IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 17, Issue 4, Ver. I (July – Aug. 2015), PP 118-123
www.iosrjournals.org

Twitter Sentiment Classification on Sanders Data using Hybrid Approach

Kishori K. Pawar[1], R. R. Deshmukh[2]
[1, 2]*(Department of Computer Science & Information Technology, Dr. Babasaheb Ambedkar Marathwada University Aurangabad (MS) India*

Abstract : *Sentiment analysis is very perplexing and massive issue in the field of social data mining. Twitter is one of the mostly used social media where people discuss on various issues in a dense way. The tweets about a particular topic give peoples' views, opinions, orientations, inclinations about that topic. In this work, we have used pre-labeled (with positive, negative and neutral opinion) tweets on particular topics for sentiment classification. Opinion score of each tweet is calculated using feature vectors. These opinion score is used to classify the tweets into positive, negative and neutral classes. Then using various machine learning classifiers the accuracy of predicted classification with respect to actual classification is being calculated and compared using supervised learning model. Along with building a sentiment classification model, analysis of tweets is being carried out by visualizing the wordcloud of tweets using R.*
Keywords: *Sentiment analysis, Machine Learning, Twitter, Opinion score, R packages, Wordclouds.*

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.

II. Proposed Work

Machine learning and Lexicon based, these are the common two approaches to do sentiment classification. We have used hybrid approach i.e. machine learning as well as lexicon based approach to do sentiment classification. Following is the basic flow carried out to do sentiment classification.

2.1 Data Collection

For sentiment classification we have used two corpuses of pre-labeled tweets.

2.1.1 Twitter Sentiment Corpus by Sanders

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 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. [4]

2.1.2 AFINN

Along with the corpuses we have also collected the list of positive, negative words which are useful while feature extraction. AFINN is a list of words rated each by its valence which ranges from -5 to +5. We are using AFINN-111 version in our study which contains 2477 words and phrases. [5]

2.1.3 OpinionFinder

OpinionFinder contains list of 1600 positive and 1200 negative word lists. [6]

2.1.4 Opinion Lexicons

This is a list of positive and negative opinion words for english developed by Bing Liu and Minquing Hu at the University of Illinois at Chicago. This Lexicon library contains around 6800 words. [7]

2.2 Preprocessing

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. 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.

2.3 Feature Extraction

Feature extraction is carried out using rule based learning. Following features and their respective values are being extracted through preprocessed tweets. The feature values are also utilized to calculate the sentiment score of each tweet.The following set of features has been used:

2.3.1 n-gram feature

We find useful set of unigrams after removing stopwords from the text. Those unigrams are extracted who give good information gain.

2.3.2 Lexicon Feature

Following are the lexicon features extracted from text.

- **Positive lexicons**

We have merged set of positive words from the existing positive lexicons i.e. AFINN, OpinionFinder. There are 3372 positive words considered in this work. And these set of positive lexicons are used to tag positive polarity in the tweets. Further we have evaluated opinion score of each tweet from feature values.

- **Negative Lexicons**

We have used 4787 negative lexicons in this work. Just like positive lexicons these negative lexicons are used to tag negative polarity in the tweet.

2.3.3 POS Feature

POS features are nothing but number or count of Part-of-Speech features from the tweets. These POS contains nouns, adjectives, adverbs, verbs, etc. lexicons are used to tag positive polarity in the tweets. Further we have evaluated opinion score of each tweet from feature values.

2.3.4. Microblogging Feature

- **PosSmiley**

A set of smileys (:-D, (:, :-))indicating positive sentiments along with the sentiment score is used to evaluate the PosSmiley feature value of each tweet. The set of positive smiley emoticon is collected from dataset [14]. This dataset contains 85 emoticons with polarity. We have added a set of extra smileys into the available emoticon database manually to make a generalized set of emoticons.

- **NegSmiley**

Similar to PosSmiley NegSmiley(:'(,:'-(,:() feature is used to evaluate the negative sentiments from the tweets.

- **NeuSmiley**

These are the smileys indicating neutral emotion i.e. neither positive nor negative sentiments. Some of the neutral smileys are :-\, :-/, :-O, :-|, etc.

- PosHashtag

The hashtag is a word associated with every tweet denoting overall moto, emotion behind every tweet. It starts with # symbol. Thus extracting sentiment of hashtag is vital to know sentiment of the tweet. Thus PosHashtag feature's value is scored to 1 if hashtag bears positive sentiment. Example of PosHashtag is #happymoments

- **NegHashtg**

Similar to PosHashtag, NegHashtag is used to extract negative sentiment from the tweets.

2.4 Sentiment Classification using Classifiers

We used supervised machine learning approach. Different machine learning classifiers have been used by us on our model to classify the tweets into their respective classes of sentiments. Following are the machine learning classifiers used in this study:

2.4.1. Naive Bayes

This method computes the probability of a text document is about a particular topic, using the words of the document to be classified and the estimated probability of each of these words as they appeared in the set of training documents for the topic. [8]

2.4.2 Neural networks

During training, a neural network looks at the patterns of features (e.g. words, N-grams or phrases) that appear in a document of the training set and tries to produce classifications for the document. If its effort doesn't match the set of desired classifications, it amends the weights of the connections between neurons. It replications this process until the attempted classifications match the desired classifications. [9]

2.4.3 Linear discriminant analysis

Discriminant Analysis classifies the classes into mutually exclusive and exhaustive groups using set of measurable features. LDA classifies objects (here sentiment polarity) into set of features (neutral , positive, negative).

Sentiment polarity is a dependent variable whose value can be neutral, positive, or negative. While other features like number of positive, negative hashtags etc. are independent variables. So in discriminant analysis, the reliant variable(Y) is the collection and the independent variables (X) are the object features that might describe the collection. The reliant on variable is always category (nominal scale) variable while the independent variables can be any dimension scale. [10]

2.4.4 Quadratic Discriminant Analysis

QDA is a general discriminant function with quadratic decision boundaries which can be used to classify datasets with two or extra classes. LDA has less expectedness power than QDA but it needs to estimate the covariance matrix for each classes. [11]

2.4.5 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. [12]

2.4.6 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). [14]

III. Experimental Results

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 a small subset of Sanders tweets with obtained opinion score and resultant polarity.

As explained in the proposed work the sentiment classification of the given datasets is carried out and the following accuracy results and F-score for Sanders datset and Standford Twitter Sentiment dataset are obtained in Table 2 and 3.

Confusion matrix for each sentiment classifier and its classification is evaluated. Some of those are shown in figure 1. [15]

We have plot the wordcloud (using R)of the mostly discussed words from the tweets and have an empirical study of text mining. For example in the sanders dataset the tweets are of the topic apple, twitter, google and Microsoft. Wordclouds of those respective words along with the polarity wordclouds of these words are shown in figure 2, 3. [16]

Figure 4 shows the bar plot representation of number of tweets verses the polarity classes of tweets regarding the subject 'microsoft'. And figure 5 shows the bar plots number of tweets verses the emotion in the tweets. These emotions are nothing but the extended categorization of the polarity. These polarity and emotion classifications is carried out using naivebayes classifier.This bar plot is drawn using sentiment package from R library. [17].

Table 1. Classified Tweets with their Opinion Score

	tweet	score	Class
1	RT @imightbewrong: 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/mcz9RSDf #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 InitialismsÂ http://t.co/nEqHcJsYÂ Â /via @ruhanirabin	0	Neutral
15	#google #Android #ice cream sandwich	0	Neutral

Table 2. Accuracy of Sentiment Classification Results on Sanders Data

Machine Learning Classifier	Sanders Data
	Accuracy (%)
Neural Network	88.62
QDA	86.95
SVM	**88.65**
Naive Bayes	86.98
LDA	88.39
Random Forest	**88.65**

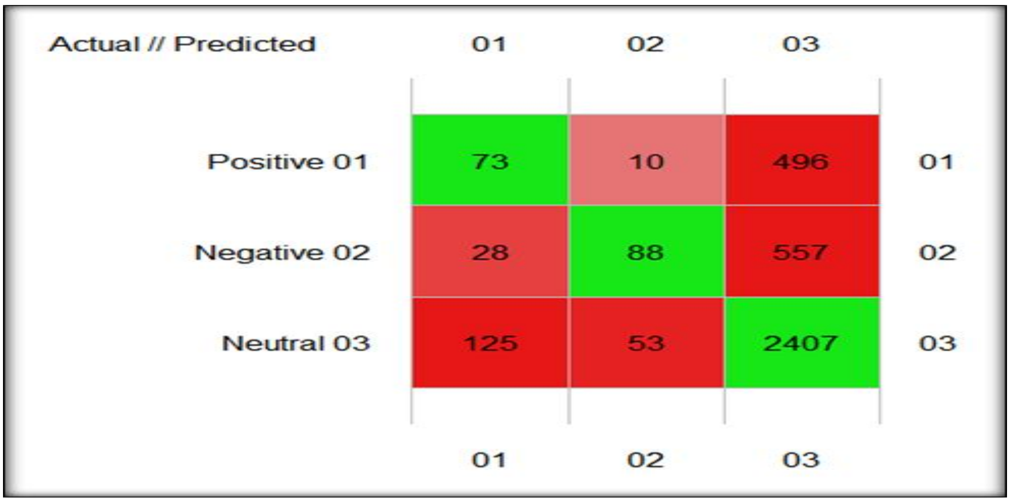

Fig 1: Confusion Matrix obtained when Naive Bayes classifier is applied on Sanders Data

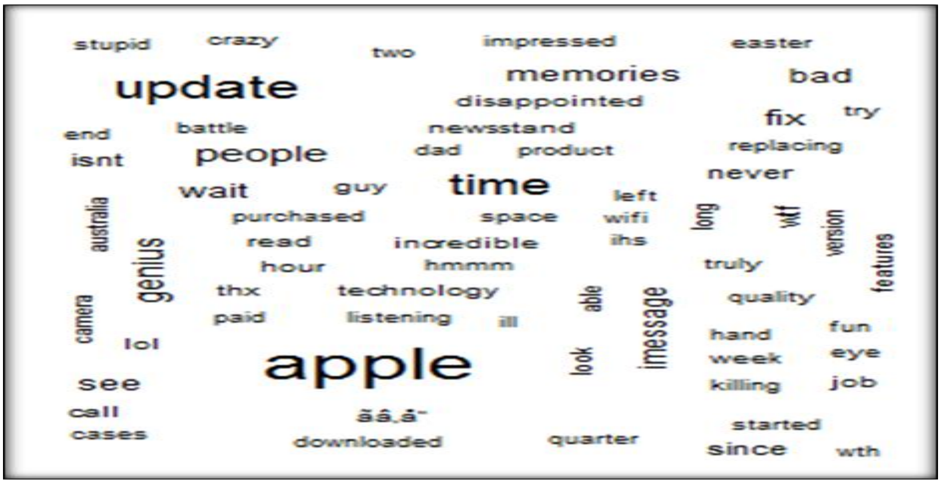

Fig 3: Wordcloud of Polarity labeled tweets on subject Twitter in Sanders Data

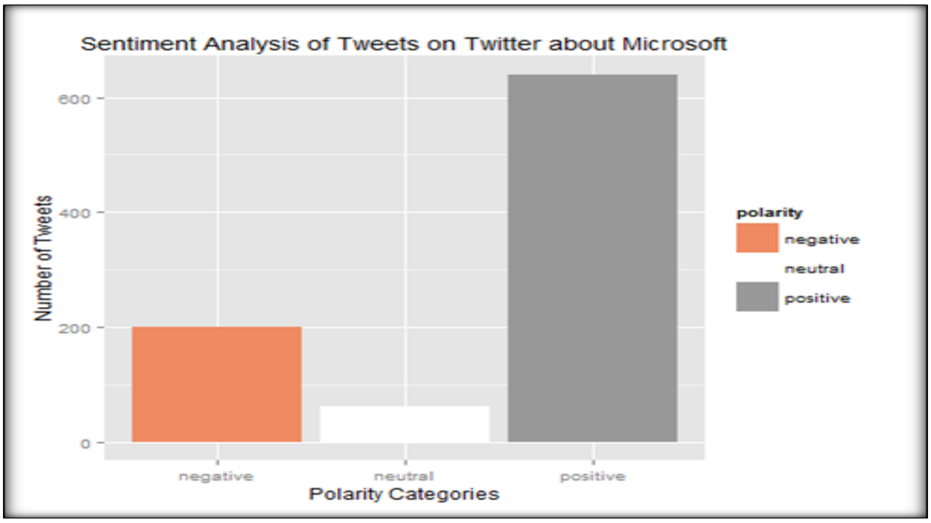

Fig 4: Bar plot of Polarity Categories vs. Number of tweets on subject Microsoft in Sanders Data

IV. Discussion

The accuracy values showed in table 2 shows that from all supervised learning classifiers neural network gives the best classification. The confusion matrix shown in Fig 1 gives the better understanding of how correctly the tweets are classified into pre-fined classes. Using various R packages we have drawn the wordclouds of our datasets. These wordclouds are useful to do empirical study. Fig 2 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 analyze 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 analyzing the wordclouds.

Extending the concept of simply drawing a wordcloud, we have also plotted the wordcloud differentiating on particular feature in Fig 3. 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.

V. Conclusion

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 dictionary based approach can be used to extract word level sentiments of the tweets, further it can give the accuracy of about 88 %. Among the machine learning classifiers used SVM and Random Forest classifiers give the highest accuracy on results. We can state that R environment is a very good framework and statistical and programming language for data mining, analysis and visualizing the results.

Acknowledgements

This review 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

Examples follow:

[1]. Kishori K. Pawar, Pukhraj Shrishrimal, R. R. Deshmukh, Twitter Sentiment Analysis: A Review, International Journal of Scientific & Engineering Research, 6(2), 2015, 957-964.
[2]. Bo Pang, Lillian Lee, Opinion mining and sentiment analysis, Foundations and Trends in Information Retrieval ,2(1-2), 2008, 1–135.
[3]. Bing Liu, Sentiment analysis and opinion mining, Synthesis Lectures on Human Language Technologies, no. 1, 2012, (1-167).
[4]. http://www.sananalytics.com/lab, sited on March 10, 2015.
[5]. http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010, sited on April 1, 2015.
[6]. http://mpqa.cs.pitt.edu/lexicons/arg_lexicon/, sited on April 1, 2015
[7]. http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#datasets, sited on April 3, 2015
[8]. http://www.computeruser.com/emoticons?name_directory_startswith=#, sited on April 28, 2015.
[9]. Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Text classification and Naive Bayes, Cambridge University Press, April 1, 2009.
[10]. 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, 3(3), March 2014.
[11]. Teknomo, Kardi (2015) Discriminant Analysis Tutorial. http://people.revoledu.com/kardi/tutorial/LDA, sited on May 28, 2015
[12]. http://www.saedsayad.com/lda.htm, sited on May 28, 2015.
[13]. https://en.wikipedia.org/wiki/Support_vector_machine, sited on May 28, 2015.
[14]. https://en.wikipedia.org/wiki/Random_forest, sited on May 30, 2015.
[15]. Ph. Grosjean & K. Denis , "Package 'mlearning'- Machine learning algorithms with unified interface and confusion matrices", Version 1.0-0, CRAN Repository,2012.
[16]. Ian Fellows, "Package 'wordcloud'- Word Clouds", Version 2.5, CRAN Repository, 2013.
[17]. Timothy P. Jurka, " Package 'sentiment' - Tools for Sentiment Analysis", Version: 0.2, CRAN Repository, 2012.