How to Cite:
Sundravadivelu, K., Thangaraj, M., & Gnanambal, S. (2022). An extensive work on comparing sentiment patterns in twitter archives between two persons. *International Journal of Health Sciences*, *6*(S7), 5170-5180. https://doi.org/10.53730/ijhs.v6nS7.13104

An extensive work on comparing sentiment patterns in twitter archives between two persons

K. Sundravadivelu
Assistant Professor, Department of Computer Science, Madurai Kamaraj University, Madurai, India
Email: svadiveluk2021@gmail.com

Dr. M. Thangaraj
Professor and Head, Department of Computer Science, Madurai Kamaraj University, Madurai, India.
Email: thangaraj-dcs@mkuniversity.org

Dr. S. Gnanambal
Guest Lecturer, Department of Computer Science, Rajadoraisingam Govt.Arts College, Sivaganga, India.
Email: gnanardm@gmail.com

Abstract---In social media starting from 2006 Twitter gets Major attention because of the text shared via Twitter causes positive or negative impacts. Hence comparing two persons twitter archives is important to analyse the similarity of words they use semantics of the reviews and finding the Trends of words. This study evaluates how the characteristics of tweets are changing over years. This study finds which words used in tweets cause a number of retweets and also returns the sentiments of tweets.

Keywords---Sentiment analysis, sentiment patterns, Tweet extraction, word usage and etc.

Introduction

Billions of people are using twitter to convey their suggestions, Sentiments about daily live news and events [1]. These tweets are analysed to know the active state of people and Society with respect to the noted issues which is helpful to analyse the frequency of words, trend of words and sentiment of tweets [2].

This type of social media text is used for the purpose of understanding the thoughts of people about a particular product or event [3]. This will be used to measure the citizens' emotions and catastrophe events.

Twitter has more than 300 million users and produces 6000 tweets per second [4]. Nowadays evaluating the tweets is mainly used to identify patterns and find the hidden inference of data [5].

Discovering similar patterns that are generated between persons with respect to things generated is one of the twitter based content applications [6]. Government utilises this type of similarity Bayesian analysis to identify the thread related contents, whereas Business people used this for recruitment and target marketing [7]. Individual persons use this task to identify similar persons.

Related work

This framework [8] evaluates the tweets into positive, negative and neutral then those are stored in a database. Then this study compares the data with respect to various classification and regression algorithms. But this paper doesn't suggest which algorithm best corresponds to classification and regression.

This paper [9] identifies the similar users for the purpose of profiling users with respect to social and security purposes. Seven attributes are used to profile users such as retweets, favourite and common hash tags, common interest, profile similarly following and followers. The returned results are evaluated by humans who use the service. In order to handle huge amounts of data this work uses map reduce approach but this work missed some important features such as account type.

This paper [10] addresses the behaviour of those persons who joined both Face book and Twitter. It is used to find the specificity of users such as choice of friends, privacy settings and activities carried out. The main drawback of this study is lack of sample size and latent variables.

This paper [11] compares the twitter contents with the news media of New York Times. The topics are discovered via LDA model then by the use of unsupervised topic modelling the tweeted text is compared with the news medium. But this framework lacks in the visualisation of twitter contents.

Proposed system:

This architecture of the proposed research study is depicted in figure 1. This contains three layers 1. Tweets extraction and wrangling layer 2. Word usage layer and 3. Sentiment patterns analysis layer. These three layers are explained briefly in the below sections.

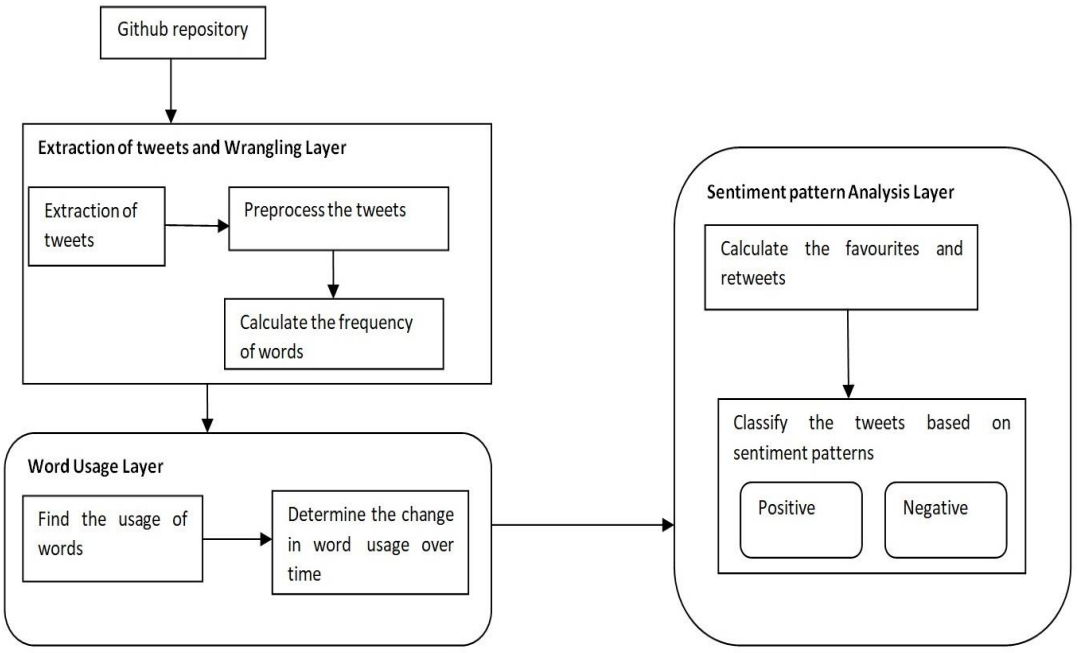


Figure 1: Architecture of the proposed research framework.

Tweets extraction and wrangling layer:

Tweets Extraction is the process of extracting tweets for analysis and which is a first task required for analysis. R programming is used as an experiment to carry out the task. The twitter dataset of two familiar popularities Elonmusk and trump is drawn from Github repository. In addition to that this layer contains two more tasks 1. Preprocess the tweets and 2. Calculate frequency of words. These two tasks are described below,

Pre-process the tweets:

The retrieved raw tweets are not pure and highly unstructured. It consist of URL's, Punctuation, hashtags and lots of redundant text which needs to be repaired to achieve better quality in analysis. First the tweets are grouped in a corpus. From the R language the packages tm and string are utilized to carry out these text mining tasks. Table 1 shows the limited list of functions that are used to clean the irrelevant data from the corpus.

No	Function name	Meaning
1	removePunctuation	Remove punctuation marks
2	removeNumbers	Remove numbers that has no value in sentiments
3	filter(stopwords(source = "stopwords-iso"))	Remove stop words such as articles, conjunction and etc using the pre defined stopwords-iso list.
4	trimws	Removal of extra white spaces
5	Str_remove(^RT)	Removal of text that are retweet

Table1: List of few functions used in pre-process

Extraction of hashtag: Twitter identifies the topic via hashtags. Hence construct a new feature on the basis of values of hashtag. After pre-processing the Tweets counts for Elonmusk and Trump are shown in figure 2.

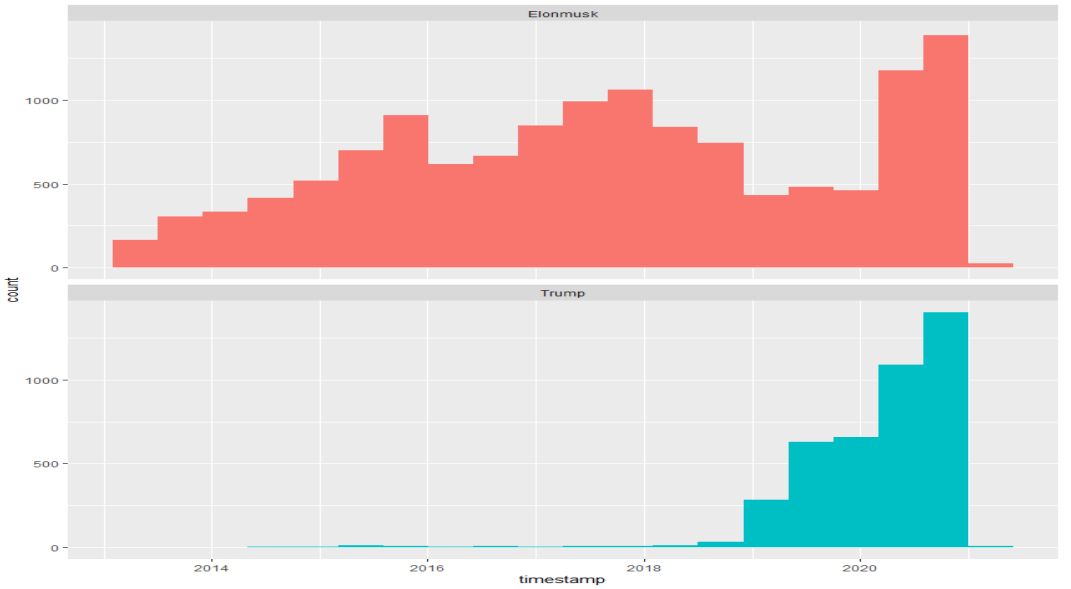


Figure 2: Twitter counts of Elon and Trump

Calculate frequency of words:

This step is used to find the frequency of word occurrences in the corpus. Corpus is grouped with respect to the person then calculates the number of times each word is used by each and every person. That returns how many words are used by each person. Later determine the frequency of each word. In order to calculate the frequency of words, pivot_wider() method is used to construct a data frame.

Word usage layer:

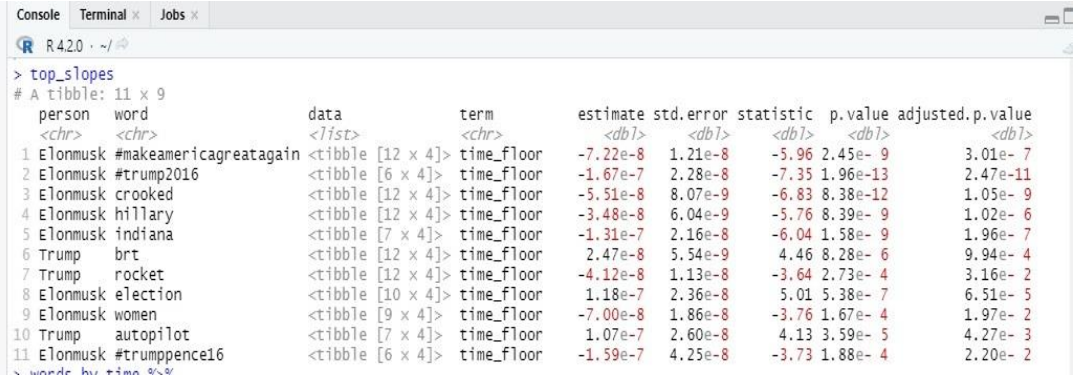
This layer focuses the words usage in tweets. It consists of two components 1. Find the usage of words and 2. Determine the change in word usage over time.

Find the usage of words:

In order to achieve precise results, str-detect () method is utilized to neglect user names from the corpus. Then the words that are used more than 20 times are taken into consideration. Then find the log-odds ratio of the word. With the adaptation of Log odds ratio, determine the words that are more or less probable to arrive from each one's account. Log odds ratio is the logarithmic form of odds ratio.

Determine the change in word usage over time:

Next task is to identify which words have changed rapidly in their respective twitter feeds over time. To carry out this task, first the time bin is created. For that purpose a time bin representing a nearly 1 month time frame is constructed. Then calculate the number of times the word is used in this timestamp bin. Finally, find the words that are used at least 30 times. Now the data frame contains a row with respect to the combination of person-word. Then find the slopes with the use of glm objects which are in a tidy package. That returns the most significant slopes and which words moderately changed their frequency in the tweets. The returned results of the slopes regarding the most significant words are presented in the figure 3.



Console Terminal Jobs

R 4.2.0 · ~/

```
> top_slopes
# A tibble: 11 x 9
   person   word                data              term       estimate std.error statistic  p.value adjusted.p.value
   <chr>    <chr>               <list>            <chr>         <dbl>     <dbl>     <dbl>    <dbl>            <dbl>
 1 Elonmusk #makeamericagreatagain <tibble [12 x 4]> time_floor -7.22e-8   1.21e-8     -5.96 2.45e- 9         3.01e- 7
 2 Elonmusk #trump2016          <tibble [6 x 4]>  time_floor -1.67e-7   2.28e-8     -7.35 1.96e-13         2.47e-11
 3 Elonmusk crooked             <tibble [12 x 4]> time_floor -5.51e-8   8.07e-9     -6.83 8.38e-12         1.05e- 9
 4 Elonmusk hillary             <tibble [12 x 4]> time_floor -3.48e-8   6.04e-9     -5.76 8.39e- 9         1.02e- 6
 5 Elonmusk indiana             <tibble [7 x 4]>  time_floor -1.31e-7   2.16e-8     -6.04 1.58e- 9         1.96e- 7
 6 Trump    brt                 <tibble [12 x 4]> time_floor  2.47e-8   5.54e-9      4.46 8.28e- 6         9.94e- 4
 7 Trump    rocket              <tibble [12 x 4]> time_floor -4.12e-8   1.13e-8     -3.64 2.73e- 4         3.16e- 2
 8 Elonmusk election            <tibble [10 x 4]> time_floor  1.18e-7   2.36e-8      5.01 5.38e- 7         6.51e- 5
 9 Elonmusk women               <tibble [9 x 4]>  time_floor -7.00e-8   1.86e-8     -3.76 1.67e- 4         1.97e- 2
10 Trump    autopilot           <tibble [7 x 4]>  time_floor  1.07e-7   2.60e-8      4.13 3.59e- 5         4.27e- 3
11 Elonmusk #trumppence16       <tibble [6 x 4]>  time_floor -1.59e-7   4.25e-8     -3.73 1.88e- 4         2.20e- 2
```

Figure 3: Returned results with respect to most significant slopes in the given timestamp for Elonmusk

Sentiment pattern analysis layer:

Sentiment pattern analysis is one of the leading natural language processing (NLP) technique applied to analysis whether tweet is positive, negative or neutral. For Sentiment analysis the texts that are retweeted are extracted from the corpus. It consists of two tasks 1. Calculate the favourites and retweets and 2. Classify the tweets based on sentiments. These two tasks are explained in below.

Calculate the favourites and retweets:

In this case count the number of times each of the words was returned. Then determine the total number of retweets for every person. Finally calculate the median number of retweets for each person and word. This can be achieved through group by () and summarise () methods of tidy text packages that calculates the median value of retweets. Favourites are the words that are highly retweeted. By the use of slice-max () the tweets that have peak median value are returned.

Classify the tweets based on sentiment patterns:

In order to analyze the popular patterns and new emotions in a text, tidytext package provides getsentiments() method which finds sentiments on the basis of nrc,Bing and AFINN lexicon. By the combination of tidytext and lexicons of textdata package, generates the repository which has word to emotion pairs. This study uses AFINN lexicon that provides negative or positive score to each word then emotion score is added by the getsentiments() method. Finally if the total score is greater than zero means positive sentiment, score less than zero shows it is negative sentiment whereas score zero represents neutral.

Performance Evaluation

The proposed study was implemented in R-4.2.1 in Pentium® Core i5 with 8 GB RAM, running on Windows 10. The results retrieved from the three layers of the proposed study are discussed in this section.

Base on word frequency:

With the use of ggplot2 package, Tweets extraction and wrangling layer generates the frequency of word occurrences in a plot which is shown in figure 4. The words closest to the line in figure 4 indicate that these words return the same frequency for both persons. The words that are remote to the line show that these words are used heavily by one person then the next person.

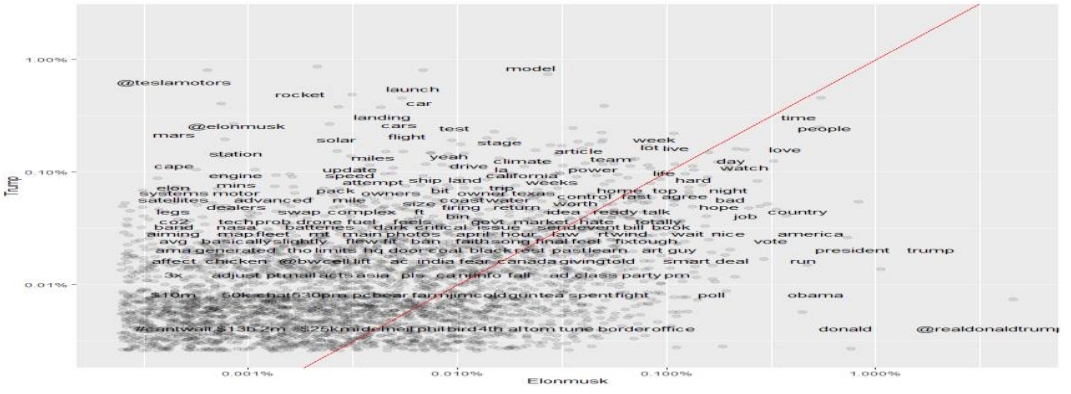


Figure 4: Frequency of word occurrences

Based on words usage:

For the purpose of determining words usage, the log odds ratio method is used. The log odds ratio for the two persons is shown in figure 5 and the figure concludes that the Elon was more active in the given timestamp in our case it is 2020 than trump.

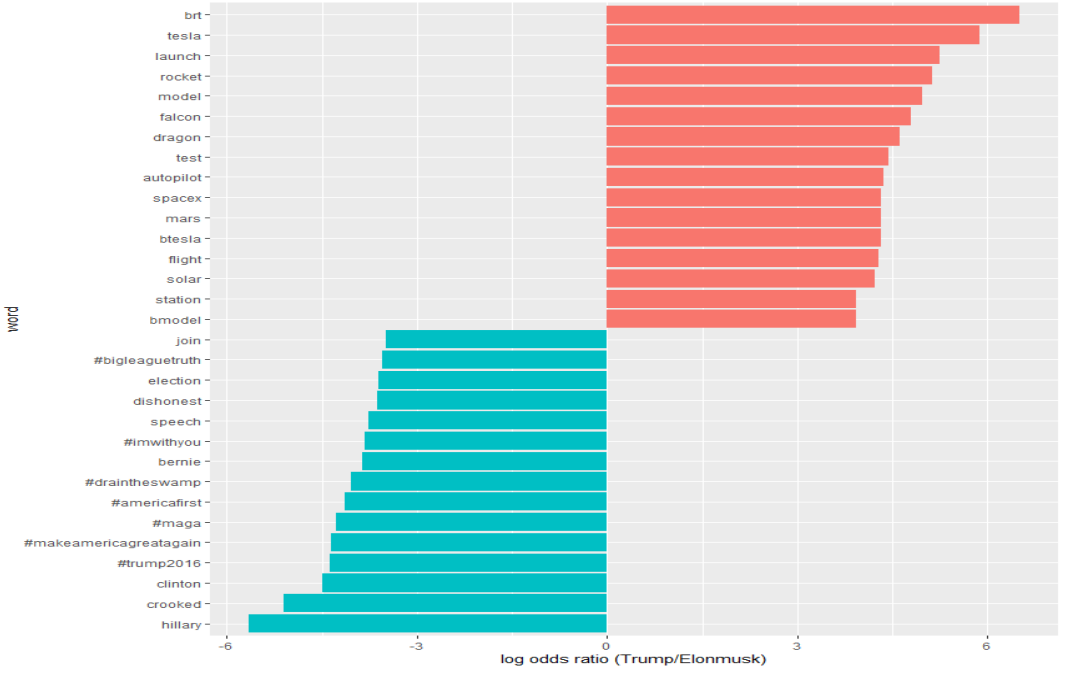


Figure 5: Log odds ratio to calculate usage of words

The next task of the word usage layer is to evaluate the changes occurred in words usage over time. The results of the Elon and Trump with respect to change in word usage are presented in figure 6 and 7. These two figures show that the change in usage of words is higher in Trump.

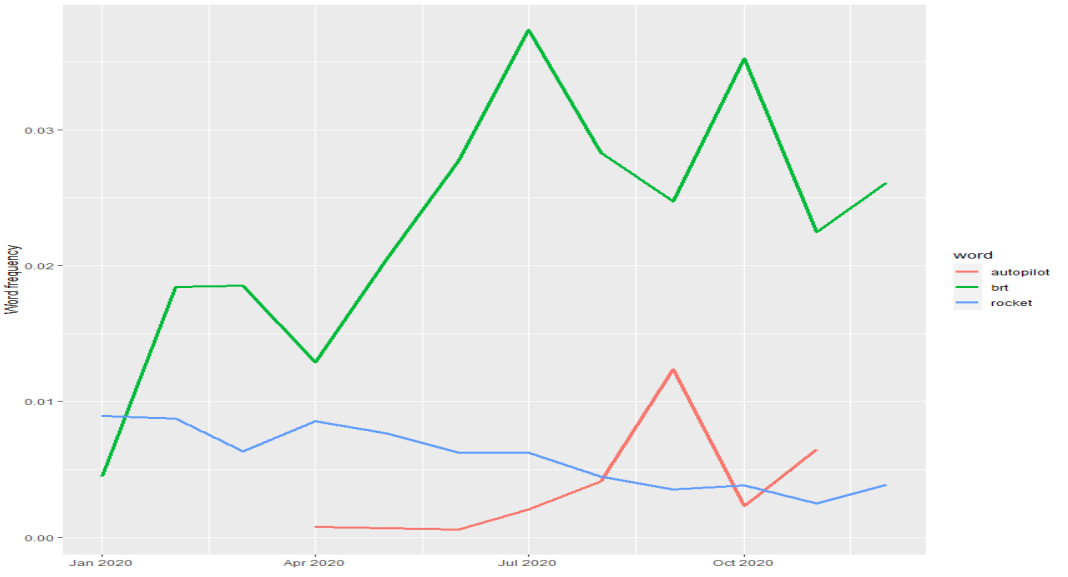


Figure 6: Change in usage of words in the timeframe Jan 2020 to December 2020 for Elonmusk.

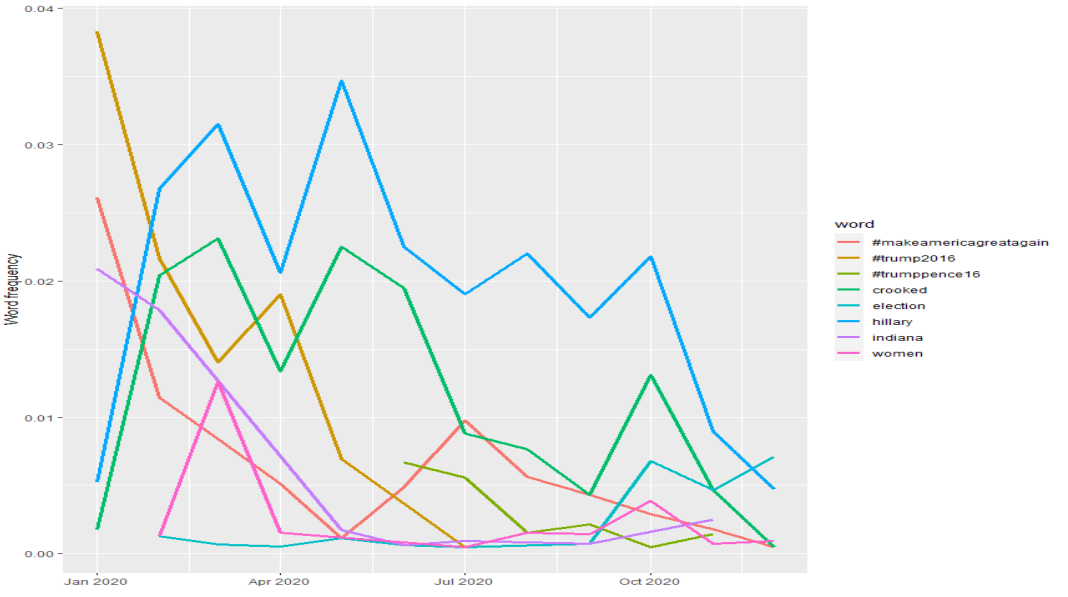


Figure 7: Change in usage of words in the timeframe Jan 2020 to December 2020 for Trump.

Based on favourites, retweets count:

During the process of sentiments analysis tasks the favourites and retweet count is taken into account. Median value of retweets containing each word for Elon and Trump is calculated which is presented in figure 8.

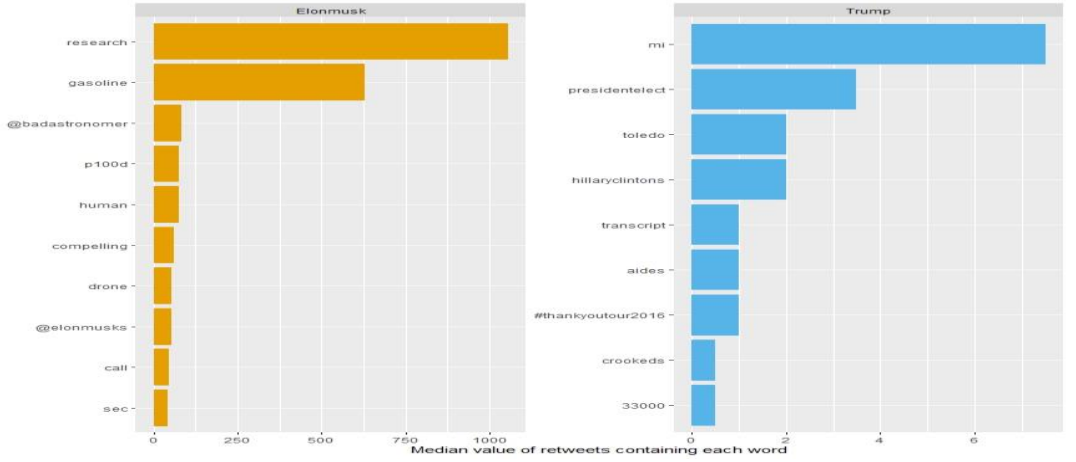


Figure 8: Median value of retweets containing each word for Elon and Trump
Final task is to evaluate the tweet sentiment patterns of both Elon and Trump. The results of sentiment classification are in figure 9.

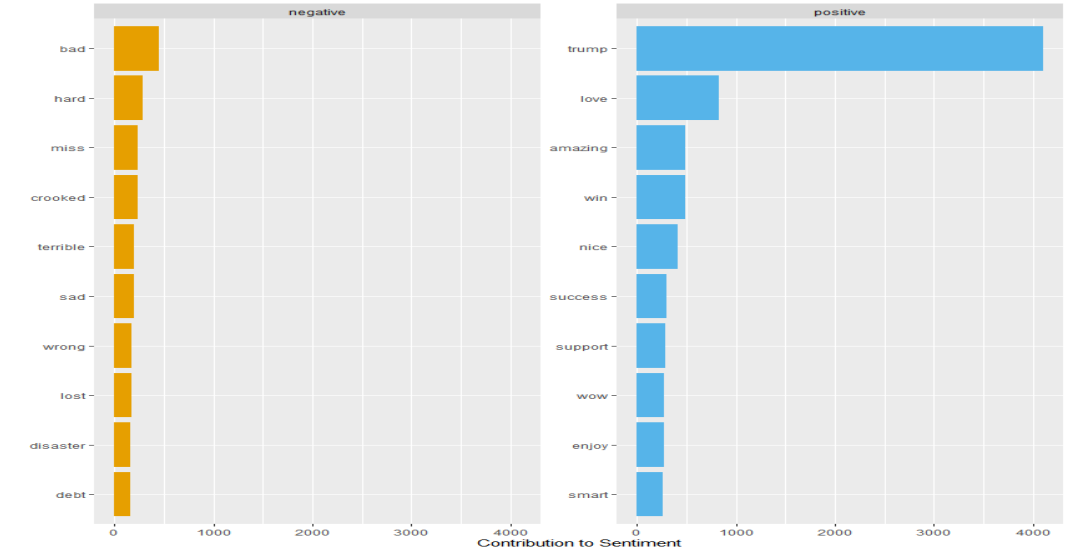


Figure 9: Results of sentiment classification

In the above figure the word trump is shown as positive which indicates the former president of US hence it falls under positive score. This proposed study next deal with TF-IDF score. TF represents the Term Frequency of the word that is how often the word occurs in the dataset. TF is calculated by,

TF = Frequency (word)/Total no of words

Another measure is the Inverse Document Frequency (IDF) which reduces the weight for frequently used words while the words that are not frequently used are increases their weight. That can be calculated by the below formula as

IDF = Log ((Total number of tweets) / (Number of tweets contains the word))

These two measures TF and IDF are combined to produce TF-IDF score to indicate the importance of a word. The TF-IDF results related to Elonmusk and trump tweets dataset is given in figure 10.

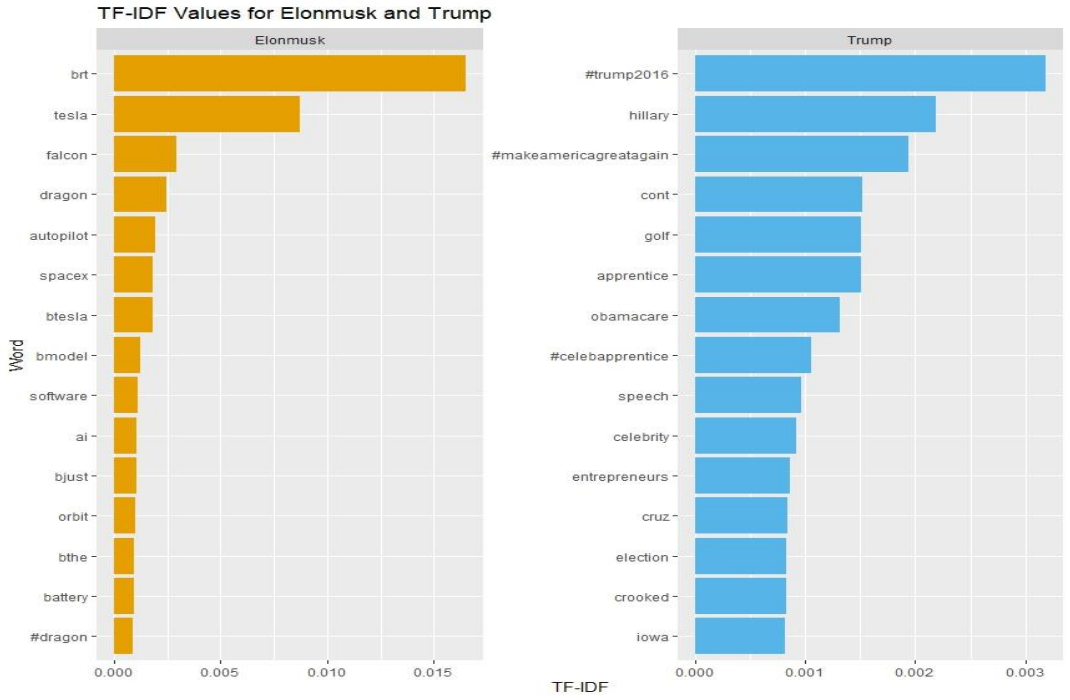


Figure 10: TF-IDF results related to Elonmusk and trump tweets

Conclusion

The comparison of word frequency shows which words are tweeted more likely and less often tweeted. It measures characteristics of tweets and trend of tweets that changes over time. This study also concludes which words cause more number of retweet and favourites. The calculated TF-IDF score will be used in future to extend the study in the aspect of neural network based classification. This study shows the significance of sentiment in people, with the adaptation of new technology future research in sentiment analysis can be improved more. In addition to that further research is carried out to handle multiple sentiment patterns.

References

[1]. H. Kwak, H. Park, C. Lee, and S. Moon, “What Is Twitter, a SocialNetwork or a News Media?” Proc. 19th Int’l Conf. World Wide Web, 2010 pages 591-600.
[2]. Srinivasan M. S., Thulasidasan Sunil and Srinivasa Srinath. “A comparative study of two models for celebrity identification on Twitter.” India: Computer Society of India, 2014. COMAD. Hyderabad, pages 57-65.
[3]. Zhang, L and Liu, B. “Sentiment Analysis and Opinion Mining” , Springer US, Boston, MA (2017) pages 1152–1161.
[4]. Razis G and Anagnostopoulos I, “Discovering similar Twitter accounts using semantics”, Eng Appl Artif Intell. 2016 issue 51 pages 37–49.
[5]. A. Krouska, C. Troussas and M. Virvou, “The effect of preprocessing techniques on twitter sentiment analysis,”, 7th InternationalConference on Information, Intelligence, Systems Applications (IISA),2016.
[6]. Noah Goodman and Daniel Light, “Coding Twitter, lessons from a content analysis of informal science”,2016. Available online at: https://cct.edc.org/sites/cct.edc.org/files/publications/AERA2016_TwISLE.pdf.
[7]. Zimmer, M., and Proferes, N. J, “A topology of Twitter research: Disciplines, methods, and ethics”, Aslib Journal of Information Management, issue 66(3), pages 250-261.
[8]. Aditya Kulkarni1 and Shubham Mhaske. “Tweet Sentiment Analysis and Study and Comparison of Various Approaches and Classification Algorithms Used”, International Research Journal of Engineering and Technology (IRJET) ,Issue: 04 ,2020. Pages 2619-2624.
[9]. Hind AlMahmoud and Shurug AlKhalifa,” TSim: a system for discovering similar users on Twitter”,Journal of Big Data, 2018, issue 5.
[10]. Francesco Buccafurri, Gianluca Lax and et.al.” Comparing Twitter and Facebook user behavior: Privacy and other aspects”, Computers in Human Behaviour, issue 52, 2015. Pages 87-95.
[11]. Wayne Xin Zhao1, Jing Jiang and et.al.” Comparing Twitter and Traditional Media using Topic Models”, European Conference on Information Retrieval ECIR 2011: Advances in Information Retrieval pp 338–349.