International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-2, July 2019

# Stemming and Lemmatization of Tweets for Sentiment Analysis using R

**Swati Sharma, Mamta Bansal**

***Abstract:*** ***In our digital India, the use of social media like twitter, blogs and various forums is growing with the rapid rate. Thus the size of the data is becoming big day by day and in the span of this type of high varied and volume data, manual analysis would be a clumsy job. So, there is an alarming rate to analyze that large amount of data to make it suitable for analysis purpose. As a most elaborate open source platform, R has immeasurable user communities that thrives and perpetuate a huge amount of text analysis packages. So, in this paper we are analyzing movie related tweets using machine learning in R.***

***Index Terms:*** ***BOW, Linear Classifier, NLP, Rule based Classifier.***

## I. INTRODUCTION

Sentiment Analysis is a category of data mining that evaluate the inclination of an individual's opinion through text mining, Natural Language Processing (NLP), text analysis, Machine Learning, lexicon based approach etc. It focusses to identify the perspective of an individual writer, speaker or the overall polarity with respect to some event. Supervised and unsupervised learning approaches can be used to classify customer's opinion. Opinion Mining is very important to classify different taxonomies as it is a very main tool for categorizing datasets on a broad scale. For example, in bloom's taxonomy, the clusters are formed to analyze the level of understanding.

There are numerous techniques to classify sentiments such as machine learning approach and lexicon based approach. Supervised and unsupervised learning are used to implement machine Learning approach whereas dictionary based and corpus based approach are used to implement lexicon based approach. Further corpus based approach uses statistical as well as semantic technique.

BOW (Bag-of-Words) representation is commonly used in machine learning approach. This method focusses on the independent words and disregards the significance of subjective and semantic information in the document. All the words in the document hold an equal importance. In sentiment analysis, the BOW description is mainly used as it results in high dimensionality of analysis space. To reduce this high dimensionality of attribute space, machine learning algorithms are used such as attribute selection technique which chooses only significant attributes by stemming the noisy and irrelevant words.

## II. METHODOLOGY

To access tweets from twitter an application is being created to get the consumer key and its secret key, access token and its token secret key for creating a handshake authorization as twitter is allowing accessing its data through this API.


**Figure 1. Twitter Application**

People's tweets about Bollywood movies i.e. Kesari and Kalank are collected using R by installing following packages:

```
>install.packages("twitteR")
>install.packages("RCurl")
>require(twitteR)
>require(RCurl)
```

A word cloud is a picture consisting of words that jointly relates to a cloudy image. The magnitude of a word reflects how significant it is or how frequently it comes in a text. An individual generally use word clouds to smoothly generate an abstract of huge amount of data. So on the basis of tweets, following word clouds are created:


**Figure 2. Word cloud of Kesari movie**

**Revised Manuscript Received on 30 July 2019.**

* Correspondence Author

**Swati Sharma,** Ph.D pursuing from Shobhit University, A.P. at M.I.E.T. ,Meerut, India

**Dr. Mamta Bansal,** C.S. , Shobhitt University, Meerut, India.

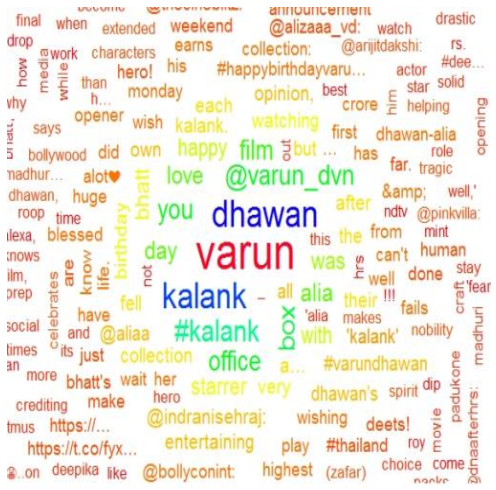


**Figure 3. Word cloud of Kalank movie**

An affinity map is a simple representative to understand and relate all the information. When we have a large amount of varied data such as facts, figures, brainstorming ideas, individual opinion, user requirements, perception and design issues. An affinity map is a category of clustering our dataset or grouping our related information. Following to word cloud, an affinity map is generated:

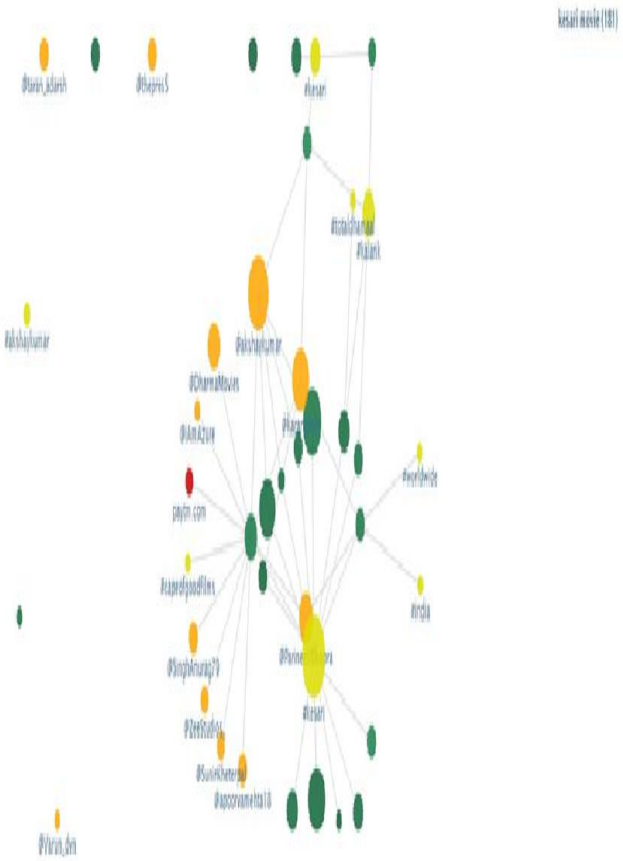

**Figure 4. Affinity map of Kesari movie**

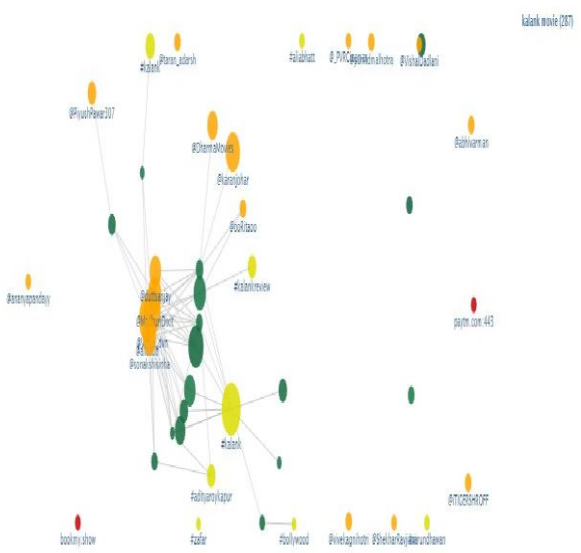

Figure 5. Affinity Map of Kalank Movie

A heat map is a realistic and graphical presentation of data wherein individual figures are contained in a graph in the form of matrix. It consists of four quadrants i.e. active, subdued, pleasant and unpleasant. Heat map is getting popular to represent large volume of data as comprehensible. Then, following heat map is also generated:

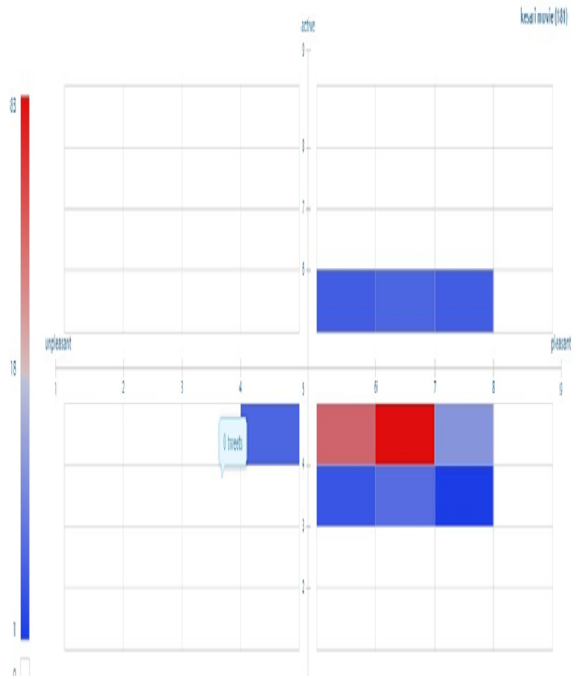


**Figure 6 Heat map of Kesari Movie**

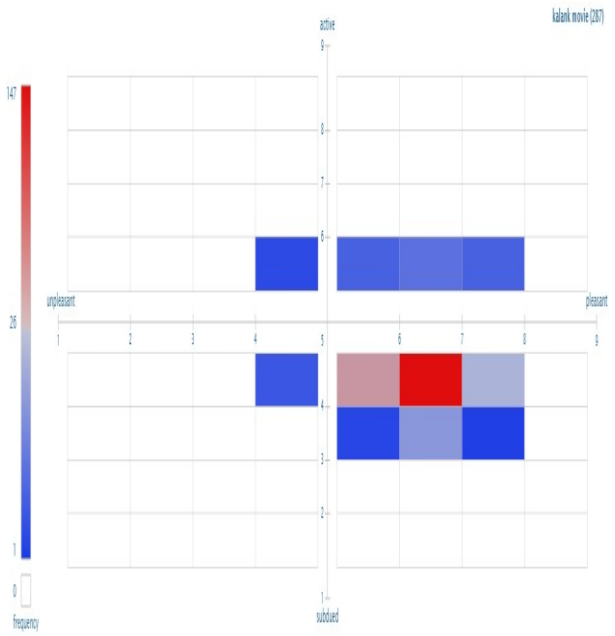


**Figure 7. Heat map of Kalank Movie**

## III. CONCLUSION

In this paper, we are summarizing movie related tweets by creating an application on twitter and accessing handshake authorization for accessing the tweets. Using stemming and lemmatization, the noisy data, stop words, punctuations, symbols and unidentified words are filtered to get data suitable for analysis purpose. Using R, we have created a word cloud, a affinity map and a heat map to showcase the sentiments of people. As this work primarily focusses on the tweets in English language, so in the near future we can summarize the tweets for bilingual text refining.

## REFERENCES

1. Prof. SudarshanSirsat, Dr.Sujata Rao, Dr.Bharti Wukkadada, 2019, "Sentiment Analysis on Twitter Data for product evaluation", IOSR Journal of Engineering .
2. Donia Gamal, Marco Alfonse, El-Sayed M.El-Horbaty and Abdel-Badeeh M.Salem, 2019, "Twitter Benchmark Dataset for Arabic Sentiment Analysis", IJMECS.
3. Abdullah Alsaeedi , Mohammad Zubair Khan, 2019, "A Study on Sentiment Analysis Techniques of Twitter Data", International Journal of Advanced Computer Science and Applications, Vol. 10, No. 2.
4. Hetu Bhavsar, Richa Manglani,2019, "Sentiment Analysis of Twitter Data using Python", International Research Journal of Engineering and Technology (IRJET).
5. Ayesha Rafique, Kamran Malik, Zubair Nawaz, 2019, "Sentiment Analysis for Roman Urdu", Mehran University Research Journal of Engineering and Technology, Vol 38, Issue 2.
6. Omer Awad Mohammed, 2019, "Translating Ambiguous Arabic Words Using Text Mining", IJCSMC.
7. Sahar Sohangir, Dingding Wang, Anna Pomeranets, Taghi M.Khoshgoftar, 2018, "Big Data : Deep Learning for financial sentiment analysis", Journal of Big Data ISSN: 2196-1115 (Online).
8. Haiyun Peng, Yukun Ma, Yang Li, Erik Cambria, 2018, "Learning multi grained aspect target sequence for chinese statemen"t.
9. Vishal Vyas, V.Uma, 2018, "An extensive study of Sentiment analysis tools and binary classification of tweets using Rapid Miner", https://www.researchgate.net/.../301408174_Twitter.
10. Mandava Geetha Bhargava, Duvvada Rajeswara Rao, 2018, "Sentiment Analysis on social media using R programming", International Journal of Engineering and Technology.
11. Tao Chen, Ruifeng Xu, Yulen He, Xuan Wang, 2017, "Improving sentiment analysis via sentiment type classification using BiLSTM-CRF and CNN Expert Systems with Applications", Volume 72, Pages 221-230.
12. Upma Kumari, Dinesh Soni, Dr.Arvind K Sharma, 2017, "A Cognitive study of Sentiment Analysis Techniques and Tools : A Survey", International Journal of Computer Science and Technology.
13. Leszek Ziora, 2016, "The sentiment analysis as a tool of business analytics in contemporary organizations", Uniwersytet Ekonomiczny w Katowicach.
14. G.Vaitheewaran, Dr.L.Arockiam, 2016, "Combining Lexicon and Machine Learning Method to enhance the accuracy of Sentiment Analysis on Big Data", International Journal of Computer Science and Information Technology.
15. Tajinder Singh, Madhu Kumari, 2016, "Role of Text Pre-Processing in Twitter Sentiment Analysis", Procedia Computer Science.
16. Pranali Borele, DilipKumar A.Borilar, 2016, "An approach to sentiment Analysis using Artificial Neural Network with comparative Analysis of Different Techniques", IOSR.
17. S.K.Bharti, B.Vachha, R.K.Pradhan, K.S.Babu, S.K.Jena, 2016, "Sarcastic Sentiment Detection in tweets streamed in real time: a big data approach", Digital Communication and Networks.
18. Devika MD, Sunitha C, Amal Ganesh, 2016, "Sentiment Analysis:A comparative study on Different Approaches", ICRTCSE.
19. Souraya Ezzat, Neamat EI Gayar, Moustafa M.Ghanem, 2012, "Sentiment Analysis of Call centre audio conversations using text classification", IJCISSIMA.

## AUTHORS PROFILE

**Swati Sharma,** B.tech (Honors.), M.Tech (Honors.), Ph.D. pursuing from Shobhit University. I am currently working in MIET Meerut as an Assistant Professor since 2010. My area of interest is data mining, Database Systems, Data Structure, Operating Systems. I had done Python certification from NPTEL. Certified in R language, gold certified. Published an article in newspaper on Sentiment analysis - an upcoming research area.

**Mamta Bansal** has been in the field of Technical Education for the last 30 years. Mamta Bansal received her Ph.D. (Computer Engineering & Information Technology) in 2013 from Shobhit University, Meerut. Currently she is working as an Professor in Shobhit University, Meerut. Her research interest includes Data Mining, Web Mining, Crawler, Cloud Computing and Programming Languages. She has about 40 papers on her credit, which have been published in International, National Journals of repute and presented in various Conferences. She has guided many B.Tech., M.Tech., MCA, M.Phil. and Ph.D. students for their dissertation, project report works.