Sentiment Analysis Using Optimized Feature Sets in Twitter Dataset Domains

Anup Haribhau Raut[1]* Dr. Rahul K. Pandey[2]

[1] PhD Student, Maharishi University of Information Technology, Lucknow

[2] PhD Guide, Maharishi University of Information Technology, Lucknow

Abstract – The purpose of this paper is verify the efficiency of Sentiment analysis and comparative analysis, while the prevalence of social networking, more and more research is being done in the area of tweet analysis. Twitter use as the media for sharing information by many people, Twitter become a valuable topic for further discussion when the wave of using Twitter as a communication tool makes sentiment analysis. In this research we attempt to present the data mining based sentiment analysis tool, it tries to compare three functions: finding positive, negative tweets, neutral tweets from information resources and sentiment analysis among Twitter tweets. Here we also do preprocessing so it helps to improve performance. Then we use machine learning algorithm. This model concentrate on examine tweets from those media sites, thus provide a way to find out technology trends in the future.

Keywords—Data Mining, Senmitemt Analysis, Positive, Negative, Neutral Tweets.

---------------------------♦-----------------------------

1. INTRODUCTION

With the increase in the popularity of social networking, micro-blogging and blogging websites, a huge quantity of data is generated. We know that the internet is a collection of networks, and the age of the internet has changed the way people express their thoughts and feelings. People are connecting with each other with the help of the internet through blog posts, online conversation forums, and many more. The quantity of information is unreasonable for a normal person to analyze with the help of naive technique. Social network analysis is a methodology mainly developed by sociologists and researchers in social psychology [3]. Social network analysis views social relationships in terms of network theory, while individual actor being seen as a node and relationship between each node are presented as an edge. Social network analysis has been defined as an assumption of the importance of relationships among interacting units, and the relations defined by linkages among units are a fundamental component of network theories. Social network analysis has emerged as a key technique in modern sociology. It has also gained a significant following in anthropology, biology, communication studies, economics, geography, information science, organizational studies, social psychology, and sociolinguistics. Afterwards, there are many scholars expanded the use of systematic social network analysis. Due to the growth of online social networking site, online social networking analysis becomes a hot research topic recently. Social network analysis (SNA) is the process of investigating social structures through the use of networks and graph theory. It characterizes networked structures in terms of nodes (individual actors, people, or things within the network) and the ties, edges, or links (relationships or interactions) that connect them. Examples of social structures commonly visualized through social network analysis include social media networks, memes spread, friendship and acquaintance networks, collaboration graphs, kinship, disease transmission, and sexual relationships. These networks are often visualized through in which nodes are represented as points and ties are represented as lines. Social network analysis has emerged as a key technique in modern sociology. It has also gained a significant following in anthropology, biology, communication studies, economics, geography, history, information science, organizational studies, political science, social psychology, development studies, sociolinguistics, and computer science and is now commonly available as a consumer tool [5].

Sentiment analysis is mainly concerned with the identification and classification of opinions or emotions of each post. Sentiment analysis is broadly classified in the two types: first one is a feature or aspect based sentiment analysis and the other is objectivity based sentiment analysis. The posts related to movie reviews come under the category of the feature based sentiment analysis. Objectivity based sentiment analysis does the exploration of the posts which are related to the emotions like hate, miss, love etc. In general, various symbolic techniques and machine learning techniques are used to analyze the sentiment from the twitter data. So in another way we can say that a sentiment analysis is a system or model that takes the documents that analyzed the input, and generates a detailed document summarizing the opinions of the given input document. In the first step pre-processing is done. In pre-processing we are removing the stop words, white spaces, repeating words, emoticons and #hash tags. To correctly classify the tweets, machine learning technique uses the training data. So, this technique does not require the database of words like used in knowledge-based approach and therefore, machine learning technique is better and faster. Several methods are used to extract the feature from the source text. Feature extraction is done in two phases: In the first phase extraction of data related to twitter is done i.e. twitters specific data is extracted. Now by doing this, the tweet is transformed into normal text. In the next phase, more features are extracted and added to feature vector. Each tweet in the training data is associated with class label. This training data is passed to different classifiers and classifiers are trained. Then test tweets are given to the model and classification is done with the help of these trained classifiers [9]. So finally, we get the tweets which are classified into the positive, negative and neutral.

Sentiment analysis has been an important topic for data mining, while the prevalence of social networking, more and more research is being done in the area of tweet analysis. Many people use Twitter as the media for sharing information, the wave of using Twitter as a communication tool makes sentiment analysis on Twitter become a valuable topic for further discussion. In this research work we attempt to present the data mining based sentiment analysis tool, it comprises three functions: sentiment analysis among Twitter tweets, finding positive, negative and neutral tweets from information resources. This tool focuses on analyzing tweets from those media sites, thus provide a way to find out technology trends in the future.

2. RELATED WORK

This section presents the review of previous methods precisely.

Chamlertwat, W (2012)

In [1], they propose a system, the Micro-blog Sentiment Analysis System(MSAS), based on sentiment analysis to automatically analyze customer opinions from the Twitter micro-blog service. The MSAS consists of five main functions to (1) collect Twitter posts, (2) filter for opinionated posts, (3) detect polarity in each post, (4) categorize product features and (5) summarize and visualize the overall results. We used the product domain of smart phone as our case study. The experiments on 100,000 collected posts related to smart phones showed that the system could help indicating the customers' sentiments towards the product features, such as Application, Screen, and Camera. Further evaluation by experts in smart phone industry confirmed that the system yielded some valid results.

T. M.; Dittman (2012)

In [2] presents seven univariate feature determination methods and gathers them into a solitary family entitled First Order Statistics (FOS) based feature choice. These seven all offer the quality of utilizing first order factual measures, for example, mean and standard deviation, despite the fact that this is the first work to relate them to each other and consider their performance contrasted and each other. In order to inspect the properties of these seven strategies, we performed a progression of similitude and classification experiments on eleven DNA microarray datasets. Our results demonstrate that by and large, each feature determination procedure will make different feature subsets when contrasted with alternate individuals from the family. However, when we take a gander at classification we find that, with one exemption, the procedures will deliver great classification results and that the strategies will have comparative performances to each other. Our suggestion is to utilize the rankers Motion to-Clamor and SAM for the best classification results and to maintain a strategic distance from Fold Change Proportion as it is reliably the most noticeably awful performer of the seven rankers.

Kouloumpis, E (2011)

In [3] they investigate the utility of semantic features for identifying the sentiment of Twitter messages. They assess the usefulness of existing lexical resources and also features that catch information about the informal and innovative language utilized as a part of microblogging. We adopt a regulated strategy to the issue; however, use existing hash labels in the Twitter data for building training data.

V; Arora (2013)

In [4] they propose utilizing two unique sets of features to alleviate the data inadequacy problem. One is the semantic feature set where we extract semantically concealed ideas from tweets and after that incorporate them into classifier preparing through insertion. Another is the sentiment-theme feature set where we extract idle points and the related subject sentiment from tweets, at that point increase the first feature space with these sentiment-points. Experimental outcomes on the Stanford Twitter Sentiment Dataset demonstrate that both feature sets beat the baseline model utilizing unigrams as it were. Additionally, utilizing semantic features equals the beforehand reported best outcome. Utilizing sentiment subject features accomplishes 86.3%

sentiment classification accuracy, which beats existing methodologies.

Hassan Khan et al.(2012)

In [5] approach incorporates thorough data pre-processing took after by supervised machine learning. They gathered labeled datasets of various spaces with the goal that machine learning won't be restricted to a specific area. To learn SVM classifier they make utilization of various preparing sets each influence SVM to learn distinctive feature sets - 1) Information gain(IG) with feature presence and 2) feature recurrence 3) Cosine closeness with feature presence and 4) feature recurrence. They found that feature presence is superior to feature recurrence.

Agarwal et. al. (2014)

In [6], found that for better outcomes utilizing machine learning approaches, discovering great features is a testing undertaking. They gave the idea of "Semantic Parser" and regarded ideas as features. They utilized the base Excess and Most extreme Relevance (mRMR) feature selection component. They utilized diverse feature sets for their grouping undertaking e.g. unigrams, bigrams, bitagged and reliance parse tree alongside their proposed plot so results can be contrasted and.

Davidov et al.,(2010)

In [7] proposed a way to deal with use Twitter client characterized hashtag in tweets as a grouping of sentiment write utilizing accentuation, single words, n-grams and examples as various feature composers, which are then consolidated into a solitary feature vector for sentiment order. They made utilization of K-Closest Neighbor procedure to allow sentiment labels by developing a feature vector for every case in the preparation and test set.

Po-Wei Liang et.al.(2014)

In [8] utilized Twitter Programming interface to gather twitter data. Their preparation data falls into three distinct classifications (camera, film, portable). The data is labeled as positive, negative and non-opinions. Tweets containing opinions were sifted. Unigram Naive Bayes model was actualized and the Naive Bayes disentangling freedom presumption was utilized. They additionally disposed of pointless features by utilizing the Common Information and Chi-square feature extraction technique. At long last, the introduction of a tweet is anticipated. i.e. positive or negative.

Khan, Jawad (2014)

In [9] they propose an administered lazy learning model using syntactic rules for the item features and opinion words extraction in subjective review sentences. In our lazy learning algorithm, i.e. K-NN with k=3 is utilized for the review sentences' classification into two classes (subjective, objective). Our experiment demonstrates that proposed method can enhance the performance of existing work as far as normal exactness, recall, and f-score for the extraction of opinion sentences and item features.

Clayton J (2014)

In [10] creator display VADER, a straightforward rule-based model for general sentiment analysis, and contrast its effectiveness with eleven commonplace condition of-rehearse benchmarks including LIWC, Once again, the General Inquirer, SentiWordNet, and machine learning focused methods depending on Guileless Bayes, Most extreme Entropy, and Bolster Vector Machine (SVM) algorithms. Utilizing a mix of subjective and quantitative methods, they first build and experimentally approve the best quality level rundown of lexical features (alongside their related sentiment power measures) which are specifically sensitive to sentiment in microblog-like settings. They at that point join these lexical features with a thought for five general rules that encapsulate linguistic and grammatical traditions for communicating and stressing sentiment force. Strangely, utilizing our tightfisted rule-based model to assess the sentiment of tweets, we find that VADER outperforms singular human raters (F1 Classification Accuracy = 0.96 and 0.84, individually), and sums up more favorably crosswise over settings than any of our benchmarks.

3. DATASET

In Sentiment Analysis of Twitter one of the major challenges is to gather a labelled dataset. For training and testing classifiers researchers have made public the following datasets.

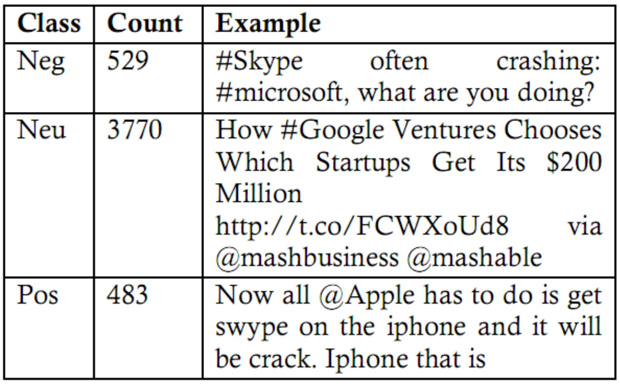
1. Twitter Sentiment Corpus

This is a collection of 5513 tweets collected for four different topics, namely, Apple, Google, Microsoft, Twitter It is collected and hand-classified by Sanders Analytics LLC. Each entry in the corpus contains Tweet id, Topic and a Sentiment label. We use Twitter-Python library to enrich this data by downloading data like Tweet text, Creation Date, Creator etc. for every Tweet id. Each Tweet is hand classified by an American male into the following four categories. For the purpose of our experiments, we consider Irrelevant and Neutral to be the same class.

Illustration of Tweets in this corpus is show in Table 1.

- **Positive** For showing positive sentiment towards the topic
- **Neutral** For showing no or mixed or weak sentiments towards the topic
- **Negative** For showing negative sentiment towards the topic
- **Irrelevant** For non-English text or off-topic comments

Table 1: Twitter Sentiment Corpus

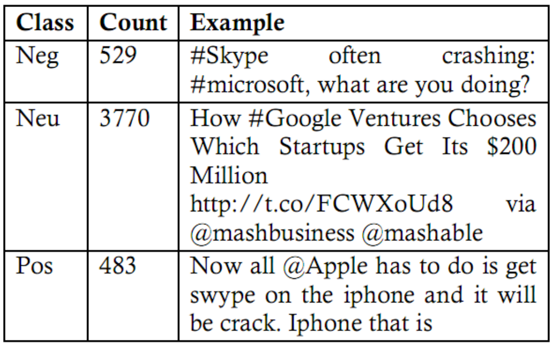


Class	Count	Example
Neg	529	#Skype often crashing: #microsoft, what are you doing?
Neu	3770	How #Google Ventures Chooses Which Startups Get Its $200 Million http://t.co/FCWXoUd8 via @mashbusiness @mashable
Pos	483	Now all @Apple has to do is get swype on the iphone and it will be crack. Iphone that is

2. Stanford Twitter

This corpus of tweets, developed by Sanford's Natural Language processing research group, is publically available. The training set is collected by querying Twitter API for happy emoticons like ":)" and sad emoticons like ":(" and labelling them positive or negative. The emoticons were then stripped and Re-Tweets and duplicates removed. It also contains around 500 tweets manually collected and labelled for testing purposes. We randomly sample and use 5000 tweets from this dataset. An example of Tweets in this corpus is shown in Table 2.

Table 2: Stanford Corpus



Class	Count	Example
Neg	529	#Skype often crashing: #microsoft, what are you doing?
Neu	3770	How #Google Ventures Chooses Which Startups Get Its $200 Million http://t.co/FCWXoUd8 via @mashbusiness @mashable
Pos	483	Now all @Apple has to do is get swype on the iphone and it will be crack. Iphone that is

4. PROPOSED METHODOLOGY

In this section, the proposed method architecture and algorithms are presented.

Sentiment Analysis in twitter is very difficult because of its short length. Presence of emoticons, slang words and incorrect spellings in tweets forced to have a preprocessing venture before feature extraction. There are different feature extraction methods for gathering important features from content which can be connected to tweets moreover. Be that as it may, the feature extraction is to be done in two stages to extricate pertinent features. In the first stage, twitter specific features are extracted. At that point these features are expelled from the tweets to make ordinary content. After that, again feature extraction is done to get more features. This is the thought utilized as a part of this paper to create an efficient feature vector for dissecting twitter sentiment. Since no standard dataset is available for twitter posts of electronic gadgets, we made a dataset by gathering tweets for a specific timeframe. By doing sentiment analysis on a specific space, it is conceivable to identify the influence of area information in picking a feature vector. Different classifiers are utilized to do the classification to find out their influence in this specific space with this specific feature vector.

We propose the novel machine learning based sentiment analysis algorithm for twitter dataset. The sentiment analysis method is composed of three main steps such as pre-processing, feature extraction and classification. After the pre-processing steps in which misspelling and other errors are removed, the two main types of features extraction are performed. In first category twitter specific features are extracted such as tag, special keyword, and presence of negation, emoticon, and number of positive keywords, number of negative keywords, number of positive hash tags and number of negative hash tags. In second feature, the bag-of-words (BoW) method is used. The hybrid feature vector is created for each twitter message. After feature extraction, the ensemble classifier is designed for classification purpose.

5. SYSTEM ARCHITECTURE

The proposed system contains various phases of development. A dataset is created using twitter posts of movie reviews. As we know that tweets contain slang words and misspelling. So, we perform a sentence level sentiment analysis on tweets. This is done in three phases. In a first phase preprocessing is done. Then Feature vector is created using relevant features. Finally, using different classifiers, tweets are classified into positive, negative and neutral classes.

We introduce a model which collects tweets from social systems administration destinations and in this manner give a perspective of business intelligence. In our framework, there are two layers in the sentiment analysis tool, the data processing layer and sentiment analysis layer. Data processing layer deals with data collection and data mining, while sentiment analysis layer utilizes an application to introduce the result of data mining.

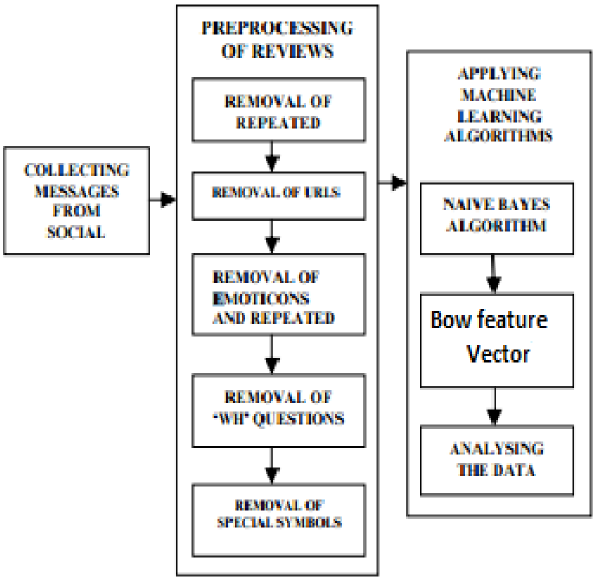

COLLECTING MESSAGES FROM SOCIAL
PREPROCESSING OF REVIEWS
REMOVAL OF REPEATED
REMOVAL OF URLS
REMOVAL OF EMOTICONS AND REPEATED
REMOVAL OF 'WH' QUESTIONS
REMOVAL OF SPECIAL SYMBOLS
APPLYING MACHINE LEARNING ALGORITHMS
NAIVE BAYES ALGORITHM
Bow feature Vector
ANALYSING THE DATA

Figure 1 System Architecture

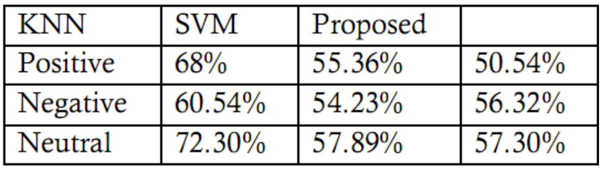
The period of Internet has changed the way people express their perspectives. It is presently done through blog entries, online talk forums, item review websites and so forth. People rely on this client created content. When somebody needs to purchase an item, they will look into its reviews online before making a choice. The measure of client created content is too expansive for an ordinary client to investigate. Along these lines, to computerize this, different sentiment analysis procedures are utilized. Symbolic procedures or Information base approach and Machine learning strategies are the two fundamental systems utilized as a part of sentiment analysis. Learning based approach requires a vast database of predefined emotions and an efficient information portrayal for identifying sentiments. Machine learning approach makes utilization of a training set to build up a sentiment classifier that classifies sentiments. Since a predefined database of whole emotions isn't required for machine learning approach, it is preferably less difficult than Information base approach. In this undertaking, we utilize machine learning method for classifying the tweets.

6. RESULTS AND DISCUSSION

In this section we discus result of proposed method with KNN and SVM classifier.

Now here we discuss the precision analysis show blow table show comparative value of each categories of tweet like positive, negative and neutral with respect to the each technique. In figure 2 x axis show tweet category with classification technique and y axis shows the percentage value of precision.

Table 3 Precision Analysis



KNN	SVM	Proposed	
Positive	68%	55.36%	50.54%
Negative	60.54%	54.23%	56.32%
Neutral	72.30%	57.89%	57.30%

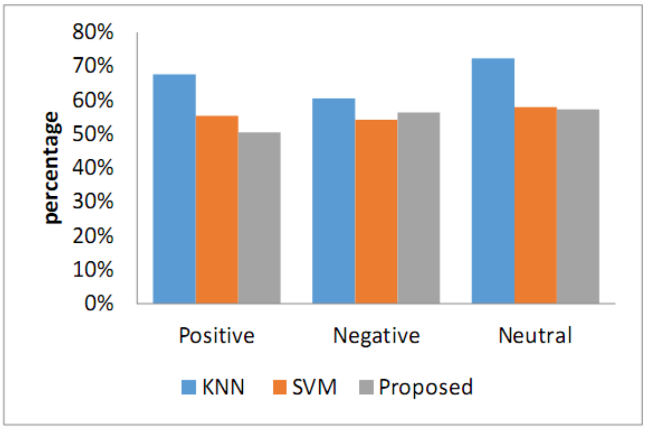
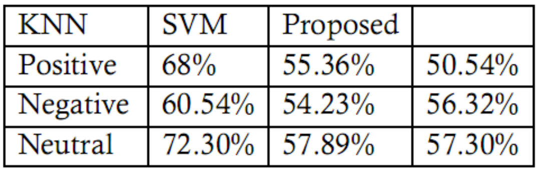


Figure 2 Precision analysis

Now here we discuss the recall analysis show blow table show comparative value of each categories of tweet like positive, negative and neutral with respect to the each technique. In figure 3 x axis show tweet category with classification technique and y axis shows the percentage value of precision.

Table 4 Recall Analysis



KNN	SVM	Proposed	
Positive	68%	55.36%	50.54%
Negative	60.54%	54.23%	56.32%
Neutral	72.30%	57.89%	57.30%

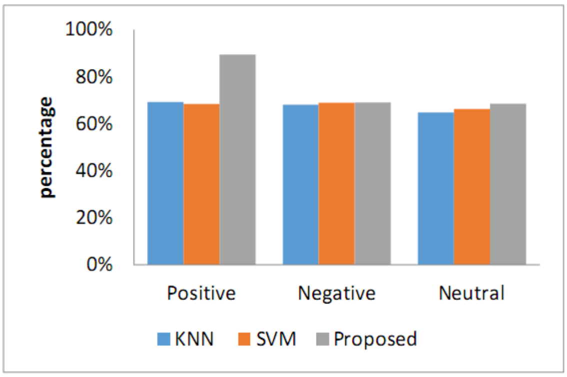


Figure 3 Recall analysis

Now here we discuss the F-score analysis show blow table show comparative value of each categories of tweet like positive, negative and neutral with respect to the each technique. In figure 4 x axis show tweet category with classification technique and y axis shows the percentage value of precision.

Table 5 F-score Analysis

KNN	SVM	Proposed	
Positive	68%	55.36%	50.54%
Negative	60.54%	54.23%	56.32%
Neutral	72.30%	57.89%	57.30%

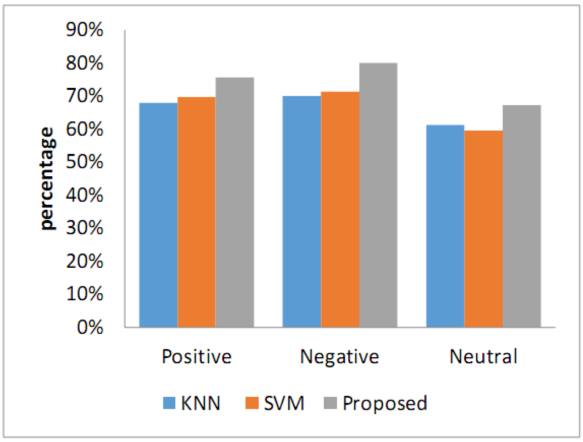


Figure 4 F-score analysis

CONCLUSION

In this paper we discus result of sentiment analysis using optimized feature sets. The detailed comparative analysis of such techniques is presented in paper. The proposed method classified the tweets in positive, negative and neutral sentiments whereas much of the literature in this field is associated with 2-way classification. The work of proposed model has gone through pre-processing stage, features generation stage and classifiers learning stage. The analytical evaluation of proposed model is done in terms of precision, recall and f-measure. The comparative observations positive are taken against the SVM and KNN methods. The comparative results show that the proposed model has improved the accuracy and f-measure of tweet class prediction.

Future work is doing same job with for Hindi tweets there are many users on Twitter that use primarily Hindi language. The approach discussed here can be used to create a Hindi language sentiment classifier.

REFERENCES

Dawn Chamlertwat, W.; Bhattarakosol, P.; Rungkasiri, T.; and Haruechaiyasak, C. (2012). Discovering consumer insight from twitter via sentiment analysis. J. UCS 18(8): pp. 973–992

Hutto, Clayton J., and Eric Gilbert (2014). "Vader: A parsimonious rule-based model for sentiment analysis of social media text." Eighth International AAAI Conference on Weblogs and Social Media.

Khan, Farhan Hassan, Usman Qamar, and Saba Bashir (2016). "A semi-supervised approach to sentiment analysis using revised sentiment strength based on Senti Word Net." Knowledge and Information Systems.

Khan, Jawad, and Byeong Soo Jeong (2016). "Summarizing customer review based on product feature and opinion." Machine Learning and Cybernetics (ICMLC), 2016 International Conference on. IEEE, 2016

Khoshgoftaar, T. M.; Dittman, D. J.; Wald, R.; and Fazelpour, A. (2012). First order statistics based feature selection: A diverse and powerful family of feature selection techniques. In Proceedings of the Eleventh International Conference on Machine Learning and Applications (ICMLA), pp. 151–157. ICMLA.

Kouloumpis, E.; Wilson, T.; and Moore, J. (2011). Twitter sentiment analysis: The good the bad and the omg! ICWSM 11: pp. 538–541.

Neethu M., S. and Rajashree R. (2013). Sentiment Analysis in Twitter using Machine Learning Techniques" 4th ICCCNT 2013, at Tiruchengode, India. IEEE – 31661

Pablo Gamallo, Marcos Garcia (2014). "Citius: A Naive-Bayes Strategyfor Sentiment Analysis on English Tweets", 8th InternationalWorkshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland,Aug 23-24 2014, pp. 171-175.

Saif, H.; He, Y.; and Alani, H. (2012). Alleviating data sparsity for twitter sentiment analysis. CEUR Workshop Proceedings (CEUR-WS. org).

V.; Arora, I.; and Bhatia, A. (2013). Fast and accurate sentiment classification using an enhanced naive bayes model. In Intelligent Data Engineering and Automated Learning–IDEAL 2013. Springer. pp. 194–201.

Corresponding Author

Anup Haribhau Raut*

PhD Student, Maharishi University of Information Technology, Lucknow