



INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Monitoring the Keyphrase of tweets using R

Reema Gupta

Assistant Professor

Department of Computer Science

Govt PG College for Women, Sector-14, Panchkula

Abstract—There is a great challenge how to manage big Data. As data storage and processing in a way that it becomes available to be consumed is really a tough task. Because data cannot be load into our traditional database system for analysis and storage, it leads to high cost and much time. Big data is described by 5 Vs (Volume, Variety, Velocity, Variability, Veracity). There are variety of tools available for big data analysis. R is one of the most popular statistical analysis tool. It works on Windows, Linux. R consists of numerous ready to use algorithms, formulae for analysis purpose.

Keywords—*Analysis, Veracity, Package, Tweets, R*

INTRODUCTION

Big data analytics include collection of data from various sources, mainly from social media, IOT, cloud computing, internet, databases and to process it in a way that it can be consumed by analysts [1]. Data analysis is important for a better decision making and also a challenging task as it includes data in unstructured, semi structured and structured form. Big data analysis faces lot of problems as data contain more ambiguity, errors, incomplete due to which storage, fault tolerance, and quality matters [2].

Big data analytics and statistics help in long term decisions. Big data is defined from 5 perspective also known as 5V's.

1. Velocity
2. Volume
3. Variety
4. Veracity
5. Value

Velocity deals with the speed of accumulation of data generation. Volume is related to size. A variety deal with the nature of data i.e. Unstructured, Semi Structured and Structured. Veracity refers to the uncertainty in data and quality of data. Value is useful data as data is very big in quantity and useless until turn into a useful form [3]. In traditional analysis, statistical calculations consume more time. So, we need to use tools and techniques. There are many tools and techniques used for big data analysis. One of the General purpose programming language used for statistical analysis is R. R is a open source software which also provides the visual analytics and having strong graphical and statistical tool embed. R facilitates data management processes such as Transformations, subsetting and cleaning[4]. R consists of numerous ready to use algorithms, formulae for analysis purpose. It allows user to create data Output in the form of plots, graphs, and diagrams. R is also used by Google, facebook, Microsoft etc. for analysis.

This paper presents analysis conducted using R tool on Twitter-tweets.

R AND RSTUDIO

R is one of the most popular statistical analysis tools. It works on Windows, Linux. R consists of numerous ready to use algorithms, formulae for analysis purpose. It's a programming language used for data manipulations, statistical analysis, and data visualization. R is an open source and free. R includes Conditional statements, recursive functions, and input/output commands, built in graphical tools for visualization[5][7] . Thousands of packages are available. To use a package, you must first install it and then load it

Packages are installed using

```
install.packages(name of package)
```

Loading of package is done using

```
library(name of package)
```

The R workspace consists of all the data objects you've created or loaded during your R session. When you quit R by either typing q() or exiting out of the application window, R will prompt you to save your workspace.

PROPOSED WORK

The methodology used to extract the tweets and further analyzed using R is described step by step as shown in figure 1. Pre-processing is a first step plays a very important role in text mining techniques and applications.

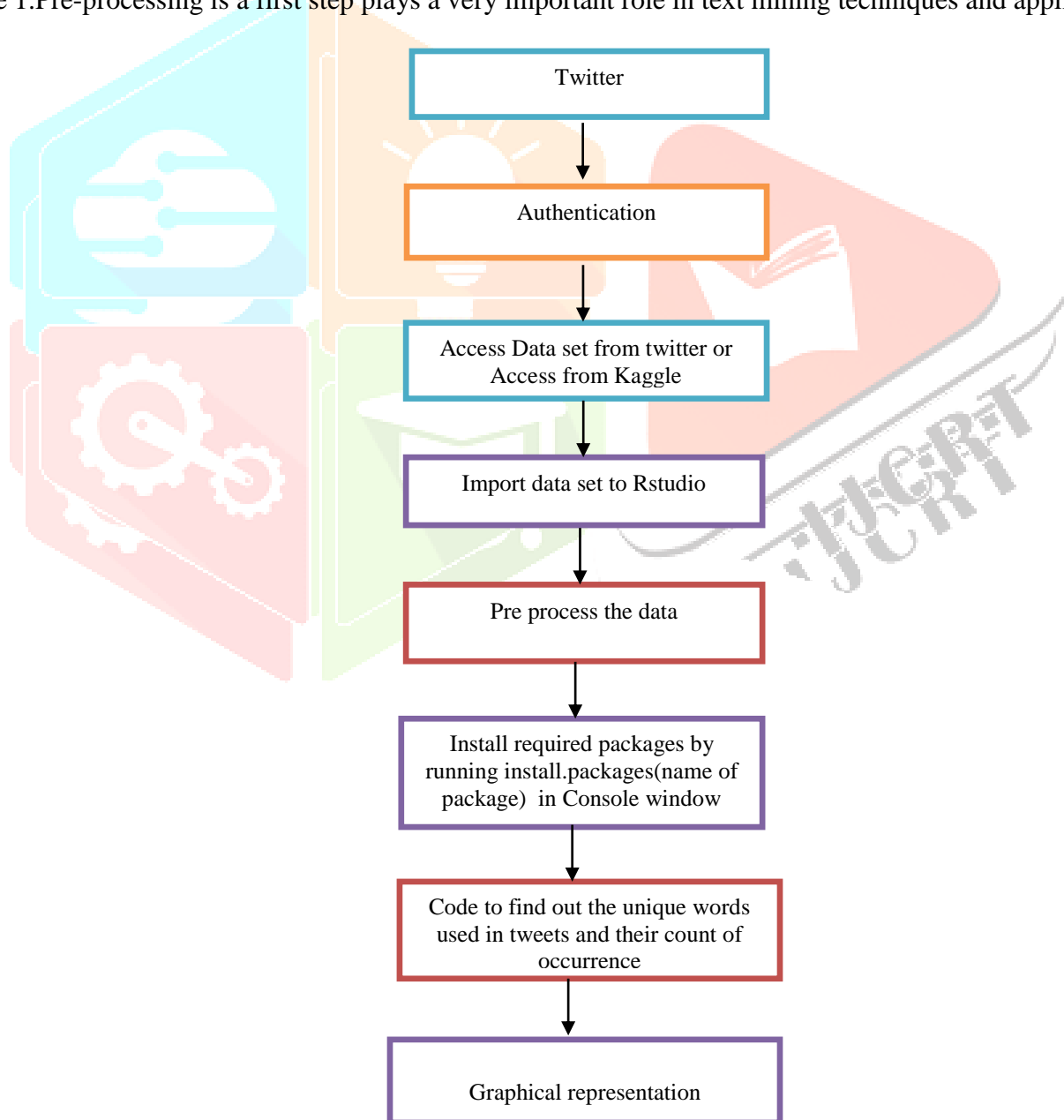


Figure 1. Process to analyze the tweets

RESULTS

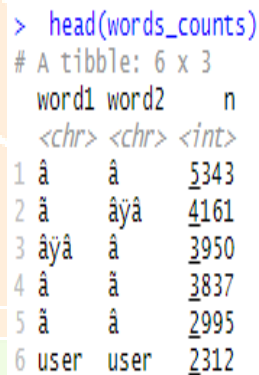
In this section the words appeared in tweets, are extracted for further analysis[6]. The word counts based on their occurrences have been shown in figures.

Following packages are used to extract the results.

```
library(twitteR)
library(rtweet)
library(syuzhet)
library(ggplot2)
library(dplyr)

library(tidyr)
library(tidytext)
library(widyr)
library(devtools)
```

1. Head(word_counts)

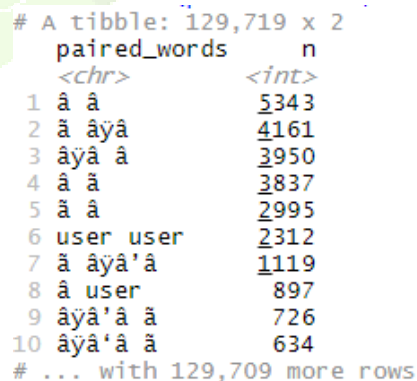


```
> head(words_counts)
# A tibble: 6 x 3
  word1 word2   n
  <chr> <chr> <int>
1 â     â     5343
2 ä     äyâ   4161
3 äyâ   â     3950
4 â     ä     3837
5 ä     ä     2995
6 user  user   2312
```

Figure 2. Word counts appeared single time

Head displays top 6 records with number of count of appeared words.

2. When a pair of words come together most of the time



```
# A tibble: 129,719 x 2
  paired_words   n
  <chr>         <int>
1 â â           5343
2 ä äyâ         4161
3 äyâ â         3950
4 â ä           3837
5 ä ä           2995
6 user user      2312
7 ä äyâ'â       1119
8 â user         897
9 äyâ'â ä        726
10 äyâ'â ä        634
# ... with 129,709 more rows
```

Figure 3. Word counts appeared in pairs

When a pair of words come together most of the time in the tweets (data set taken)

3. When three words repeat maximum number of times

```

paired_words      n
<chr>             <int>
1 ă âyâ â         3165
2 ă â â           2600
3 âyâ â â         2323
4 â â âyâ         1889
5 â â â           1436
6 â â â           1274
7 user user user   989
8 â â user         716
9 ă âyâ'â â       659
10 âyâ'â â âyâ'â  477
# ... with 190,064 more rows
    
```

Figure 4. Word counts appeared three times

When three words repeat more time it is counted with count displayed.

4. When two pairs(4 words) occur together

```

paired_words      n
<chr>             <int>
1 ă âyâ â â       1813
2 âyâ â â âyâ     1739
3 â â âyâ â       1473
4 â â â â         1129
5 ă â â â         1093
6 â â â â         1063
7 user user user user 464
8 ă âyâ'â â âyâ'â 461
9 âyâ'â â âyâ'â â 241
10 i am thankful for 221
# ... with 212,148 more rows
    
```

Figure 5. Word pairs occurred together

When 4 words appeared maximum number of times displayed with number of times of its repetition.

5. Count of unique words appeared in tweets

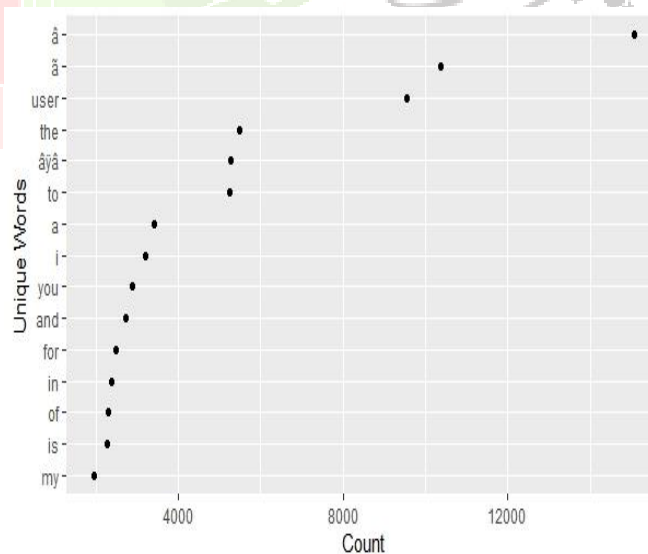


Figure 6. Unique words appeared in tweets

This graph indicates the unique words on Y-axis appeared in tweets and their count on X-axis.

Word Network: Tweets using the hashtag when count>250

Text mining twitter data Using R

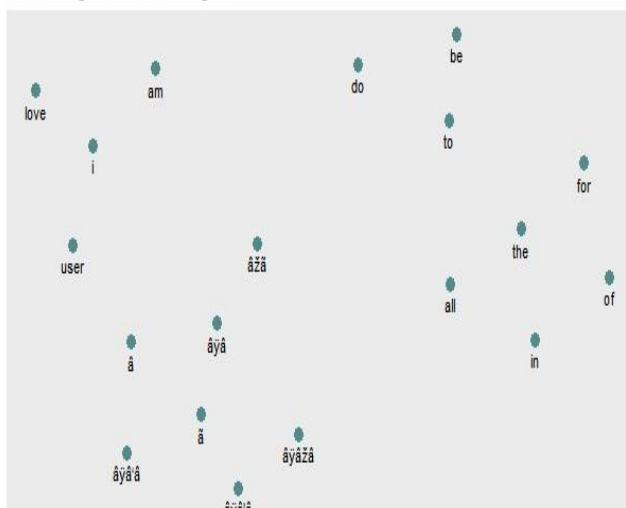


Figure 8. Words count>250

❖ When $N > 500$

This (Word network) shows the no of words in the network of tweets having count more than 500.

Word Network: Tweets using the hashtag when count>500

Text mining twitter data Using R

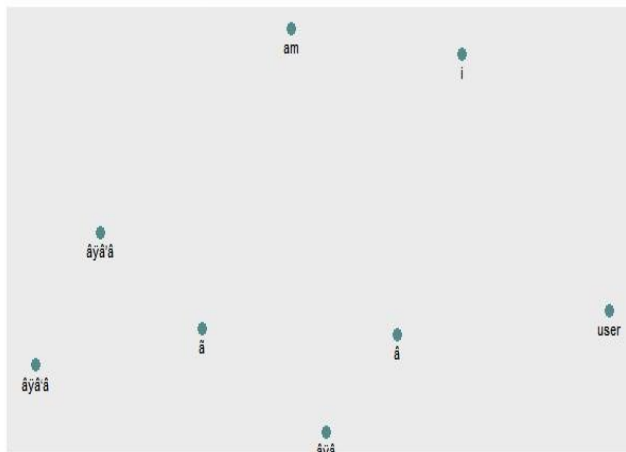


Figure 9. Words count>500

❖ When $N > 1024$

This (Word network) shows the no of words in the network of tweets having count more than 1024.

Word Network: Tweets using the hashtag when count>1024

Text mining twitter data Using R

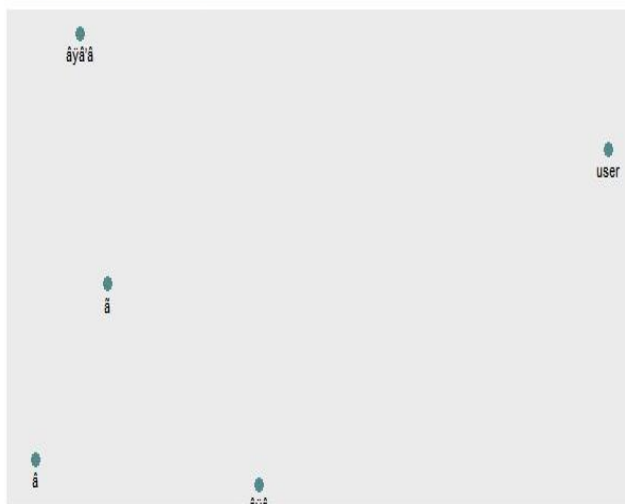
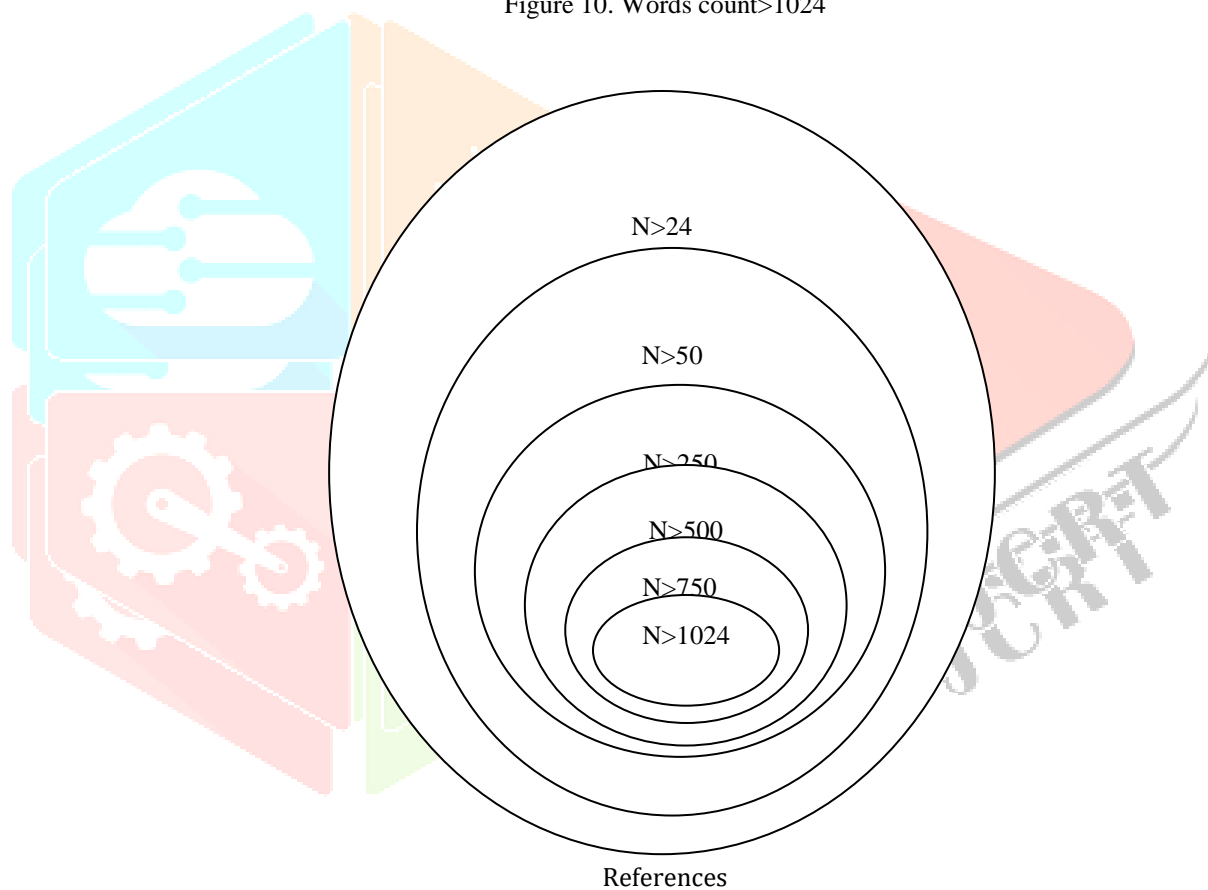


Figure 10. Words count>1024



References

Figure 11. Venn Diagram

$$N>24 \geq N>50 \geq N>250 \geq N>500 \geq N>750 \geq N>1024$$

The number of words with cpunt more than 24 are highest as compared with the one with 50 count and same it is compared and in the last number of words with count 1024.

References:

- [1] R. Khobragade and L. H. Patil, "Facebook Data Mining and Sentiment Analysis Using R Language," vol. 9, pp. 16–20, 2019.
- [2] S. Patil, "Big Data Analytics Using R," *Int. Res. J. Eng. Technol.*, pp. 2395–56, 2016.
- [3] C. Paper, U. Tochukwu, and E. Internationale, "Big data statistics with R," *Researchgate.Net*, no. August, 2015.
- [4] G. Ostrouchov, "Programming with Big Data in R Why R?: Programming with Data," 2016.
- [5] Savita and N. Verma, "A Review Study on Big Data Analysis Using R Studio," *Int. J. Eng. Technol. Manag. Res.*, vol. 6, no. 6, pp. 129–136, 2020.
- [6] Q. Zhang, Y. Wang, Y. Gong, and X. Huang, "Keyphrase extraction using deep recurrent neural networks on twitter," *EMNLP 2016 - Conf. Empir. Methods Nat. Lang. Process. Proc.*, pp. 836–845, 2016.
- [7] Kouloumpis, E., Wilson, T., & Moore, J. (2021). Twitter Sentiment Analysis: The Good the Bad and the OMG!. *Proceedings of the International AAI Conference on Web and Social Media*, 5(1), 538-541. <https://doi.org/10.1609/icwsm.v5i1.14185>
- [8] S. Bhuta, A. Doshi, U. Doshi and M. Narvekar, "A review of techniques for sentiment analysis Of Twitter data," *2014 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT)*, Ghaziabad, India, 2014, pp. 583-591, doi: 10.1109/ICICT.2014.6781346
- [9] K. Sailunaz and R. Alhajj "Emotion and Sentiment Analysis from twitter text" *Journal of Computational Science* Vol 26, September 2019.
- [10] Dang, Nhan Cach, María N. Moreno-García, and Fernando De la Prieta. 2020. "Sentiment Analysis Based on Deep Learning: A Comparative Study" *Electronics* 9, no. 3: 483.
- [11] Gao Y, Xie Z, Li D. Electronic Cigarette Users' Perspective on the COVID-19 Pandemic: Observational Study Using Twitter Data. *JMIR Public Health and Surveillance* 2021;7(1):e24859