A Study on Sentiment Analysis using Tweeter Data

Suprajha S
UG[BE] Scholars
Department of Computer Science & Engineering
BMSCE, Bangalore

Yogitha C
UG[BE] Scholars
Department of Computer Science & Engineering
BMSCE, Bangalore

Architha J Sanghvi
UG[BE] Scholars
Department of Computer Science & Engineering
BMSCE, Bangalore

Dr. H S Guruprasad
Professor and Head
Department of Computer Science & Engineering
BMSCE, Bangalore

Abstract

Twitter is a popular micro blogging service where users create status messages (called "tweets"). Tweets are short messages with a maximum length of 140 characters. The distinguishing characteristics of tweets are hashtags. Hashtags are used for logically grouping tweets and searching them. Authors of those messages write about their life, share opinions on variety of topics and discuss current issues. As more and more users post about products and services they use, or express their political and religious views, micro blogging websites become valuable sources of people's opinions and sentiments. Such data can be efficiently used for marketing or social studies. As a result, there has been a tremendous need to design methods and algorithms which can effectively process a wide variety of text applications.
Keywords: Tweeter, Sentiment Analysis, Text Data, Classifier

I. INTRODUCTION

The problem of text mining has gained increasing attention in recent years in recent years because of large amounts of text data, which are created in a variety of social network, web and other information-centric applications. Unstructured data is the easiest form of data which can be created in any application scenario. With the explosive growth of social media - reviews, forum discussions, blogs, social networks on the Web, individuals and organizations are increasingly using public opinions in these media for their decision making.

- An enormous amount of content is generated in twitter each day.
- Twitter has users from varied backgrounds ranging from students, companies, celebrities, politicians etc. Thus, by using Twitter, the sentiments of varied groups can be captured.

Sentiment analysis systems are being applied in almost every business and social domain because opinions are central to almost all human activities and are key influencers of our behaviors. There has been considerable research done. The following is a review of all relevant previous work.

II. LITERATURE SURVEY

Glivia et. al. [1] assesses the usefulness of twitter hashtags in sentiment analysis. They analyzed 10,173,382 tweets related to the Brazilian Presidential elections in 2010. They analyzed these tweets and observed that the positive behavior of the tweeters across time was in accordance to the hypothesis that hashtags sentiments match the overall population sentiment. They also verified that the information propagation in twitter follows a cascade model where people make their decisions consciously or not, based on someone else's sentiments and choices. Alec Go et. al. [4] introduced a method for classifying twitter messages. First, the query term is normalized so that the query term by itself is not biased. Positive and negative tweets are collected by using ":) / :-) "And ":(/ :-(". 80,000 positive and 80,000 negative tweets were collected as training set. The tweets are then preprocessed. The emoticons are then stripped from the training data because emoticons have a negative impact on the accuracies of SVM and Maximum Entropy classifiers but little effect on Naive Bayes classifier. They explored the usage of unigrams, bigrams, unigrams and bigrams, and Parts of Speech features. The following were the accuracies observed using different classifiers.

Table – 1
Accuracies observed using different classifiers

	Naive Bayes	*Maximum Entropy*	*SVM*
Unigram	*81%*	*80.4%*	*82.9%*

Bigram	*Not useful*	*Not useful*	*Not useful*
Unigram and Bigram	*82.7%*	*82.7%*	*81.6%*
Parts of speech	*81.5%*	*81.9%*	*80.4%*

Apoorv Agarwal et. al. [8] have built models for two classification tasks: a binary task of classifying sentiment into positive and negative classes and a 3-way task of classifying sentiment into positive, negative and neutral classes. Experimentation is done with unigram model, feature based model and tree kernel based model. For the tree kernel based model they designed a new tree representation for tweets. They used the unigram model for sentiment analysis for Twitter data, as their baseline. Their experiments show that a unigram model is indeed a hard baseline achieving over 20% over the chance baseline for both classification tasks. Their feature based model that uses only 100 features achieves similar accuracy as the unigram model that uses over 10,000 features. Their tree kernel based model outperforms both these models by a significant margin. They also experimented with a combination of models: combining unigrams with features and combining features with the tree kernel. Both these combinations were found to outperform the unigram baseline by over 4% for both classification tasks.

Alexander Pak et. al. [6] used Twitter API to collect twitter datasets in three categories namely positive, negative and neutral. A corpus analysis is performed by checking the distribution of word frequencies in the corpus. The plot was found to obey Zipf's law and hence confirmed a proper characteristic of the collected corpus. The corpus analysis is done in one more method. The parts of speech of each word are tagged using TreeTagger (Schmid, 1994) and interpreted the difference of tag distributions between sets of text (positive, negative, neutral or subjective, objective). The sentiments are classified using both Naive Bayes and SVM classifiers. Naive Bayes was found to perform better. Two Bayes classifiers were trained, one based on n-gram and another based on parts of speech distribution. To increase the accuracy some common n-grams are discarded that do not strongly indicate any sentiment. The best performance was achieved when a bigram was used. Oshini Goonetilleke et. al. [21] address the issues around Big data nature of twitter and the need for new data management which limits the use of existing systems. For this reason, the system basically involves components like Focused Crawling which is used in effective retrieval and better coverage (twiiterecho is an open distributed crawler for twitter), pre-processing of tweets can be done using Tokenization and Stemming. Twitterzombie is a platform used for gathering the data and analyzing. To provide scalability and efficiency of processing large amounts of data, Tred Miner can be used for real time analysis of tweets. Tweeql allows the resulting tweets to be collected in batches and then stores them in relational data base for further identification of sentiments. Basically, one general platform is presented for twitter using similar systems. Martha Arias et. al. [23] have focused on study of hashtag level of sentiment classification. The main aim of this task is to automatically generate the overall sentiment for a given hashtag in twitter. It is found that out of 0.6 million randomly selected tweets 14.6% contained at least one hashtag. A two stage SVM classifier is used to determine the polarity of sentiment and finally, it is found that performance can be invariably increased by Boosting loopy belief propagation with accuracy prediction up to 77.72%. A simple baseline approach is developed on the results of tweets using simple voting strategy where a binary value was given to corresponding positive, negative and neutral tweets.

Po-Wei Liang et. al. [3] use Twitter API to collect twitter data. The training data falls in three different categories (camera, mobile phone and movie). The data is labeled as positive, negative and non-opinions instead of utilizing data that contains emoticons to identify the sentiment. This is because emoticons are not always consistent with the sentiment. Pre-processing is done on the considered tweets. In the next step, the tweets containing opinions are filtered. This is done by using Naive Bayes classifier on the training set. Unigram Naive Bayes model is implemented and the Naive Bayes simplifying independence assumption is employed. In the next step, useless features are eliminated by using the Mutual Information and Chi square feature extraction method. Next, the orientation of an opinion sentence is predicted. i.e. positive or negative. It is found that tweets are classified as opinionated and non-opinionated with a 76.8 %. Using feature selection, an accuracy of 96.6% is obtained. The test data set is classified with an accuracy of 90.17% with the focus on only positive and negative data. Another training set is generated using emoticons to denote positive or negative tweets. Using the emoticon trained data set, the tweets are classified with an accuracy of 58.65%. So, it was concluded that using the emoticons to collect training data is not always accurate. Bowick et. al. [14] deals with distant supervision for topic identification of tweets and how this can be applied to analyzing Presidential Job Approval ratings from Twitter data.

1) The authors obtained seven months of tweets which contained around 476 million tweets labeled with usernames and timestamps, collected through the Twitter API.
2) Tweets were aligned with polling data using their timestamps.
3) Two trend lines for approval and disapproval ratings were created. The positive and negative sentiment scores were compared against these two trends.
4) Their results outperform previous work on Presidential Job Approval prediction (O'Connor 2010). They presented two novel approaches to the domain: a coupled distantly supervised system, and a topic-neutral baseline, both of which outperformed previous results.

Spencer et. al. [25] has presented a web based tool Sentimentor to classify live Twitter data into positive negative and objective tweets. It has an interface that allows users to easily analyze the word distributions and pictorial representation of the sentiments in tweets. Twitter API is used for data extraction process. The collected tweets were pre-processed. The POS tagging of each word was done and unigrams and bigrams were extracted. The Naive Bayes algorithm was used for classification with POS tags, unigrams and bigrams as features. It is found that the best accuracy of 52.31 % was obtained by using bigram without POS tagging. The use of POS tags has had a negative effect on the accuracy probably because of the use of summation of POS tags across a phrase rather than considering binary occurrences.

Barbosa et. al. [7] have presented effective sentiment detection approach for Tweets. They have used emoticons as noisy labels. The performance can be accredited to the following features: (1) An abstract representation of tweets is used instead of tweets as such and (2) the data source provides labels that are beneficial. The main disadvantage is when tweets have contradicting emotions. Modha et. al. [30] have developed a tool for automatic sentiment analysis for unstructured data. It talks about Big Data, its scope and challenges and opinion mining in big data. Instead of following the traditional approach of considering only subjective sentences and ignoring the objective ones, the paper considers both of them to be equally important. Two sets are considered - a document set D= {d1, d2, ..., dN}, and a sentence set S= {S1, S2, ..., Sn}. In the first step, sentences or sentences of documents are classified into two categories Opinionated and Non Opinionated, regardless whether it is subjective or objective. In the second step, opinionated sentences are classified as subjective sentences and Objective sentences. And in the third step, subjective sentences are classified into positive, negative or neutral sentences. For complex sentences, context and semantic orientation are attached. Finally, the fourth step involved classifying objective sentences into positive, negative or neutral.

Stephan Winkler et. al. [18] have presented an Ensemble modeling approach for sentimental analysis using Machine learning algorithms. The approach presented relies on analysis of words found in sentences and formation of Heterogeneous models (i.e. Binary as well as Multi classification) that are calculated using machine learning methods. Applied machine learning techniques used are decision trees, random forest, neural networks and K-nearest neighbor and are used with boosting algorithm to increase the accuracy of prediction rule. For classifying sentiments based on positive and negative, Gaussian process is used. SVM's were used to select the best models from a set of models which can be increased by 60.4% and the ratio of wronged samples can be decreased up to 7.2% only if the threshold values is set to 1.0. Tamilselvil et. al. [22] have considered basically two domains stock market and movie box revenue for which a decision tree is to be presented called summary tree. The study of sentiment analysis is done in three stages namely collection, cleaning and pre-processing. For the removal of irrelevant data, a probabilistic model called Latent Dirchlet Allocation is considered. Three datasets for stock market are considered. The stock market domain used the nonlinear models (SVM's and neural networks) along with sentiment indices did really predictions well whereas with movie box revenue with no sentiment indices did not do any specific predictions at all. A large volume of tweets along with SVM'S are considered with other linear models to increase the forecasting of predictions. Xiaolong Wang et. al. [24] have performed a linguistic analysis on collected data and discovered sentiments of it. A proposed sentiment classifier is built that is able to determine the sentiments in document. Here, basically the emoticons are used to indicate the user's moods. They used SVM's and CRF learners to classify sentiments, SVM and Naive Bayes were able to obtain at least 70% accuracy together. Bayesian learning and turning opinion mining was used as models for sentiment classifiers. It is found that use of sentiment topics provided better predictions than semantic features. The sentiment topics provide accuracy up to 80.2% when compared with Naive Bayes used along with sematic topic features. It is concluded that sentiment topics give more accurate predictions even with less features.

Theresa et. al. [2] have observed that using a prior priority alone for classifying contextual polarity of phrases gives an accuracy of only 48%. They analyzed that 76% of the errors resulted from using words with non-neutral prior polarity being used in neutral contextual polarity in the phrases. So, a two-step approach is proposed to classify such words. The first step used is a Neutral-Polar classifier which tries to determine if an instance is neutral or polar in context. Four types of learning algorithms, Boosting, Memory based learning, Rule learning and Support Vector learning are used. The performance of the neutral-polar classifier is compared based on two baselines. The first baseline uses just word tokens. The second baseline uses both word tokens and polarity. The second step is the Polarity Classification where the polarity of all clues identified as polar in step one are classified. Mandel et. al. [9] examined the response to the natural disaster Hurricane Irene on Twitter. They collected about 65,000 tweets over a span of 2 weeks and grouped them by location, gender in order to analyze them. Based on the level of concern, a sentiment classifier was trained to categorize the messages. First, the tweets were manually classified, then the classifier was trained on that data and then this classifier was used to label unlabeled data. The paper finds that the number of tweets from a particular region is directly proportional to the time hurricane hits that region and level of concern on the days leading up to the hurricane's arrival is dependent on the region. Pak et. al. [12] presented a novel approach to collect a dataset with positive and negative sentiments, and a collection of objective texts (with no sentiments). This method allowed them to collect negative and positive sentiments such that no human effort was needed (and probably won't be needed) for classifying the documents. Objective texts were also collected automatically. A statistical linguistic analysis of the collected corpus was performed. The collected corpus was used to train a sentiment classifier. Experimental evaluations on a set of real microblogging posts were conducted to prove that this technique is efficient and performs better than previously proposed methods.

Luo et. al. [13] explains the challenges and an efficient technique to mine opinions from Twitter messages. Spam and wildly varying language makes opinion retrieval within Twitter challenging. They developed a unique ranking function which used social features and opinionated features of tweets for better opinion retrieval. They found out that their method optimized the BM25 baseline and the VSM baseline by improving MAP by 56.82% and 33.75% respectively. Their ranking model could achieve comparable performance with a method using manually tagged tweets. The experimental results showed their approach is still effective for opinion retrieval with TREC Tweets2011 dataset. Davidov et. al. [15] experimented with semi-supervised sarcasm identification on two different data sets consisting of 5.9 million tweets and 66000 product reviews from Amazon. A robust algorithm called SASI is used for recognition of sarcasm, to experiment with the tweets and reviews from Amazon dataset. Evaluating in various ways and with different parameters they achieved high precision on both datasets. The authors intend to design and develop a sarcasm recognition system for review ranking, summarization systems and brand monitoring systems. Sentiment analysis requires a dictionary or lexicon for efficient classification of tweets. Rao et. al. [16] discuss about building an emotional dictionary for efficient sentiment analysis of news on twitter. Each word in the news can be compared to the words in the emotional dictionary and then can be classified according to the sentiments. They presented various algorithms like Acc@1 algorithm and AP algorithm for constructing a word level and topic level emotional dictionary. This approach is different from previous methods because it is automatic, language independent and unlimited in volume. Three pruning strategies are developed for effective refining of the emotional dictionary. One drawback of their method was that it is less effective on news headlines. Shawn O'Banion et. al. [19] presented a novel approach in mining preference data from natural language experience in social media. The approach considered the voting decision of twitter users in 2012 US election. Statistical analyses model were used to predict election outcomes and campaign decision. The main goal was to know the preferences about the people who never expressed their opinions. All models were compared against 2 simple baselines. The first predicted Obama if the user used the #Obama2012 or #Romney2012, the second predicted if the user is friends with official @BrackObama twitter or @Romney. On the day of elections the baselines performed surprisingly well.

Pablo et. al. [5] presented a family of Naive Bayes classifiers for detecting the polarity of English tweets. Two different Naive Bayes classifiers have been built namely Baseline (trained to classify the tweets as positive, negative and neutral), and Binary (makes use of a polarity lexicon and classifies as positive and negative. Neutral tweets are not considered). The features considered by the classifiers are Lemmas (nouns, verbs, adjectives and adverbs), Multiword, and Polarity Lexicons from different sources and Valence Shifters. The training data set of tweets is obtained from SemEval Organization-2014 and additional annotated tweets from external sources. Many combinations of the above mentioned strategies and features are implemented. It is also concluded that performance is best when binary strategy is used with multiword and valence shifters features. Marc Egger et. al. [17] have investigated about text based User-Generated-Content which is useful and relevant for corporate companies. They have considered the processes along three stages collection, analysis (positive, negative or neutral) and visualization, along with top down approach is used beginning with techniques operating on document and drilling down to lower levels of information extraction. Here topic modeling techniques such as Probabilistic Latent Semantic Analysis or Latent Dirchilet allocation are considered, which can be used to uncover abstract topics within document. Construction of Decision trees is found to be more Complex for abbreviation dictionaries for this Naïve Bayes can be used to detect the end of sentence. To map each word of text onto parts of speech tagger Distance based approach is used.

Kumar et. al. [26] retrieve twitter data using Twitter API. They preprocess the tweets and add weightage according to the number of exclamation marks and the adjectives, verbs and adverbs are tagged in each tweet. Emoticons are replaced by their polarity. Adjectives and negative words are taken into account to calculate the polarity of the whole phrase. Polarity of the tweet is calculated based on a formula. The system proposed had characteristics of perceiving the sentiments in tweets.

Kouloumpis et. al. [27] performed sentiment analysis using Twitter hashtags (e.g., #bestfeeling, #newphone, #androidwhat) to identify positive, negative, and neutral tweets. They used three different corpora of Twitter messages in their experiments- a hash tagged data set, an emoticon data set and a manually annotated data set ISIEVE. Their goal for these experiments was two-fold. First, they wanted to evaluate whether the training data with labels derived from hashtags and emoticons is useful for training sentiment classifiers for Twitter. Second, they wanted to evaluate the effectiveness of the tweets after pre-processing for sentiment analysis in Twitter data. It was concluded that the experiments on twitter sentiment analysis show that part-of-speech features may not be useful for sentiment analysis in the microblogging domain. Davidov et. al. [28] proposed a supervised sentiment classification framework which used data from Twitter. By utilizing 50 Twitter tags and 15 smileys (☺,☹, :-o) as sentiment labels, this framework avoids the need for labor intensive manual annotation, allowing identification and classification of diverse sentiment types of short texts. The paper also evaluated previously developed frameworks for sentiment classification and showed that their framework successfully identified sentiment types of sentences that were untagged. The quality of the sentiment identification was also confirmed by actual manual classification and cross validation. Dependencies and overlap between different sentiment types represented by smileys and Twitter hashtags was also explored. In this study, four different feature types (punctuation, words, n-grams and patterns) for sentiment classification were used and the contribution of each feature type for this task was evaluated. Narr et al., [29] examine a language independent sentiment classifier. The tweets in four languages, English, French, German and Portuguese are considered. Emoticons as labels are used to collect training data because they believed that in short messages like tweets, the emoticons are often consistent with the overall sentiment of the tweet. Naive Bayes classifier is used from NTLK Natural Language Processing Toolkit. 10000 tweets are used for testing which was manually annotated using the help Mechanical Turk workers into positive, negative, neutral or irrelevant. They used various combinations of the classifier. With Unigram classifier an accuracy of 81.3% was achieved in English tweets. Best accuracies observed for the four languages are: English - 81.3%, French - 74.9%, German - 79.8% and Portuguese - 64.9%.

Krushikanth et. al. [10] intend to apply data mining tools to generate interesting patterns for predicting box office performance of movies using data collected from multiple social media and web sources including Twitter, YouTube and the IMDb movie databases. The box office predictions are based on the factors derived from a historical movie database, count of the Twitter followers and sentiment analysis of the comments of YouTube viewers'. The movies are identified as Hit, Neutral or Flop using Weka's K-Means clustering tool specifically Weka's J48. Burnap et. al. [11] built models that predicted information flow size and survival on Twitter following the terrorist event in Woolwich, London (2013). Information flow is the propagation of information over time posted on Twitter via the action of retweeting. After several trial and errors, they decided to use the zero-truncated negative binomial (ZTNB) regression method. Cox regression technique was used to model survival because it best estimates proportional hazard rates for independent measures. From about 427,330 tweets they analyzed that the sentiment expressed in the tweet is statistically predictive of both size and survival of information flows. Other press related information that the people were subjected to also was a significant predictor of size. Using the results from this paper, the authors intend to test the predictive efficiency of this model on other cases that exhibited similar characteristics. Makazhanov et. al. [20] study the problem of predicting the political preference of users on twitter and showing the predictions of users interacting with other parties. The paper evaluated the work on Alberta state elections and showed that model performed on the basis of F-measure and Sentiments. The goal is to predict which party, if any user is likely to vote for, for this a party is trained with binary classifier and it provides corresponding confidence value. The polarity of sentiments can be identified using Supervised and Unsupervised methods. They added a new feature to know about two kinds of users (i.e. spammer and non-spammers). It is predicted that logistic regression did better than SVM's when spammers were included who had lower F-measure.

III. CONCLUSION

The amount of text data generated by Tweeter is increasing. The sentiment analysis on this text data is important in many aspects. This paper studies the different approaches for semantic analysis for tweeter data.

ACKNOWLEDGEMENT

The work reported in this paper is supported by the college through the TECHNICAL EDUCATION QUALITY IMPROVEMENT PROGRAMME [TEQIP-II] of the MHRD, Government of India.

REFERENCES

[1] Glivia A. R. Barbosa, Wagner Meira Jr, Ismael S. Silva, Raquel O. Prates, Mohammed J. Zaki, Adriano Veloso, "Characterizing the Effectiveness of Twitter Hashtags to Detect and Track Online Population Sentiment", ACM SIGCHI Conference on Human Factors in Computing Systems, Juried Works-in-Progress, May 2012.
[2] Theresa Wilson, Janyce Wiebe, Paul Hoffmann, "Recognizing Contextual Polarity: An Exploration of Features for Phrase-Level Sentiment Analysis", Journal of Computational Linguistics archive, Volume 35, Issue 3, Sept 2009, pp 399-433, DOI:10.1162/coli.08-012-R1-06-90.
[3] Po-Wei Liang, Bi-Ru Dai, "Opinion Mining on Social Media Data", IEEE 14th International Conference on Mobile Data Management, Milan, Italy, June 3 - 6, 2013, pp 91-96, ISBN: 978-1-4673-6068-5, http://doi.ieeecomputersociety.org/10.1109 /MDM. 2013.73
[4] Alec Go, Richa Bayani, Lei Huang, "Twitter Sentiment Classification Using Distant Supervision", Journal on Processing, 2009, pp1-6.
[5] Pablo Gamallo, Marcos Garcia, "Citius: A Naive-Bayes Strategy for Sentiment Analysis on English Tweets", 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland, Aug 23-24 2014, pp 171–175.
[6] [6] Alexander Pak, Patrick Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining", 7th conference on International Language Resources and Evaluation (LREC'10), Valletta, ISBN: 2-9517408-6-7, May 2010.
[7] Luciano Barbosa, Junlan Feng, "Robust Sentiment Detection on Twitter from Biased and Noisy Data", 23rd International Conference on Computational Linguistics: Posters, Pages 36-44, 2010.
[8] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, Rebecca Passonneau, "Sentiment Analysis of Twitter Data", Workshop on Languages in Social Media, 2011, pp 30-38, ISBN: 978-1-932432-96-1.
[9] Benjamin Mandel, Aron Culotta, John Boulahanis, Danielle Stark, Bonnie Lewis, Jeremy Rodrigue, "A Demographic Analysis of Online Sentiment during Hurricane Irene", 2nd Workshop on Language in Social Media, pp 27-36, 2012.
[10] Krushikanth R. Apala, Merin Jose, Supreme Motnam, C.-C. Chan, Kathy J. Liszka, Federico de Gregorio, "Prediction of Movies Box Office Performance Using Social Media", IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2013, pp 1209-1214ISBN: 978-1-4503-2240-9, DOI: 10.1145/2492517. 2500232.
[11] Pete Burnap, Matthew L. Williams, Luke Sloan, Omer Rana, William Housley, Adam Edwards, Vincent Knight, Rob Procter, Alex Voss, "Tweeting the terror: modelling the social media reaction to the Woolwich terrorist attack", Social Network analysis and Mining, Springer 2014, DOI 10.1007/s13278-014-0206-4.
[12] Alexander Pak, Patrick Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining", 7th conference on International Language Resources and Evaluation (LREC'10), Valletta, Malta, 2010, ISBN 2-9517408-6-7
[13] Zhunchen Luo, Miles Osborne, Ting Wang, "An effective approach to tweets opinion retrieval", Springer Journal on World Wide Web, Dec 2013, DOI: 10.1007/s11280-013-0268-7.
[14] Micol Marchetti-Bowick, Nathaneal Chambers, "Learning for Microblogs with Distant Supervision: Political forecasting with Twitter", 13th Conference of the European Chapter of the Association for Computational Linguistics, 2012, Pages 603-612, ISBN: 978-1-937284-19-0.
[15] Davidov, Tsur, Rappoport, "Semi- Supervised Recognition of Sarcastic Sentences in Twitter and Amazon", 14th Conference on Computational Natural Language Learning, 2010, pp 107-116, ISBN: 978-1-932432-83-1.
[16] Yanghui Rao, Jingsheng Lei, Liu Wenyin, Qing Li, Mingliang Chen, "Building emotional dictionary for sentiment analysis of online news", Journal on World Wide Web, Volume 17, Number 4, pp 723-742, June 2014, DOI 10.1007/s11280-013-0221-9.
[17] Marc Egger, Andre Lang, "Brief tutorial on how to generate user generated content", German Journal on Artificial Intelligence, pp 53-60, Feb 2013.
[18] Stephan Winkler, Susanne Schaller, Viktoria Dorfer, Michael Affenzeller, Gerald Petz, Michał Karpowicz, "Data-based prediction of sentiments using heterogeneous model ensembles", Journal on Soft Computing, July 2014, pp 1-12.
[19] Shawn O'Banion, Larry Birnbaum Knight Lab, "Using Explicit Linguistic Expressions of Preference in Social Media to Predict Voting Behavior", IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Niagara Falls, Canada, Aug. 25- 28 2013, ISBN: 978-1-4503-2240-9, pp 207-214, http://doi.ieeecomputersociety .org.
[20] Aibek Makazhanov, Davood Rafiei, "Predicting Political Preference of Twitter Users", International Conference on Advances in Social Networks Analysis and Mining, Niagara Falls, Ontario, Canada, 25-28 August 2013.
[21] Oshini Goonetilleke, Timos Sellis, Xiuzhen Zhang, Saket Sathe, "Twitter analytics: A Big data management perspective", ACM SIGKDD Explorations Newsletter - Special issue on big data archive, Volume 16, Issue 1, June 2014, Pages 11-20, DOI: 10.1145/2674026. 2674029.
[22] A. Tamilselvi, M. ParveenTaj, "Sentiment Analysis of Micro blogs using Opinion Mining Classification Algorithm", International Journal of Science and Research (IJSR), Volume 2, Issue 10, October 2013, pp 196-202, ISSN (Online): 2319-7064.

[23] Marta Arias, Argimiro Arratia, Ramon Xuriguera, "Forecasting with Twitter Data", ACM Transactions on Intelligent Systems and Technology (TIST) - Special Section on Intelligent Mobile Knowledge Discovery and Management Systems and Special Issue on Social Web Mining archive, Volume 5, Issue 1, December 2013, Article No. 8, DOI: 10.1145/2542182.2542190.
[24] Xiaolong Wang, Furu Wei, Xiaohua Liu, Ming Zhou, "Topic Sentiment Analysis in Twitter: A Graph-based Hashtag Sentiment Classification Approach", 20th ACM international conference on Information and knowledge management, Oct 2012, pp 1031-1040, DOI:10.1145/2063576.2063726.
[25] James Spencer, Gulden Uchyigit. Sentimentor: Sentiment Analysis of Twitter Data, The 1st International Workshop on Sentiment Discovery from Affective Data (SDAD 2012) pp 56-66, Bristol, UK.
[26] Akshi Kumar, Teeja Mary Sebastian, "Sentiment Analysis on Twitter", International Journal of Computer Science Issues, Volume 9, Issue 4, No 3, July 2012, pp 372-378, ISSN (Online): 1694-0814.
[27] Efthymios Kouloumpis, Theresa Wilson, Johanna Moore, "Twitter Sentiment Analysis: The Good, Bad and the OMG!", 5th International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21, 2011, pp 538-541.
[28] Dmitry Davidov, Oren Tsur, Ari Rappoport, "Enhanced Sentiment Learning Using twitter Hashtags and Smileys", 23rd International Conference on Computational Linguistics, Biejing, pp 241-249, Aug 2010.
[29] Sascha Narr, Michael Hulfenhaus, Sahin Albayrak, "Language-Independent Twitter Sentiment Analysis", KDML workshop on knowledge discovery, data mining and machine learning, Dortmund, Germany, Sep 2012.
[30] Jalaj S Modha, Gayatri S Pandi, Sandip J Modha, "Automatic Sentiment Analysis for Unstructured Data", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 12, Dec 2013, pp 91-97, ISSN: 2277 128X.