet which is a publicly available lexical resource.

Lexicon based approach

Lexicon based approach depends on finding the opinion lexicon which is used to analyze the text. Two methods are used in this approach.

Corpus based approach begins with a seed list of opinion words and then finds their opinion words in a large corpus to help in finding opinion words with context specific orientations. This could be done by statistical or semantic methods. It is unsupervised learning as it does not require prior training in order to classify the data. The second approach is dictionary based which depends on finding opinion seed words and then searches the dictionary of their synonyms and antonyms.

Lexicon based approach [14] 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 list of words and expressions used to express people’s subjective feelings and opinions. For e.g., start with positive and negative word lexicons, analyze the document for which sentiment need to find. Then if the document has more positive lexicons, it is positive otherwise it is negative.

Antonio Moreno- Ortiz, Chantal Perez Hernandez [8] used lexicon based approaches to Sentiment Analysis (SA) using sentitext. Sentitext is a web-based, client server application written in C++ (main code) and Python

DOI 10.17148/IJARCCE

(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. A rule based domain independent method of sentiment classification at the sentence level was proposed by A. Khan et al. [9]. Sentences were classified into subjective and objective and their semantic scores were checked using SentiWordNet. Whole sentence structure, contextual information and word sense disambiguation is considered to calculate the final weight of each individual sentence.

Hybrid approach

Sentiment classification can be improved by combining both machine learning and the lexicon based approaches which are used by few research techniques.

Zhang et al. [10] explore an entity-level sentiment analysis approach to the Twitter data. An augmented lexicon-based method is employed for this. 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 documents 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 lexicon-based method. They achieved accuracy of 85.4%.

Namita Mittal, Basant Agarwal et. al [11] introduces a hybrid approach for automatically classifying the sentiment of Twitter messages. The proposed approach consists of three stage hierarchy. First of all a tweet is labelled according to emoticons it have, then labelling tweet using pre-defined list of strong positive and strong negative words and finally using subjectivity lexicon and probability method. Further, various cascading and hybrid methods are proposed based on subjectivity lexicon and Probability based method. In addition to this, effect of discourse relations is investigated at the pre-processing step.

Akshil Kumar, Teeja Mary Sebastian [12] proposed and investigated a paradigm to mine the sentiment from a popular real-time microblogging service, Twitter. They expounded a hybrid approach using both corpus based and dictionary based methods to determine the semantic orientation of the opinion words in tweets.

Farhan Hassan Khan, Usman Qamar [13] 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.

IV. COMPARISON OF DIFFERENT TECHNIQUES

According to most of the researchers report Machine learning is high accurate when compared to other techniques. Supervised learning generally requires a large training data which is very expensive. That is the main limitation of this method. Better sentiment analysis performance can be achieved by lexicon based approach which is best suitable for short text. The main advantage of hybrid approach is that it uses a lexicon/ learning combination to attain high accuracy from s powerful machine learning algorithm and stability from lexicon based approach.

Different approaches used for sentiment analysis has different accuracy gain which is presented in Table 1.

TABLE 1SUMMARY OF SENTIMENT ANALYSIS USING TWITTER MESSAGES

Paper	**Technique**	**Accuracy**
A. Shrivatava, S. Mayor and B. Pant [3]	SVM	70.5%
Murat C. Ganiz, Dilara Torunoglu [6]	Naïve Bayes (NB)	Highly accurate
Farhan Hassan Khan, Usman Qamar [13]	ML and Lexicon	85.7%
Zhang et al. [9]	ML and lexicon	85.4%
Namita Mittal, Basant Agarwal [11]	Hybrid approach-subjectivity lexicon and probability	71.12%
Akshil Kumar, Teeja Mary Sebastian [12]	Hybrid approach-corpus and dictionary method	80%

V. CONCLUSION AND FUTURE SCOPE

Twitter offers an unprecedented opportunity to create and employ theories that search and mine for sentiments. The data can be mine and useful knowledge information can be analyzed through sentiment analysis process. Various methods which show the impact and applications of sentiment analysis using twitter were discussed in this paper. Different techniques can be combined to overcome their individual drawbacks and enhance the sentiment analysis performance.

A real time sentiment analysis tool can be developed in order to compare the performance with the application like Tweet Feel, Twendz, and Sentiment140. Supervised learning algorithms can be used to further increase their accuracy.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

REFERENCES

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

[2] T.K Das, D.P Acharjya, M.R Patra, " Opinion mining about a product by analysing public tweets in twitter," IEEE Proceedings of International Conference on Computer Communication and Informatics (ICCI-2014), Jan 03-05, 2014, Coimbatore, India

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

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

[5] Shailesh Kumar Yadav, "Sentiment Analysis and Classification: A survey", International Journal of Advance Research in Computer Science and Management Studies, Volume 3, Issue 3, March 2015..

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

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

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

[9] 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.

[10] 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 lLaboratories, 2011.

[11] Namita Mittal, Basant Agarwal, Saurabh Agarwal, Shubham Agarwal, Pramod Gupta, "A Hybrid Apprach for Twitter Sentiment Analysis," Proceedings of ICON-2013: 10th International Conference on Natural Language Processing, pp: 116-120, 2013, Noida, India.

[12] Akshi Kumar, Teeja Mary Sebastian, " Sentiment Analysis on Twitter," IJCSI International Journal of Computer Science Issues, Vol 9, Issue 4, No 3, July 2012.

[13] Farhan Hassan Khan,"TOM: Twitter Opinion mining framework using Hybrid Classification scheme, Decision Support Systems" (2014).

DOI 10.17148/IJARCCE