International Journal of Advance Research in Science and Engineering
Volume No.06, Issue No. 10, October 2017
www.ijarse.com
IJARSE
ISSN: 2319-8354

Sentiment Analysis on Twitter

Arti Bansal

Computer Application,Guru kashi University,(India)

ABSTRACT

With the growing popularity of social sites on web, people have begun to express their opinions on a wide variety of topics on Twitter and other similar services. Sentiment analysis is mainly concerned with identifying and classifying emotions that are expressed within a text. Twitter sentiment analysis often becomes a difficult task due to slang words and misspellings. Daily we counter new words which make it very difficult. Twitter restricts the length of a tweet up to 140 character. In this paper, I discuss about the problems in sentiment analysis on Twitter and various approaches which are used by different researchers and their results.

I. INTRODUCTION

The popularity of microblogging stems from its distinctive communication services such as portability, immediacy and ease of use, which allow users to instantly respond and spread information with limited or no restrictions on content. Twitter is currently the most popular and fastest-growing microblogging service, with more than 140 million users producing over 400 million tweets per day-mostly mobile as of June 2012. Twitter enables users to post status updates, or tweets, no longer than 140 characters to a network of followers using various communication services.

Tweets have reported everything from daily life stories to latest local and worldwide events. Twitter content reflects real-time events in our life and contains rich social information and temporal attributes. Monitoring and analyzing this rich and continuous flow of user-generated content can yield unprecedentedly valuable information.

Online social media sites (Facebook, Twitter, YouTube, etc.) have revolutionized the way we communicate with individuals, groups, and communities and altered everyday practices (Boyd and Ellison 2007). Several recent workshops, such as semantic analysis in social media (Farzindar and Inkpen 2012), are increasingly focusing on the impact of social media on our daily lives. For instance, Twitter has changed the way people and businesses perform, seek advice and create "ambient awareness" and reinforced the weak and strong tie of friendship. Unlike other media sources, Twitter messages provide timely and fine-grained information about any kind of event, reflecting, for instance, personal perspectives, social information, conversational aspects, emotional reactions, and controversial opinions.

1.1 Twitter

Twitter was developed in March 2006 by JACK DORSEY, NOAH GLASS, BIZ STONE and EVAN WILLIANNS and launched in July 2006. Twitter headquarter is established in San Francisco,California(U.S.).

Twitter is a free social networking microblogging service that allow only registered users to tweet and unregistered users can only read them. Tweets have a 140 character length. Social media such as Facebook or

Twitter provide important platform where user can share their opinion on a particular topic. The default setting for Twitter is public unlike facebook. Anyone can follow anyone on public Twitter. The hashtag which acts like a meta tag,is expressed as #keyword. Twitter uses an open source web framework called Ruby On Rails(ROR).

1.2 Sentiment Analysis

Sentiment analysis or sentiment mining or polarity mining or opinion mining is concerened with analysis of text containing opinion and emotions. It is aprocess where the dataset cosists of emotions or attitudes which takes the way a human thinks. Sentiment analysis aims to understand the opinions expressed on text and classify them into different catageories like positive, negative or neutral. Sometimes one tweet expressed some mixture of emotions like postive and negative. Example :- I like choclate but it is bad for teeth. This tweet have both positive and negative view.

Sentiment analysis of Twitter requires a great effort from the classifier because tweets are short in length and leading to huge ambiguity. Tweets have 140 character length and can be used more informally with slangs and special characters. Sentiment analysis is a very popular research area even before Twitter.

II. RELATED WORK

Aggarwal C.C. and Zhao P. (2013) design graphical models for representing and processing text data by using the concept of distance graphs which represents the documents in terms of distance between words. This can provide rich information about the behavior of the underlying data. It provides a graphical paradigm which turns into to be an effective text representation for processing. After that they analyze their approach with a large number of different classification and similarity search application.

Burns A. (2016) reviews the origin of twitter. He traces the origin and gradual development of the platform and outlines some of the key contemporary uses of Twitter. He define the synchronous communication between multiple participants who are digitally co-present Twitter have launched in march 2006, initially influenced by SMS but the limitation of Twitter messages not more than140 characters. Retweets were commonly proceeded by “RT@ username” to acknowledge that the messaged send from username. In 2015, company also introduced new feature “quoted tweet” that generates a URL linking to the original tweet page on the Twitter website. In 2007, Hashtags are suggested for brief keyword preceded by the hash symbol. Twitter posts generally text based, further addition to Twitter insertion of images, videos and links to other types of content by URL pointing to the location. Twitter is particularly well suited to the rapid dissemination and subsequent discussion and evaluation of news report. Twitter as a back channel to broadcast contents or live contents, from popular entertainment through sports to conference. The ecosystem of third party developers and service providers which has emerged around Twitter constitutes a further node in this network.

Cheng S.*et.al.* (2013) analysis the big data problems and give the strong side of solving big data problems by Swam Intelligence. They have survey about the potential of Swarm Intelligence in big data analytics. They analyses mainly three properties of big data which are high dimensionality of data, the dynamical change of data and multi objective problems. Big data may contain many kind of unstructured or semi structured data. The problems of big data can be solved by SI. SI is based on population of individuals is a collection of nature

inspired searching techniques. There exists many SI algorithm among them ACO, which was originally designed for discrete optimization problem and PSO, which was originally designed for continuous optimization are most commonly used. In Swarm Intelligence algorithm, there are several solutions exist at the same time. The premature convergence may happen due to the solution getting clustered together too fast.

Duric A. and Song F. (2012) describes sentiment analyze based on feature selection methods from Lexicon based approaches where the set of feature are generated by humans. Traditionally, text classification seeks to classify a document by topic but SA deals with opinions about topics. They approached the task of feature selection by using content and syntax model, known as HMM-LDA to separate the entities in a review document. HMM-LDA models entities and modifiers as long range dependencies, allowing us to separate words into semantic and syntactic classes. They proposed feature selection schemes achieved competitive results in our experiments for document polarity classification. They minimize the impact by separating the semantic class from syntactic classes and as a result, removing some of the neutral features that present in the baseline schemes.

Elloumi W. *et.al.* (2014) presents a novel approach by introducing PCO, which is modified by algorithm to improve the performance of TSP(Travelling Salesman Problem). They mainly focus on modifying ACO using two operations: first by adjusting the parameter Q_0, which relates to both exploitation and exploration in ACO. Second, escaping the trap b reinitializing Q_0 in a way of exploration. They research a real ants behavior for reaching out the food. Traditionally ACO is used for discrete optimization while PSO is for continuous optimization. When hybrid ACO and PSO, PSO supervised ACO it solves continuous optimization problems. They work on 1000 iterations on PSO-m-ACO and coded in MATLAB. When they compared proposed approach with traditional findings, they converge rapidly to a minimum as the number of particles is increased in the Swam. So they prove that necessity of hybridization used between PSO and ACO.

Gautam G. and Yadav D. (2014) proposed the analysis of performance by machine learning approaches and Word net. In the proposed approach, Twitter dataset is created and preprocess the data by removing repeated words and punctuations. Before preprocessing, data becomes raw and it is difficult to handle. So after preprocessing, data efficiency is increased. Then feature extraction method, extracts feature from the dataset by unigram model. It discards the preceding the preceding and successive word occurring with the adjective in the sentences. After that these features are classified by using machine learning approaches such as Naïve Bayes, Maximum entropy, SVM. After that classification they used semantic analysis derived from the WorldNet database. This database is of English words which are linked together. When two words are close to each other than they are like synonym. It is helpful to show the polarity of sentiment for the user. They use Python and Natural language kit to train the Naïve Bayes, Maximum Entropy and SVM. The performance is measured on the basis of recall, precision and accuracy and WorldNet have high accuracy.

Ismail H.M. *et.al.* (2016) compare the performance of different machine learning classifier for twitter sentiment analysis. For this analysis STS (Stanford Testing Documents) dataset is used. They analyze unigram as well as bigram as feature spaces. They analyze TF representation of data set. They evaluate the performance of multinomial NB, Bernoulli NB and SVM in sentiment mining. They choose WEKA for evaluating the performance of the selected classifiers. The overall accuracy for bigram datasets. Training time for unigrams

dataset is in general less than bigrams. Multinomial NM produced the best results with frequency unigram dataset. Unigram as a form of representing dataset feature proved to be more effective in the context of Twitter sentiment analyzes as they produce less sparse dataset.

Kontopoulous E. *et.al.* (2013) proposed the deployment of original ontology based techniques towards a more efficient sentiment analysis of Twitter posts. In this approach, posts are not simply categorized by sentiment score; instead receive a sentiment grade for each distinct notion in the post. An ontology means describe the relation among the terms of a specific domain. If any sentences have two sentiments then it gives doubtful results. In this approach, first domain ontology is created and then sentiment analysis applied on this by using two formal approaches FCA and Ontology Learning. The proposed architecture, given the higher observed recall ratios, appears to perform evidently better than the custom built system method.

Kouloumpis E. *et.al.* (2011) evaluate the usefulness of existing lexical resources as well as features that capture information about the informal and creative language used in microblogging companies such as Twitrratr, TweetFeel and social mention are just a few who are advertise Twitter sentiment analysis as one of their services. They used unigram and bigram and included features used in sentiment analysis. Finally, they include features to capture some of the more domains specific language of microblogging using Hashtags to collect training data did prove useful, as data collected based on positive and negative emotions. So which method produces the better training and whether two sources of training data are complementary may depend on the type of feature used.

Lima A.C. and Castro L.N. (2012) proposed an automatic sentiment classifier for twitter messages. Sentiment analysis tasks can be used as one such feedback mechanism. This task corresponds to classifying a text according to the sentiment that the writer intend to transmit. A classifier mainly requires a pre classified data to determine the class of new data. The sample is pre classified manually, making the process time consuming and reducing its real time applicability for big data. They used TV shows for Brazilian stations for benchmarking and captured in a 24hr interval and fed into system. The proposed technique achieved an average accuracy of 90%. The automatic sentiment analysis reduces human intervention and complexity and cost of whole process.

Lin Y.S. *et.al.* (2014) Measured the similarity between two documents. They compute the similarity between two documents with respect to a feature and divide the task into the three cases, the feature appears in both the documents, the feature appears in one of the document, the feature appear in none of the documents. As the similarity increases as the difference between two involved feature values decreases. In the last case, the features have no contribution to the similarity. They measure in several text applications, including K-NN based single and multi label classification, k-means clustering and HAC. They used mainly three data sets webkb, reuters-8 and RCV1. For webkb, the randomly selected training documents are used for training and testing documents are used for testing. They mainly focus on textual features. The experimental results could depend on applications domains, feature formats and classification clustering algorithm.

Maharani W. (2013) proposed the method to analyze sentiments through lexical based and machine learning approaches. They classify opinion contained tweets using two methods. First, machine learning approaches are applied such as SVM, ME, MNB, K-NN and based on outcome, lexicon based approach is applied. Based on the result of system tests and analysis it can be concluded that scoring result using lexical database for Indonesia

language able to classifying opinion into positive and negative. Model based approach with Machine learning produce better accuracy rate than lexical based approach. The accuracy rate using Machine learning approach is depended on the dataset which become the training data and determine varied parameter for each method.

Mane S.B. *et.al*. (2014) proposed the sentiment analysis using the Naive Bayes approach and a Hadoop cluster for distributed processing of textual Twitter post data. Hadoop solve problems which had lot of data for processing which doesn't fit into tables. Twitter data being unstructured can be best stored using Hadoop. HDFS (Hadoop Distributed File System) has a high throughput access to application and is suitable for application with large amount of data. In this paper, they mainly focus on speed of performing analysis than its accuracy. They performed sentiment analysis on big data which is achieved by dividing tasks in modules with Hadoop. They remove stop words, convert unstructured data into structured data and emotions symbols converted into words. The overall accuracy of this paper is determined by time required to access from various modules. They use emotions but the use of Hashtags to determine the context of tweet is not done but the Hashtags are frequently used in twitter. So it is necessary to analysis the Hashtags. With this limitation the accuracy is found to it is be 72.27%.

Nethu M. S. and Rajsree (2013) analysis the Twitter data about electronic product using Machine Learning approach. They present a new feature vector for classifying tweets as positive, negative. They create a dataset using automatically Twitter API and split into training set and test set. Then preprocess the tweets by removing URL, avoiding misspellings and slang words after that create the feature vector in two steps. Firstly, twitter specific features are extracted and then these features are removed from tweets and again feature extraction is done as it is done on normal text. After creating feature vector, Naive Bayes, SVM, Maximum entropy and ensemble classifier are used for classification using Matlab simulator. Then the performance of this classifier is analysis on the basis of precision, recall, accuracy. All these classifier have almost similar accuracy for the new feature vector for electronic product domain.

Pennacchioti M. and Maria A.P. (2011) classify the user on the basis of profile, messaging behavior, linguistic content of message and social network information. In the profile feature, length of the user name, number of numeric and alphanumeric characters in the user name etc. are analyzed. In messaging behavior, identify the users who rarely post tweets but have many followers tend to be information seekers, while user who often post Url in their tweets are most likely information providers. In this paper mainly three classification tasks are done, detecting political affiliation, detecting a particular ethnicity and identifying Starbucks Fans. They presented a generic model for user classification in social media and provide extensive quantitative and qualitative analysis.

Silva Nadica F.F. *et.al.* (2014) analysis the sentiments in tweets by using the ensemble of Naive Bayes, SVM, Random forest and logistic regression. They show that the use of ensembles of multiple classifier combined with score obtained from Lexicons, can improve the accuracy of tweet sentiment classification. They investigate different representation of tweets that take Bag-of-words and feature hashing into account. They combine multiple classifiers to generate a single classifier. They conducted experiments on WEKA platforms to run ensemble of classifier. By considering different combinations of Bag of Words, feature hashing and lexicons, we can evaluate the potential of ensemble to boost classification accuracy. Ensemble obtained from BoW and

lexicon has provided the best results. In contrast to other approaches, very good classification accuracy rates were obtained even for small sample sizes.\

III. CONCLUSION

Twitter have the problem of slang words and misspell words while sentiment analysis has done. In this paper, I review many papers for related problem and various authors try to solve this problem with various approaches. Mainly machine learning and optimization techniques are used and categorize the tweets in positive or negative category. But still there is till no appropriate solution find out for properly categorize the tweets.

REFERENCES

[1] Aggarwal, Charu C., and Peixiang Zhao. "Towards graphical models for text processing." *Knowledge and information systems* 36, no. 1 (2013): 1-21.

[2] Bruns, Axel. "Real-Time Applications (Twitter)." Handbuch Soziale Praktiken und Digitale Alltagswelten (2016): 1-9.

[3] Cheng, Shi, Yuhui Shi, Quande Qin, and Ruibin Bai. "Swarm intelligence in big data analytics." In *International Conference on Intelligent Data Engineering and Automated Learning*, pp. 417-426. Springer, Berlin, Heidelberg, 2013.

[4] Duric, Adnan, and Fei Song. "Feature selection for sentiment analysis based on content and syntax models." *Decision Support Systems* 53, no. 4 (2012): 704-711.

[5] Elloumi, Walid, Haikal El Abed, Ajith Abraham, and Adel M. Alimi. "A comparative study of the improvement of performance using a PSO modified by ACO applied to TSP." *Applied Soft Computing* 25 (2014): 234-241.

[6] Gautam, Geetika, and Divakar Yadav. "Sentiment analysis of twitter data using machine learning approaches and semantic analysis." In *Contemporary computing (IC3), 2014 seventh international conference on*, pp. 437-442. IEEE, 2014.

[7] Ismail, Heba, Saad Harous, and Boumediene Belkhouche. "A Comparative Analysis of Machine Learning Classifiers for Twitter Sentiment Analysis."*Research in Computing Science* 110 (2016): 71-83.

[8] Kontopoulos, Efstratios, Christos Berberidis, Theologos Dergiades, and Nick Bassiliades. "Ontology-based sentiment analysis of twitter posts." *Expert systems with applications* 40, no. 10 (2013): 4065-4074.

[9] Kouloumpis, Efthymios, Theresa Wilson, and Johanna D. Moore. "Twitter sentiment analysis: The good the bad and the omg!." *Icwsm* 11, no. 538-541 (2011): 164.

[10] Lima, Ana Carolina ES, and Leandro Nunes de Castro. "Automatic sentiment analysis of Twitter messages." In *Computational Aspects of Social Networks (CASoN), 2012 Fourth International Conference on*, pp. 52-57. IEEE, 2012.

[11] Lin, Yung-Shen, Jung-Yi Jiang, and Shie-Jue Lee. "A similarity measure for text classification and clustering." *IEEE transactions on knowledge and data engineering* 26, no. 7 (2014): 1575-1590.

[12] Maharani, Warih. "Microblogging sentiment analysis with lexical based and machine learning approaches." In *Information and Communication Technology (ICoICT), 2013 International Conference of*, pp. 439-443. IEEE, 2013.

[13] Mane, Sunil B., Yashwant Sawant, Saif Kazi, and Vaibhav Shinde. "Real time sentiment analysis of twitter data using hadoop." *IJCSIT) International Journal of Computer Science and Information Technologies* 5, no. 3 (2014): 3098-3100.

[14] Neethu, M. S., and R. Rajasree. "Sentiment analysis in twitter using machine learning techniques." In *Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on*, pp. 1-5. IEEE, 2013.

[15] Pennacchiotti, Marco, and Ana-Maria Popescu. "A Machine Learning Approach to Twitter User Classification." *Icwsm* 11, no. 1 (2011): 281-288.

[16] Silva, Nadia FF, Eduardo R. Hruschka, and Estevam R. Hruschka. "Tweet sentiment analysis with classifier ensembles." *Decision Support Systems* 66 (2014): 170-179.