International Journal of Research Available at https://edupediapublications.org/journals
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 05 Issue 12
April 2018

IMPLEMENTATION OF PREDICTIVE MODELS AND TEXT CLASSIFICATION ON TWITTER DATA

Arun.K[1], Dr A. Srinagesh[2]

[1]Research scholar in Computer Science and Engineering Department, Acharya Nagarjuna University.
[2]Associate professor in Computer Science and Engineering Department, RVR&JC.

Abstract:

Twitter is the social news and network blogging service, "what happening now in the world and what people are talking about right now". It has Monthly on average 330 million active users(1).It is one of the best platform to share views in the form of tweets and re-tweets. If we process the Twitter data set we can get sentiment analysis of the people who are participating. People are very interesting about the future predictions of the current leading issues and talks in the twitter regarding any issues. In this paper we prepared a prediction model for sentiment analysis using machine learning model, that prediction model generates number of positives, negatives and neutrals from the data set. In this way more data sets are predicted for the sentiments. New model is designed for future predictions on the sentimental numerical data by applying the regression models, and that model is applicable to test data in getting the future prediction. In this paper we design a sentiment prediction model for GST and future prediction. Here the data set is Goods and service tax (GST) of the India, collected from the twitter for one month, clean the GST data set and then predict the sentiment using naive bakes supervised prediction algorithm. Finally we applied regression machine learning algorithms to find and asses the near future sentiments of the public on GST, if we compare the regression methods, we can observe that linear regression model is the best fit model for the GST twitter data to predict near future sentiments of the twitter users.

Keywords

Sentiment analysis, Naive Bayesian classifier, regression algorithms, machine learning algorithms, NLP.

1.INTRODUCTION

Twitter is moving very close to the people from top level celebrities to a very common people. It bridges the communication gap between leading celebrities to public in the world. Monthly twitter active users in India are 26.3 million(2), in the world 330(1) millions up to third quarter of 2017.

GST was introduced by the Indian government with effect from 1 July 2017, and later GST was revised many times. The government of India maintains transparency to the people through all types of media including twitter. GST has two types of plans; these are CGST (central GST) and SGST (state GST). In India twitter users have been drastically increased. The people share their opinions in the twitter about the GST updates. Sharing of different types of their opinions liken queries about GST, problems of GST, benefits of GST, criticism and support on GST. At the same time Indian government is very transparent to the nation and the government addresses solutions and benefits to the public queries and issues.

Sentiment analysis is the best prediction technique to analyze public opinions from the twitter GST data sets. There are three types of models to implement sentiment analysis namely Machine Learning Approach (ML), Lexicon Based Approach (LB) and Hybrid approach. Linguistic features are used in machine learning approach and these are applied for well known ML algorithms The Lexicon based approach is driven by an opinion lexicon, which is nothing but collection of pre-compiled opinion terms. It is mainly divided into two main approaches s– the dictionary based approach and the corpus based approach which can be used in semantic and statistical methods.
The hybrid approach combines the above two ML and LB approaches (3).

Government needs public feedback about the policy. It is very interesting that before and after the execution of government policies, they acquire mass feedback from the public(4).

People think about the GST policy future and at the same time government of India is also thinking the future sentiments of the people and forecasting the sentiment data set to predict the future data sets by using the regression algorithms(5). Regression analysis is used to model the relationship between a response variable and one or more predictor variables(6). The prediction variables depend on the response variables in order to get the future values.

Finally the model is designed for future forecasting on the GST twitter data is the complex application, R language is

used to design the supporting environment, with its more availability of packages up to 10,000(7). Packages support the text, data, graphical and all types of operations.

2. RELATED WORK

A system called MoodLens, is the first system for sentiment analysis of Chinese tweets in Weibo. In MoodLens, 95 emoticons are mapped into four categories of sentiments, i.e. angry, disgusting, joyful, and sad, which serve as the class labels of tweets(8).

How social media content can be used to predict real-world outcomes. In particular, the chatter from witter.com to forecast box-office revenues for movies(9).

Hotspot detection and forecast using sentiment analysis and text mining approaches. First, create an algorithm to automatically analyze the emotional polarity of a text and to obtain a value for each piece of text. Second, this algorithm is combined with K-means clustering and support vector machine (SVM) to develop unsupervised text mining approach(10).

Twitter data to forecast the outcome of the 2015 UK General Election, While a number of empirical studies to date have demonstrated striking levels of accuracy in estimating election results using this new data source, there have been no genuine i.e. pre-election forecasts issued to date(11).

A forecast is merely a prediction about the future values of data. However, most extrapolative model forecasts assume that the past is a proxy for the future, Regression analysis is a statistical technique to analyze quantitative data to estimate model parameters and make forecasts(5).

Public opinion derived from polls with sentiment measured from analysis of text from the popular micro blogging site Twitter, and also predict future movements in the polls.(12) . This is modelled by ARIMA and auto regressions methods to getting the accurate future value predictions.

The utility of linguistic features is, detecting the sentiment of Twitter messages. We evaluate the usefulness of existing lexical resources as well as features that capture information about the informal and creative language used in micro blogging. We take a supervised approach to the problem, but leverage existing hash tags in the Twitter data for building training data(13).

3. SYSTEM MODEL

This model contains two stages, one is generating the sentiment analysis, and second one is predicting the forecast sentiments. These are explained in the following.

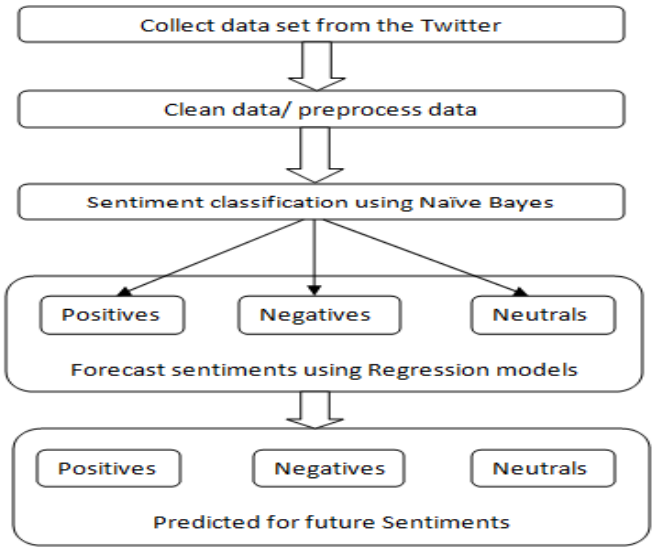

Figure 1: Model for sentiments analysis and future prediction sentiments.

3.1 Collect data set from the Twitter:-

Getting API key from the twitter:-

The user needed to register in the twitter developer for setting up the API. This API is required for authentication to search tweets from a third party application. It uses an industry standard process called OAuth. OAuth creates the handshake between twitter and R using something called as “Consumer Key” and “Consumer Secret”(14).

Once API key is created, the following keys are generated to the user

1. Consumer Key (API Key)
2. Consumer Secret (API Secret)
3. Access Token
4. Access Token Secret.

These are used to access the stream data from the twitter.

Data set from the twitter:-

Data sets are collected from twitter in to the required format for the text analysis.

The user required to generate the request of the twitter service provider by using hash tags of the current topics, titles of the current topics through the API key.

Data set which is generated by the twitter is redirected to one of the target format.

3.2 Data Prepossessing or Data Cleaning process:-

Removing URL links in the corpus: - Twitter comments contains URL links along with the data, links are not needed in the text analysis, so it is better to remove the URL links from the twitter comments.

International Journal of Research Available at https://edupediapublications.org/journals
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 05 Issue 12
April 2018

Removing numbers: -Numbers are not required in text analysis or sentiment analysis, so we have to remove the numbers from the twitter data to refine the tweets.

Replacing the negative mentions :- some words are using for the negative mentions like don't, can't, won't these are not processed in the text analysis, replacing the negative mentions are transforming into useful words without changing the meaning, like "do not", "can not", "would not". This can improve the efficiency of the sentiment analysis(14).

Removing stop words :- stop words refers to a very common word in a language, such as "the","is", and "at". These words are not required in the sentiment analysis. Such types of words are removed from the text data.

3.3 Sentiment classification using Naive Bayesian classification:-

Naive Bayesian classification is one of the best supervise machine learning algorithm for the sentiment analysis and text mining application. The sentiment analysis provides two types of information; one is emotional sentiments and the other is sentiment polarity.

These are calculated by using the naive Bayesian classification method.

Emotional Sentiments :- Emotional sentiment analysis contains seven classifications, such as:

1. Joy, 2.Anger, .Fear, 4.Sadness, 5.Surprise, 6.Disgust, 7.unknown.

All these measures are calculated according to the sentiment scores which are generated by the naive Bayesian model.

Classifies the emotion of a set of texts using a naive bayesian classifier trained on Carlo Strapparava and Alessandro Valitutti's emotions lexicon(15).

Sentiment Polarity :- The polarity of the sentiment analysis was classified in three class objects 1. Positive, 2.Negative, 3. Neutral.

The Naive Bayesian classification is a supervised machine learning, which is used for polarity based sentiments(16). Classifies the polarity of a set of texts using a naive bayes classifier trained on Janyce Wiebe's subjectivity lexicon.

Naive Bayesian is based on the Bayes theorem with independent assumptions between predictors.

Naive Bayesian classifier: $P(c|x)=p(x|c)p(c)/p(x)$

Where $p(c|x)$ is the posterior probability of the class given predictor,

$P(c)$ is the prior probability of the class,

$P(x|c)$ is the likelihood which is the probability of predictor given class.

$P(x)$ is the probability of the predictor.

Here $p(c|x)$ is used to find the class label of a given word, which is already trained on the janyce wiebe's subjectivity lexicon, depends on that each and every word which is classified, then that trained model is applied on the new word to classify.

3.4 Forecast Sentiments using Regression Models:-

Once sentiments are obtained from the twitter data set then predict the forecasting future sentiments like positive, negative and neutral. Here apply some regression methods on the numerical values of the sentiments.

Forecasting sentiments contains two step processes.

Step 1: modelling: Train the regression algorithm for the sentiment data set.

Step 2: Testing: Apply the trained algorithm on the test sentiment data set. Then the result of the Testing is the final prediction of the future sentiments.

This two step process is very common for the regression applications. With the numerical data set train the regression model, and then apply the model for test data.

Here some prediction algorithms are compared for the effective future predictions.

1. Linear Regression,
2. Super Vector machine Regression,
3. Poisson Regression,
4. Random Forest Regression.

In this analysis all Algorithms are trained with the sentiment data set, then again testing is applied for Test data set. All models are not common predicted values, only one of the methods is the suitable for the prediction.

For finding best fit model, we can use proximity measures.

1. Manhattan distance, 2. Euclidean distance, 3. Correlation.

Both Hamming distance and Euclidean distance are derived from the common Minkowski distance metric (17).

If q=1 then Minkowski becomes Manhattan equation, otherwise q=2 Euclidean is the result.

Euclidean $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$

Manhattan $\sum_{i=1}^{k}|x_i - y_i|$

Minkowski $\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$

These can provides the similarity and dissimilarity of the model and test result in the predictions. And one more measure for the proximity is correlation.

4. RESULTS AND EXPERIMENTATION

The predicting future sentiment analysis is divided into two parts.

1. Sentiment analysis from the twitter data sets using the classification methods.
2. Predicting Future sentiments using regression models.

Data sets:-

International Journal of Research Available at https://edupediapublications.org/journals
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 05 Issue 12
April 2018

Here we collect one month of the GST data set from the twitter platform, from 01-09-2017 to 30-09-2017, this data set is used to model the system. From 01-10-2017 to 10-10-2017 we will once again collect the data sets from the twitter which is used for test data sets.

Sentiment analysis from the twitter data set:-

Here we applied two types of sentiment analysis methods 1. Emotional Sentiment analysis and 2. Sentiment polarity. Sentiments are calculated for one month of the data, and the sentiment results are collected in one more data set files.

Here one sample output presented for the sentiment analysis.

S.no	Date of tweets	Total	Positive	Negatives	Neutrals
1	07-09-2017	8286	4775	2628	883

Table 1: Sentiment poalrity for one data set.

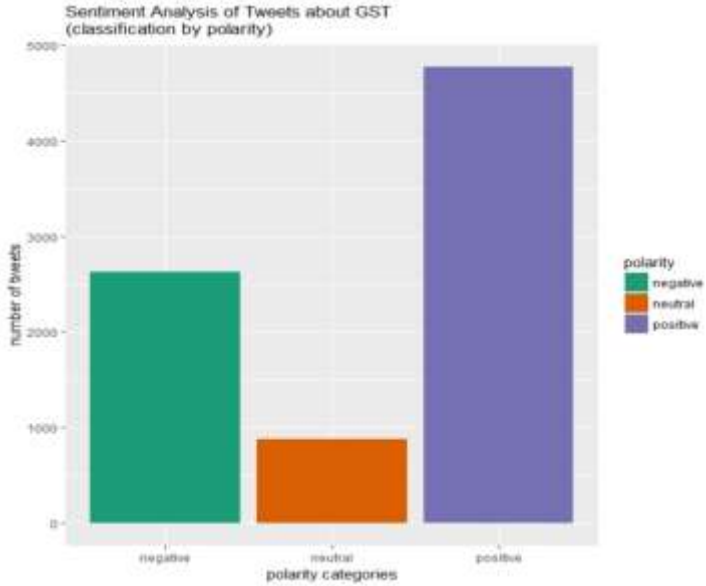

Figure 2: Sentiment polarity

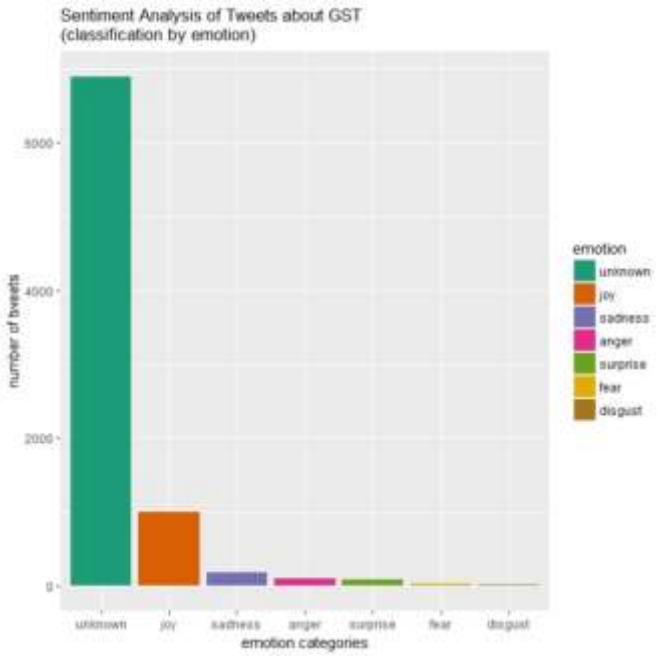

figure 3: Emotional sentiments

In this way all the data sets from 01-09-2017 to 30-09-2017, sentiments are calculated. In the flowing figure we can observe the sentiments per day in September-2017 month. Total count is specifying the total number of tweets generated from the twitter. Positive, negative, neutral are the responses to the GST.

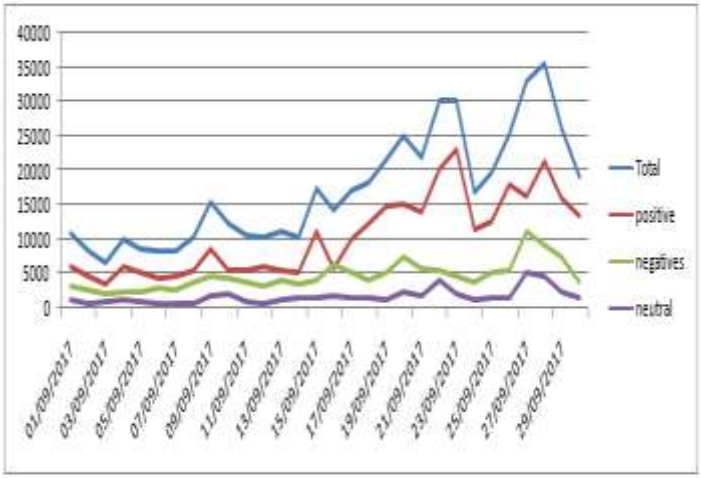

Figure 4: September-2017 Sentiments on GST

Number of tweets increases gradually, mostly positive responses are recorded.

When government official address any information to the public about GST modified policies either it is favour or not, if it is favour then positives increase otherwise negative may increases. Especially on 16-09-2017 negative comments are more than the positive comments.

International Journal of Research Available at https://edupediapublications.org/journals
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 05 Issue 12
April 2018

In the following graph we can observe the percentages of the tweets.

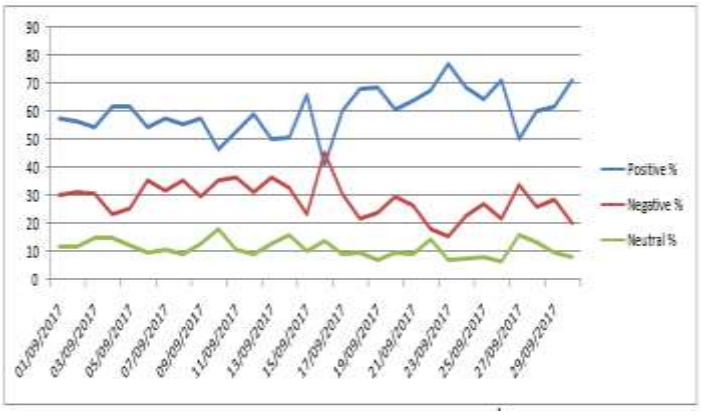

Figure 5: September-2017 Sentimental percentage on GST.

Indian people are giving on average 60% to 70% positive tweets on the GST, 20% to 30% as the negative comments, and 10% to 15% comments for neutrals.

Predict future sentiments:-

To design the model using regression algorithms we considered one month sentimental data set as a sample input.

Once the model is acceptable model, then model is applied to the test data to predict future sentiments. Here we are applied input data to the four regression models and generated models are verified according to the parameters of the regression model. In this experiment model is designed for only positives.

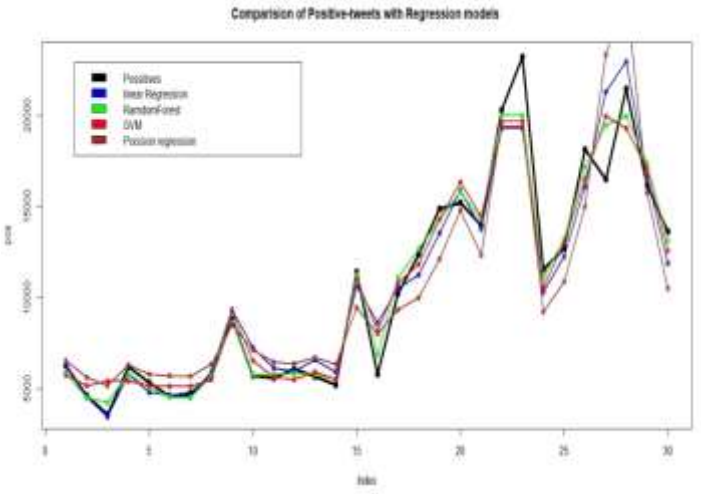

Figure 6: September Sentiments modelling and comparison

Positives are the actual sentiment values. There are four types of models that are designed

These are 1. Linear regression, 2. Random Forest, 3. Super vector machine Regression, 4. Poisson Regression.

These results are compared with proximity measure Manhattan distance for the positive, negatives, neutrals.

Ranking as per the less error or less distance.

S.NO	Name of the Regression	Manhattan Distance (positives)	Manhattan Distance (Negatives)	Manhattan Distance (Neutral)	Rank
1	Linear Regression	0	0	0	1
2	Random Forest	158.045	675.9574	660	3
3	Super Vector machine	234.924	5994.464	199	4
4	Poisson Regression	2.999	0	0	2

Table 2: Comparison regression models

In this model design as per the graph and statistics, Linear Regression conations absolute zero Manhattan Distance in all the cases.

Second step in the prediction is applied on the test data set of the twitter GST data. Test data contains only ten days data from 01-10-2017 to 10-10-2017. The Twitter GST data contains only counts of the tweets per day, but there is no classified or no sentimental data. Now we can apply the regression models which are already trained. The same regression methods are applied on the test data to get the unknown positive, negative, neutrals. Here we applied for only positive predictions.

Date	Total	Actual positives	Random Forest	Linear Regression	Poisson Regression	Super vector machine
01-10-2017	19582	13819	13254	12227	10823	12986
02-10-2017	17971	11914	12036	11139	9879	11687
03-10-2017	19030	12181	13105	11854	10496	12566
04-10-2017	26683	16126	17233	17022	16255	17416
05-10-2017	30823	19439	20064	19818	20595	19672
06-10-2017	43258	25645	20005	28216	41922	35855
07-10-2017	47453	28675	20005	31049	53281	41235
08-10-2017	41698	28776	20005	27162	38346	15477
09-10-2017	29872	19185	20014	19176	19505	19306
10-10-2017	23567	16632	15559	14918	13603	15469

Table 3: Predicted positive tweets form 1-10-2017 to

10-10-2017.

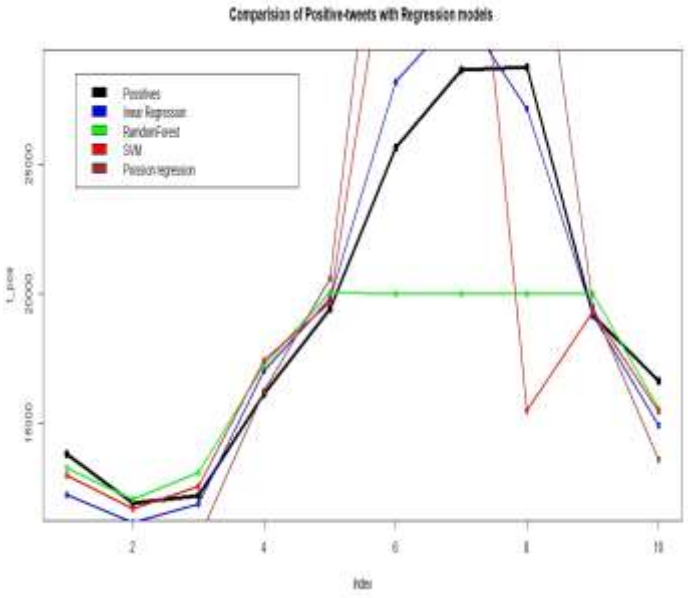

Figure 7: October Sentiments predicted by regression models

From above graph we can compare the new predicted positive sentiment values by actual positive sentiment values. All regression models predicted different values.

S.NO	Name of the Regression (Positives)	Euclidean Distance	Manhattan Distance	Correlation	Rank
1	Linear Regression	4687.992	189	0.981	1
2	Random Forest	13731.4	21112	0.875	3
3	Super Vector machine	20143.77	9277	0.735	2
4	Poisson Regression	31441.49	42313	0.954	4

Table 4: comparison of predictions models

Here linear regression result is very close to the actual positive sentiments in the test prediction. Linear regression test model has very less in Manhattan Distance, Euclidean distance and correlation when compare with remaining regression models.

Hence Linear regression prediction values are the best predicted values in this test prediction, but in the model, random forest got the best fit model. When we compare the model and test cases linear regression is the best model to predict future sentiments of the GST sentiment.

5. CONCLUSION

Predicting future sentiments in the GST from twitter data set, initially sentiments are classified as positive, negative, neutrals by using the one of the Naive Bayesian supervised machine learning method. Data set have one month and ten days data, some days got very less tweets and some day's very high number of tweets was generated and for all the days sentiments calculated. Predicting the future sentiments was processed by applying the regression models. In the result comparisons of regression model, linear regression model gives the best predicted values. Hence with this model we can predict the future sentiments.

6. REFERENCES

1. Twitter: number of active users 2010-2017: *htstps://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/*
2. India: number of Twitter users 2017: *https://www.statista.com/statistics/381832/twitter-users-india/*
3. Liu B. Sentiment Analysis and Opinion Mining. Synth Lect *Hum Lang Technol. 2012 May 23;5(1):1–167.*
4. Imran M, Castillo C, Diaz F, Vieweg S. *Processing Social Media Messages in Mass Emergency. ACM Comput Surv 2015 Jun 26 47(4):1–38*
5. Guerard JB. Regression Analysis and Forecasting Models. *Introduction to Financial Forecasting in Investment Analysis. New York, NY: Springer New York; 2013*
6. Regression Analysis | Examples of Regression Models | *Statgraphics : http://www.statgraphics.com/regression-analysis*
7. CRAN now has 10,000 R packages: *http://blog.revolutionanalytics.com/2017/01/cran-10000.html*
8. Zhao J, Dong L, Wu J, Xu K. MoodLens. In: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining- KDD '12 . New York, New York, USA: ACM Press; 2012. p. 1528.*
9. Asur S, Huberman BA. Predicting the Future with Social Media. In: *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. IEEE; 2010, p. 492–9.*

e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 05 Issue 12
April 2018

10. Li N, Systems DW-D support, 2010 U. *Using text mining and sentiment analysis for online forums hotspot detection and forecast. Elsevier .*
11. Burnap P, Gibson R, Sloan L, Southern R, Studies MW-E, 2016 U. 140 characters to victory?: *Using Twitter to predict the UK 2015 General Election. Elsevier.*
12. O 'connor B, Balasubramanyan R, Routledge BR, Smith NA. From Tweets to Polls: *Linking Text Sentiment to Public Opinion Time Series.*
13. Kouloumpis E, Wilson T, Icwsm JM-, 2011 U. *Twitter sentiment analysis:The good the bad and the omg! aaai.org*
14. Setting up Twitter API to work with R | SAP Blogs: *https://blogs.sap.com/2014/03/16/setting-up-twitter-api-to-work-with-r/*
15. Valitutti A, Strapparava C, Stock O. Lexical Resources and Semantic Similarity for *Affective Evaluative Expressions Generation. In 2005. p. 474–81.*
16. Medhat W, Hassan A, Korashy H. Sentiment analysis algorithms and applications: *A survey. Ain Shams Eng J. 2014 Dec 1;5(4):1093–113.*
17. Dissimilarities between Data objects, *introduction to data mining, pang-ning tan, page(69,70).*