Advances in Computational Sciences and Technology
ISSN 0973-6107 Volume 10, Number 7 (2017) pp. 2173-2182

http://www.ripublication.com

# A Survey for Text Analytics Using R Language

**Neeti Sangwan**
*Assistant Professor ,Maharaja Surajmal Institute of Technology, New Delhi, India.*

**Vinita Malik**
*Information Scientist, Central University of Haryana, Mahendergarh, Haryana, India.*

**Sunesh**
*Assistant Professor, Maharaja Surajmal Institute of Technology, New Delhi, India.*

**Sukhdip Sangwan**
*Assistant Professor, D.C.R.U.S.T, Murthal, Haryana, India.*

**Abstract**

This research work aims to study Text Analytics by mainly focusing on "Sentiment Analysis" and text mining approaches by following an algorithm for the development of results based on degree of truth. The algorithm that has been followed is implemented using R Language. In the present modern day internet savvy world, there has been rapid advancement in micro-blogging sites and services, along with social media websites. These arenas have sprung up to play an important role in relaying an individuals' opinion on a variety of topics. They have also become platforms where ones' opinions and views on certain issues (which may also include current hot topics and trends pertaining to a variety of domains), or a product or service, or about an individual and/or an organization may come forth. It has therefore become imperative to gathering of data to analyze and study the general public opinion with respect to the subject on which the data is being gathered. This collection of huge cache of data then begets various data refinement techniques – doing away with fillers and in the end leaving us with invaluable keywords, which are of utmost importance. Opinion analysis is then employed, which recognizes and distinguishes the

positive and negative reviews from the relatively neutral ones for the concerned entity. Sentiment analysis is employed to factor in the level of positivity or negativity associated with an opinion. The reviews and opinions of some individuals may not always be true, and may be heavily tilted towards or away from the concerned entity. These reviews may chart a completely different course in statistical prediction models and are therefore a threat to the validity of the generated predictions. It is therefore imperative to identify the fake and spam reviews and root them out. After fake review detection and removal is achieved, statistical models, namely, CART and Random Forest are applied to analyze the data and chart prediction trends pertaining to the entity. In this paper, we aim to show how we have tackled all the steps mentioned above in text analytics. We are using twitter as a platform and the various tweets pertaining to a topic as the necessary raw data required for text analytics.

**Keywords:** Text Analytics, Sentiment Analysis, Opinion Analysis, CART Modelling, Predictive Modelling, Twitter Analysis.

## INTRODUCTION

Social Media is increasingly becoming a part of everyday life. People nowadays don't shy away from sharing what's on their mind with their compatriots and followers via the social platform. The large number of tweets on twitter being posted each second stand as a testimony for the same. Twitter has thus emerged as a valuable source of information in understanding what the general population thinks on a certain subject. In order to analyze the prevailing opinion among the populace, tweets are being fetched from twitter and subjected to text analytics.

## LITERATURE SURVEY

**Jyoti Nandimath** et al[1] has studied in recent past, consumers or viewers before buying a product goes through a review, though the increase of e-commerce has increased it has also increased the number of counterfeiting comments. Thus, the need of today is to eliminate inappropriate and fraudulent reviews from a site or an application. The opinion mining process could be done on the texts as well as on the sentences. Various connecting opinion analyzing algorithms and opinion integration algorithms are involved in Opinion mining and sentiment analysis. This paper focuses on opinion integration algorithm, and identifies the fake comments, detects spam based on evidence classifier. In the above research paper three stages are implemented namely, Data Acquisition, Data Processing and Decision Unit. The first step of the architecture works on collecting the data or Big Data. On the collected data, other functionalities are done to avoid the fake content or redundant material. In the second

layer of the architecture, filtration and load balancing is implemented. Here, filtration is used to simplify the content whereas load balancer divides the contents into segments and assigns them to various processing servers. The data in the end is free of useless reviews and has associated keyword assigned to it. Lastly, Decision Unit, here the data is collected and the results from each modal are processed against all and then combine, organizes, and stores these results in NoSQL database. Hence, in this paper spam and fake review detection system is developed.

**Veena Dubey** et al[2] found the target is of eliminating irrelevant content through different techniques. A system is evolved by using customer reviews; this is known as Opinion Mining.Given an arrangement of reviews our task involves features identification of any item on which clients have communicated their assessment. Techniques like data mining and NLP in order to mine the features have been used. Every component is partitioned into positive and negative reviews. To decide the sentiment orientation, basically there are the three subtasks which are mentioned below:

1. Identify an arrangement of adjectives regularly used to express opinions using NLP method, which are called opinion words.
2. For each opinion word semantic orientation is determined.
3. Decide the opinion or sentiment orientation for each sentence by generating a summary out of the discovered information.

The research paper talks about opinions of a user on iPhone, the views represent positive, negative, objective, feelings and holder of the review. After disintegrating the word corpus document level sentiment analysis is implemented. The classification is defined as a supervised learning technique with three labels: positive, negative and neutral. Further, a range from 1 to 5 stars is also used to understand the review. There are spammers (individual, group) who give negative or fake reviews regarding a product resulting into negative publicity. To manually label the reviews as spam is humanly not feasible hence it has been tried to judge through AI techniques but it is also hard. If the spammer uses algorithm to build a review say, innocent reviews it becomes difficult for machine methodologies too. Framing an effective set of features is the significant activity of sentiment classification. **Bharat R.Naiknaware** et al[3] knew that social media monitoring has been increased in the past. In addition to that, e-commerce has also played its role. With huge indulgence of people in social media the research paper mines the reviews, comments, status of Twitter. Twitter API with R tools and packages has been used in the above mentioned research paper. There are 3 types of Levels of Sentiment Analysis namely; Document level analysis, which analyse the document reviews wholly. Sentence level analysis, a single sentence is focussed upon and Aspect level analysis also called as Feature analysis this technique is better than the other two as it does not look for a sentence or document moreover an entity is considered, thus making this type of level fine-grained. The source of data used is of

micro blogging sites, blogs, revie (Patil M.S., 2003)ws, and data sets. In the concerned research paper, the data is incorporated from Micro Blogging site and the Level of

Sentiment Analysis used is Sentence level analysis. This is done to check the credibility of government schemes. The steps followed are:

1. Collection of people tweets
2. Pre-processing
3. Feature selection
4. Sentiment word identification
5. Sentiment polarity identification
6. Sentiment classification and
7. Analysis of reviews

**Md. Daiyan** et al [4] explained that sentiment analysis or opinion mining involves a scheme to collect and classify opinions about a product in order to track the humor of the people for a certain product. On the various e-commerce sites people buy products using internet and gives their feedback or opinion about the product. Authors in this paper provided the overview on Sentiment analysis or Opinion Mining. Authors discussed the various problems and opportunities in the field of sentiment analysis or opinion mining along with their raising factors. Factors are as follows:

1. The development of machine learning methods in natural language processing and information retrieval.
2. The available datasets for training in machine learning algorithms.
3. Realization of the fascinating intellectual challenges and commercial and intelligence applications that the area offers. Various steps followed are:
    a) Getting Data
    b) Tweets Preprocess
    c) Feature Extraction
    d) Formation of Feature Vector
    e) Feature Extraction from Vector

Classical sentiment classification fails to find out the likes and dislikes of the reviewers or opinion holder. It classifies the review document only as positive or negative. But a positive document of any product does not signifies that the reviewer hold positive opinion on all the aspects of the product. Similarly, negative opinion document does not ensure that reviewer has negative opinion on all aspects. In an evaluative document,

the reviewer writes both positive and negative aspects of the object, although the general sentiment on the object may be positive or negative. In order to obtain detailed review, feature-based opinion mining is required which contains two main tasks:

1. Object features identification.
2. Determining Opinion Orientation.

**M.S.Patil** et al[5] explained review spamming or opinion spamming i.e. illegal activities like fake reviews writing. It may mislead readers by producing false positive or false negative opinion to entities to promote them or to ruin their reputations. Opinion spam is of many kinds like fake comments, fake reviews, fake blogs. Authors explained mainly three types of the spam:

- False Opinion: reviews that contain false opinions on products.

  i.e. positive spam review that provide fake positive opinion of a product for promotion or negative spam review which includes reviews that provides false negative opinion on a product to damage product's reputation.
- Review on Brands: These provide reviews on the manufacturer seller or brand.
- Non-Reviews: this type of reviews have no opinions. These are not affected by reader of reviews. It affects the automated mining systems only. In this paper authors discussed the review on brand spam detection. These reviews are on the brand, manufacturer or seller of product and not on product. It is required to find features in reviews by using feature selection algorithms to identify the spam. Feature selection algorithms are of two types: subset selection and feature ranking. Subset selection algorithms are used to find the set of all the possible combinations of features of provided data. Authors explained the use of decision tree for decision making on reviews on brands.

**Sushant Kokate** et al[6] discussed the difficulty in finding review spam or to recognize fake reviews. Authors also found many unexpected rules to identify unusual review patterns, to analyze a review data set and to indicate spam activities. The technique is domain independent. Using the technique, to analyze an Amazon.com review dataset and found many unexpected rules and rule groups which can indicate spam activities. The proposed system initially user enters the name of the movie for obtaining the reviews given by the different reviewers or customers. After entering the name of the movie, API fetches the website of movie review and fetch all the reviews of the movies providing by the websites. After that clustering algorithm is implemented for clustering the reviews in the groups. After completing the process of clustering, the ARFF file is generated, this ARFF file contains the features required for detecting the original reviews and instances of the above attributes. This ARFF contains number of attributes like is question mark present in the review, Capital word in review, polarity, links,

comparison, etc. This ARFF file given as a input to the classifier, used J48 classifier for the detecting the reviews. Training and testing process are done by the J48 classifier. After completing the process of classification, fake and truthful reviews are detected. These reviews now qualify for the further checking for Brand (Kokate Sushant, 2015)Spam detection. From this type of review removing stop words is necessary, after that this review and putting for the stemming. This reduces the document to a certain level. Now with remaining keywords, checking the support count and comparing it with pre decided Threshold Value. Words with support count more than the threshold value will be considered as Brand Spam. Result may retain certain words which cannot be labelled as Brand and it wholly depends on the user or person to judge that through Active Learning. In the result section are discussed the results obtained by the system for detecting fake and truthful reviews given by the users. diagram or charts are shows the number of reviews of user.

**Qingxi Peng** et al[7] discussed that the sentiment analysis techniques to detect review spam. Authors proposed the method to calculate score for sentiments from natural language text using parser and analyse the relationship between spam reviews and sentiment score. Various rules are generated using the relationship between spam reviews and sentiment score. Lastly, in order to find out spam reviews and spam store more efficiently, authors combined the time series with these rules. Detection method provided in the paper is better than existing methods as shown in the experimental results. Mainly three activities are discussed to achieve the better results. Firstly, shallow dependency parser is used to generate sentiment lexicon for sentiment score computation. Then various discriminate rules is generated. Finally, time series method to find spam reviews is established.

## EXPERIMENTAL WORK & RESULTS

Text Analytics in itself is a broad paradigm. As such, in the project ensued, numerous steps had to be followed.

First of all, a web application had to be created and a user friendly interface developed as shown in Figure 1. A user is able to interact with an application only via the means of the user interface (UI) put in place. An application is incomplete without an UI, and generation of the same is therefore the first and foremost requirement.

For the project, Twitter is the platform necessary to supply the raw data. Tweets will serve the purpose of raw data for text analytics.

The web application developed requires just two parameters in order to function – keyword and the number of tweets a user wants to go through.

Next comes the collection of data. Once the keyword and number of tweets have been entered, the next step is establishing a link between our developed web app and twitter.

The keyword henceforth will be treated as our concerned entity.

The application goes through twitters' tweets, searching for the entered keyword and picking up a tweet if there is a mention of the keyword within the tweet. It picks as many tweets as the number specified initially at the beginning.

The identified tweets are then saved in a file. The computed file is sent for processing.

After the data collection process is completed, the application concerns itself with the cleaning and refinement of the collected data. A large corpus of words is pre-fed in the application. Words present in this corpus are first and foremost removed from the collected tweets. Next punctuation marks are removed. After that, conversion of the remaining text to lower case is done. The tweets are then made free of any stopwords. Ultimately, the tweets are stemmed, so any part which doesn't add value like in differentiation, we only differ for sentiment, so we stem it to differ.


**Figure 1:** Proposed Model's Home Screen

Once the cleansed data is obtained, it is subjected to opinion analysis. Opinion lexicons are loaded, which are a list of positive and negative opinion and sentiment words of English language.

Next, we apply sentiment analysis. A score is assigned to each sentiment detected, the score value ranges from -5 to +5. A negative value indicated a negative outlook radiated by the word, and lower the score, lower is the negativity indicated. Same goes for positive outlook, which are assigned a positive score, higher the value, higher is the positive outlook. A score of 0 indicated neutrality.

Each and every word in a tweet has an associated sentiment score. A tweet may contain words who individually may radiate a positive, negative or a neutral outlook. The next process is the generation of score dataframe, in which the sentiment score of the words

present in a tweet is added up and the sum compiled. This sum may come to be a positive or a negative number or even 0. A positive score for the entire tweet indicated favorable outlook. A negative score indicates a bleak approach, while a score of 0 indicates indecisiveness on the topic concerned.

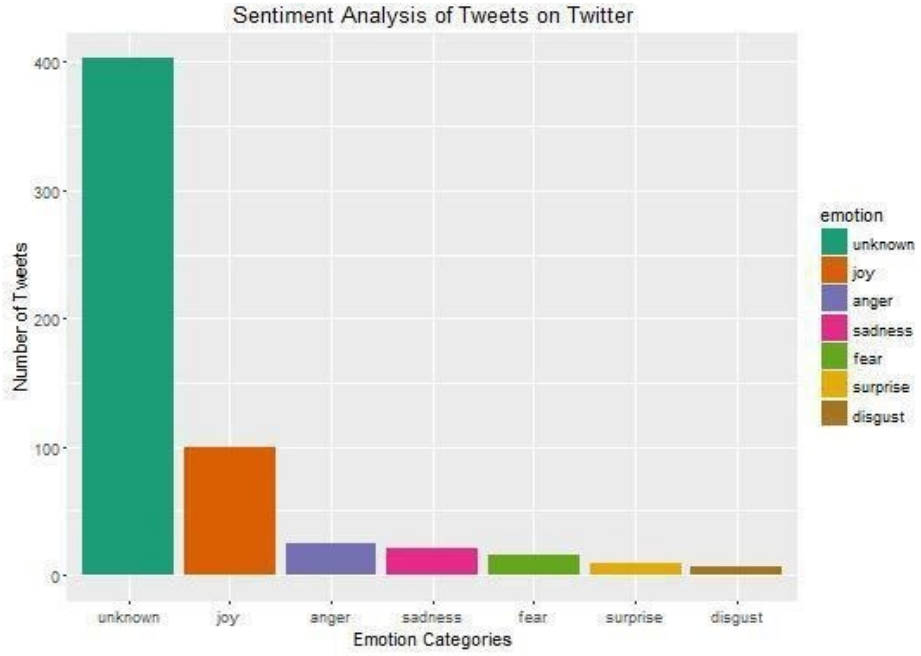


**Figure 2:** Graphical analysis of emotions

An important step after opinion and sentiment analysis is the detection of fake tweets and their removal. First the content is reviewed for -- various lexical features like part-of-speech n- grams, word n-grams and other lexical features, similarity of reviews in style and content from different reviewers, and for any semantic inconsistencies.

Next, abnormal behavior are reviewed. Public data is taken from reviews Web sites like reviewer id, time of posting, frequency of posting, past of the product, time taken to post a review, and many more.

Next comes the analysis part. Frequencies of the words present in all tweets fetched and stored in a document term matrix. Words with frequencies below 20 are removed.

Then a sparse matrix is created and removal of sparse terms below 0.995 frequency. The sparse terms are added and presented as columns. The dataset is split into train and test sets in the ratio of 70:30. A CART Model is created on the train set and the test set is validated.

Same process is applied on the dataset in case of Random Forest. A baseline model is created and both then comparison is done for accuracy. The final results are shown in Figure 2 and Figure 3.

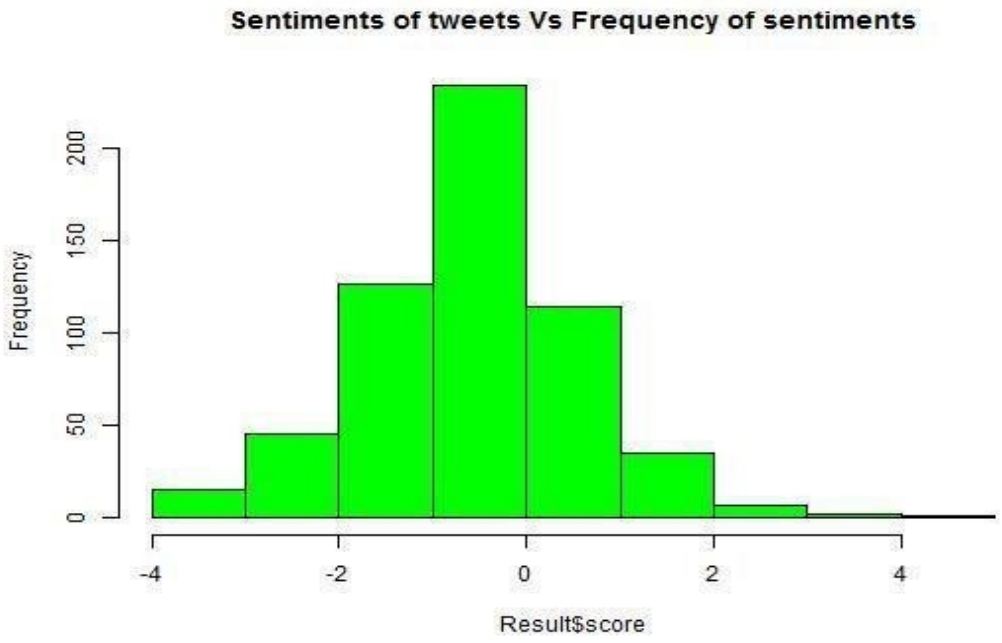


**Figure 3:** Histogram showing end Results

## CONCLUSION

Text Analytics is the field pertaining to gathering of data to analyze and in-depth study analysis of the general public opinion with respect to the subject on which the data is being gathered. This paper has dwelled upon this field, employing various techniques and procedures for the same. For this project, twitter was used as a platform and the various tweets pertaining to a topic as the necessary raw data required for text analytics.

First and foremost, the user interacts with the application with the help of application User Interface, wherein they are required to enter a keyword and the number of tweets the application scrolls through. Next, based on the data entered in the fields, data gathering is done which leads to the process of data cleaning and its refinement, leaving behind only important words and removing all redundant ones. The remaining parts of the collected tweets are subjected to systematic process of Opinion Analysis and Sentiment Analysis. A sentiment score is assigned to each word and finally to the entire tweet which tells of the outlook depicted by that tweet. After these events, successful identification and removal of fake tweets is necessary, else they can derail future prediction models in favour of or against the concerned entity. In the end, the output

result is analysed and prediction models of CART and Random Forest applied to chart future trends. A baseline model is created and both then comparison is done for accuracy.

## REFERENCES

[1] Daiyan Md., T. S. (2015). A Literature Review on Opinion Mining and Sentiment Analysis . *International Journal of Emerging Technology and Advanced Engineering* , 262-280.

[2] Dubey Veena, G. D. (2016). Sentiment Analysis Based on Opinion Classification Techniques: A Survey . *International Journal of Advanced Research in Computer Science and Software Engineering* , 53-58.

[3] Kokate Sushant, T. B. (2015). Fake Review and Brand Spam Detection using J48 Classifier . *International Journal of Computer Science and Information Technologies*, 3523-3526.

[4] Nandimath Jyoti, K. B. (2017). Efficiently Detecting and Analyzing Spam Reviews Using Live Data Feed. *International Research Journal of Engineering and Technology (IRJET)* , 1421-1424.

[5] Patil M.S., B. A. (2003). Review on Brand Spam Detection Using Feature Selection . *International Journal of Advanced Research in Computer Science and Software Engineering*, 744- 748.

[6] Peng Qingxi, Z. M. (2014). Detecting Spam Review through Sentiment Analysis . *JOURNAL OF SOFTWARE*, 2065-2072. [7]R.Naiknaware Bharat, K. S. (2016). Sentiment Analysis of Indian Government Schemes Using Twitter Datasets. *IOSR Journal of Computer Engineering (IOSR-JCE)* , 70-78.