Proc. of Int. Conf. on Advances in Computer Science and Application, CSA

Twitter Sentiment Classification on Sanders Data by using Machine Learning Approach

Kishori. K. Pawar [1] and R. R. Deshmukh [2]
[1, 2] Department of Computer Science & IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, Maharashtra, India
Email: akishoripawar@gmail.com, Email[2]: ratnadeep_deshmukh@gmail.co.in

***Abstract*—Sentiment analysis is very perplexing field of social data mining. The tweets give peoples' opinions. In this work, we have used pre-labeled Sanders Analytics data. We used machine learning approach and RTextTools. The steps involved in this works are creation of corpus, DTM, container, training data using ML algorithm and classification using training models. Comparative study of accuracy on various classifiers & Empirical study of collected tweets is being carried out using wordcloud of tweets using R packages.**

***Index Terms*— Sentiment analysis, Machine Learning, Twitter, RTextTools, Wordcloud.**

I. INTRODUCTION

Keeping the opinions, views, sentiments on social media is a general trend nowadays. These views can be of a company, consumer products, person, customer services and anything. Thus social media data like tweets about a topic contains huge amount of information. These tweets are useful to consumers as well as manufacturer if it is about certain products or brands. Tweets can also be used for public advantage in a democracy if tweets say about a person or party. Extracting the polarity or sentiments from the tweets is challenging task due to natural language complexity, dense form of tweets, slang words and short forms of words etc. [1]

Sentiment Analysis is popular text mining which identify and extract subjective information into various polarity classes. Thus the result of sentiment analysis and classification can be used in strategic, managerial, and operational decision making. [2]

As sentiment classification is about extracting opinions, they are mainly surrounded to a topic, to which user labels as positive, neutral or negative [3]. Thus it is necessary to find about the topic on which a user want to comment.

Social Media Data is a big data to process. Thus learning the patterns in the data is the easy and efficient way for the knowledge extraction. Hence we have used various machine learning classifiers to classify the tweets into positive, negative and neutral tweets.

The steps involved in this works are creation of corpus, then creating document term matrix, then creating an object container, then training the data using machine learning algorithm and finally the classification of the data using training models into classes.

II. LITERATURE REVIEW

Sentiment mining is a part of data mining which process the Electronic text and tag the words into three

Grenze ID: 02.CSA.2015.4.19

categories that is positive, negative and neutral. Different techniques are used for sentiment analysis, classification and summarization. Different techniques used for sentiment summarization are Data mining, classification of Text, Information Retrieval and Summarization of Text shows general structure of sentiment analysis. Sentiment analysis can be achieved at various levels, the levels are: Phrase Level, Aspect Level, Sentence Level, Document Level, Natural Language Processing. Depending upon nature of use level of Sentiment analysis is selected [4].

Sentence level classification is deals with the considering polarity of each sentence. Document level classification can also be applied to sentence level classification to classify the sentences in polarity. Here also we have to consider the subjectivity and objectivity of the sentence. Subjective sentences contain words related to particular domain. Single sentence contains single opinion about single domain. Complex sentence are also commented in reviews. In such case sentence level classification cannot be useful. Sentence level classification is deals with the positive, negative and neutral sentiments. Sentence level classification is deal with the subjectivity classification. For Example, "I brought Canon Camera last week. At initial stage everything was good. The pictures were high quality and clearer, although it was bit bulky. Then it stops working today". The first sentence contains no opinion as it simply states a fact. All other sentences express implicit and explicit opinions. The last sentence "Then it stops working today" is objective sentence but current used methodology cannot express opinion for the above sentences even it carry negative sentiment or undesirable sentiment [5].

Machine learning methodology consist of supervised, unsupervised and semi supervised categories .Each category is again sub divided as shown in figure 2.1. Supervised Learning methodology predicts attribute classes on the basis of given set of training values. It contains training and testing dataset. Training dataset is smaller which contain same attributes as testing dataset. It is more efficient and accurate. A training dataset created model test on test corpora contains the same attributes but no predicted attribute. Accuracy of model checked that how accurate it is to make prediction. Classification is a supervised learning used to find the relationship among attributes. Prediction hit rate is used to measure the accuracy of extracted rules that how true they are to make prediction by applying on test data [6].

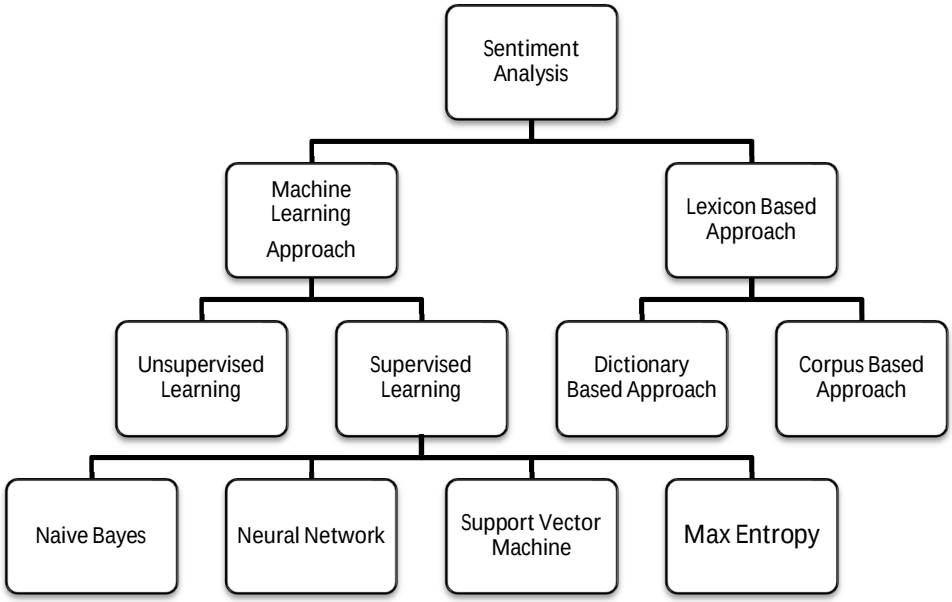

Figure 1: Techniques used in sentiment analysis

It contains different methodology like SVM (Support Vector Machine), NB (Naïve Bayes), ME (Maximum Entropy), Decision Tree, etc.

Liu et al. (2012) propose an approach called emoticon smoothed language model (ESLAM) which is able to integrate data labeled with noisy labels with manually annotated tweets. The authors first train a language model for 570 positive, 654 negative and 2503 neutral tweets of the Sanders corpus. Then they use noisy

emoticon data collected via the Twitter API to smooth those language models. For classification they pick the class of the model yielding the highest score. In the preprocessing step they insert generic placeholders for usernames, digits and URLs, remove stopwords, conduct stemming and lowercasing and remove retweets and duplicates from the dataset. Additionally, they distinguish links pointing to popular picture and video websites from other links. They find that their emoticon smoothed language model performs better than using a fully supervised language model and report an accuracy of 82.5% as their best result for polarity classification and 79.5% for subjectivity classification. In an evaluation of the effect of using manually labeled data the authors find that it significantly improves classification performance compared to using only data labeled with noisy labels [7].

Pang et al. (2002) employ three machine learning methods to classify movie reviews into two classes: positive and negative. Neutral reviews are omitted in the study, which made the problem less complicated. Dave et al. (2003) perform various supervised learning classification (Rainbow algorithm, SVMs, naive Bayes) on product reviews from websites C|net and Amazon.com. They are able to obtain satisfactory results for the review classification task through the choice of appropriate features and metrics [8].

Sakaki et al. (2010) use a probabilistic spatiotemporal model to build an autonomous earth quake reporting system in Japan using Twitter users as sensors. As an application, they construct an earthquake reporting system in Japan. Frequent seismic and the significant number of people using Twitter in the country enable them to spot an earthquake by monitoring feeds on Twitter with high probability. System reports earth movements rapidly and noticed users. The notification message is delivered much faster than the information broadcasted by the JMA. No papers written in the field of sentiment analysis and text mining use information extracted from Twitter to measure customer satisfaction. The specificity of the language used on Twitter imposes many challenges on the researcher in terms of data preparation for the analysis, which was not addressed by previous work. Such as, this Master's thesis aims to fill that gap [9].

The ideology of implementation of this system used in this project is based on the underlying principles developed in [10] where the tweets were classified using unigram vectors.

Riya Suchdev et al. [11] partition the obtained dataset into training and testing dataset. Polarity detection of test tweets is carried out in three steps: Preprocessing (elimination of the slang words, misspellings and other faults), Feature vector creation and Sentiment classification using classifiers. In this way tweets are classified into positive, negative and neutral classes.

Twitter Microblog Mining is performed by Eman M.G. Younis [12] in four phases. Data access, data cleaning, and data analysis. Lexicon based approach is used to classify tweets. Twitter API is used to extract tweets using R. Using tm package in R data cleaning is performed, it includes removal of punctuation, spaces, stop words, URLs and stemming. The third phase includes finding association rules, frequent terms, and finding overall polarity of tweet. Tweet is scored using a sentiment score function. In the last phase of visualization, wordcloud and bar is plotted, it shows the frequency of words.

Alexander Pak et al. [13] did sentiment analysis in three phases: Corpus collection, Corpus analysis , Training the classifier [Feature extraction, Classifier, Increasing accuracy] .Based on happy emotions [:-), :), etc.] and sad emotions [:-(, :(, etc.] Alexander Pak et al. had presented a method to collect a corpus with positive and negative sentiments, and a corpus of objective texts. Word frequencies of the corpus is calculated and plotted in corpus analysis phase. Subjectivity or objectivity of tweets is calculated using the proportion of presence of POS tags in the tweets. Third most important phase is training phase. N-gram feature is used in feature extraction phase. The steps to extract n-grams from twitter post is as follows: i) Filtering (Removal of URL, Retweets, user name, and emotions), ii) Tokenization (segmenting the text by splitting it by punctuation marks and spaces), iii) Removing stopwords (removal of a, an, the), iv) Constructing n-grams (making n-grams of consecutive words). After extracting feature, they build a sentiment classifier using the multinomial Naive Bayes classifier. The naive bayes classifier is trained using two features i.e. POS and n-gram. Finally they tested the classifier on Twitter posts and performed sentiment classification.

Akshi Kumar et al. [14] uses lexicon based approach to classify tweets into positive, negative and neutral tweets. The first phase is collection of tweets using twitter API. Then preprocessing which includes removal of removal of retweets, URLs, replacing emotions with their sentiment polarity, removal of punctuations, and use of POS tagger. Preprocessing phase is succeeded by scoring phase. Each tweet is scored in scoring module using linear regression module and WordNet. Finally if the sentiment score is less than zero the tweet is considered as negative, if it is zero it is classified as neutral and if score is positive tweet is considered as positive.

III. MOTIVATION

The micro-blogging platform Twitter offers a high potential for sentiment and opinion mining due to its large world-wide user base and its open data access models. People use Twitter to share their thoughts and opinions in real-time making it perfect for marketing research and quick incident reaction.

IV. METHODS

A. Random Forest

Random forests operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random forests correct for decision trees' habit of overfitting to their training set. Random Forests grows many classification trees. To classify a new object from an input vector, put the input vector down each of the trees in the forest. Each tree gives a classification, and we say the tree "votes" for that class. The forest chooses the classification having the most votes (over all the trees in the forest). [15]

B. Support Vector Machine

The first step is feature selection – the unsupervised identification of a reasonably small set of features in which the essential information content of the input data is concentrated. The second step is the classification where the feature domains are assigned to individual classes. Given a set of training examples, each marked for belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on. [16]

C. Neural networks

During training, a neural network looks at the patterns of features (e.g. words, phrases, or N-grams) that appear in a document of the training set and attempts to produce classifications for the document. If its attempt doesn't match the set of desired classifications, it adjusts the weights of the connections between neurons. It repeats this process until the attempted classifications match the desired classifications. [17]

D. Bagging

It is a name derived from "bootstrap aggregation", was the first effective method of ensemble learning and is one of the simplest methods of arching. The meta-algorithm, which is a special case of the model averaging, was originally designed for classification and is usually applied to decision tree models, but it can be used with any type of model for classification or regression. The method uses multiple versions of a training set by using the bootstrap, i.e. sampling with replacement. Each of these data sets is used to train a different model. The outputs of the models are combined by averaging (in case of regression) or voting (in case of classification) to create a single output. Bagging is only effective when using unstable (i.e. a small change in the training set can cause a significant change in the model) nonlinear models. [18]

E. MAXENTROPY

The Max Entropy classifier is a probabilistic classifier which belongs to the class of exponential models. Unlike the Naive Bayes classifier that we discussed in the previous article, the Max Entropy does not assume that the features are conditionally independent of each other. The MaxEnt is based on the Principle of Maximum Entropy and from all the models that fit our training data, selects the one which has the largest entropy. The Max Entropy classifier can be used to solve a large variety of text classification problems such as language detection, topic classification, sentiment analysis and more.

V. METHODOLOGY

We have followed following steps to do sentiment analysis.

A. Corpus Creation

We have used Twitter Sentiment Corpus version 0.2 in this work. These are 5500 hand-classified tweets on 4 topics. These tweets are labeled as positive, negative, neutral and irrelevant. Among which 1786 irrelevant

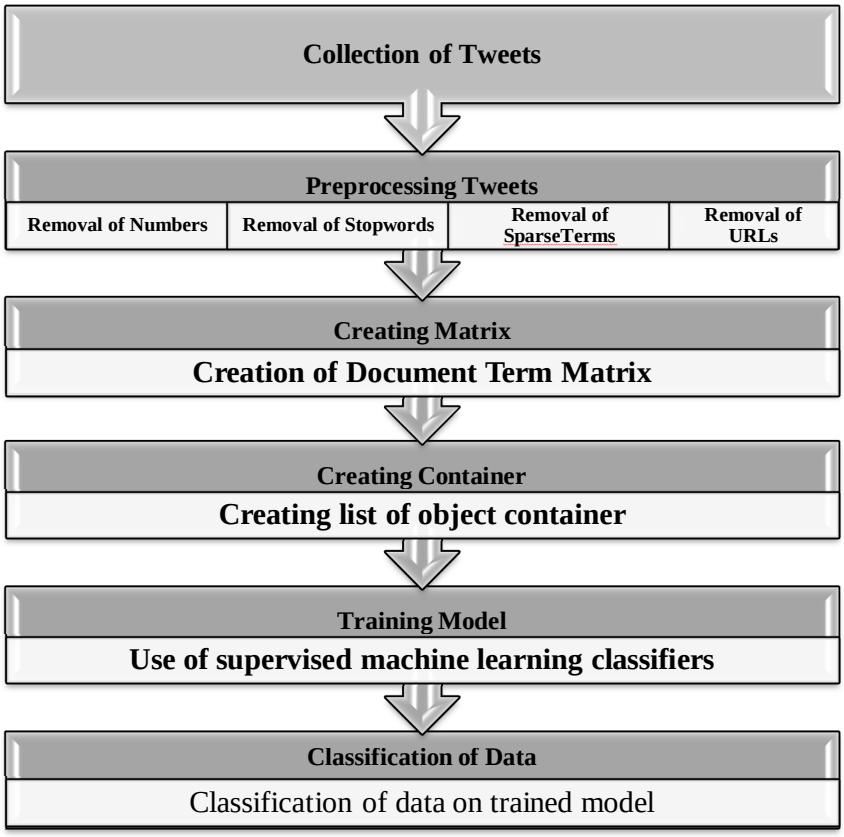

Figure 2: Workflow for Sentiment Analysis

tweets are not considered in this work because they are irrelevant to the topic and they are not in English language. In this corpus, there are 570, 654, 2503, positive, negative and neutral tweets respectively. [19]

B. Preprocessing tweet

In this phase all the text data is cleansed off. All unnecessary white spaces, tabs, newline character is removed from the text. The URLs from the tweets are removed. The RT tag mentioned before every re-tweeted tweet is removed. All punctuations, numbers are also removed from the tweets. Unnecessary sparse terms are removed. The stopwords are removed from the tweets. All text is converted to lowercase to have consistent messages. Stemming is performed on each word of tweet.

C. Creating Matrix and Container

Here, the initial step is to generate a document term matrix. This is being carried out using R's tm package. The matrix is then partitioned into a container, which is essentially a list of objects that will be fed to the machine learning algorithms in the next step. The output is of class matrix_container and includes separate train and test sparse matrices, corresponding vectors of train and test codes, and a character vector of term label names. a container object is created, which holds all the objects needed for further analysis.

D. Train the Model

As we are using supervised classification, we can use algorithms to train their data. The data must be labelled with the corresponding class name. The first 2600 tweets will be used to train the machine learning model, and the last 1127 tweets will be set aside to test the model.

E. Classification

Based on the learning of trained model, the data are classified. After that, the classification is summarized and performance evaluation is being carried out, where we can check efficiency of classifiers.

F. Cross validation

Users may use n-fold cross validation to calculate the accuracy of each algorithm on their dataset and determine which algorithms to use in their ensemble. We can perform n-fold cross validation. While deciding the appropriate n-fold it is important to consider the total sample size so that enough data are in both the train and test sets to produce useful results.

VI. Test Results

The tables are designed to have a uniform style throughout the paper? It does not matter how you choose to place the inner lines of the table, but we would prefer the border lines to be of the style shown in Table 1. For the inner lines of the table, it looks better if they are kept to a minimum. The caption heading for a table should be placed at the top of the table.

TABLE I. ESIMATED RESULTS ON SANDERS DATA

Classifiers	Accuracy	PRECISION	RECALL	FSCORE
SVM	0.754421	0.546667	0.493333	0.513333
BAGGING	0.737807	0.546667	0.46	0.486667
FORESTS	0.75888	0.486667	0.486667	0.483333
TREE	0.67154	0.396667	0.35	0.33
NNETWORK	0.648179	0.343333	0.4	0.35
MAXENTROPY	0.844883	0.446667	0.49	0.456667

TABLE II. CLASSIFIED TWEETS WITH THEIR OPINION SCORE

	tweet	score	Class
1	RT @iknightbewrong: I'm OVER people bitching about the #iPhone4S... I think it's the smartest phone I've ever had and I'm very happy. :) Way to go @Apple!	2	Positive
2	What a fantastic service I've been given by Malcolm and Dom at Manchester's @apple Store! Thank you guys!! :-)	3	Positive
3	Google Earth Helps Locate Salmonella Hotspots http://t.co/mcc9RSDf #google	1	Positive
4	Not impressed much with the new Android update. But good signs: a readable font, emphasis on design, and less nerdiness. #google	3	Positive
5	#iCloud set up was flawless and works like a champ! To the Cloud @Apple	3	Positive
6	shit, shit, shit. IOS5 update ate all my apps, data and media just like @apple said it would. This is going to take some time to rebuild.	-3	Negative
7	@bisquiat @Apple the upgrade just slows down my phone so much, it's stuck half the time. uch. thankfully no other damage. sucks for you :(	-6	Negative
8	@FishMama: If you made a purchase, just wait for the @apple survey! hate going b/c of the bad #custserv	-3	Negative
9	@APPLE WOW @MOTOROLA Just crushed your dreams....	-1	Negative
10	iTunes is @Apple's worst product. Worse than the #Newton or the hockey puck mouse. It's utterly painful to use.	-4	Negative
11	thinking thinking thinking thinking #dotnet #asp #microsoft	0	Neutral
12	#HEUTE - #Microsoft #Office #2010 Home & Student Product Key Card [1 User] - statt 149â‚¬,~ nur 89,99â‚¬,~ - http://t.co/hbKnmgdE	0	Neutral
13	Karate kid, skittles and cranberry juice. Goodnight #twitter	0	Neutral
14	Top 50 #Twitter Acronyms, Abbreviations and Initialisms Â http://t.co/nEqHcJsY Â /via @ruhanirabin	0	Neutral
15	#google #Android #ice cream sandwich	0	Neutral

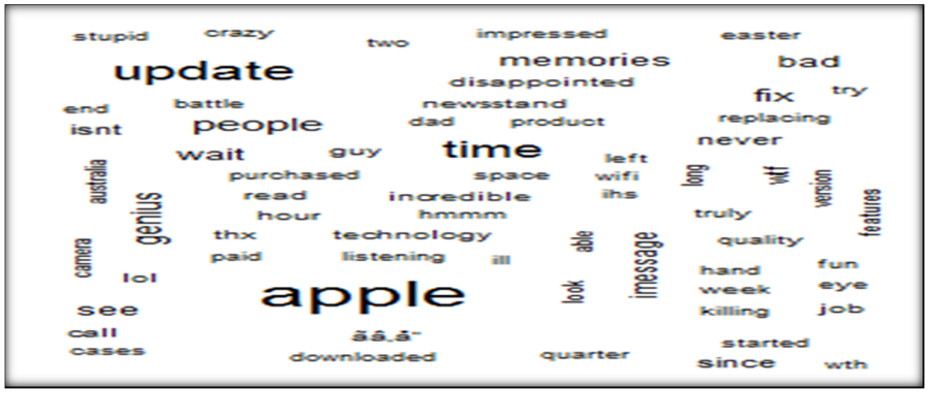

Figure 3: Wordcloud of tweets which are on subject Apple in Sanders Data

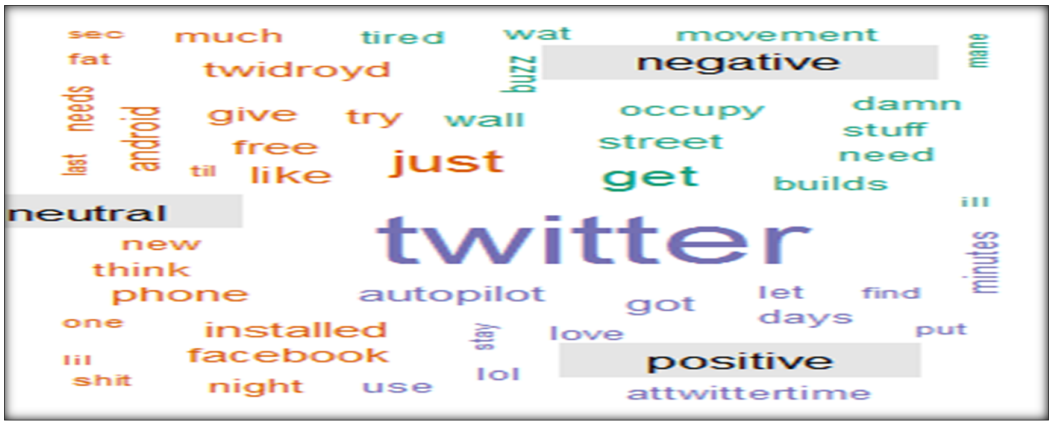

Figure 4: Wordcloud of Polarity labeled tweets on subject Twitter in Sanders

VI. Summarization

The twitter sentiment classification is carried out giving the resultant opinion score for each tweet. If the score of the tweet is greater than 0 it is considered to be positive, if it is less than 0 it is considered to be negative and if it is zero it is considered as neutral. Table 1 shows the Accuracy, Precision, Recall, F-score of individual classifier on sanders data. Table 2 shows a small subset of Sanders tweets with obtained opinion score and resultant polarity.

Using various R packages we have drawn the wordclouds of our datasets. These wordclouds are useful to do empirical study. Fig 3 shows the wordcloud of tweets relating to subject apple. Thus this wordcloud is useful to study most frequent words discussed in the text. Thus as these tweets are the reviews regarding apple company, we can analyse that the mostly discussed words about apple are: product, features, update, version, download, technology, camera etc. Along with these, the words like incredible, genius, purchased, impressed, and started give us the positive feedback in the tweets. In contrast words like killing, stupid, disappointed, ill, battle, crazy give the negative feedback about the product. Thus overall opinion about the product or service can be drawn using analysing the wordclouds.

Extending the concept of simply drawing a wordcloud, we have also plotted the wordcloud differentiating on particular feature in Fig 4. Here the wordcloud is plotted on tweets on the subject twitter. The wordcloud is divided into sentiment classes (here positive, negative and neutral) each section gives the frequently occurring words belonging to respective class. For example, in this wordcloud ‘love’ word occurred in positive sentiment section, ‘damn’ word occurred in negative sentiment section while the words like ‘Facebook’, ‘phone’, ‘android’ occurred in neutral sentiment section. In the same way we have drawn wordclouds on different subjects and we can have analysis of discussion made in the tweets. From the graphical representation of bar plot in Fig 4 and 5 we can summarize that in the domain of ‘Microsoft’ there are 650 positive, 75 neutral and 200 neutral tweets present in the dataset. And in the domain of ‘Google’, there are 150 tweets showing the emotion of joy, 50 showing sad, 24 of surprise, 11 of anger, 10 of fear, 5 of disgust,660 unknown. Similar to the above results we have did the graphical analysis on tweets of the topic Twitter, Google, Microsoft, Apple.

VII. Conclusions

Sentiment Classification is being carried on Sanders Analyst Data, resulting the classes of positive, negative and neutral classes. From the experimental results and discussion we came to the conclusion that machine learning approach can be used to extract word level sentiments of the tweets, further it can give the highest accuracy of about 84.48 % using Max Entropy Text Classifier. We can state that R environment is a very good framework and statistical and programming language for data mining, analysis and visualizing the results.

This work has been carried out on a small set of tweets. To do peer analysis the same methodology has to apply on big collection of tweets. Only machine learning features are utilized in this work, in future we would like to do sentiment analysis using machine learning features.

ACKNOWLEDGMENT

This work has been partially supported by Department of Computer Science and IT Dr. Babasaheb Ambedkar Marathwada University, Aurangabad. The views expressed here are those of the authors only.

REFERENCES

[1] Kishori K. Pawar, Pukhraj Shrishrimal, R. R. Deshmukh, “Twitter Sentiment Analysis: A Review”, International Journal of Scientific & Engineering Research (957-964), Volume 6, Issue 4, April-2015.
[2] Bo Pang, Lillian Lee, “Opinion mining and sentiment analysis”, Foundations and Trends in Information Retrieval Vol. 2, No 1-2 (2008) 1–135.
[3] Bing Liu, Sentiment analysis and opinion mining, Synthesis Lectures on Human Language Technologies (1-167), no. 1, 2012.
[4] Sowmya Kamath S, Anusha Bagalkotkar, Ashesh Khandelwal, Shivam Pandey, Kumari Poornima, “Sentiment Analysis Based Approaches for Understanding User Context in Web Content”, International Conference on Communication Systems and Network Technologies (CSNT), 978- 0-7695-4958-3/13, 2013.
[5] Keisuke Mizumoto, Hidekazu Yanagimoto and Michifumi Yoshioka, “Sentiment Analysis of Stock Market News with Semi-supervised Learning”, IEEE Computer Society, IEEE/ACIS 11th International Conference on Computer and Information Science, p.325-328, 2012.
[6] Ayesha Rashid, Naveed Anwer, Dr. Muddaser Iqbal, Dr. Muhammad Sher, “A Survey Paper: Areas, Techniques and Challenges of Opinion Mining”, IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 6, No 2, November 2013.
[7] Bing Liu. “Sentiment analysis and opinion mining”. Synthesis Lectures on Human Language Technologies, 5(1):1-167, 2012.
[8] F. Sebastiani, “Machine learning in automated text categorization”, ACM Computational Survey, 34(1):1-47, 2002.
[9] T. Sakaki, M. Okazaki, and Y. Matsuo. “Earthquake shakes twitter users: real-time event detection by social sensors”. 19th international conference on World Wide Web, 2010.
[10] Timothy P. Jurka, Loren Collingwood, Amber E. Boydstun, Emiliano Grossman and Wouter van Atteveldt (2012). RTextTools: Automatic Text Classification via Supervised Learning. R package version 1.3.9. http://CRAN.R-project.org/package=RTextTools
[11] Riya Suchdev, Pallavi Kotkar,Rahul Ravindran, "Twitter Sentiment Analysis using Machine Learning and Knowledge-based Approach", International Journal of Computer Applications (0975 – 8887),Volume 103 – No.4, October 2014.
[12] Eman M.G. Younis,"Sentiment Analysis and Text Mining for Social Media Microblogs using Open Source Tools: An Empirical Study", International Journal of Computer Applications (0975 – 8887), Volume 112 – No. 5, February 2015.
[13] Alexander Pak, Patrick Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining", International Conference on Language Resources and Evaluation (17-23), LREC 2010, May 2010.
[14] Akshi Kumar, Teeja Mary Sebastian, "Sentiment Analysis on Twitter", IJCSI International Journal of Computer Science Issues, (1694-0814), Vol. 9, Issue 4, No 3, July 2012.
[15] https://en.wikipedia.org/wiki/Random_forest, sited on May 30, 2015.
[16] https://en.wikipedia.org/wiki/Support_vector_machine, sited on May 28, 2015.
[17] Miss. Vidya Alone, Mrs .R.B.Talmale, “Message Filtering Techniques for On-Line Social Networks: A Survey”, International Journal of Application or Innovation in Engineering & Management, Volume 3, Issue 3, March 2014.
[18] Arching (adaptive reweighting and combining) is a generic term that refers to reusing or selecting data in order to improve classification.
[19] Sanders, Niek J. "Twitter Sentiment Corpus." Sanders Analytics. Sanders Analytics LLC., 2011. Web. 16 Nov. 2013.