Design and implementation of Sentiment Analyzer for top Engineering colleges in India using online Twitter data

Neelam Shukla[1]

[1]*Student, Department of Computer Science and Information Technology, Sam Higginbottom University of Agriculture, Technology and Sciences, Naini, Allahabad (U.P.), India*

Abstract - *People opinions and reviews available to us are one of the most critical factors to decide our views and select the brand, product or service. Social media has created many ways for people to raise their voice, beliefs or opinion. Sentiment Analysis is a case of natural language processing which could identified the perception of people about any specific product, service or issue etc. The Sentiment Analysis tool is to function on a series of expressions for a given item based on the quality and features. Twitter is a micro-blogging site where people share their views and thoughts about any issues or topics, these data shall be useful for sentiment analysis however presents a challenge for analysis because of its humongous and disorganized nature. This paper carried out a study and analysis based on online twitter data available in term of people's opinions regarding top engineering colleges in India by developing the Sentiment analyzer. A probabilistic model classifier such as Decision Tree, K-NN, Naïve-Bayes or Maximum Entropy shall be used for spelling correction and classification. In this dissertation Maximum Entropy classifier has been used to develop the sentiment analyzer.*

Key Words: **Sentiment Analysis, Machine Learning, Opinion Mining, Natural Language Processing, Twitter, Decision Tree, K-NN, Naïve-Bayes & Maximum Entropy.**

1. INTRODUCTION

Social Media has got the increased interest among researchers as it is one of the most significant information exchange technology. People views on social media to express their feeling about various products or services, discussions on current issues, complain and feedbacks for products which they used in daily life are very informative and shall be used for-

- Scientific surveys for a social or political purpose.

- Improve product/ services by companies and product owners.

- Organizations like educational institutions know their popularity and enhance their quality etc.

Social media such as micro-blogging are very much involved in everybody's life now days. People shares their views, thoughts, information's etc. across the world free of cost using only internet connections. Twitter is one such well known micro-blogging site getting around 500 million tweets per day [Twitter website]. Each user has a daily limit of 2,400 tweets and 280 characters per tweet (excluding Chinese, Japanese, and Korean which have limit of 140 characters). [Twitter website]. Users tweet every day about various subjects like products, services, day to day activities, places, personalities etc.

Twitter is hugely valuable resource from which data can be extracted in form of useful informations by using text mining tools for sentiment analysis to know the perception of people (Positive, Negative or Neutral) regarding any topic, issue or product etc.

2. RELATED WORKS

A lot of works have been put in this area and there is an immense prevalence and surveys of items and administrations offered by various associations. Most similar works have been carried out by mainly Nehal Mamgain, Ekta Mehta, Ankush Mittal & Gaurav Bhatt (2016), Md Shoeb & Jawed Ahmed, (2017) and Neethu M. S. and Rajasree R., (2013). Nehal Mamgain, Ekta Mehta, Ankush Mittal & Gaurav Bhatt (2016) carried out Sentiment Analysis of Top Colleges in India Using Twitter Data published in IEEE Transaction on Computational Techniques in Information and Communication Technologies. Md Shoeb & Jawed Ahmed, (2017) also carried out Sentiment Analysis and Classification of Tweets Using Data Mining published in IRJET Volume: 04 Issue: 12. Neethu M. S. and Rajasree R., (2013), carried out Sentiment Analysis in Twitter using Machine Learning Techniques published in 4th IEEE International Conference on Computing, Communications and Networking Technologies, pp. 1-5, Tiruchengode, India. In this work maximum entropy classifier model has been used for sentiment analysis which has not been used in these studies.

3. PROCEDURE

To develop the analyzer following procedure or algorithm has been implemented in system and coding has been done based on these. These are briefly described in the following sections.

3.1 Fetching Twitter Data:

In this algorithm, tweets shall be fetched from Twitter using twitter API, based on keywords i.e. College Names in this case. Further fetched Tweets shall be stored in the database for further processing. For fast processing and simple design provision has been made for maximum 100 nos. of tweets which shall be fetched for keyword entered and stored in database.

3.2 Preprocessing:

In this algorithm, the tweets which are foreign made to database from the twitter API, these tweets comprise of pointless words, whitespaces, hyperlinks and unique characters. First we have to do separating process by evacuating every single superfluous word, whitespaces, hyperlinks and extraordinary characters. In this study, twitter data concerning three of the top colleges in India was obtained in JSON format. Unique tweets referring to IIT-Kanpur (IIT-K), IIT-Delhi (IIT-D) and IIT-Madras (IIT-M) were extracted in order to reduce the bias of user opinions, eliminate redundant data and minimize the frequency of tweets which may be spam or fake reviews. The tweets also provide information about the user, location, time-zone. In order to segregate the user opinion from user information, preprocessing was performed on the tweets. Removal of URLs, repeated letters in sequence which occurred more than twice with two of the same letter, ASCII escapes sequences for Unicode characters, uninformative symbols and some but not all punctuations from the tweets was performed in order to sustain emoticons in the tweet. Expansion of SMS lingo, emoticons and abbreviations in net speak has been performed in order to include user opinions fitted rigidly under the constraint of 280 characters by referencing a slang dictionary which contains roughly 5,200 slang words and incorporates about 270 emoticons. The preprocessing steps aim to begin the feature extraction process and start extracting bags of words from the samples. One of the main focuses is to reduce the final amount of features extracted. Indeed, features reduction is important in order to improve the accuracy of the prediction for both topic modeling and sentiment analysis. Features are used to represent the samples, and the more the algorithm will be trained for a specific feature the more accurate the results will be. Hence, if two features are similar it is convenient to combine them as one unique feature. Moreover, if a feature is not relevant for the analysis, it can be removed from the bag of words.

Lower uppercase letters, the first step in the preprocessing is to go through all the data and change every uppercase letter to their corresponding lowercase letter. When processing a word, the analysis will be case sensitive and the program will consider "data" and "Data" as two totally different words. It is important that, these two words are considered as the same features. Otherwise, the algorithms will affect sentiments which may differ to these two words. For example, on these three sentences: "data are good", "Awesome data", and "Bad Data". The first and second sentences both contain "data" and are positive, the third sentence contains "Data" and is negative. The algorithm will guess that sentences containing "data" are more likely to be positive and those containing "Data" negative. If the uppercases had been removed the algorithm would have been able to guess that the fact that the sentence contains "data" is not very relevant to detect whether or the sentence is positive. This preprocessing step is even more important since the data are retrieved from Twitter. Social media users are often writing in uppercase even if it is not required, thus this preprocessing step will have a better impact on social media data than other "classical" data.

Remove URLs and user references, Twitter allows user to include hashtags, user references and URLs in their messages. In most cases, user references and URLs are not relevant for analyzing the content of a text. Therefore, this preprocessing step relies on regular expression to find and replace every URLs by "URL" and user reference by "AT_USER", this allows to reduce the total amount of features extracted from the corpus. The hash tags are not removed since they often contain a word which is relevant for the analysis, and the "#" characters will be removed during the tokenization process. Remove digits, digits are not relevant for analyzing the data, so they can be removed from the sentences. Furthermore, in some cases digits will be mixed with words, removing them may allow to associate two features which may have been considered different by the algorithm otherwise. For example, some data may contain "iphone", when other will contain "iphone7". The tokenization process, which will be introduced later.

Remove stop words, In natural language processing, stop words are often removed from the sample. These stop words are words which are commonly used in a language, and are not relevant for several natural language processing methods such as topic modeling and sentiment analysis. Removing these words allows reducing the amount of features extracted from the samples. As we process, each of the tweets, we keep adding words to the feature vector and ignoring other words. Let us look at the feature words extracted for the tweets.

Table:1 Positive Feature Words

POSITIVE TWEETS	FEATURE WORD
Iit, kanpur, location is very good	" Very Good",
Iit, delhi, fantastic campus	"Fantastic"
Iit, madras, Superb lab	"Superb"

Table:2 Neutral Feature Words

NEUTRAL TWEETS	FEATURE WORD
Iit, kanpur, college will open on 21.	" open",
Iit, delhi, share some picture	"picture"
Iit, madras, located in Chennai	"located"

Table:3 Negative Feature Words

NEGATIVE TWEETS	FEATURE WORD
Iit, kanpur, horrible food in mess	"horrible"
Iit, delhi, worst air quality	"Worst"
Iit, madras, frequently crime happens	"crime"

The entire feature vector will be a combination of each of these feature words. For each tweet, if a feature word is present, we mark it as 1, else marked as 0. Instead of using presence/absence off feature word, we may also use the count of it, but since tweets are just 280 chars, I use 0/1. Now, each tweet shall become as a bunch of 1s and 0s and based on this pattern, a tweet is labeled as positive, neutral or negative.

Given any new tweet, we need to extract the feature words as above and we get one more pattern of 0s and 1s and based on the model learned, the classifiers predict the tweet sentiment. It's highly essential to understand this point.

In my full implementation, method of distant supervision has been used to obtain a large training dataset. This method is detailed out in Twitter Sentiment Classification using Distant Supervision.

3.3 Self-learning and Word Standardization System:

In this algorithm, first we have to instate the word reference (first emphasis dictionary). In the lexicon for the most part we have to introduce the positive, negative nonpartisan and things. Every single huge datum and information mining ventures in view of the prepared information, without prepared information (introduction of words).So instatement of the prepared information is vital. In the self-learning framework, we are doing word institutionalization, here we are not considering past, present and future status of the words, just we are thinking about the word.

3.4 Sentiment Classification:

In this algorithm, preprocessed tweets are brought from the database one by one. In the first place we require check one by one watchword whether that catchphrase is thing are not, if thing we will expel it from the specific tweet. After that the rest of the watchwords checked with assessment compose, regardless of whether that catchphrases are certain opinion or negative conclusion or impartial feeling. The rest of the watchwords in the tweet which does not has a place with any of the supposition will be relegated transitory conclusion in light of the more check of positive, negative and impartial. In the second cycle if the remaining word crosses the limit of positive, negative or nonpartisan, that watchword forever included as development in the lexicon. At long last opinion of the tweet is recognized in light of the positive, negative and impartial words in the specific tweet. This part depicts about the prerequisites. It determines the equipment and programming prerequisite that are needed for software to keeping in mind the end goal, to run the application appropriately. The Software Requirement Specification (SRS) is clarified in point of interest, which incorporates outline of this exposition and additionally the functional and non-practical necessity of this dissertation.

Maximum Entropy Classifiers: Average rate at which information is produced by stochastic source of data is defined as information entropy. The Max entropy classifier falls under the category of exponential model. The features are conditionally independent of each other in Naïve Bayes classifier but in Max Entropy we do not assume that the features are conditionally independent of each other. The Max Entropy selects the one which has the largest entropy from all the models that fit our training data so it is based on the principle of Maximum entropy. Different text classification problems such as language detection, topic classification, sentiment analysis and more can be solved by the Max Entropy classifier. Maximum Entropy classifier is used for classification when we don't know anything about the prior distributions and when it is unsafe to make any such assumptions. In classification task when we can not assume the conditional independence of the features, use of Max entropy classifier gives best result. In text classification problems it is better to the use of Max entropy classifier because in text classification problems features are usually words which obviously are not independent. Estimation of parameters of model are called the optimization problem, due to this Max entropy requires more time to train as compare to naïve bayes. So this model gives robust result and in terms of CPU and memory consumption it is competitive.

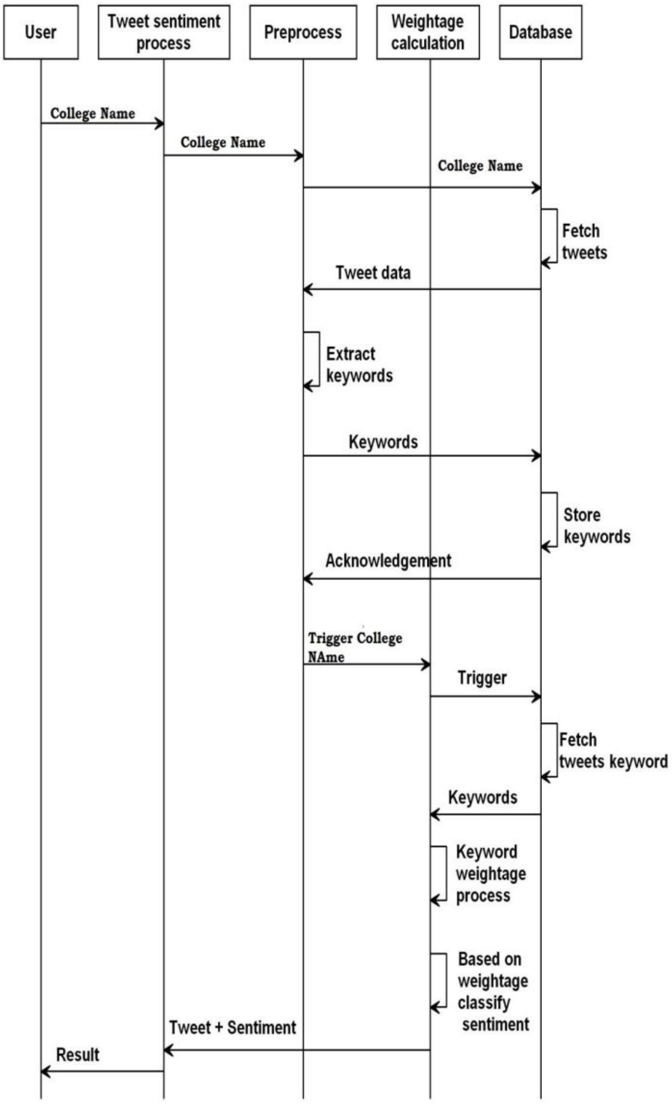


Figure:1 Sequence Diagram

Theoretical Background of Maximum Entropy: The Maximum Entropy classifier uses search techniques for maximizing the performance of classifier, to find a set of parameters. Maximum Entropy uses search techniques rather than using probabilities to set the model's parameters. It uses the set of parameters that maximizes the total likelihood of the training corpus, which is defined as

P (Feature) = Σx|in|corpous P (label(x) / features (x)

Where P (label/features) is defined as-

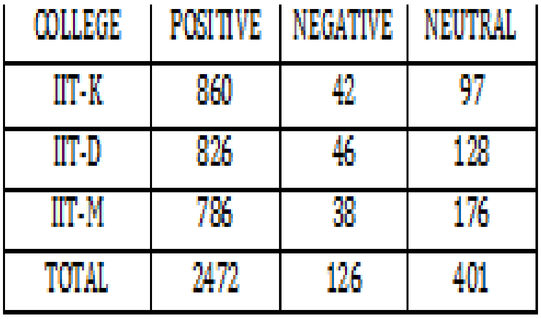
P(label/features) = P(label,features)/ Σlabel P (label,features)

For maximizing the likelihood of training set, there is no way to directly calculate the model parameters, because of complex interactions between the effects of related features.

4. RESULTS AND DISCUSSION

From the developed system data fetched from twitter for IIT-Kanpur, IIT-Delhi and IIT-Madras have been used for further study & analysis. The dynamic tweets fetch from twitter during the time period of 11 Oct to 20 Oct 2018 on daily basis i.e. 10 Sets of data have been used for Sentiment Analysis and for ROC Curve. (*Undecided tweets added in Neutral which has no impact on analysis). The result shows that IIT-Kanpur has the most positive sentiment while IIT-Madras has minimum negative sentiment. The accuracy and precision of system has been determined by calculating TP, TN, FP, FN, TPR (Sensitivity), TNR & FPR (1-Specificity) and based on the same ROC curves have been plotted .

Table:4 Sentiment Analysis Table



COLLEGE	POSITIVE	NEGATIVE	NEUTRAL
IIT-K	860	42	97
IIT-D	826	46	128
IIT-M	786	38	176
TOTAL	2472	126	401

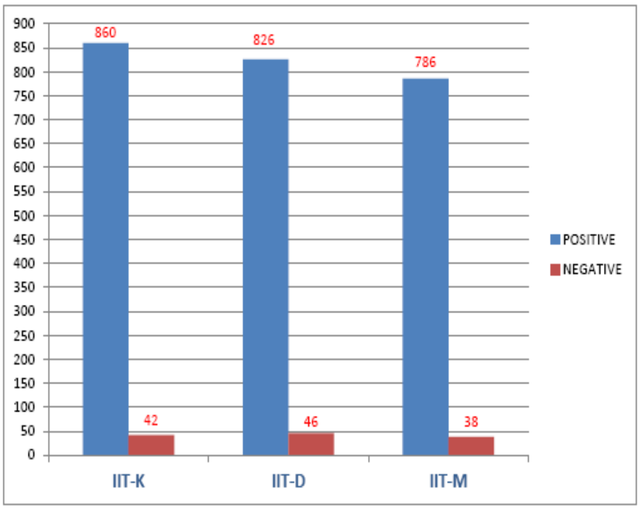


Figure:2 Sentiment Analysis Graph Obtained

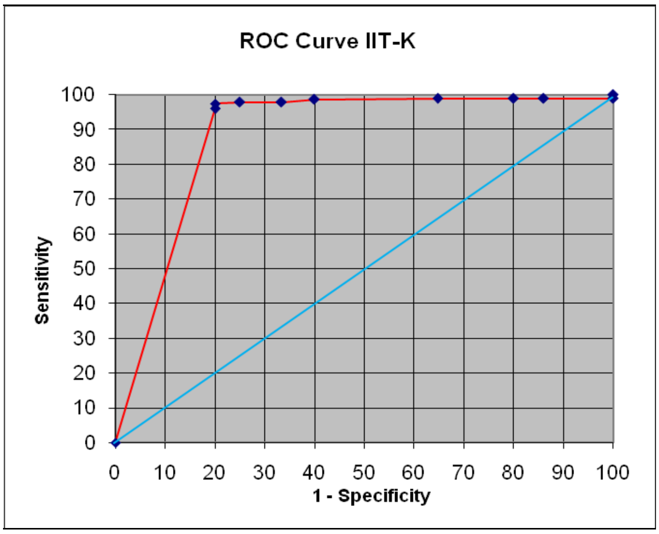


Figure:3 ROC Curve for IIT-Kanpur

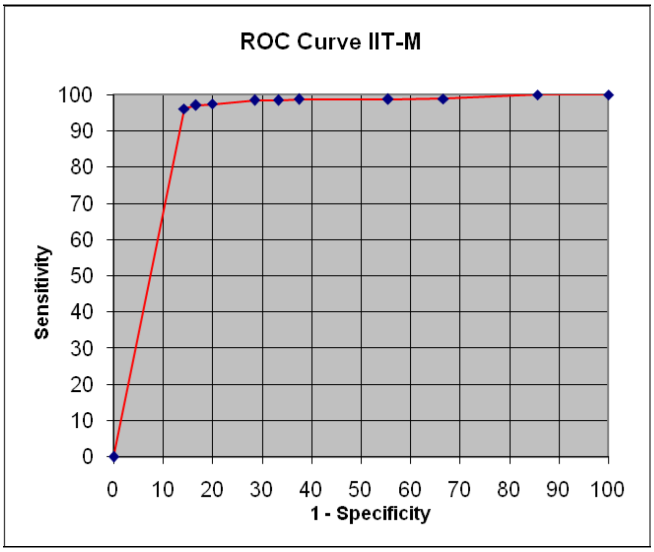


Figure:4 ROC Curve for IIT-Mumbai

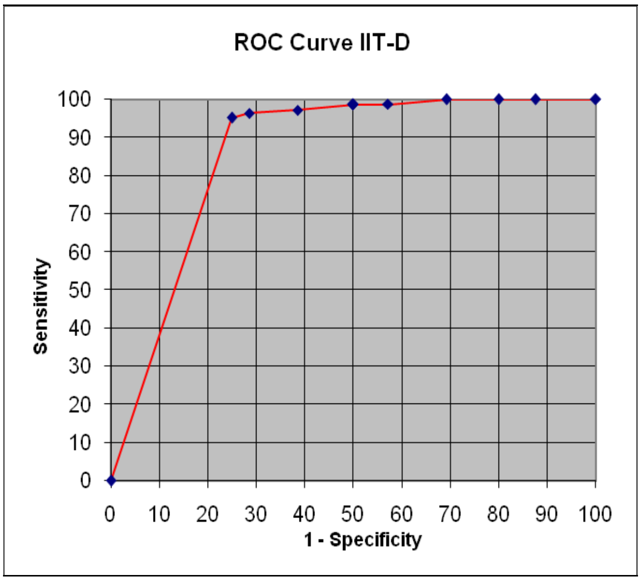
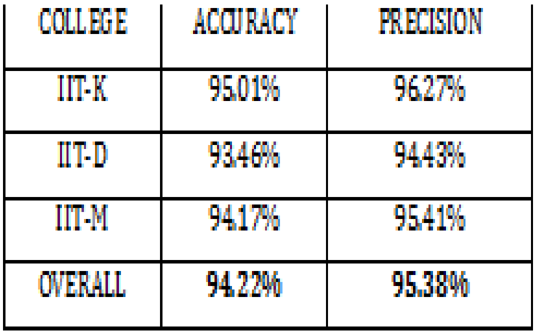


Figure:5 ROC Curve for IIT-Delhi

Table:5 Precision and Accuracy Result



COLLEGE	ACCURACY	PRECISION
IIT-K	95.01%	96.27%
IIT-D	93.46%	94.43%
IIT-M	94.17%	95.41%
OVERALL	94.22%	95.38%

5. CONCLUSION

Sentiment Analysis is a very important tool in data mining which shall utilized huge unstructured information available online in different social networking and micro blogging sites for different research or survey. Like for improvement or betterments of products or services. Today's many of great startups and businesses are the results of opinion mining. In this work one of the analysis has been performed by developing the analyzer for college sentiment analysis which may also be helpful for other fields with some modification in algorithm & coding. However sentiment selection, classification and calculations are as yet an open research field. Most of the classifications used in sentiment analysis for giving best result are Decision Tree, K-NN and NaïveBayes, Maximum Entropy, and Support Vector Machines. In which Maximum Entropy method has been selected for this work which shall be considered as one of the best classifier.

6. FUTURE SCOPE OF WORK

While this study and work it has been observed that the neutral counts of tweets have not been taken care in account which might be give some strength to the sentiment analysis. Also as a trend it has been extended to globalists in all fields so tweets in other foreign languages may also be taken care and in India it shall be extended for Hindi and other regional languages also. One more characteristic which may open a door in this research area is to consider opposition views or tweets wisely and segregate these accordingly. Like if we take views of users on government polices by online sentiment analysis, opposition parties always prone to oppose these. So while online sentiment analysis views or tweets of these institutions and persons shall be suitably analyze that how to tackle with these and make a framework for such type of sentiment analysis.

7. ACKNOWLEDGEMENT

I express profound sense of gratitude to my supervisor, Dr. Tulika, Department of Computer Science and Information Technology, Sam Higginbottom University of Agriculture, Technology and Sciences, Naini, Allahabad, for her motivation, feedback on research questions, experiments and the writing and presentations that followed, philosophical discussions, and the wisdom conveyed. I show herewith appreciation for her continuous help and support to Dr. Tulika, a great mentor, for her patience and enthusiasm, and It is due to her excellent brilliant ideas, able guidance and directions that I could grow as a M. Tech. scholar and produce this work. I heartily thank Dr. Tulika for giving me adequate of time to educate me in and developing the idea. I am thankful to HOD (CS&IT) & SAC members for their valuable observation on this work and their guidance to make this success. I am also thankful to and fortunate enough to get constant encouragement, support and guidance from all Teaching staffs of CS&IT which helped me in successfully completing this work.

REFERENCES

[1] AA Medhat, Ahmed Hass & Hoda Korashy (2014), Sentiment analysis algorithms and applications: AWal survey, Ain Shams, Engineering Journal 5, 1093–1113.

[2] Abakar K. A. A., Yu C., (2014) Performance of SVM based on PUK kernel incomparison to SVM based on RBF kernel in prediction of yarn tenacity, Indian Journal of Fibre and Textile Research, vol. 39,pp. 55-59.

[3] Ali Hasan, Sana Moin, Ahmad Karim and Shahaboddin Shamshirband, (2018). Machine Learning- Based Sentiment Analysis for Twitter Accounts Math. Comput. Appl.

[4] Arora D., Li K.F. and Neville S.W.,(2015) Consumers' sentiment analysis of popular phone brands and operating system preference using Twitter data: A feasibility study, 29th IEEE International Conference on Advanced Information Networking and Applications, pp. 680-686, Gwangju, South Korea.

[5] Bahrainian S.-A., Dengel A.(2013), Sentiment Analysis and Summarization of Twitter Data", 16th IEEE International Conference on Computational Science and Engineering, pp. 227-234, Sydney, Australia.

[6] Bespalov D., Bai B., Qi Y., and Shokoufandeh A.,(2011) Sentiment classification based on supervised latent n-gram analysis, 20th ACM international conference on Information and knowledge management, pp. 375-382, New York, USA.

[7] Choi C., Lee J., Park G., Na J. and Cho W., (2013)Voice of customer analysis for internet shopping malls, International Journal of Smart Home: IJSH, vol. 7, no. 5, pp. 291-304.

[8] Dos Santos C. N. and Gatti M.,(2014) Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts, 25th International Conference on Computational Linguistics, pp. 69-78, Dublin, Ireland.

[9] Girma H. (2009), A Tutorial on Support Vector Machine, Center ofExperimental Mechanics, University of Ljubljana.

[10] Hsu C.-W., Chang C.-C., Lin C.-J.,(2014) A practical guide to support vectorclassification, National Taiwan University, Taipei.

[11] https://drive.google.com/file/d/0B0ChLbwT19XcOVZFdm5wNXA5ODg/view (SentiWordNet_3.0.0.tgz)

[12] https://en.wikipedia.org/wiki/Twitter

[13] https://www.ravikiranj.net/posts/2012/code/how-build-twitter-sentiment-analyzer/

[14] https://nlp.stanford.edu/software/classifier.shtml stanford-classifier-2018-02-27.zip)

[15] J. Bradford DeLong (1988), "Productivity Growth, Convergence, and Welfare:

Comment,"AmericanEconomic Review 78: 5 (December), pp. 1138-1154.

[16] Jotheeswaran J. and Koteeswaran S., (2015)Decision Tree Based Feature Selection and Multilayer Perceptron for Sentiment Analysis, Journal of Engineering and Applied Sciences, vol. 10, issue 14, pp.5883-5894.

[17] Kanakaraj M., Guddeti R M.R., (2015), Performance Analysis of Ensemble Methods on Twitter Sentiment Analysis using NLP Techniques, 9th IEEE International Conference on Semantic Computing,pp. 169-170, Anaheim, California.

[18] Karatzoglou A., Meyer D., Hornik K.,(2006), Support vector machines in R.,Journal of Statistical Software, vol. 15, issue: 9.

[19] King R. A., Racherla P. and Bush V. D., (2014)What We Know and Don't Know about Online Word-of-Mouth: A Review and Synthesis of the Literature, Journal of Interactive Marketing, vol. 28, issue 3, pp. 167- 183.

[20] Koto F. and Adriani M., (2015), A Comparative Study on Twitter Sentiment Analysis: Which Features are Good?, Natural Language Processing and Information Systems, Lecture Notes in Computer Science vol. 9103, pp.453-457.

[21] Manning C. and Raghavan P., (2008), Introduction to information retrieval, New York: Cambridge University Press.

[22] Md Shoeb & Jawed Ahmed, (2017), Sentiment Analysis and Classification of Tweets Using Data Mining, IRJET Volume: 04 Issue: 12

[23] Ministry of Human Resource Development, http://mhrd.gov.in/statist

[24] Multilayer perceptron,https://en.wikipedia.org/wiki/Multilayer_perceptron

[25] NCSS.com

[26] Neethu M. S. and Rajasree R., (2013), Sentiment Analysis in Twitter using Machine Learning Techniques, 4th IEEE International Conference on Computing, Communications and Networking Technologies, pp. 1-5, Tiruchengode, India.

[27] Nehal Mamgain, Ekta Mehta, Ankush Mittal & Gaurav Bhatt (2016), Sentiment Analysis of Top Colleges in India Using Twitter Data by, IEEE Transaction on Computational Techniques in Information and Communication Technologies

[28] Ng A.Y., Jordan M. I., (2002), On Discriminative vs. Generative classifiers: Acomparison of logistic regression and naive Bayes, Advances in Neural Information Processing Systems vol. 14, pp. 841- 848.

[29] Nielsen F.A., Making sense of micro posts, Finn Årup Nielsen blog,

https://finnaarupnielsen.wordpress.com/tag/sentiment analysis/

[30] Pak A. and Paroubek P.(2010), Twitter as a Corpus for Sentiment Analysis and Opinion Mining, 7th International Conference on Language Resources and Evaluation, pp. 1320-1326, Valletta, Malta,

[31] Pierre FICAMOS, Yan LIU (2016), A Topic based Approach for Sentiment Analysis on Twitter Data, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 12, 2016

[32] Posting a tweet, https://support.twitter.com/articles/15367-posting-atweet

[33] Rajendran S., Kalpana B., (2011), A Comparative Study and Choice of an Appropriate Kernel for Support Vector Machines., International Journal of Soft Computing and Engineering (IJSCE), vol. 1, issue: 5.

[34] Ravikiran Janardhana, (2012) how to build a twitter sentiment analyzer ? May 08, 2012, 05:27 PM

[35] Rupal Singh & Divakar Yadav (2018),Sentiment analysis for social network a review, ICRISE 2018

[36] Salazar D. A., Velez J. I., Salazar J. C., (2012), Comparison between SVM and Logistic Regression: Which One is Better to Discriminate?, Colombian Journal of Statistics, Special Issue Biostatistics, vol. 35, no. 2, pp. 223-237.

[37] Segaran T. and Hammerbacher J., (2009), Beautiful Data: The Stories behindElegant Data Solutions, Beijing: O'Reilly.

[38] Sentiment Analysis https://en.wikipedia.org/wiki/Sentiment

[39] Shahheidari S., Dong H., Bin Daud M.N.R.(2013), Twitter sentiment mining: A multidomain analysis, 7th IEEE International Conference on Complex, Intelligent and Software Intensive Systems, pp.144-149, Taichung, Taiwan.

[40] Socher R., et al (2013), Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank, Conference on Empirical Methods in Natural Language Processing, Seattle, Washington.

[41] Twitter Usage/Company Facts, https://about.twitter.com/company

[42] W. J. Krzanowski and D. J. Hand (2009), A Review of: "ROC Curves for Continuous Data. (Book style with paper title and editor)