**Abstract**

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.

**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 analyse the prevailing opinion among the populace, tweets are being fetched from twitter and subjected to text analytics.

**Literary Survey**

[1]. This research paper has tried to study review spam and spam detection since all reviews given by clients using a product were not given with genuine aim, which misleads companies taking decisions considering the quality of their product.

[2]. This research paper has aimed to tackle the problem of sentiment polarity categorization, one of the fundamental problems of sentiment analysis. It also gives insight into the future work of into the future work of sentiment analysis.

[3]. Spam reviews may mislead customers by promoting or demoting a target store. This research paper aims to identify spam reviews through sentiment analysis.

[4]. This paper introduces a writing study on sentiment analysis and the methodologies adopted for the same. Focus has been given on various opinion classification techniques, which may be performed on any dataset.

[5]. In this research paper, work is proposed for detecting reviews on brand spam detection using feature selection.

[6]. Overview on Sentiment Analysis along with methods, tools and dataset used by researchers with their accuracy has been given in this research paper.

**Experimental Work**

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. 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.

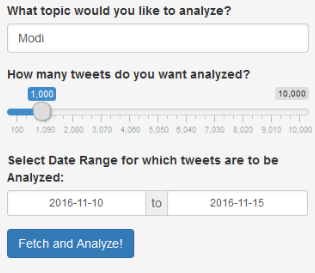


Figure 1

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.

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.

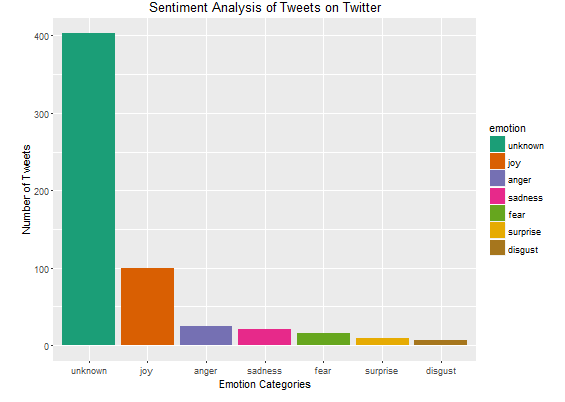
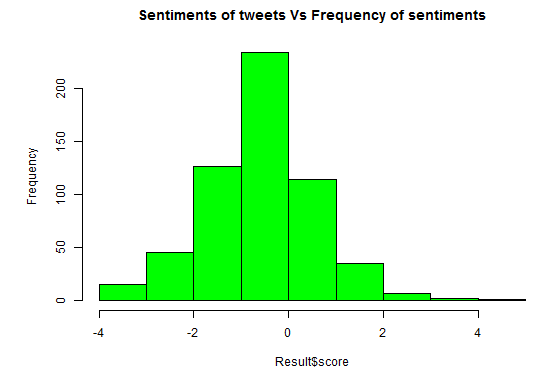


Figure 2

An important step after opinion and sentiment analysis is the detection of fake tweets and their removal. First the content is reviewed for -- lexical features such as word n-grams, part-of-speech n-grams, and other lexical attributes, Content and style similarity of reviews from different reviewers, and for any semantic inconsistencies.  
Next, abnormal behaviour are reviewed. Public data is taken from Web sites, such as, reviewer id, time of posting, frequency of posting, past reviews 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. Figure 3

**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. SushantKokate and Bharat Tidke – Fake Review and Brand Spam Detection using J48 Classifier. In: (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (4) , 2015, 3523-3526. Department of Computer Engineering, Flora Institute of Technology, Pune, India
2. Xing Fang and Justin Zhan – Sentiment Analysis using product review data. In: Fang and Zhan *Journal of Big Data* (2015) 2:5, DOI 10.1186/s40537-015-0015-2. Department of Computer Science, North Carolina A&T State University, Greensboro, NC, USA.
3. QingxiPeng and Ming Zhong – Detecting Spam Review through Sentiment Analysis. In: JOURNAL OF SOFTWARE, VOL. 9, NO. 8, AUGUST 2014. State Key Lab of Software Engineering, Wuhan University, Wuhan, China
4. VeenaDubey and DharmendraLal Gupta – Sentiment Analysis based on Opinion Classification Techniques : A Survey, In: International Journal of Advanced Research in Computer Science and Software Engineering, Vol 6, Issue 6, June 2016. Department of CSED, K.N.I.T, Sultanpur, Uttar Pradesh, India.
5. M.S.Patil and A.M.Bagade – Review on Brand Spam Detection using Feature Selection, In: International Journal of Advanced Research in Computer Science and Software Engineering, Vol 3, Issue 9, Sept 2013. Pune University, Pune, India.
6. Md. Daiyan, Dr S.K. Tiwari, Manish Kumar, M. AftabAlam – A Literature Review on Opinion Mining and Sentiment Analysis, In: Internation Journal of Emerging Technology and Advanced Engineering, Vol 5, Issue 4, Apr 2015. Institute of Computer Science & IT & Department of Mathematics, Magadh University, Bodh Gaya, Bihar, India & Department of Computer Science & IT, Singhania University, Pacheri Bari, Rajasthan, India.