A Survey on Twitter Sentiment Analysis with Various Algorithms

Purva Mestry
BE Comp. Engg.
VIVA Institute of Technology,
Virar, India

Shruti Joshi
BE Comp. Engg.
VIVA Institute of Technology,
Virar, India

Sonal Mehta
BE Comp. Engg.
VIVA Institute of Technology,
Virar, India

Ashwini Save
PhD Comp. Engg.
D. J. Sanghvi college of Engineering,
Mumbai, India

ABSTRACT
The era of social networking has increased the amount of data generated by the user. People from all over the world share their opinions and thoughts on the micro-blogging sites on daily basis. Twitter is one of the most widely used micro-blogging site where people share their reviews in the form of tweets. The short and simple nature of the tweets makes it easier to use and analyze. The tweets also provide a richer and more varied content of opinions and sentiments about the latest topics. Sentiment is the feeling or attitude towards something and sentiment analysis is analyzing or studying about the various reviews given by people. The process of Sentiment Analysis tends to understand these opinions and categorize them into positive, negative, neutral.

Keywords
Social networking, micro-blogging, Twitter, sentiment analysis

1. INTRODUCTION
Sentiment analysis is a category of natural language processing for tracking the mood or review of the public about a particular product or topic. Now a days large quantity of data is available on internet, data mining is applied to collect knowledge from the data in many domains. Users express their opinions on day-to-day basis about various services or products using micro-blog posts, review sites etc. For example, while considering the movies the classification of sentiment is mainly to find out the opinion about the viewer or customer and how the director or producer can do better in their next movie. Discovering the sentiments manually from a huge data is very difficult. The other vital issue faced is that the data available may be unstructured. So, getting the sentiments from the huge unstructured data is very difficult and automatic classification of sentiment is at great demand. This automatic classification can be very useful for sentiment analysis in various applications like analysis of news sites on the Web, marketing report survey, opinion mining, system recommendation, and summarization of opinions. Classification of Sentiment can be represented as the training a classifier problem using reviews shown with the help of polarity i.e. positive, negative and neutral sentiment. This is extremely useful for both for the producers as well as the consumers to know what public think about a service or product.

Sentiment analysis over various micro-blogs faces several new challenges due to the typical short length and irregular order of such type of content. Following are some challenges faced in sentiment analysis of Twitter feeds [1]:

i. Named Entity Recognition (NER): NER is a technique to extract entities such as organization, people, and locations from tweets.

ii. Anaphora Resolution: This process resolves the issue of what a noun phrase and pronoun are defined to. For example, "Amit went to Arijit Singh's concert then for a long drive, it was amazing". For a machine, "it" is an unknown parameter.

iii. Parsing: The process that identifies a subject and object of the sentence.

iv. Sarcasm: The process of using irony for conveying contempt.

1.1 Sentiment Analysis and Opinion Mining
The study of people's point of view or emotions towards a product or an event is "Sentiment Analysis". Sentiment analysis helps to track the reputation of product or services in general. Sentiment classification can be at sentence level or document level. Document level classification needs to filter out the sentences that doesn't contain opinion words before classifying it into positive or negative. The method for classifying the phrases first extracts the opinionated text, then estimates the positions of these texts in the phrases and finally positive or negative value is assigned to the given phrase.

1.2 Twitter
There are almost 111 micro-blogging sites today over the internet. These micro-blogs are actually social media that that the people use to share their posts. Among the 111 micro-blogs, twitter is one of the most popular sites. Twitter lets the people post tweets (message) of 148 characters in length Micro-blogging websites are social media that helps users to make short and frequent posts. As one tweet only consists of 148 characters, it makes the process of sentiment analysis easier.

2. SENTIMENT CLASSIFICATION TECHNIQUES
The main twitter sentiment classification techniques are Support Vector Machines (SVMs), Naïve Bayes Classifier, Fuzzy logic, Baseline Model, Feature vector approach, and Hybrid approach.

2.1 Support Vector Machines (SVM)
Support Vector Machines are widely used algorithms since 1960s. SVMs are the set of algorithms that are used for pattern recognition. SVM algorithms are well-known and powerful classification learning tool. Different model can be built using SVMs. Three models that can be developed using these algorithm for machine learning are unigram Model,

feature based model and tree kernel model. Feature extraction can be done using these models.

2.2 Baseline Model

In baseline model, initially the preprocessing steps are carried out study the polarity frequencies of unigrams, bigrams, and trigrams in the training data set. Three probability score is given to each token: Neutral probability, Positive probability and negative probability. A feature vector is then created for the tokens that can differentiate the tweet's sentiment effectively. Before the probability values are calculated the infrequent words are filtered out, this serves the baseline model. For example, Emotion Determiners present with value 1 indicates its presence in the text and 0 indicates its absence in the text. Various features are appended to this model after building it.

2.3 Fuzzy Logic

Fuzzy logic is used to draw and retrieve sentiments from a text or document. Fuzzy logic uses the concept of reasoning that gives results in approximation rather than exact results. Fuzzy logic is useful for managing such approximate information. The numerical score of sentence is evaluated between the range from 0 to 1.

2.4 Corpus and Dictionary Based

In dictionary based approach, first the seed words of opinions are searched and then it looks for their synonyms and antonyms. Some of the opinion words are listed manually. The list is then expanded by searching into popular or well-known corpora like WordNet. The strength of polarity is also listed in the dictionary for each word. The corpus based approach is use to find opinion words with context-specific orientations which depends on syntactic patterns.

3. DIFFERENT RESEARCHES ON SENTIMENT ANALYSIS

3.1 Sentiment Analysis on Twitter Data

V. Sahayak, V. Shete and A. Pathan [3] proposed in 2015 about "Sentiment Analysis on Twitter Data" suggested the hybrid approach that classifies the tweets from twitter dataset in sentiment categories like positive , negative, and neutral. The two techniques used in this approach are corpus based, dictionary based sentiment classification, and it includes POS for polarity features and tree kernel to avoid monotonous features. The feature extractor & different machine learning classifier are explained and used in this methodology. The machine learning classifiers are Naïve Bayes, Support Vector Machines (SVM), and Maximum Entropy. These classifiers are used to make three models for feature extraction process namely unigram model, tree kernel model, feature based model. They developed the process of sentiment analysis of tweets, which contains three sections. First section is data extraction, which helps to extract opinion words from tweets. Second section does preprocessing of all the extracted words, which includes emoticons handling, filtration, tokenization, removal of stop words, n-gram construction. Third section classifies the sentiments using machine learning classifiers. This section works in two steps: 1. Model construction. 2. Model usage to check accuracy of classification. When complexity of emoticons and opinions increases, it becomes difficult for this approach to give right answer. This can be a drawback. For example, "The product was awesome but the services were gruesome". In this case, this approach may get confused for the result of sentiment.

3.2 Sentiment Analysis using Fuzzy Logic

Md. A. Haque &T. Rahman[1] proposed in 2014 "Sentiment Analysis with the help of Fuzzy Logic" by ranking the review in terms of positive and negative is the ranking perspective and it is achieved using fuzzy logic. The need of sentiment analysis is based on the two sectors i.e. classification of documents according to the orientation of sentiments such as positive and negative, other sector is gathering information by identifying the subjective or objective (SO) polarity of the comment or post, identifying the positivity or negativity (PO) polarity of comment or post and by identifying the degree of PN-polarity in terms of good, better or best. The tool to determine the polarity of lexical (the sentence is converted into sequence of tokens) is SentiWordNet. This gives numerical score to token range from 0 to 1. By having the values of the post and the weights,the result can be computed by calculating the weighted and arithmetic mean, from that percentage of the individual sentiment(subjective &objective) is declared and by using concept of normalization the results are present in better way.

3.3 Sentiment Analysis on Twitter

A.Kumar and T. M. Sebastian[5] proposed in 2012 about "Sentiment Analysis on Twitter" developed hybrid approach using dictionary based and corpus based method to calculate semantic score of the opinion words in tweets. This approach uses the features like capitalization, emoticons, etc. while preprocessing the tweets. The approach more focuses on opinion words which must be a combination of adjectives and verbs. In this hybrid approach, dictionary based method is used to find semantic score of adverb, and verb. And corpus based method is used to find sematic score of adjectives. The list of adverbs and verbs along with their sematic sterngths ranging from -1 to1 are taken into consideration while calculating semantic score of adverbs and verbs. The varying semantic strenghts of words provides high accuracy while handling multiple opinions and emoticons. For example, "very good" will get more semantic strength than "good" . The negation handling is also achieved by this approach and it is accurate. The linear equation which is the highly focused part of this approach is used to calculate overall semantic score of single tweet. This linear equation calculates semantic score of each tweet more accuratly by considering uppercase tweets, repeated letters, exclamation marks, emoticons, adjective group, verb group. According to semantic score of tweet, tweet is classified into three categories namely positive, negative, neutral.

3.4 A Fuzzy Logic Based on Sentiment Classification

J.I.Sheeba and Dr.K.Vivekanandan[5] proposed in 2014 "A Fuzzy Logic Based on Sentiment Classification" which says that,fuzzy logic is a type of probabilistic logic and it deals with reasoning that is approximate rather than fixed. Fuzzy logic is used for dealing with heterogeneous or vague information. Traditional logic may have many values but fuzzy logic can have values that range from 0 to 1. The input to the fuzzy logic is from sentiment classification step. The weights are assigned for each word. Based on the weight of the word, the "Threshold" value is calculated. The threshold value is calculated based on the average of each listed word. Finally, the list of positive, negative and neutral words are listed which is greater than or equal to the Threshold value. The algorithm is Fuzzy C-means algorithm use in this method which gather all the same words to reduce the emotions list and group the emotions based on the cluster centroid. The

topic of particular category which is input is identified by the framework. The main motive of the method is to return the reduced and accurate emotions list. The five major steps are:

i. To classify the both implicit and explicit emotions includes data processing and sentiment classification.

ii. Apply fuzzy logic for sentiment classification.

iii. Implement Fuzzy C-means algorithm.

iv. To generalize author classification i.e identification and characterization using POS(Part of Speech) tagger or Qtag Tool.

v. Finally identify the topic.

The evaluation of the result is done by considering metrics in term of quality measures name as precision, recall and F-Measure.

3.5 Sentiment Analysis and Opinion Mining

Y. Sharma, V. Mangat and M. Kaur [4] proposed in 2015 about "Sentiment Analysis & Opinion Mining" that suggested various approaches based on which the sentiments can be analyzed.

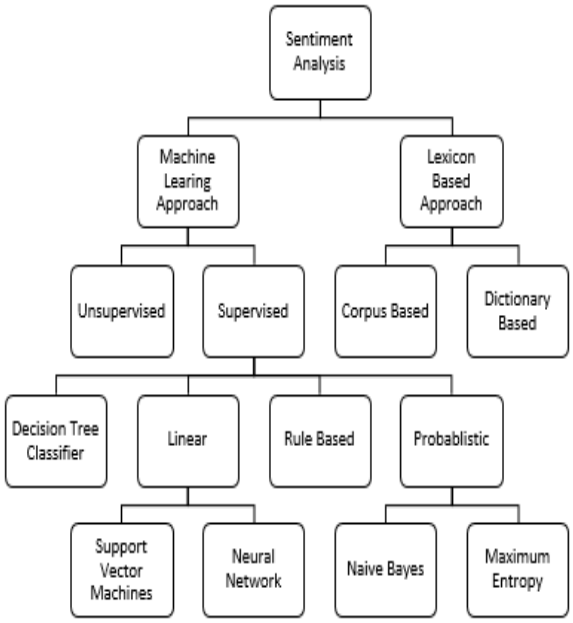

Fig. 1: Sentiment Classification Technique

The different approaches for sentiment analysis are shown in figure 1. It also includes the feature selection method, which reduces the unconnected information. It enhances the classification accuracy, and also decreases the running time of algorithm. The selection step is to remove the target, stop words, URL, & stemming. The two broad ways in which the sentiment analysis is done are Machine learning approach and Lexicon based approach. Machine learning approach learns from the previously generated results whereas the Lexicon based approach is usually fixed and gives approximated results.

3.6 A Hybrid Approach for Twitter Sentiment Analysis

N. Mittal & B. Agarwal [7] proposed "A Hybrid Approach for Twitter Sentiment Analysis" which is a three stage hierarchical model for sentiment extraction, in first stage the emoticons are labeled, then tweets are assigned sentiments using pre-defined lists of words with polarity and finally based on subjectivity of lexicon, the proposed probability based method assign the weight to all the tokens. The lexicons are weighted using various approaches like SentiWordNet, proposed probability based method, SentiWordNet (SWN) then probability based method, probability based method then SentiWordNet or Hybrid approach. The accuracy measured for hybrid approach was comparative higher than all the approaches i.e. 72.563 %. Hybrid method uses both SWN and probability based method to calculate the polarity of the token. Hence proposed hybrid approach improves the sentiment classification accuracy.

3.7 Opinion Mining of Real Time Twitter Tweets

A. Shrivatava, S. Mayor and B. Pant [8], proposed "Opinion Mining of Real Twitter Tweets," In this proposed system, a tweet puller is developed which automatically fetch the public opinion on a topic and using SVM the opinions are classified into positive, negative and neutral. First, tweets are collected using twitter API then creating domain specific dictionary. Extracting all the tweets from Twitter when server is connected is done by tweet puller. Using classification tool to generate threshold frequency for each feature and generate a text file. Text file is input to LIBSVM tool, which is proposed to provide accurate rate for testing the classification.

3.8 Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis

L. Zhang, R. Ghosh, M. Dekhil, M. Hsu, and B. Liu [9], proposed "Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis". The system is a new entity-level sentiment analysis approach for Twitter, which is done using lexicon based method, the input preprocessed tweets are analyzed and categorize into sentence type detection, coreference resolution, using opinion rule aggregate opinions are formed which is input to train sentiment classifier that is Learning-based method and finally the extract opinionated tweets are classified. Coreference resolution gives the closest entity. For example, "Amit went to Arijit Singh's concert then for a long drive, it was amazing". "it" can be resolved by considering the closest entity that is "long drive". This system gives high precision, recall, and F-score.

3.9 Twitter Sentiment Analysis The good, the bad and the neutral

Ayushi Dalmia [12] proposed "Twitter Sentiment Analysis The good, the bad and the neutral!". In this system, the lexicon based feature is further augmented by tweet specific features. The system includes English dictionary, acronym, and emoticon dictionaries. The preprocessing includes tokenization, removing non-English tweet, replace emoticons, remove URL, remove target mentions, remove punctuations from hash tags, handling sequences of repeated characters, removing numbers, removing nouns and prepositions, removing stop words, handle negative mentions and expand acronyms. After preprocessing & feature extraction, the tweets are feed into a classifier; it concludes that SVM gave the best performance. Hence, by building supervised system which merge lexicon based feature with tweet related features classify the tweets into 3-way classification-positive, negative and neutral.

3.10 Mining Sentiments from Tweets

A. Bakliwal, Piyush Arora, Senthil Madhappan, Nikhil Kapre, Mukesh Singh, Vasudeva Varma [10] proposed "Mining Sentiments from Tweets". In this system, the method for sentiment analysis is used on Stanford dataset & Mejaj dataset with achieves 88% accuracy. The feature vector approach is used to form feature using unigrams, bigrams, hash tags (#), targets (@), emoticons, special letters, & semi-supervised SVM classifier. The feature is distinguishing into Twitter specific and NLP feature. This approach is very useful when the user has to extract maximum information out of small content. It includes Emotion and Punctuations handling, spell correction, stemming, stop word removal using unigram model, noun indication, and finally score of the tweet is based on all the factors.

4. COMPARATIVE STUDY

This comparative study mainly based on above mentioned sentiment classification techniques. Table 1 shows studies done in different research papers on various classifiers and percentage accuracy given by those classifiers in year 2011 to 2015.

Table 1: Mining Techniques and their accuracy in different research papers

Studies	Mining Techniques Used	Performance (Accuracy)
V. Sahayak, V. Shete , A. Pathan[3]	Naïve Bayes Classifier or Support Vector Machines (SVMs).	Better accuracy
Md. Ansarul Haque1 , Tamjid Rahman[2]	Fuzzy logic	Moderate accuracy
J.I.Sheeba, Dr.K.Vivekanandan[6]	Fuzzy logic	85%
A. Kumar, T. M. Sebastian[4]	Hybrid approach with Corpus based and dictionary based method	80%
N. Mittal, B. Agarwal[7]	Hybrid approach Subjectivity lexicon and probability	71.12%
A. Shrivatava, S. Mayor and B. Pant[8]	SVM	70.5%
L. Zhang et al. [9]	Hybrid approach	85.4%
A. Bakliwal et al. [10]	Baseline method on Stanford Dataset	87.2%
	Feature vector approach on Stanford Dataset	87.64%
A. Dalmia, M. Gupta, V. Varma [12]	Baseline Model	59.83%
P. Chikersal, S. Poria, E. Cambria [11]	SVM	71.5%
	SVM with a rule-based layer.	72.3%

From above study, Baseline method & Feature Vector Approach method on Standford Dataset found to be better mining technique for sentiment analysis. This system is improvised version of SVM (Support Vector Machine) that is semi supervised SVM classifier. It includes certain rules which can enhance the analysis mechanism by handling emoticons and punctuation, spell correction, stemming & stop word removal using unigram. In future, such system should be made to detect the problems like Sarcasm, irony, humor, mixed feeling, Named Entity Recognition (NER), Anaphora Resolution, conflicting signals, deliberate spelling mistakes, and Parsing. The Support Vector Machine has better accuracy when ruled based approach is used with it. The system should distinguish the text by analyzing whether the information is an opinion or just a fact. The domain sentiment analysis can be very efficient if amalgamated with other domains like fuzzy logic, speech recognition, and Artificial Intelligence.

5. CONCLUSION

Many of the organizations are putting their efforts in finding the best system for sentiment analysis. Some of the algorithms give good results but still many more limitations in these algorithms. As the twitter users are increasing day by day and the posts shared by the people are short messages (tweets) it can be very useful to analyze its data set. There are many techniques developed to do sentiment analysis but the problems sarcasm is still not solved. The traditional way is very complex and time consuming but the recent approaches mentioned in this survey paper are quite simpler and efficient. More researches are done using SVM classifier, and also improvising its efficiency by introducing new rules and solving parsing problem. Fuzzy logic helps sentiment analysis provide efficient results as it is based on reasoning on the approximate values. Sentiment analysis when used with fuzzy logic helps to take decisions effectively but sometimes it may differ from the real time values. Future work may combine many different types of techniques to overcome individual's limitations, benefit from each other's merit, and measure the performance of classification technique.

6. REFERENCES

[1] Geetanjali S. Potdar, Prof R. N. Phursule, "A Survey Paper on Twitter Opinion Mining", International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064, Volume 4 Issue 1, January 2015, pp. 19-21.

[2] Md. Ansarul Haque1,Tamjid Rahman 2, "Sentiment Analysis By Using Fuzzy Logic", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol. 4,No. 1,February 2014, pp. 33-48.

[3] Varsha Sahayak, Vijaya Shete, Apashabi Pathan, "Sentiment Analysis on Twitter Data", International Journal of Innovative Research in Advanced Engineering (IJIRAE), Issue 1, Volume 2, January 2015, pp. 178-183.

[4] Yakshi Sharma, Veenu Mangat, And Mandeep Kaur, "Sentiment Analysis & Opinion Mining", Proceedings Of 21 St Irf International Conference, 8 Th March 2015, Pune, India, Isbn: 978-93-82702-75-7, Pp. 35-38.

[5] Akshi Kumar and Teeja Mary Sebastian, "Sentiment

Analysis on Twitter", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 4, No 3, July 2012, pp. 372-378.

[6] J.I.Sheeba and Dr.K.Vivekanandan, "A Fuzzy Logic Based On Sentiment Classification", International Journal Of Data Mining & Knowledge Management Process (Ijdkp) Vol.4, No.4, July 2014, pp. 27-44.

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

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

[9] 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.

[10] Akshat Bakliwal, Piyush Arora, Senthil Madhappan, Nikhil Kapre, Mukesh Singh, Vasudeva Varma, "Mining Sentiments from Tweets", 3rd Workshop on Sentimentand Subjectivity Analysis (WASSA), Report No: IIIT/TR/2012/-1,July 2012.

[11] Prerna Chikersal, Soujanya Poria, and Erik Cambria "SeNTU: Sentiment Analysis of Tweets by Combining a Rule-based Classifier with Supervised Learning", Proceedings of the 9th International on Workshop Semantic Evaluation (SemEval 2015), Denver, Colorado, June 4-5, 2015, pp. 647-651.

[12] Ayushi Dalmia, Manish Gupta_, Vasudeva Varma, "IIIT-H at SemEval 2015: Twitter Sentiment Analysis The good, the bad and the neutral!", Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Denver, Colorado, June 4-5, 2015, pp. 520-526.

[13] Efthymios Kouloumpis, TheresaWilson, Johanna Moore, "Twitter Sentiment Analysis: The Good the Bad and the OMG!" Proceedings of the International AAAI Conference on Weblogs and Social Media, 2011, pp. 538-541.

[14] Maqbool Al-Maimani, Naomie Salim, Ahmed M. Al-Naamany, "Semantic and Fuzzy Aspects of Opinion Mining", Journal of Theoretical and Applied Information Technology Vol. 63 No.2, 20th May 2014, pp. 330-342.

[15] Alexander Pak, Patrick Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining", pp. 1320-1326.