International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 4 Issue: 4
ISSN: 2454-4248
46 – 51

Micro-blogging Sentimental Analysis on Twitter Data Using Naïve Bayes Machine Learning Algorithm in Python

K.Kaviya [1], K.K.Shanthini [1], Dr.M.Sujithra [2]
[1] Msc.Software Systems , Coimbatore Institute of Technology, cbe
[2] Assistant Professor, Coimbatore Institute of Technology, cbe
Kaviya14998@gmail.com, shnthnikk@gmail.com

ABSTRACT:-As the increase of social networking, people started to share information through different kinds of social media. Sentiment Analysis, a Natural Language processing helps in finding the sentiment or opinion hidden within a text. Sentiment analysis is an approach to analyze data and retrieve sentiment that it embodies. Twitter sentiment analysis is an application of sentiment analysis on data from Twitter (tweets), in order to extract sentiments conveyed by the user. It is an important source of decision making and can be extracted, identified, evaluated from the online sentiments reviews. The main goal of the proposed framework is to connect on Twitter and search for the tweets that contain a particular keyword and then evaluate the polarity of the tweets as positive and negative. In order to select the best features Chi Square test is used and Naïve Bayes classifier is used for training and testing the features and also evaluating the sentimental polarity. The proposed system is implemented using Python.

Keywords— *Feature selection, Naïve Bayes classifier, Sentiment analysis, Twitter, Python*

*****

1. INTODUCTION

Sentimental analysis is determined by classifying the sentence into positive, negative and neutral according to the percentages. There are many social streams in which sentimental analysis can be done. People share their experiences, views, knowledge with the world by blogs, social sites and twitter etc. So it's necessary to predict what the individual think about certain things like, in politics-what's the view about ruling party, opinion about the new product, opinion about certain people, suggestions about the people, industry, movies etc. The proposed work is taken to predict the opinions of the user on what he/she tweeted. On Twitter, users are allowed to share their opinions in the form of tweets, using only 140 characters. This leads to people compacting their statements by using slang, abbreviations, emoticons, short forms etc. Along with this, people convey their opinions by using sarcasm and polysemy.

In this paper, a system which collects Tweets from social networking sites, able to do analysis on those Tweets and thus provide some prediction of business intelligence. Results of trend analysis will be display as tweets with different sections presenting positive, Negative and neutral. The proposed work addresses this problem by retrieving posts, performing pre-processing on the data, and analyzing the data using machine learning techniques to classify them by sentiment as either neutral, positive, or negative. Investigates different methods for pre-processing the micro-blogging posts and use various Naive Bayes classification and feature selection methods to determine the best approach. Sentiment Analysis is a field that is growing fairly rapidly. 81 percent of Internet users have done online research on a product at least once, meaning every year there are more articles targeting different text domains over years, where the reviews represent around the 49.12% of the articles.

2. APPROACHES FOR SENTIMENTAL ANALYSIS

The goal of Sentiment Analysis technique is to classify the tweets as positive or negative, what an opinion document expresses. Sentiment Classification is mainly divided into two different approaches

(i) Lexicon- based Approach

The Lexicon-based approach uses a collection of positive and negative sentiment terms and can be divided into corpus-based and dictionary based-approach. Unsupervised approach, but in this case it could use a dictionary with antonyms and synonyms of opinionated words and phrases with their respective sentiment orientation.

(ii) Machine learning Approach

Machine Learning approach uses machine-learning algorithms, as regular text classification algorithm. The Machine learning classifiers are divided into supervised learning and unsupervised learning. Machine learning approach is supervised as it involves use of feature extraction and training the model using feature set and some dataset. Determines the number of classes we can have and supervised learning can cluster or classify the data which treats the testing phase in providing less complexity.

3. SENTIMENT CLASSIFICATION

Sentiment classification is a process of dividing the target unit based on which the sentiments can be predicted. There are three levels of sentiment classification.

 ISSN: 2454-4248
46 – 51

- Sentence level
- Document level
- Aspect level

3.1 Sentence Level: Classifies sentiment expressed in each sentence. If the sentence is subjective it classifies it in positive or negative opinions. The sentence determines whether each sentence expresses a positive or negative.

3.2 Document Level: Here the classification is done throughout the document. Document level it is possible to classify whether a whole users opinion expresses a positive or negative sentiment. For example, given a product/service review, it is possible to determine whether the review expresses an overall positive or negative opinion about the product/service.

3.3 Aspect Level: Classify the sentiments with respect to the specific aspects of entities. Users can give different opinions for different aspects of the same entity. It yields very fine grained sentiment information which can be useful for applications in various domains.

4. PROPOSED METHODOLOGY

The methodology is to train and test the data by undergoing process like pre-processing, feature extraction, feature selection, training, testing and result.

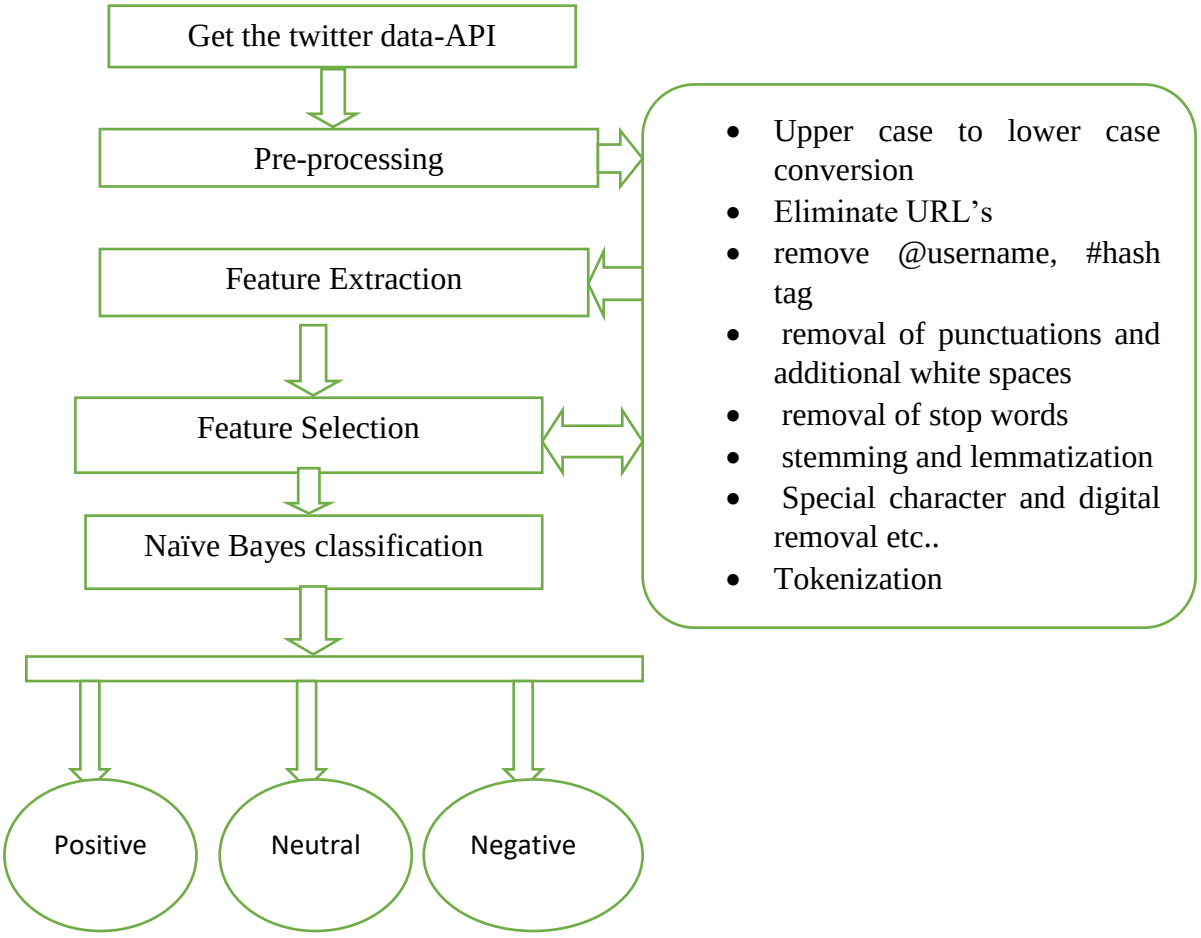


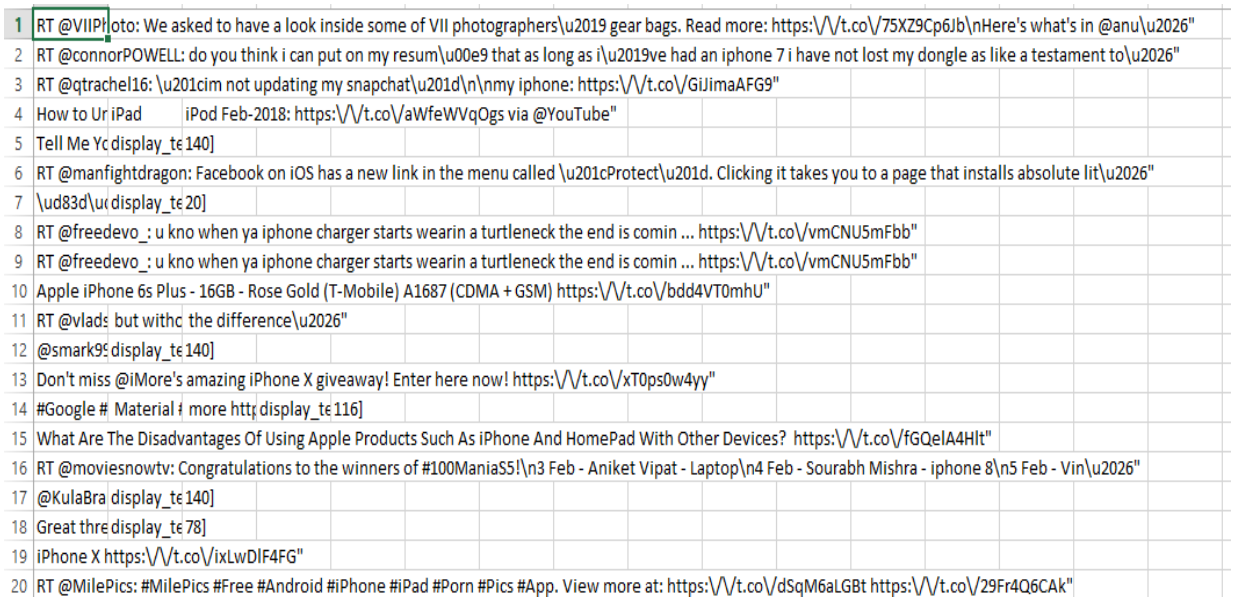
Figure 1. Proposed Methodology

4.1 COLLECTION OF RAW DATA

Tweets are collected through API. English tweets are filtered and tweet about particular product say iPhoneX is collected and stored in the file. Steps involved here is

- Collecting twitter API keys
- Connecting to twitter streaming API and downloading the data
- Saving the data in the file

 ISSN: 2454-4248
46 – 51



1	RT @VIIPhoto: We asked to have a look inside some of VII photographers\u2019 gear bags. Read more: https:\/\/t.co\/75XZ9Cp6Jb\nHere's what's in @anu\u2026"
2	RT @connorPOWELL: do you think i can put on my resum\u00e9 that as long as i\u2019ve had an iphone 7 i have not lost my dongle as like a testament to\u2026"
3	RT @qtrachel16: \u201cim not updating my snapchat\u201d\n\nmy iphone: https:\/\/t.co\/GiJimaAFG9"
4	How to Ur iPad iPod Feb-2018: https:\/\/t.co\/aWfeWVqOgs via @YouTube"
5	Tell Me Yc display_te 140]
6	RT @manfightdragon: Facebook on iOS has a new link in the menu called \u201cProtect\u201d. Clicking it takes you to a page that installs absolute lit\u2026"
7	\ud83d\uc display_te 20]
8	RT @freedevo_: u kno when ya iphone charger starts wearin a turtleneck the end is comin ... https:\/\/t.co\/vmCNU5mFbb"
9	RT @freedevo_: u kno when ya iphone charger starts wearin a turtleneck the end is comin ... https:\/\/t.co\/vmCNU5mFbb"
10	Apple iPhone 6s Plus - 16GB - Rose Gold (T-Mobile) A1687 (CDMA + GSM) https:\/\/t.co\/bdd4VT0mhU"
11	RT @vlads but withc the difference\u2026"
12	@smark9! display_te 140]
13	Don't miss @iMore's amazing iPhone X giveaway! Enter here now! https:\/\/t.co\/xT0ps0w4yy"
14	#Google # Material ! more htt display_te 116]
15	What Are The Disadvantages Of Using Apple Products Such As iPhone And HomePad With Other Devices? https:\/\/t.co\/fGQelA4HIt"
16	RT @moviesnowtv: Congratulations to the winners of #100ManiaS5!\n3 Feb - Aniket Vipat - Laptop\n4 Feb - Sourabh Mishra - iphone 8\n5 Feb - Vin\u2026"
17	@KulaBra display_te 140]
18	Great thre display_te 78]
19	iPhone X https:\/\/t.co\/ixLwDIF4FG"
20	RT @MilePics: #MilePics #Free #Android #iPhone #iPad #Porn #Pics #App. View more at: https:\/\/t.co\/dSqM6aLGBt https:\/\/t.co\/29Fr4Q6CAk"

Figure 2: Tweets Collection

4.2 PRE-PROCESSING

The process of removing noise and pre-process the tweets are called as pre-processing. The main goal of this step is to make the data readable by machine and to reduce the ambiguity in feature extraction.

FILTERING

- Removal re-tweets: There may be a possibility of tweets which appear again and again. Those tweets should be removed.
- Case Conversion: The Tweets tweeted in upper case does not provide any additional specification while training the dataset.
- Removal of #hashtags and @username: Removing "@username" via regex matching or replace it with generic word AT_USER.and hashtags with the exact same word without the hash. E.g.-#boycott force with 'boycott force'.
- Removal of punctuations and additional whitespaces: Punctuations and white spaces at the starting and ending of the tweets should be removed and also replace multiple white spaces with the single white space.
- Tokenization: Tokenization is the act of breaking the sentence into keywords. Here breaking into words which give polarity. Tokens themselves can also be separators. The list of tokens becomes input for further processing such as parsing or text mining. Tokenization is useful both in linguistics and in computer science, where it forms part of lexical analysis.
- Stop words removal: Stop words that don't affect the meaning of the tweet are removed (for example and, or, still etc.). WEKA machine learning package can be used for this purpose, which checks each word from the text against a dictionary.
- Stemming: Stemming algorithms work by cutting off the end or the beginning of the word. This cutting can be successful in some occasions, but not always and that is why there are some limitations. E.g. the word "studies", "studying" into "study".

4.3. FEATURE EXTRACTION:

A feature is a piece of information that can be used as a characteristic which can assist in solving a problem. The quality and quantity of features is very important as they are important for the results generated by the selected model. Selection of useful words from tweets is feature extraction.

- **Unigram features: One** word is considered at a time and decided whether it is capable of being a feature.
- **N-gram features: More** than one word is considered at a time.
- **External lexicon: Use** of list of words with predefined positive or negative sentiment.

Frequency analysis is a method to collect features with highest frequencies. Further, some of them are removed due to the presence of words with similar sentiment (for example happy, joy, ecstatic etc.) and created a group of these words. Along with this affinity analysis is performed, which focuses on higher order n-grams in tweet feature representation. Here we use bigram feature extraction method to extract the features.

 ISSN: 2454-4248
46 – 51

4.4 FEATURE SELECTION

The feature selection techniques treat the documents either as group of words (Bag of Words (BOWs)), or as a string which retains the sequence of words in the document. BOW is used more often because of its simplicity for the classification process. The most common feature selection step is the removal of stop-words and stemming .The most frequently used feature selection methods are

- Point wise mutual Information
- Chi-square
- Latent Semantic Indexing

4.4.1 CHI-SQUARE

Let n be the total number of documents in the collection, pi(w) be the conditional probability of class i for documents which contain w, Pi be the global fraction of documents containing the class i, and F(w) be the global fraction of documents which contain the word w.

4.5. MACHINE LEARNING METHODS

Machine learning, a branch of artificial intelligence is a scientific discipline concerned with the design and development of algorithms that allow computers to evolve behaviours based on empirical data, such as from sensor data or databases. A learner can take advantage of examples (data) to capture characteristics of interest of their unknown underlying probability distribution. Data can be seen as examples that illustrate relations between observed variables. A major focus of machine learning research is to automatically learn to recognize complex patterns and make intelligent decisions based on data; the difficulty lies in the fact that the set of all possible behaviours given all possible inputs is too large to be covered by the set of observed examples (training data).

The machine learning approach is used for predicting the polarity of sentiments based on trained as well as test data sets. It can use supervised and unsupervised methods. The machine learning uses a supervised approach when there is a finite set of classes (positive and negative). This method needs labelled data to train classifiers. In a machine learning based classification a training set is used by an automatic classifier to learn the different characteristics of documents, and a test set is used to validate the performance of the automatic classifier. The unsupervised methods are used when it is difficult to find labelled training documents.Unsupervised approaches to document-level sentiment analysis are based on determining the semantic orientation (SO) of specific phrases within the document. If the average SO of these phrases is above some predefined threshold the document is classified as positive, otherwise it is deemed negative.

1	AT_USER i see you still have a delightful hard on for obama the man youull never match in grace dignityu
2	rt AT_USER matt schlapp and cpac are getting ready for another exciting event big difference from those days when president obamu
3	rt AT_USER obama sits alone in a classroom after meeting for hours with parents of sandy hook victims nntrump sits at a discu
4	rt AT_USER note that muellerus indictments have swiftly moved trump from denying russian interference to conceding it and from defenu
5	rt AT_USER just now broke up with my girlfriend so now ium looking for someone with the following attributesn kindn lovingn beautiu
6	AT_USER itus comical how hard and long mueller is having to look for some trumprussia connection mostly catchingu
7	rt AT_USER it never ceases to amaze me how successful you have been making yourself so small petty and banal with your tweets youru
8	rt AT_USER barack obama on common sense gun laws listen to this listen to every word of this
9	rt AT_USER he did he tried to warn us mcconnell blocked it then he expelled diplomats then putin expelled american diplomats fu
10	rt AT_USER anybody remember when judge napolitano was suspended amp nearly fired by fox for reporting that barack obama went uoutsideu
11	rt AT_USER obamaus other guy was scrubbed too
12	AT_USER AT_USER AT_USER yet obama didnut do anything significant about it just being more ucflexibleud wmedvedevdisplay_text_range
13	rt AT_USER it never ceases to amaze me how successful you have been making yourself so small petty and banal with your tweets youru
14	rt AT_USER sara carter nunes drills senior obama officials on bogus steele dossier
15	rt AT_USER obama appointed a leading physicist as wh science adviserngw bush also appointed a leading physicist as wh science adviseru
16	rt AT_USER president trump is continuing to blame former president obama for not doing enough to deter russian interference in the elecu
17	rt AT_USER for years obama aggressively minimized the threat posed by putin amp russia dems and obama constantly talked about russiau
18	rt AT_USER AT_USER why wont journalists do more reporting on the huge role AT_USER played and his threats to obama he nu
19	rt AT_USER this is how our institutions amp checks and balances broke down under obamaus thuggish regime
20	rt AT_USER moscow hacked your opponents you encouraged it moscow called for your opponentsu imprisonment so did you moscow stoku

Figure 3: Pre-rocessed Tweets

 ISSN: 2454-4248
46 – 51

Naïve Bayes Classifier

It is an approach particularly suited when the dimensionality of the inputs is high. It is used to predict the probability for a given words to belong to a particular class. It is used because of its easiness in both during training and classifying steps. Pre-processed data is given as input to train input set using Naïve Bayes classifier and that trained model is applied on test to generate either positive or negative sentiment. The Bayes theorem is as follows.

$$P(X/H) = (P(H/X)\ P(H))/P(X)$$

Where X- Tuples, H-Hypothesis, P(H|X) represent Posterior probability of H conditioned on X

i.e. the Probability that Hypothesis holds true given the value of X, P(H) represents Prior probability of H
i.e the Probability that H holds true irrespective of the tuple values .P(X/H) represents posterior probability of X conditioned on H. i.e the probability that X will have certain values for a given hypothesis ,P(X) represents prior probability of X
i.e the probability that X will have certain values.

The proposed system understands whether the tweet is positive or negative based on the dictionary methods of score. An experiment result of accuracy is evaluated using following information retrieval matrices. Accuracy is the performance evaluation parameter and it is calculated by number of correctly selected positive and negative words divide by total number of words present in the corpus. The formula is given as below.

$$Accuracy = \sum True\ Positive + \sum True\ Negative / \sum Total\ number\ of\ words.$$

Where True positive is number of tweets recognized as positive and true negative is number of tweets recognized as negative respectively. The performance metrics used to evaluate the classification results are precision, recall and F-measure.

5. TWITTER SENTIMENT ANALYSIS WITH PYTHON

Python is a high level, interpreted programming language, popular for its code readability and compact line of codes. It uses white space inundation to delimit blocks. Python provides a large standard library which can be used for various applications for example natural language processing, machine learning, data analysis etc. The pre-processing in Python is easy to perform due to functions provided by the standard library. Some of the steps are given below:

- Converting all upper case letters to lower case.
- Removing URLs: Filtering of URLs can be done with the help of regular expression
- Removing Handles (User Reference): Handles can be removed using regular expression - @(\w+) .
- Removing hashtags: Hashtags can be removed using regular expression - #(\w+).
- Removing emoticons: We can use emoticon dictionary to filter out the emoticons or to save the occurrence of them in a different file.
- Removing repeated characters.

6. DATSETS

- **Stanford Twitter Sentiment Corpus (STS)**

This dataset consists of 60,000 tweets randomly selected from the Stanford Twitter Sentiment corpus (STS). The original dataset contained 1.6 million general tweets, and its test set of manually annotated tweets consisted of 177 negative and 182 positive tweets. In contrast to the training set which was collected based on specific emoticons, the test set was collected by searching Twitter API with specific queries including product names, companies and people.

- **Obama-McCain Debate (OMD)**

The Obama-McCain Debate (OMD) dataset was constructed from 3,238 tweets crawled during the first U.S. presidential TV debate in September 2008. This resulted in a set of 1,081 tweets with 393 positive and 688 negative ones. Due to the relative small size of this dataset, and the lack of a test set, we opted for a 5-fold cross validation approach instead.

 ISSN: 2454-4248
46 – 51

	A	B	C	D	E
1		text	class	prob	
2	0	Congrats @sundarpichai well deserved! Proud moment.	pos	5.1	
3	1	Congratulations @sundarpichai. My best wishes for the new role at @google.	pos	6.1	
4	2	The choice is ultimately between diplomacy and war. Iran's nuclear program accelerates if Congress ki	pos	5.3	
5	3	The movie #drive is seriously the worst movie I've ever seen.	neg	4.6	
6	4	News of another attack in Kabul is very saddening. Such attacks are unpardonable & deeply distressin	neg	4.7	
7					

Figure 2: Classified Tweets

7. SENTIMENT ANALYSIS CHALLENGES.

- Identifying subjective parts of text
- Domain dependence
- Sarcasm Detection
- Explicit Negation of sentiment
- Order dependence
- Entity Recognition
- Building a classifier for subjective vs. objective tweets.
- Handling comparisons.
- Applying sentiment analysis to Face book messages.

8. SENTIMENT ANALYSIS APPLICATIONS

- Reviews from Wesites
- Sub-component Technology
- Business Intelligence
- Applications across Domains
- Smart Homes
- Commerce
- Politics
- Sports Events

9. CONCLUSION

The main objective of the proposed work is to develop a framework for sentiment analysis classification in a real-time environment using Twitter data as the source of content. Twitter sentiment analysis, a category of opinion mining focuses on analyzing the sentiments of the tweets as positive, negative and neutral sentiments. The tweets have gone through the steps like data collection, text pre-processing, feature detection, sentiment classification. The features of the tweets are selected based on chi-square method and naïve bayes classifier is used to classify the tweets as positive, negative and neutral. The proposed work is implemented using python. Experimental results support the user in decision making process in their day today life.

10. REFERENCES

[1] A.Pak and P. Paroubek. “Twitter as a Corpus for Sentiment Analysis and Opinion Mining". In Proceedings of the Seventh Conference on International Language Resources and Evaluation, 2010, pp.1320-1326

[2] Bhumika Gupta and Monika Negi, Kanika Vishwakarma, Goldi Rawat, Priyanka Badhani,”Study of Twitter Sentiment Analysis using Machine Learning Algorithms on Python”,International Journal of Computer Applications (0975 – 8887) Volume 165 – No.9, May 2017

[3] Walaa Medhat , Ahmed Hassan , Hoda Korashy,” Sentiment analysis algorithms and applications-A survey”, Ain Shams Engineering Journal (2014) 5, 1093–1113

[4] S.Siddharth, R.Darsini, Dr. M. Sujithra, “Sentiment Analysis on Twitter Data Using Machine Learning Algorithms in Python”, International Journal of Engineering Research in Computer Science and Engineering (IJERCSE) Vol 5, Issue 2, February 2018

[5] Riya Suchdev ,Pallavi Kotkar ,Rahul Ravindran ,Sridhar Swamy, ,” Twitter Sentiment Analysis using Machine Learning and Knowledge-based Approach”, International Journal of Computer Applications (0975 – 8887) Volume 103 – No.4, October 2014 .

[6] I.Hemalatha, Dr. G.P.S.Varma, Dr. A.Govardhan,” Automated Sentiment Analysis System Using Machine Learning Algorithms”, International Journal of Research in Computer and Communication Technology, Vol 3, Issue 3, March- 2014

[7] Dipak R. Kawade, Dr.Kavita S. Oza,” Sentiment Analysis: Machine Learning Approach”, Dipak R. Kawade et al. / International Journal of Engineering and Technology.