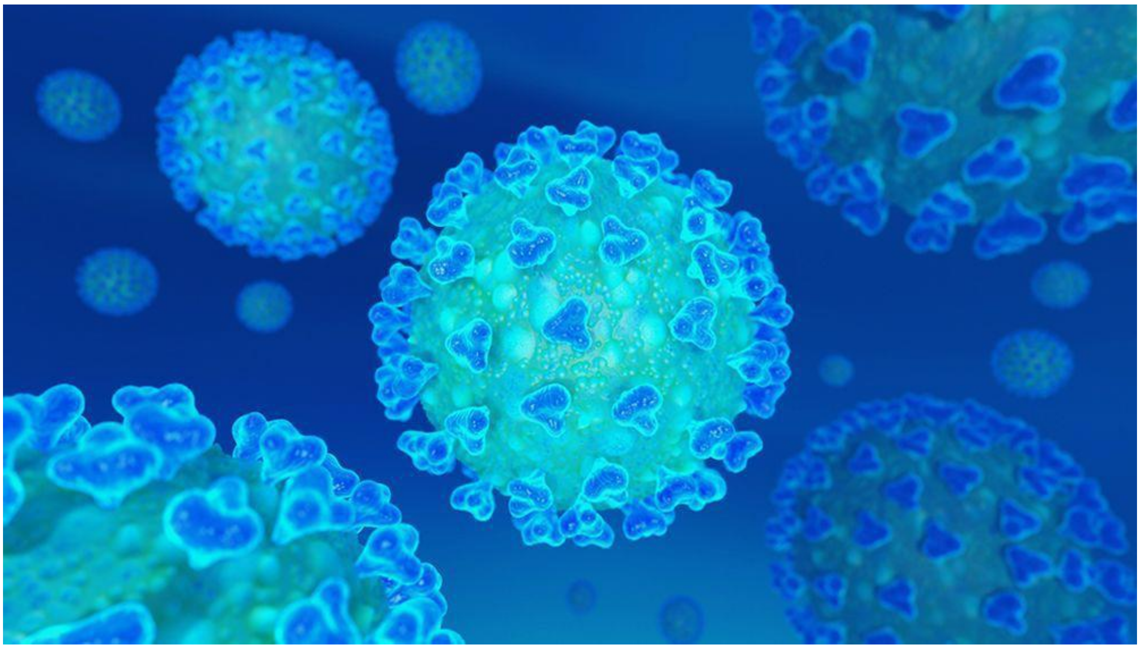
Natural Language Processing: Hands-on

COVID-19 Tweet Analysis & Classification



Lucía Gómez Osuna

Master HMDA (EIT Health) | 30.01.2022

Table of contents

Lucía Gómez Osuna
MSc HMDA (EIT Health)

30.01.2022

1. PROBLEM DESCRIPTION

COVID-19 is an infectious disease, caused by SARS-CoV-2, that has caused over 500,000 million deaths around the world[1] since it emerged in December 2019, as well as lasting health problems in those who have survived the illness. Outbreaks of this disease have been observed worldwide, forcing countries to declare severe regulations such as complete lock-down in order to control its fast rate of spread.

Over the past two years, social media platforms like Twitter have been repeatedly used to spread information, news and users' personal opinions regarding the current pandemic we are living. Even after the development of several vaccines against the disease, COVID-19 keeps being one of the "hottest" and trending topics nowadays. Therefore, this project aims to offer an analysis of the human sensations about this epidemic extracted from the information found in Twitter using natural language processing (NLP) techniques.

2. METHODOLOGY

All experiments in this investigation will be performed using the dataset "Coronavirus Tweets", obtained from Kaggle[2]. This dataset is composed of a collection of tweets from 2020 extracted from Twitter where users talk and express their opinion about COVID-19. These tweets have been manually tagged and labelled according to the mood or "sentiment" they reflect: "Extremely Negative", "Negative", "Neutral", "Positive" and "Extremely Positive".

The dataset file contained two CSV files of COVID-19 tweets, yet this project will only use the "Corona_NLP_train.csv" in order to carry out text classification on the tweet data due to problems in downloading the other file. First, an initial data exploration will be carried out to assess the quality of the data and then any necessary transformations will be performed in order to prepare a final integrated dataset to be used to learn and build a machine learning model based on the decision tree algorithm using the pre-processed data and then its performance in predicting the sentiment of the COVID-19 tweets will be evaluated using a test set.

During the analysis, the most frequent terms will be identified in the tweets of each sentiment and visualized by means of bar graphs and word clouds. The tweet data will be converted into a corpus and then into a term document matrix, where tweets will be separated into individual words through tokenization. This will be finally converted into a data frame, extracting the most frequent words in the tweets, discarding the sparse terms to only contain the most relevant terms for tweet sentiment prediction with the generated decision tree model.

3. ANALYSIS OF RESULTS

3.1 Data exploration

After importing and loading the required libraries and data sets, the first thing to do was to analyse the quality of the data. Preprocessing was performed using RStudio, which revealed the presence of null values in the column of "Location", but since this variable was irrelevant in the analysis, it was discarded, and only the columns of "OriginalTweet" and "Sentiment" were kept for this investigation.

Next, a plot was constructed to identify the number of tweets for each "Sentiment" class, and since less tweets were classified as the more extreme sentiments, the "Extremely Negative" and

[1] World Health Organization (WHO). 2022. *Coronavirus disease (COVID-19)*. [online] Available at: <https://www.who.int/news-room/q-a-detail/coronavirus-disease-covid-19> [Accessed 12 January 2022].

[2] Kaggle.com. 2022. *Coronavirus tweets NLP - Text Classification*. [online] Available at: <https://www.kaggle.com/datatattle/covid-19-nlp-text-classification?select=Corona_NLP_train.csv> [Accessed 12 January 2022].

“Negative” tweets were grouped together and classified as “Negative” as a whole for ease, and the same was applied for “Extremely Positive” and “Positive” tweets.

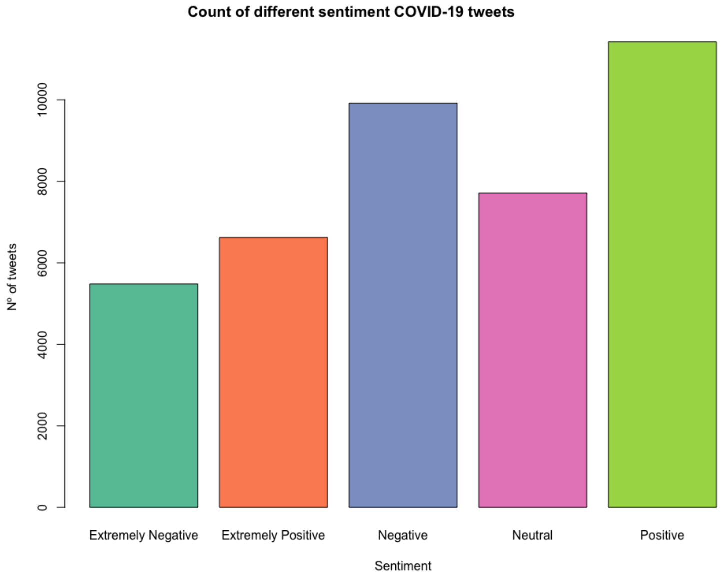


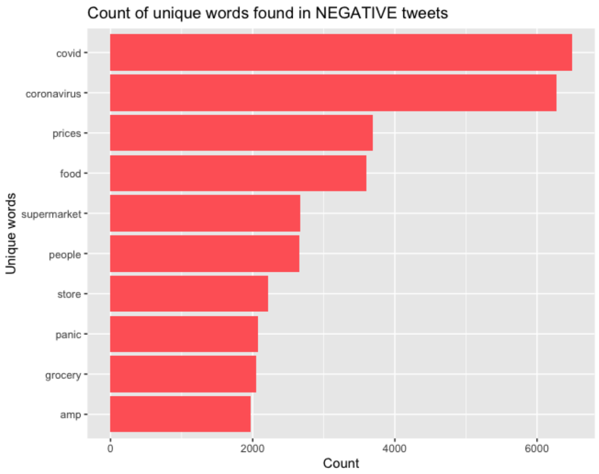
Figure 1 – *Bar plots of frequency of tweets for each Sentiment.*

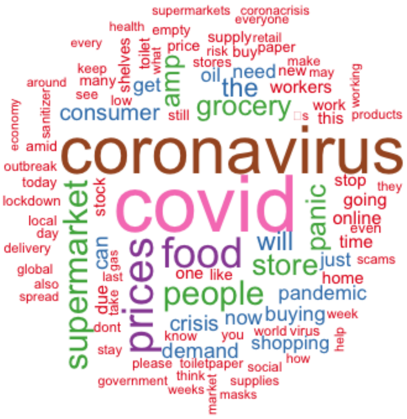
3.2 Data preparation

Since anyone can tweet online, tweets often contain a lot of gibberish and irrelevant terms. Therefore, URLs, digits, hashtags and mentions were removed during tweet preprocessing and a new column called “CleanTweet” was created for the tweet data. Similarly, stop words from the *tidytext* package were removed, and then tokenization was performed to separate tweets into individual words for further analysis.

3.3 Most frequent terms

Bar charts and word clouds were created to represent the most frequent words in the tweets for each of the sentiments, as we can see below. In the word clouds, the larger the size of the word, the higher its frequency in the corresponding sentiment tweets.





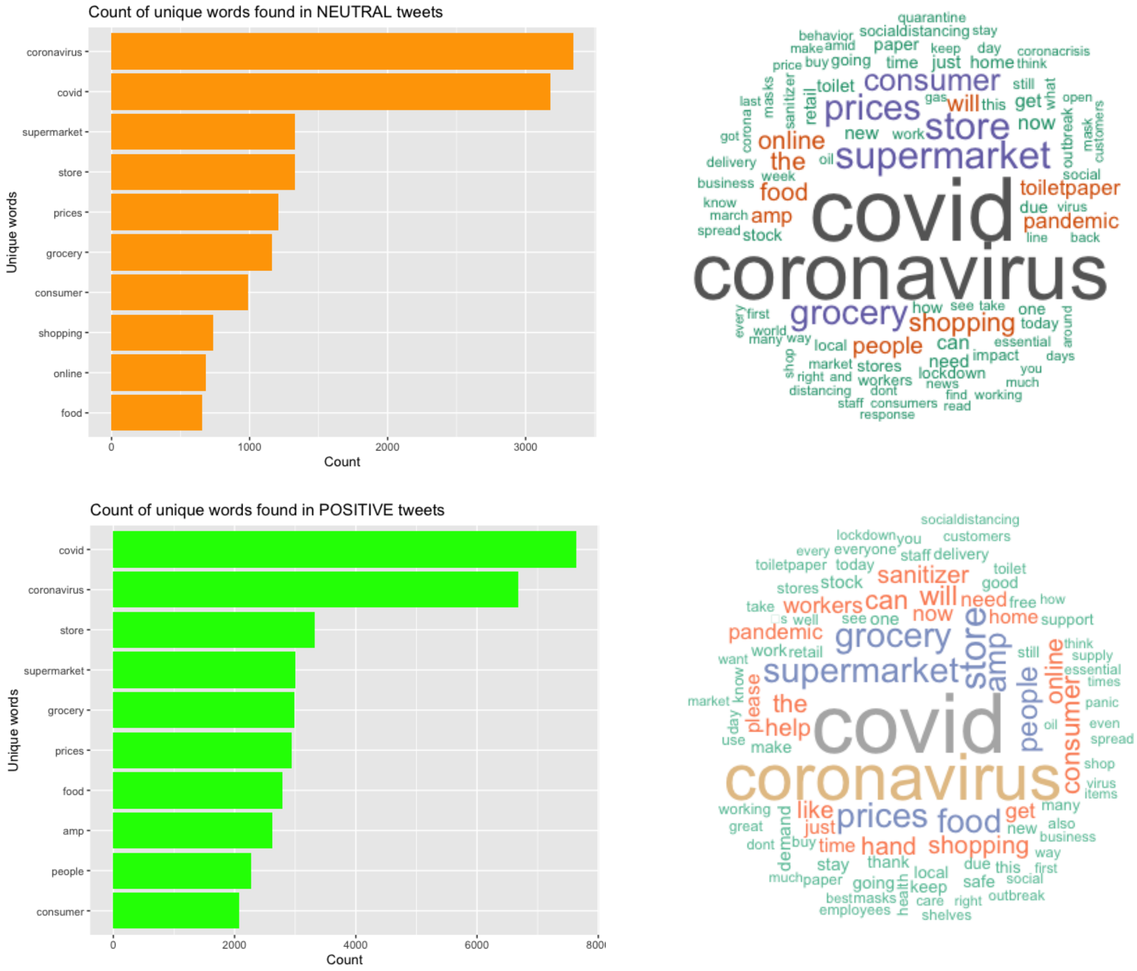


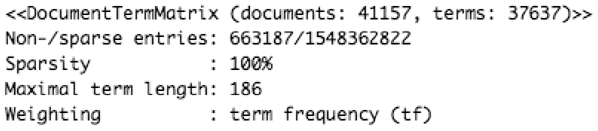
Figure 2 – Bar plots of frequency of unique words in Negative, Neutral and Positive tweets.

Overall, the charts and word clouds revealed that the three main classes of Sentiment contained very similar words in their tweets, as besides the reasonable top words of “coronavirus” and “covid”, the most common terms in all cases included “store”, “supermarket”, “prices” and “food”, referring to the surge in demand and supermarkets’ shortages that characterized the beginning of the pandemic worldwide with the coronavirus outbreak.

Regarding word clouds, we can see some differences with the appearance of terms like “panic” and “crisis” in the Negative tweets, while Positive ones were associated to terms such as “care”, “safe” and “home”. On the other hand, words in Neutral tweets were less differentiable, as these can be of any topic and address more general concepts, so no characteristic terms were found for this class.

3.4 Term document matrix

A corpus, which is a collection of text documents, was then created from the CleanTweets of the data set, and this was used to generate the term document matrix based on term frequency weighting. The resulting matrix displayed many zeroes, explaining its high sparsity, hence sparse terms were consequently removed.

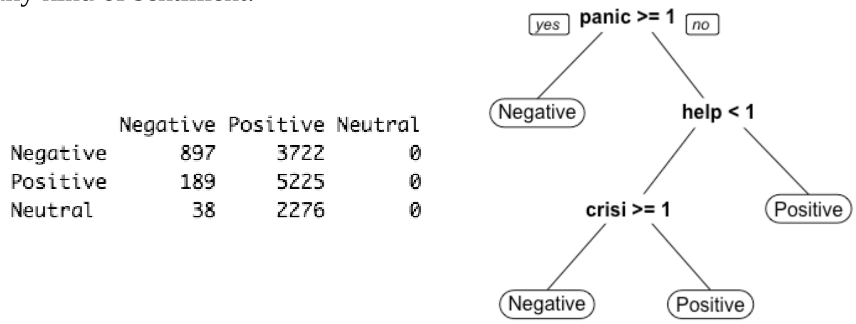


```
<<DocumentTermMatrix (documents: 41157, terms: 37637)>>
Non-/sparse entries: 663187/1548362822
Sparsity           : 100%
Maximal term length: 186
Weighting          : term frequency (tf)
```

Figure 3 – Summary of term document matrix for COVID-19 tweets.

3.5 Modelling and evaluation

Finally, the decision tree model generated selected the word “panic” as its root node, implying that this term obtained the highest gain ratio, since as the analysis above revealed, “panic” was the main word differentiating Negative tweets from others and thus had a high relevance in the classification of tweets. The model also suggested that if the tweet contained the word “help”, it would more be likely to reflect a Positive sentiment, while if it contained “panic” or “crisis”, it would be classified as a Negative tweet, which supports what we found in our analysis above. On the other hand, the decision tree was not able to distinguish Neutral tweets, which could be due to the fact that this class of tweets tended to be composed of more general terms, and hence could represent any kind of sentiment.



```
          Negative Positive Neutral
Negative       897     3722       0
Positive       189     5225       0
Neutral         38     2276       0
```

Figure 4 – *Resulting decision tree model and corresponding confusion matrix for tweet classification.*

However, the model achieved a very low accuracy of 37.41%, as it missed Neutral tweets, yet this could be due to the high similarity in words across tweets of different sentiments observed in the word clouds and bar charts generated. Future work could focus in refining the sentiment analysis to identify more distinguishable and unique words for each sentiment class to aid in prediction, and try data augmentation, gathering more tweets about COVID-19 to increase our collection of associated words for our model to identify clear relationships.

4. CONCLUSION

Overall, this investigation highlights the potential applications of NLP techniques in situations where language plays a main role, allowing to associate terms and words to particular behaviours and sentiments. Despite its low accuracy, the results of the model obtained are in accordance with the findings of our preliminary text analysis of tweets, yet further refining is required to make correct predictions on the sentiment portrayed by users tweeting about COVID-19.

The coronavirus has appeared in our lives to stay, thus it is crucial to understand how it affects society and influences behaviour, such as mental health, which can be extracted from individuals’ freedom of expression in social networks, in order to manage and derive solutions for the current pandemic we are living.

The code in R for this project as well as a README.md file describing how to run it are available in GitHub at: https://github.com/LuciaGomez88/NLP-COVID-19-Tweets-Analysis

5. REFERENCES

- Kaggle.com. 2022. *Coronavirus tweets NLP - Text Classification*. [online] Available at: <https://www.kaggle.com/datatattle/covid-19-nlp-text-classification?select=Corona_NLP_train.csv> [Accessed 12 January 2022].
- World Health Organization (WHO). 2022. *Coronavirus disease (COVID-19)*. [online] Available at: <https://www.who.int/news-room/q-a-detail/coronavirus-disease-covid-19> [Accessed 12 January 2022].