Advances in Computer Science and Information Technology (ACSIT)
p-ISSN: 2393-9907; e-ISSN: 2393-9915; Volume 5, Issue 2; April-June, 2018, pp. 86-91

http://www.krishisanskriti.org/Publication.html

Sentiments Analysis through Recent Tweets for "Agriculture" in India

Ekta Hooda[1] and Akshay Hooda[2]

[1]Research Scholar, Dept. of Mathematics & Statistics CCS HAU-Hisar (Haryana)
[2]L M Thapar School of Management Thapar Institute of Engg. & Technology Patiala (Punjab)
E-mail: [1]ektahooda@gmail.com, [2]akshayhooda@gmail.com

Abstract—*People express their views and opinion on various social and political issues through the online social media such as Twitter, Instagram, Facebook etc. Opinion and sentiments expressed through these sites reflect their reactions on the current topics of national and international importance. In the present study, sentiment analysis of recent tweets for the topic 'Agriculture' has been studied in context of India. For this purpose, 1000 recent tweets for the theme 'Agriculture' were retrieved from a radius of 300kms around the three cities (Delhi, Pune and Hyderabad) covering major part of India. Frequency barplots, and word-clouds of the highly frequent terms have been formed for understanding current pattern for agriculture. Sentiment analysis of the tweets has also been performed for understanding minds of intellectual Indians regarding agriculture. Analysis has been done using R software and its associated libraries.*

1. INTRODUCTION

Nowadays people prefer to express their views and opinion on current social and political issues through internet. These views and opinions are mainly expressed through social media websites and spread quickly among the large proportion of the internet and mobile users. In India, millions of people are using social media networking sites like Twitter, Facebook, Instagram and Google Plus to express their emotions, views and opinions. Due to the ever increasing interaction among the individuals on the internet, the social media websites are generating a large volume of sentiment rich data in the form of tweets, blog posts, comments, likes etc. Twitter is among the top social networking forums that provide a quick and easy means for people to express their views. It is an important source to know about the public opinion and sentimental analysis. For each tweet it is important to determine the sentiment of the tweet, whether is it positive, negative, or neutral. Moreover, tweets are limited to 140 characters in length and hence the emotion or feelings on particular topic remain consistent.

The data generated through social media websites is vast and unstructured. However, it contains useful information which can be used for future planning. Government agencies and multinational companies are making use of these data for their policy decisions and improvements in their offerings. The users of social media websites also get an opportunity to connect the business persons and *vis-vis* the companies get an opportunity to know their customers more closely. Customers are also taking advantage of social media websites to check reviews of a particular product from some specific brand. This is the one of the reasons that companies are making use of sent the sentiment analysis techniques to understand the mindset of the customers.

Sentiment analysis is the process that converts views and opinions expressed in tweets and database sources with the help of Natural Language Processing (NLP), Kharde and Sonawane (2016). It categorizes the different opinions in text into three categories that are positive, negative and neutral. Sentiment Analysis is also known as subjective analysis, appraisal extraction & opinion mining.

Several papers have been published in recent years on sentiment analysis of twitter data and news articles. Most of these articles have been devoted to methodological aspects or analysis of twitter data on camera, movie, mobile etc. Agriculture sector is directly linked with the world's food security where crop forecasting is considered to be a valuable tool for framing policies to address food security and poverty.

Agriculture is the back bone of the developing countries including India. However, not much work has been done on analysis and understanding of public sentiments on the very important issue like agriculture and government policies. In this study we have collected and analyzed data from twitter.com related to the recent tweets for the theme "Agriculture" from a radius of 300 kms from Delhi, Pune and Hyderabad. We are interested in the analyzing public opinions on the agriculture theme because India is an agriculture based economy and present Indian government is exploring every nook and corner for doubling the income of farmers and 2022 and improving the state of agriculture in the country.

2. LITERATURE

Sentiment analysis has been used as a Natural Language Processing task at many levels of granularity. Starting from a document level classification task (Turney, 2002), it has been used at the phrase level (Agarwal *et al*.,2009). Pang and Lee, (2008) described the existing techniques and approaches for an opinion-oriented information retrieval and reviewed work on opinion mining and sentiment analysis. Studies on sentiment analysis of Twitter data have also been undertaken by Go *et al*. (2009), Barbosa and Feng (2010) and Bermingham and Smeaton (2010). Pak and Paroubek (2010) proposed a model to classify the tweets as objective, positive and negative. They also created a twitter corpus by collecting tweets using Twitter API and automatically annotating those tweets using emoticons. Davidov *et al*. (2010) proposed an approach to utilize Twitter user- defined hastags in tweets as a classification of sentiment type using punctuation, single words, n-grams and patterns as different feature types, which are then combined into a single feature vector for sentiment classification. Liang and Dai (2013) used twitter data on three different categories (camera, movie, mobile) and labeled the data as positive, negative and non-opinions.

Whissel(1989) examined the Dictionary of Affect in Language, one of many acceptable routes to the measurement of emotion, in terms of research conducted with the dictionary and in terms of a framework of meta-measurement. WordNet, an electronic lexical database, edited by Fellbaum(1998) is considered to be the most important resource available to researchers in computational linguistics, text analysis, and many related areas.

Haussler(1999) introduced a method of constructing kernels on sets whose elements are discrete structures like strings, trees and graphs. The method can be applied iteratively to build a kernel on an infinite set from kernels involving generators of the set. Klein and Manning (2003) showed that un-lexicalized PCFGs can achieve high parsing accuracies when training trees are annotated with additional information.

3. MATERIAL AND METHODS

Three recent target datasets on the theme "Agriculture" of 1000 tweets each were extracted using twitteR package in R studio in the first week of September, 2018. The tweets were retrieved using the 'geocode' function of R by choosing Delhi, Pune and Hyderabad as the central locations and choosing a radius of 1000 kms for each, assuming to cover majority of India. The retrieved dataset contained data regarding the actual text of the tweet, date and time of creation, tweet ID, screen name of the username, retweet count and the latitude-longitude of the source location of the tweet. R Studio GUI has been used for the analysis and the packages used are namely, twitterR, ROAuth, plyr, Stringr, tm(for text mining), wordcloud and httr. A twitter application was created that provided an interface to connect R studio console with twitter.com using Twitter API. This API facilitates analysis of tweets posted by users and for extraction of tweets and underlying meta-data. The required tweets were extracted using searchTwitter function in R Studio. The tweet datasets were highly unstructured and hence were cleaned for numbers, symbols, strip of white spaces, URLs, punctuations, emoticons etc. This made the unstructured data usable for further analysis. After satisfactory cleaning, the datasets were converted into a Term-Document Matrix (TDM) to further list down the terms against their occurrences in the documents or tweets. The TDM was then used to form the wordcloud, barplot etc. of the frequent terms. Sentiment analysis was done via the nrc_sentiment_library to get the sentiment scores in the data.

4. RESULTS AND DISCUSSION

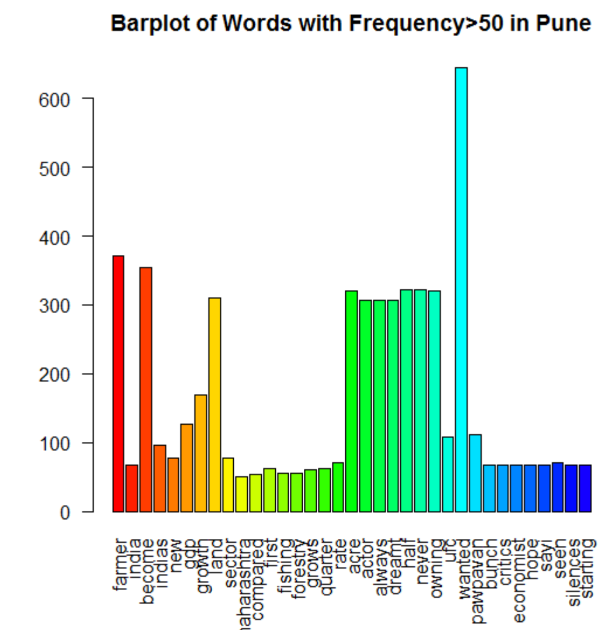

Figure 1

The bar plot diagram for the most frequent terms(for those with frequency greater than 50) for the Agriculture dataset of 1000 tweets of Pune city shows the most frequent terms to be 'wanted', 'farmer' and 'become' closely followed by 'acre', 'actor', 'always', 'dreamt', 'half' etc. with almost similar frequencies which can be attributed to a recently trending tweet of Hyderabad-based Pawan Kalyan where he mentioned that he always dreamt of owning half an acre of agriculture land and wanted to be a farmer instead of an actor. Some of the other trends in the area were regarding the 8.2% GDP and the forestry and fishery sector growing at 5.3% according to the recently disclosed data.

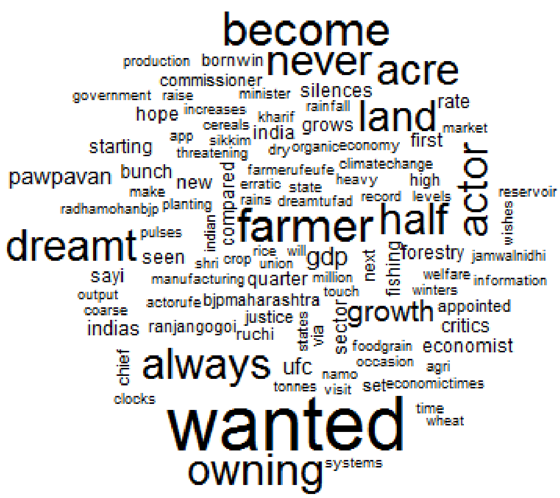

Figure 2

The wordcloud obtained from the most frequent terms also visualises the similar with the word size increasing with the increase in the frequency of the trending word.

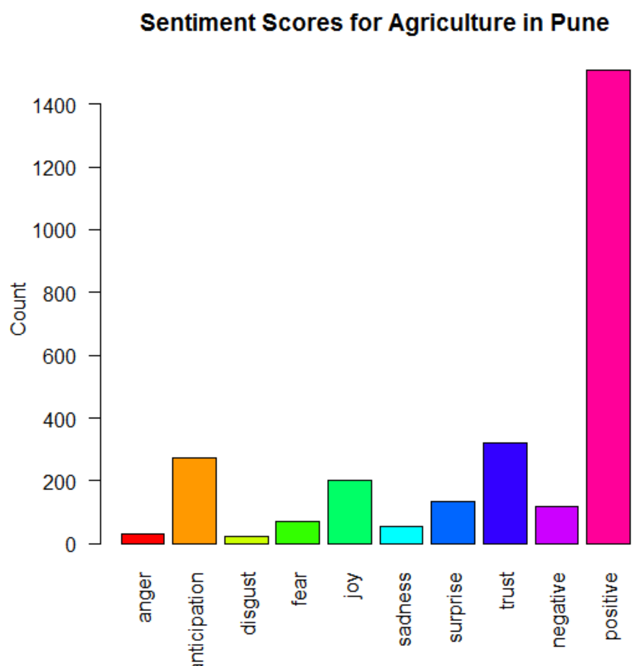

Figure 3

The sentiment analysis scores for the "Agriculture" dataset shows a highly positive overall sentiment amongst the public keeping in view the recent data with a count of above 1400 positive sentiments and a slight above 150 for the negative sentiment. Additionally, when sentiments are further classified in 8 categories, the trust sentiment takes a lead alongwith anticipation and joy, all of which are positive sentiments. Surprise and fear also follows but the positive sentiments prevail over a much larger part of the overall sentiments of the Pune-based dataset.

The second dataset based on Hyderabad comprising of 1000 tweets also shows a very similar trend as in Pune. As expected, this is also based on the tweet by Pawan Kalyan with even higher frequencies as the tweet is Hyderabad-based. This implies that the trend for the particular tweet originated during the period of the data extraction has spread to the Pune region. Another trending tweet is regarding mining companies claiming their support to all round development of surrounding villages in terms of health, education and women. Also, there is another trend caused the NCBN being invited by the United Nations to deliver a keynote address at an event, titled 'Financing Sustainable Agriculture: Global Challenges and Opportunities' to be organised on the sideline of the annual UN General Assembly in New York on September 24. In addition to these, the spurt regarding the GDP figures can be witnessed in Hyderabad as well.

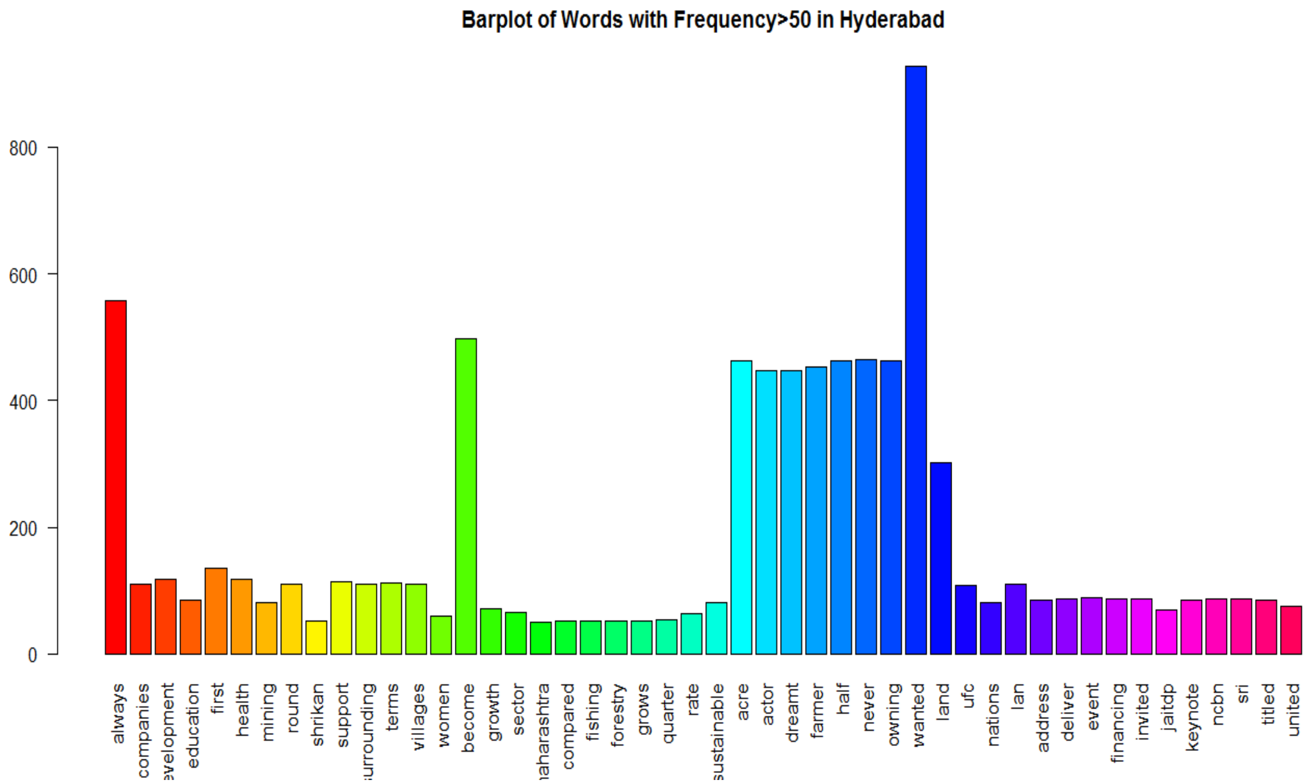

Figure 4

Similarly, the wordcloud also displays the trending words ‘wanted’, ‘dreamt’, ‘always’, ‘farmer’, ‘half’, ‘actor’, ‘never’, ‘owning’ etc sorted with increasing size with increase in the respective frequencies.

Figure 5

Like Pune, the sentiment scores for the theme of ‘Agriculture’ in Hyderabad show a huge positive sentiment with a count of nearly 1300 positive sentiments against a count of barely 50 negative sentiments as per the retrieved data. The positive sentiments trust and anticipation have a good count followed by joy. The negative sentiments of sadness, fear, anger and disgust have realtively lower frequencies indicating that there is sense of trust in agriculture as a profession, particularly due to the trending tweet where a leading actor mentioned his dream to actually become a farmer.

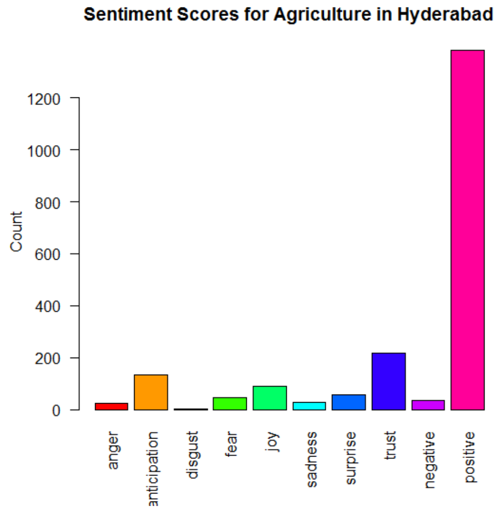

Figure 6

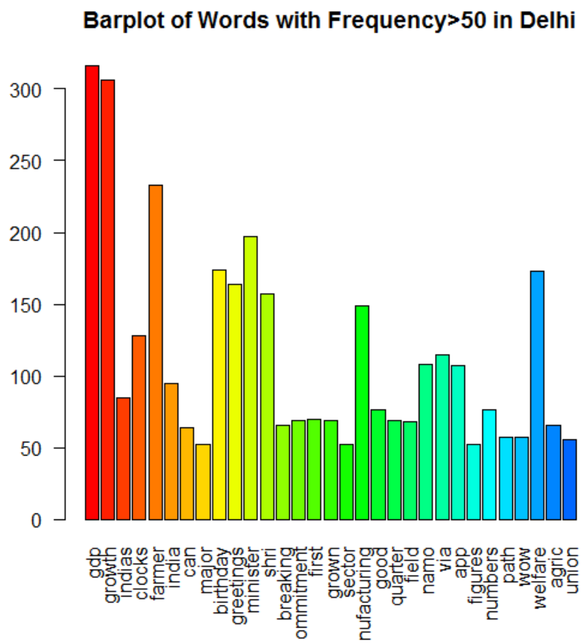

Figure 7

The Delhi-based dataset displays a relatively lower count with the highest frequencies for 'GDP', 'growth', and 'farmer', all being less than 350 counts. The Prime Minister's speech from the Red Fort on Independence Day with a call on brotherhood with J&K became a fad on Twitter. Another trend was observed due to the criticism of the agricultural policy of the current government and ignorance of the service and manufacturing sector by the opposition. The flaring GDP figure of 8.2% for the financial year 2018-19 made its mark in Delhi as well. The wordcloud for the same is shown below:

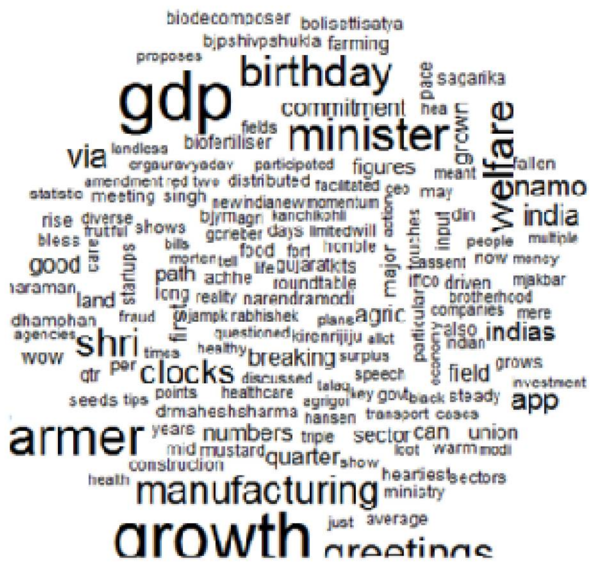

Figure 8

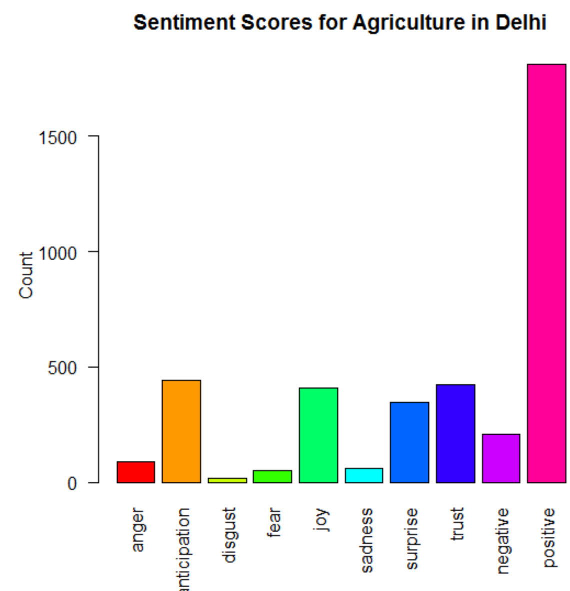

Figure 9

The Delhi-based dataset results in the highest positive sentiment count amongst the three regions with a frequency of nearly 1700 counts while the negative count lies on a mere 250. Trust, anticipation, joy and surprise sentiments(positive) show a high positive count which is justified by the public's faith in PM Modi's Independence Day promises alongwith a ray of hope brought together by the GDP figures.

REFERENCES

[1] Agarwal A., Biadsy F., and Mckeown K. 2009. Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams. *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009*), March, 24–32.

[2] Barbosa L. and Feng J. 2010. Robust sentiment detection on twitter from biased and noisy data. *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, 36–44.

[3] Bermingham A. and Smeaton A. 2010. *Classifying sentiment in microblogs: is brevity an advantage is brevity an advantage? ACM*, 1833–1836.

[4] Davidov D., Tsur O. and Rappoport A. 2010. Semi-supervised recognition of sarcastic sentences in twitter and amazon. *In Proceedings of the Fourteenth Conference on Computational Natural Language Learning, CoNLL '10*. Stroudsburg, PA, USA, 2010. Association for Computational Linguistics, 107–116.

[5] Fellbaum C. 1998. *Wordnet, an electronic lexical database*. MIT Press.

[6] Go A. Bhayani R. and Huang L. 2009. Twitter sentiment classification using distant supervision. Technical report, Stanford.

[7] Haussler D. 1999. Convolution kernels on discrete structures. Technical report, University of California at Santa Cruz.

[8] Kharde, V. A. and Sonawane, S.S. (2016). Sentiment Analysis of Twitter Data: A Survey of Techniques. International Journal of Computer Applications, 139(11) : 5-15.

[9] Klein D. and Manning C. 2003. Accurate unlexicalized parsing. *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, 423–430.

[10] Liang P. and Dai B. 2013. Opinion Mining on Social Media Data. Mobile Data Management (MDM), *2013 IEEE 14th International Conference*(2)

[11] Pang and Lee L. 2008. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, *In Proceedings of LREC*. 2(1-2):1–135.

[12] Pak A., and Paroubek P. 2010. Twitter as a corpus for sentiment analysis and opinion mining. *In Proceedings of LREC*.

[13] Turney P. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. ACL.

[14] Whissel C.W. 1989. *The dictionary of Affect in Language*. Emotion: theory research and experience, Academy press London.