Covid -19 Tweets Sentiment Analysis

Aarushi Agarwal

Executive Summary

Covid-19 has created a new normal to our way of life. As the businesses are opening, a lot of analysis is going on over some big questions – "Are people ready to move out of their houses?", "Is work from home the way it is going to be from now on?" , "What are their emotions about the change?" ,"About covid-19?" In this project, we built a model based on Naive Bayes algorithm, to analyze the sentiments of people via tweets from March till June 2020 and classified them as positive or negative. Our analysis shows us that the number of positive tweets has been consistently higher throughout the duration. Opposite to what some of the media houses are telling us, there is a lot positivity around, in terms of fighting this pandemic,being supportive towards policies and helping each other tackling one of the biggest pandemics which has forced us into our homes.

A. Introduction

Sentiment analysis is the gathering of people's views regarding any event happening in real life. The objective of this project is to classify tweets related to coronavirus as positive and negative and analyze the variation in sentiments of people with time. The tweets range from March to June 2020. The tweets are focused around the keywords #Coronavirus, #Coronaoutbreak, #COVID19. Based on the current situation, it is our assumption that the tweets will be generally negative.

This analysis can help us to come up with an agenda for the new normal. It can help businesses in making some important decisions. The use cases are infinite, ranging from inventory management which has turned out to be of particular importance and challenging during these unprecendted times, to understanding the emotions of people circling government policies during this catastrophic event.

This list could go on and on, but something common between all of these, is the fact that we are trying to see how people react in certain situations and chain of events. Something that is really important to understand now, but can be used in aspects, even when Covid-19 fades away.

B. Data Description

Two datasets have been used in this project:

1. Dataset for training the machine learning model.

A pre-classified dataset has been used from Kaggle to train the model. It has 96448 tweets, with each tweet classified as either positive or negative.

2. Tweets from March 3, 2020 and June 9, 2020.

This dataset is provided by Harvard Dataverse. It comprises of 24 csv's with coronavirus tweet ids from March 3, 2020 to June 9, 2020. The hashtags used for tweets extraction are #Coronavirus, #Coronaoutbreak and #COVID19.The tweets are from all over the world.

C.Preprocessing Data

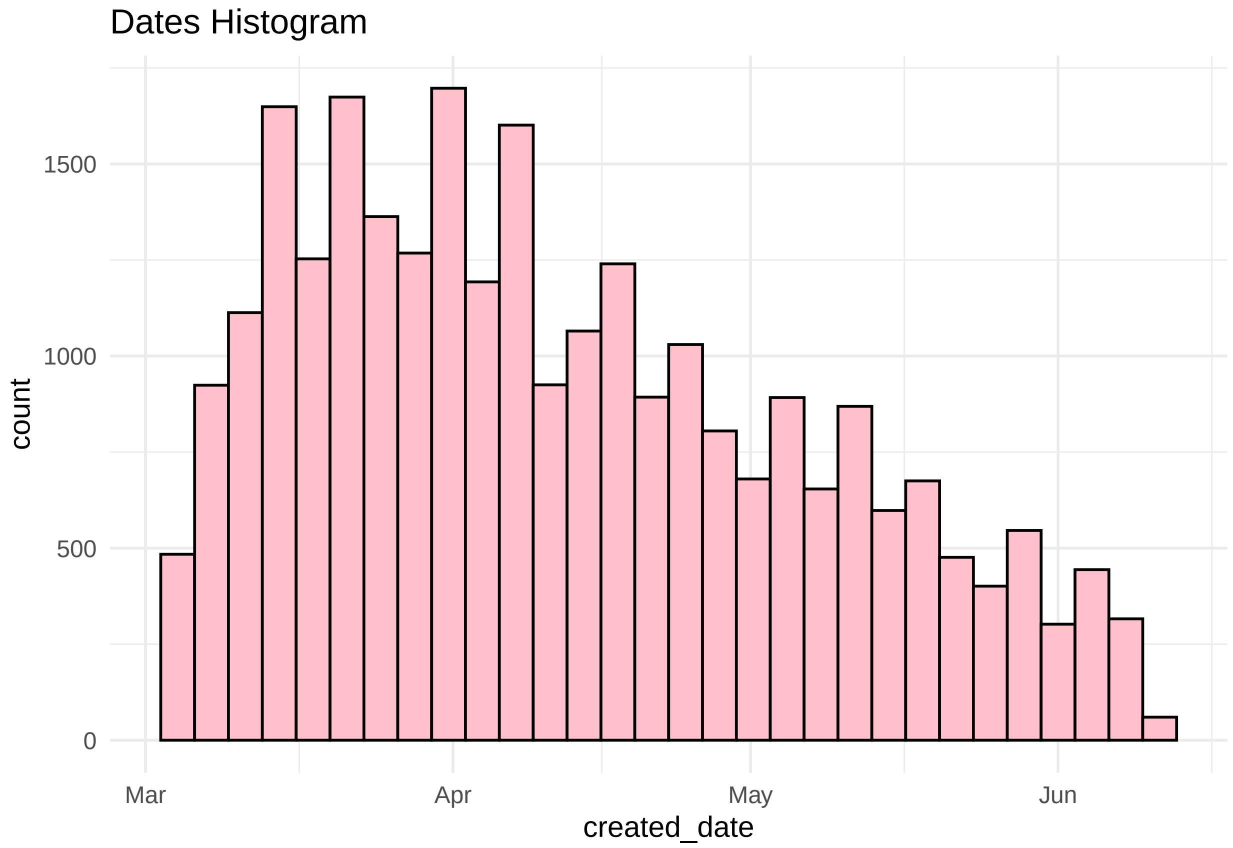
After random sampling tweets from the dataset, I boiled it down to 103,154 tweets. A tweet Hydrator application was used to get actual Twitter data(text of the tweet,created_date, names of the person, location, likes, retweet count, user popularity parameters etc.) corresponding to the random sampled tweets. This data was in json format. An online parser was used to convert the json file to a csv.

My analysis started with the parsed csv file. Starting with preprocessing the data. Initially, I selected all the tweets belonging to the English language and discarded the rest – boiling my data set to 52976 records. Further, I have removed the records where hashtags column was empty, which gave me a final dataset with 30923 records. For further analysis, I also did date formatting and tweets text formatting. Additionally, I have dropped some columns with had very few values.

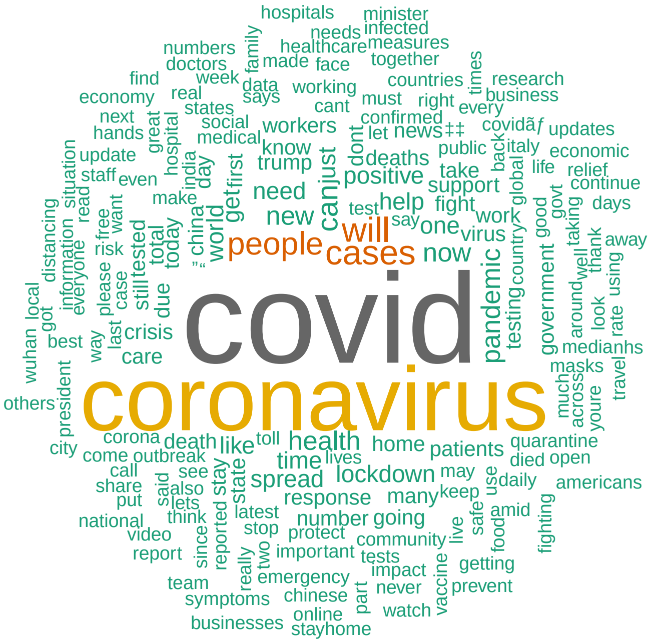
The tweets text formatting was done for both the datasets. The tweets text was first cleared from punctuation, url, special characters, upper case, digits and extra spacings. Then, I converted it into a Corpus. After converting it to a corpus, I removed english language stopwords and converted the Corpus to a Term Document Matrix. A Term Document Matrix is a mathematical matrix that describes the frequency of terms that occur in a collection of documents. Now, I have rectangular structured Dataset for further analysis.

D. Exploratory Data Analysis

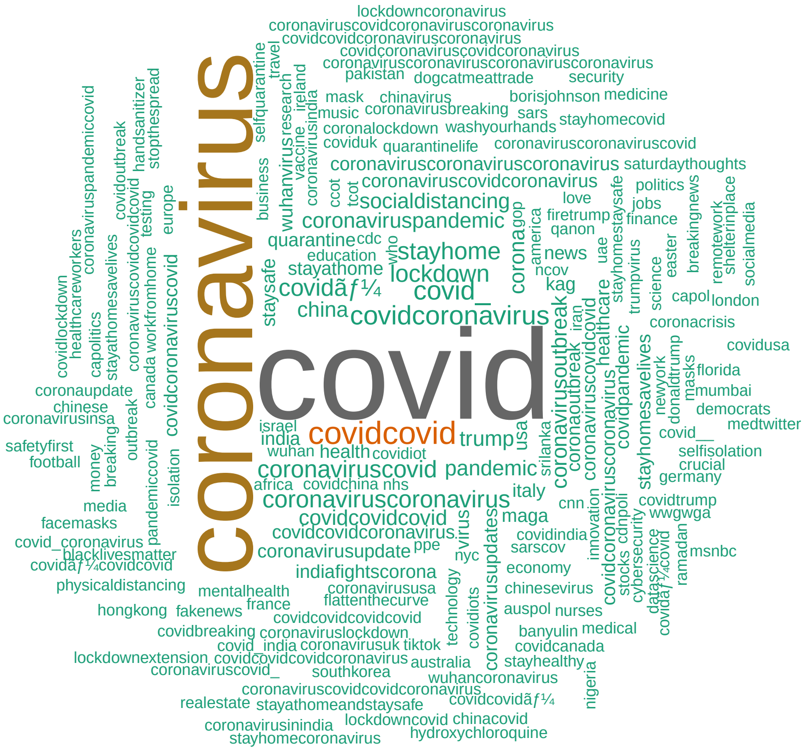
The below Histogram shows the distrubution of tweets over the time period (March-June).



Below is the Word Cloud created for the tweets text. We can see some frequent words in tweets like: Covid, Coronavirus, people, cases, will, lockdown, pandemic, etc.



Below is the Word Cloud created for the hashtags used in the tweets. Most frequent hashtags are covid and coronavirus.



E. Empirical analysis

Naive Bayes Algorithm has been used for tweets classification. It is based on Naive Bayes Rule and assumes the independence of predictor variables within each class. It handles categorical data pretty well and is computaionally efficient.

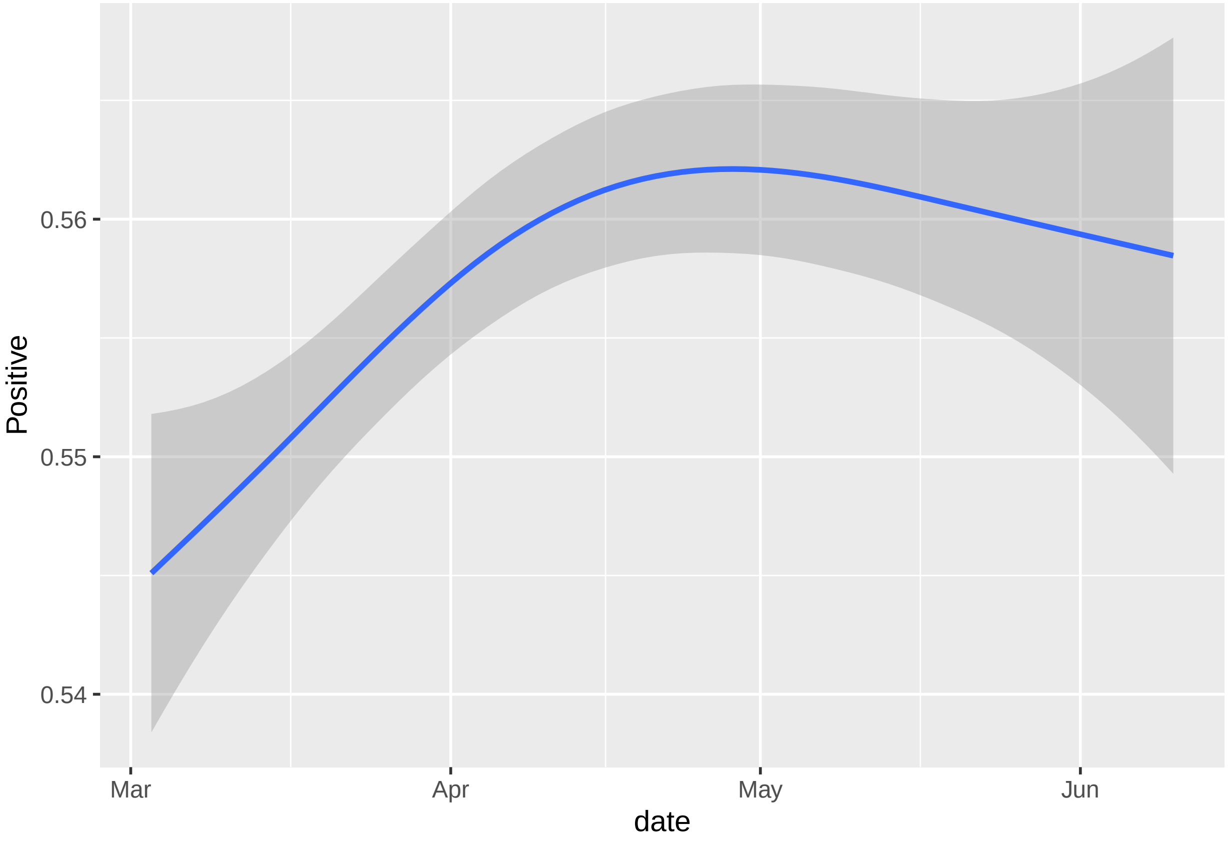
Another algorithm I considered for classification is Support Vector Machines. However, it was not a computationally efficient algorithm. It took 1.39 hours to run it once. I could not the run the tune function so i manually tried a few values for hyperparameter Cost. I got the best algorithm for cost of 2 and linear kernel.

However, I got the best accuracy for Naive Bayes Model which is 72.54%. Please see below the confusion matrix for the model.

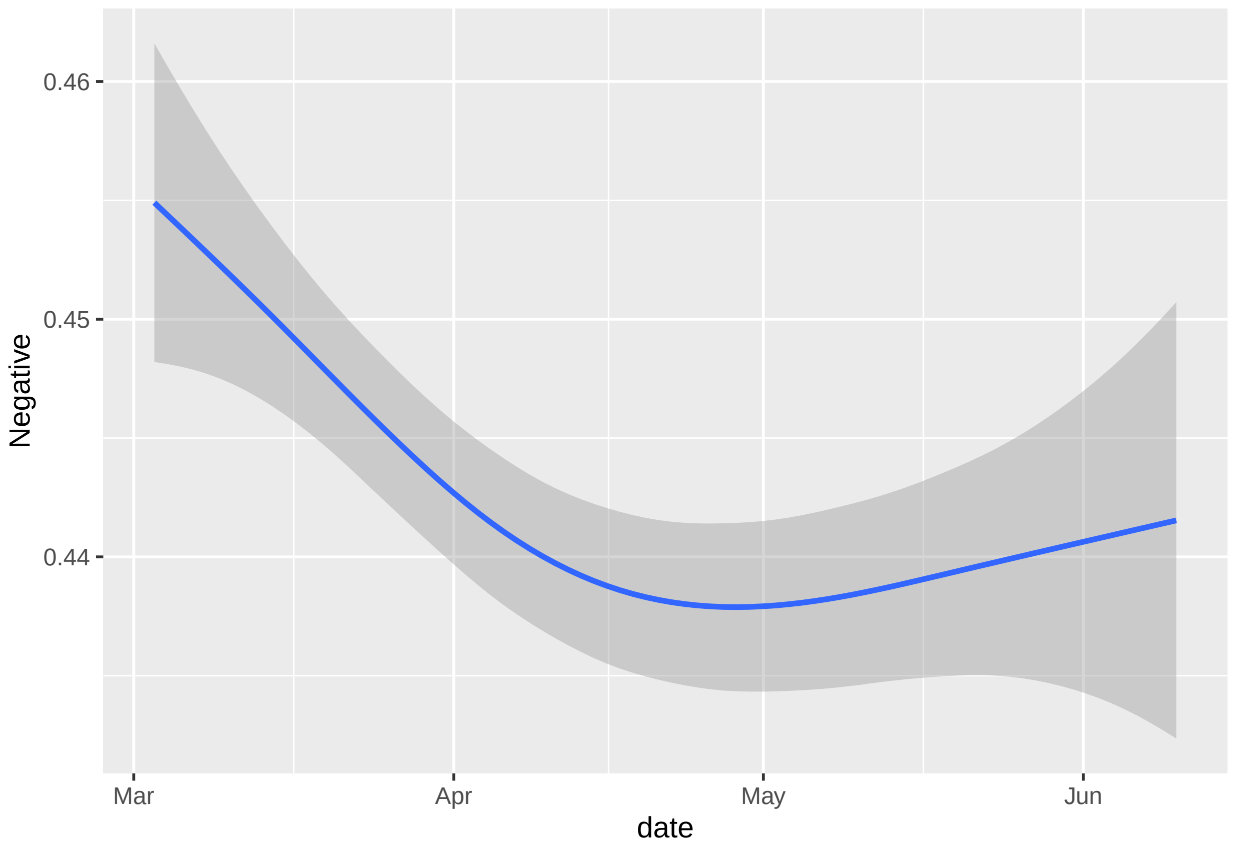
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Negative Positive
##   Negative     5288     2185
##   Positive     3094     8660
##
##                Accuracy : 0.7254
##                  95% CI : (0.7191, 0.7317)
##     No Information Rate : 0.5641
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.4348
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.6309
##             Specificity : 0.7985
##          Pos Pred Value : 0.7076
##          Neg Pred Value : 0.7368
##              Prevalence : 0.4359
##          Detection Rate : 0.2750
##    Detection Prevalence : 0.3887
##       Balanced Accuracy : 0.7147
##
##        'Positive' Class : Negative
##
```

The Below charts show the Distribution of probabilities, predicted by Naive Bayes Model, with Dates. There is an increase in positive probabilities from March till May and it goes down a little in JUne. WE can observe the vice versa in negative probabilities. These charts show us the people have been most negative in March as after that the negativity among people for coronavirus has decreased.

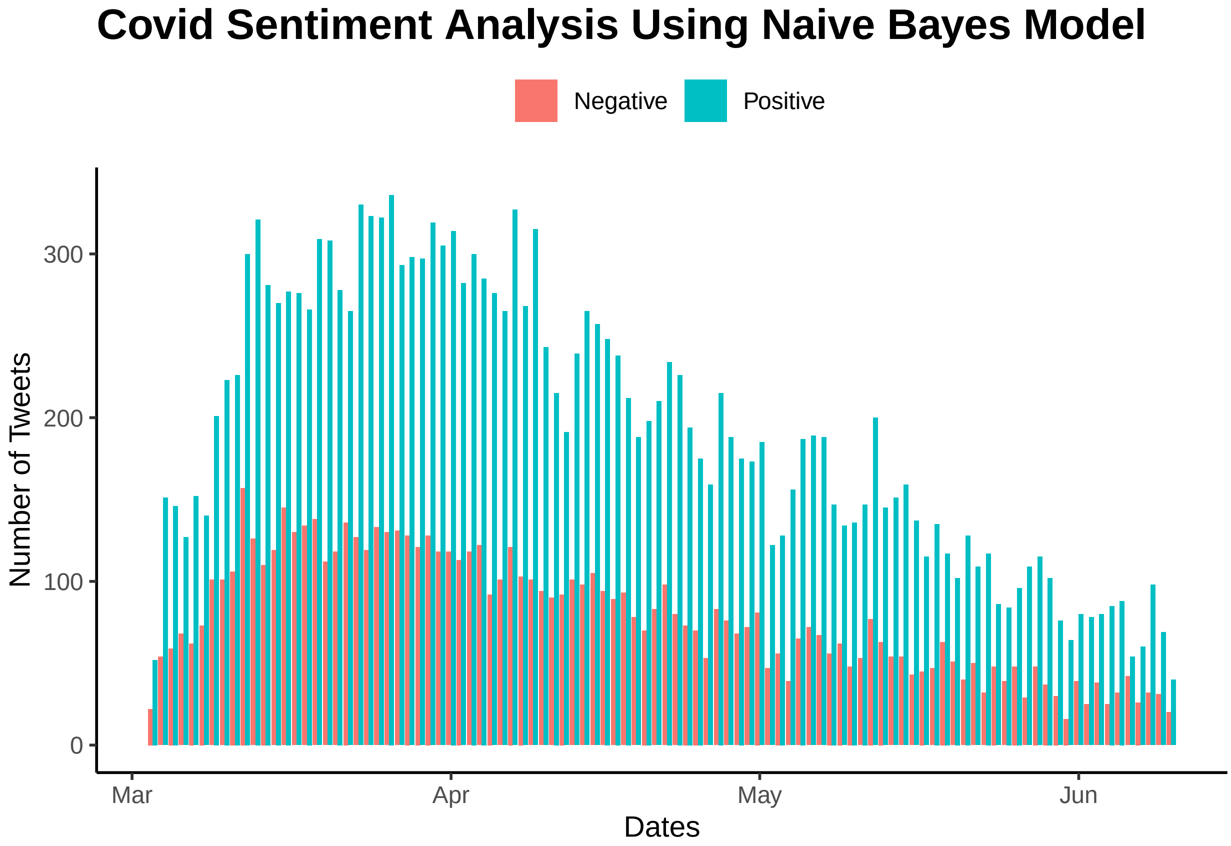
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



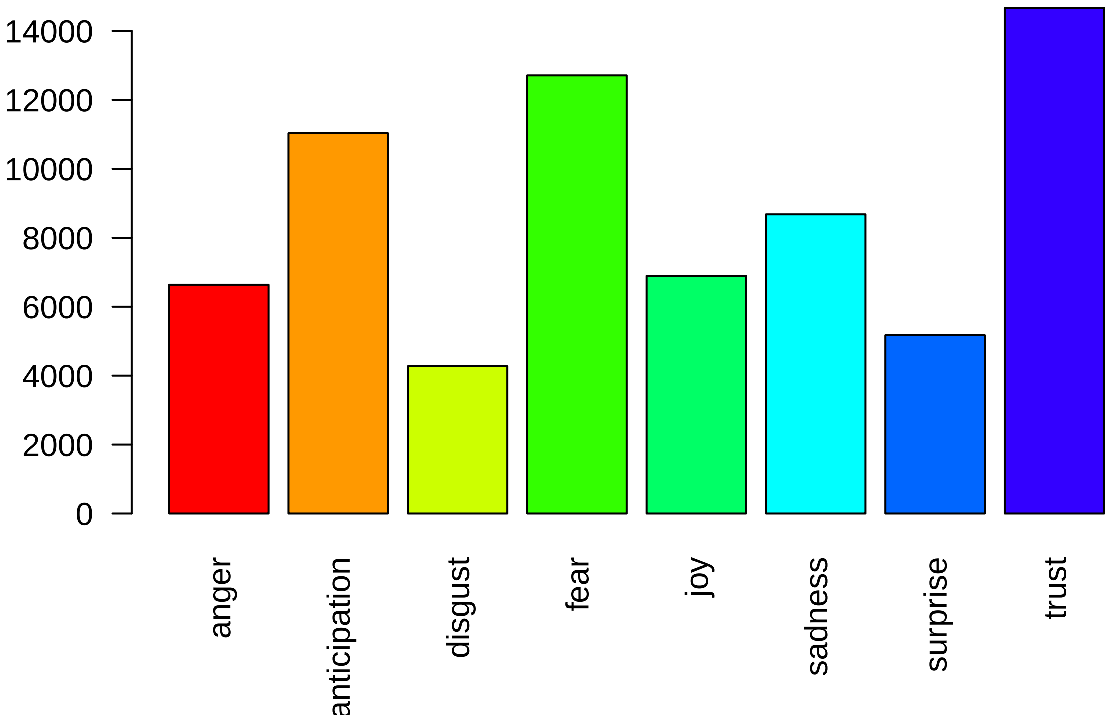
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```


0.46
0.45
0.44
Negative
Mar
Apr
May
Jun
date

Below chart shows the number of classified tweets as Positive and Negative. There is a consistent overall trend in the number of positive tweets and negative tweets throughout the duration.

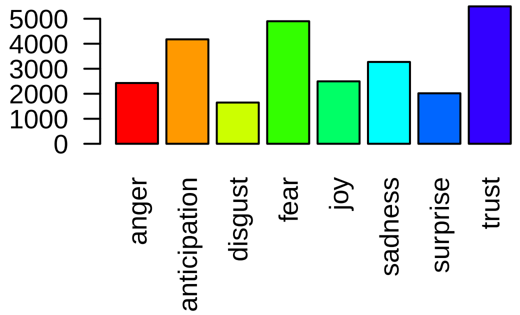


The below chart has been created using the Syuzhet Library. The Lexicon used is NRC. It gives emotions of the text passed through it. The below chart shows the overall emotions for the entire duration (March-June). The most powerful emotions come out to be of Trust, Fear and Anticipation.

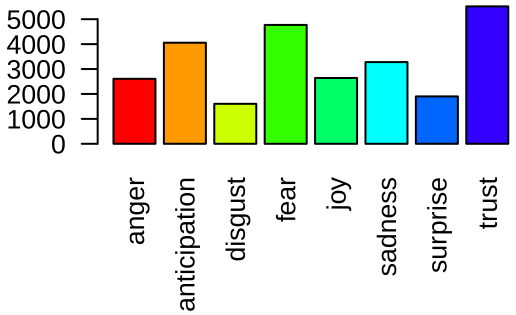


The below charts show the Emotions Separately for each month. This also shows that Trust, Fear and Anticipation are the most dominant emotions in each month.

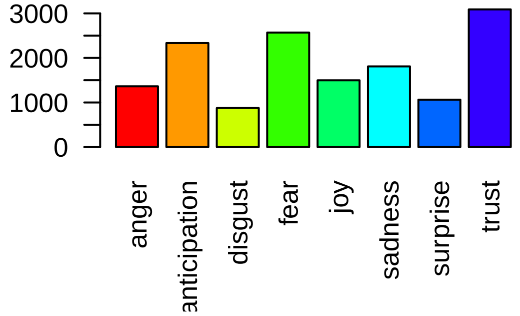
Emotions Chart for March



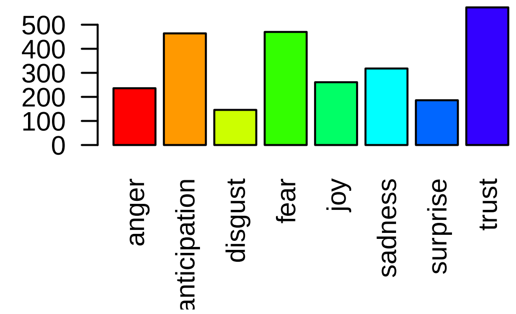
Emotions Chart for April



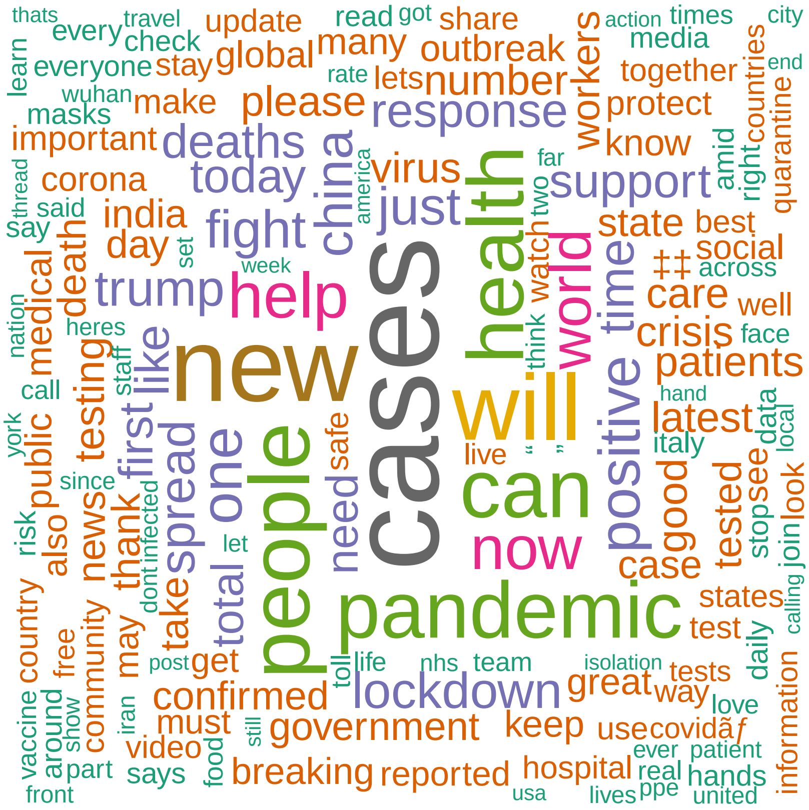
Emotions Chart for May



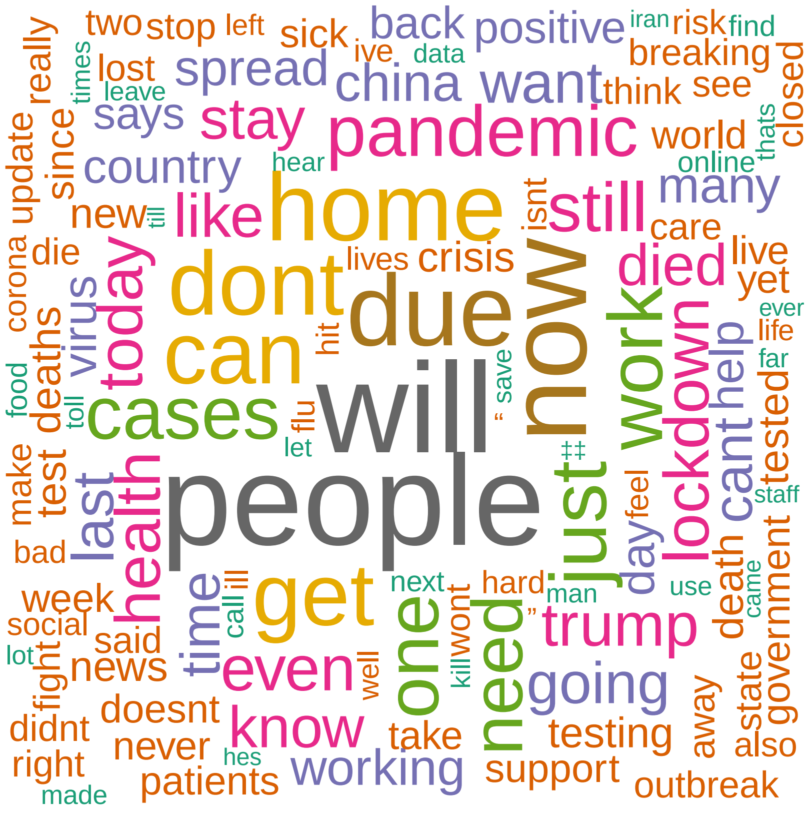
Emotions Chart for June



The Word Cloud created for Tweets classified as Positive by Naive Bayes Model. The word cloud is created by excluding coronavirus and covid words. The tweets were extracted using these words therefore, they will dominate the entire word cloud.



The Word Cloud created for tweets classified as Negative by Naive Bayes Model. The words coronavirus and covid have been excluded from the word cloud.



F. Conclusion

Contrary to the popular belief that covid-19 has stuck fear and sorrow in the life of people, it turns out that people are reciprocating in a very positive fashion, with the mindset of fighting the pandemic, being safe in this time frame, putting health as a priority. Trust Emotion is more dominant among people than Fear. This can be verified by the over all results we have had. This is really a benefiting indicator for many of the businesses around us, both in terms of moral support to sustain the current situation, but also in terms of financial and economic decisions which is largely biased by the mentality and emotions of the people.

References

1. Kerchner, Daniel; Wrubel, Laura, 2020, “Coronavirus Tweet Ids”, https://doi.org/10.7910/DVN/LW0BTB, Harvard Dataverse, V7 Link: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/LW0BTB

2. Kaggle Dataset link: https://www.kaggle.com/imrandude/twitter-sentiment-analysis

3. Online Parser: https://www.convertcsv.com/csv-viewer-editor.htm

4. https://towardsdatascience.com/twitter-sentiment-analysis-on-novel-coronavirus-covid-19-5a9f950312d8

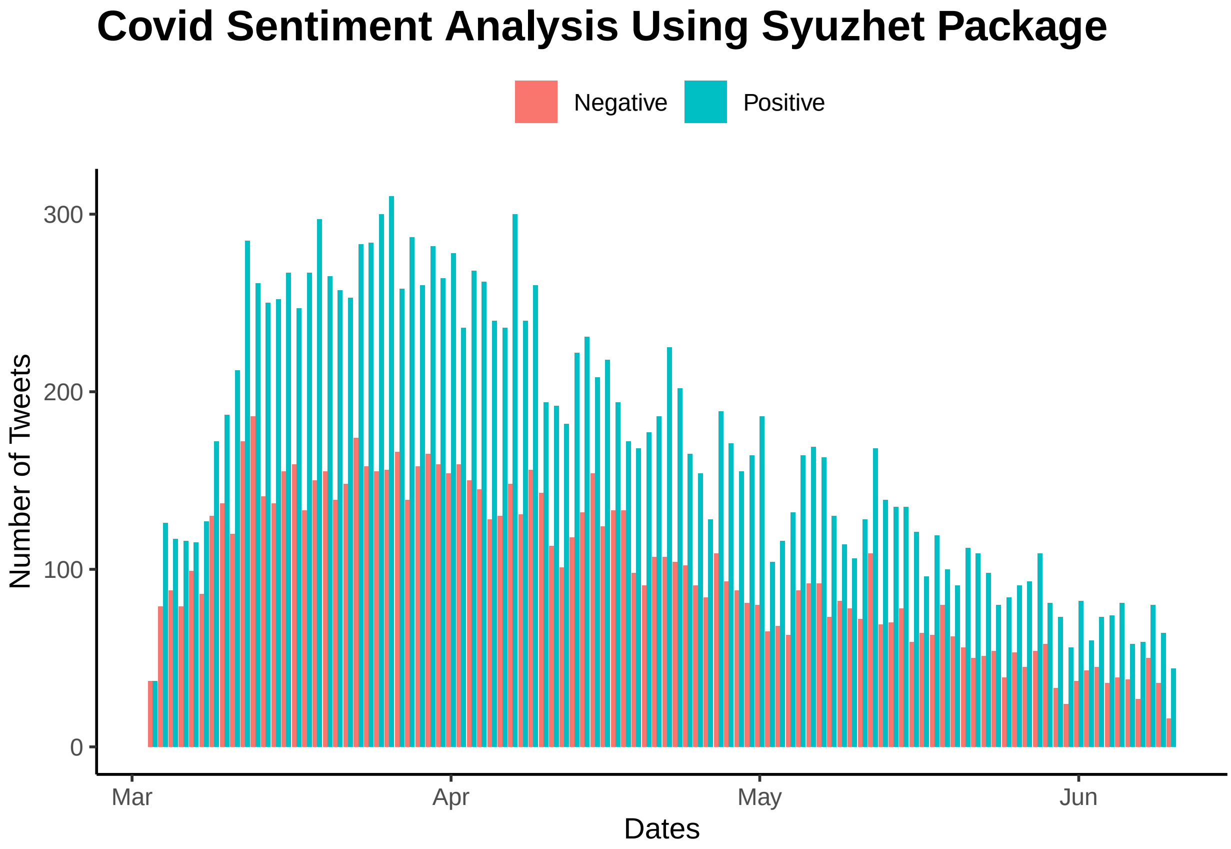
5. https://www.kaggle.com/seunowo/sentiment-analysis-twitter-dataset

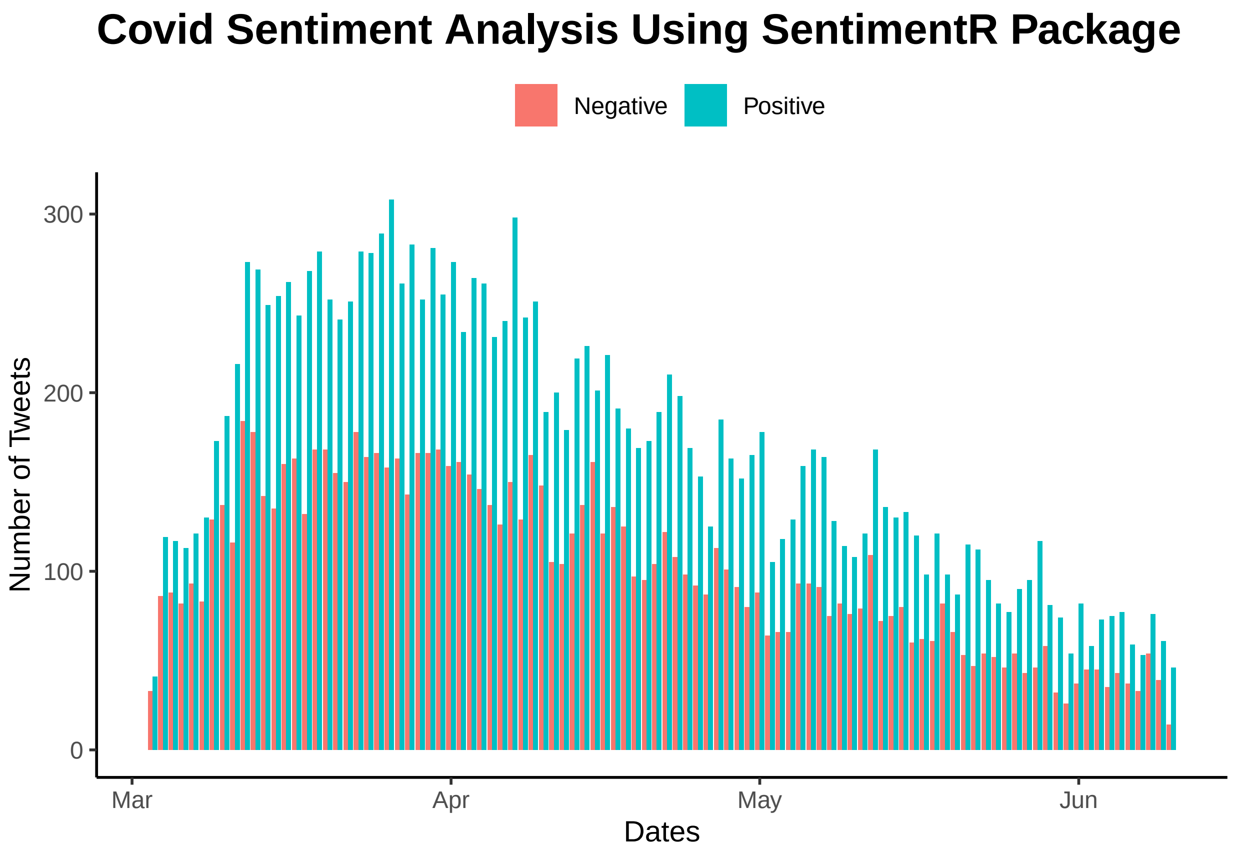
6. https://www.marsja.se/how-to-extract-time-from-datetime-in-r-with-examples/

Appendix

Below are the charts created using Syuzhet and sentimentR Package. Syuzhet Package is a custom sentiment dictionary developed in the Nebraska Literary Lab. It has a lexicon of 10748 words (Positive:3587 & Negative:7161).

SentimentR is a Natural Language Processing based library for text classification.




Covid Sentiment Analysis Using SentimentR Package
🟥 Negative
🟦 Positive
Number of Tweets
300
200
100
0
Mar
Apr
May
Jun
Dates