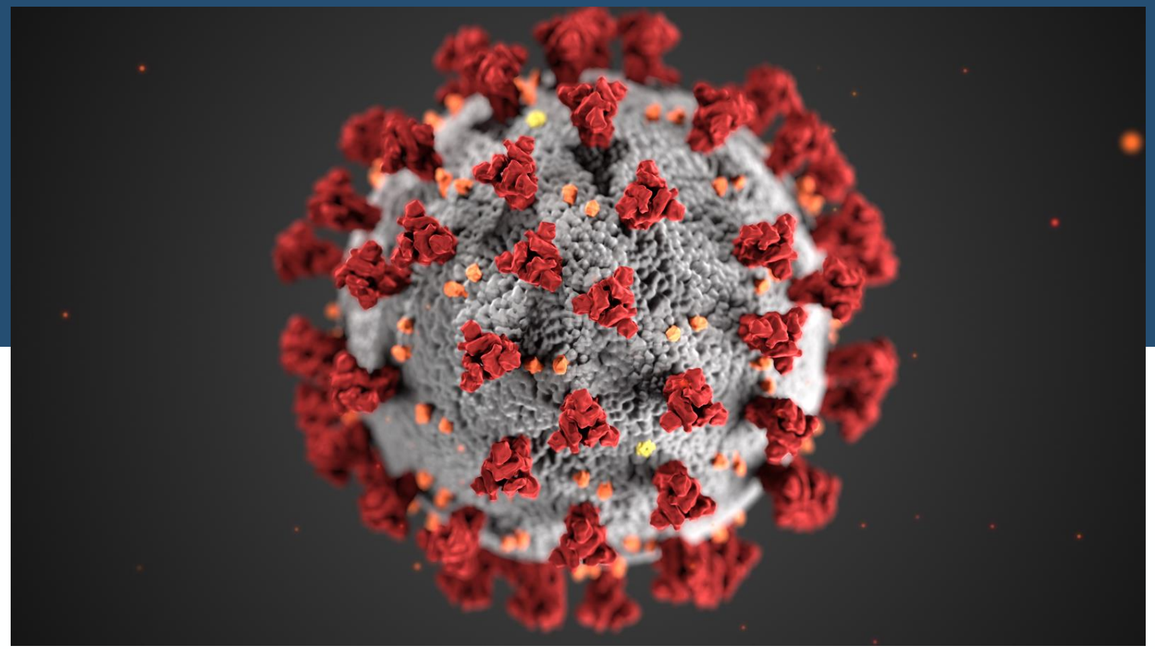
DSI-23 Capstone Project:

Analysis on COVID-19 Pfizer/BioNtech Vaccine Tweets

Jeffrey Prasetio

Contents

Background



COVID-19 is the disease caused by a new coronavirus called SARS-CoV-2.

First identified in Wuhan, China, in December 2019.

The World Health Organization (WHO) declared a pandemic on 11 March 2020.

As of 31 August 2021, more than 217 million cases and 4.51 million deaths have been confirmed.

The Pfizer/BioNtech vaccine was the first to be listed for WHO Emergency Use Listing (EUL) on 31 December 2020.

Goal

The goal is to identify potential interventions for Pfizer/BioNtech to take, in order to speed up the global vaccination progress (and increase vaccine sales)

1. Sentiment analysis on the Pfizer/BioNtech vaccine tweets, to understand the overall global sentiment.

2. Topic modelling within each sentiment, to further understand the possible reasons behind the sentiment.

The Data

Feature	Type	Description
user_name	*str*	User name of the Twitter account.
user_location	*str*	User location of the Twitter account.
user_description	*str*	User description of the Twitter account.
user_created	*date*	Date when the Twitter account was created.
user_followers	*int*	Number of the followers of the Twitter account.
user_verified	*bool*	Verification status of the Twitter account (True/False).
date	*date*	Date when the tweet was posted.
text	*str*	Content of the tweet.
hashtags	*list*	Hashtags used in the tweet.
source	*str*	Source / device where the tweet came from.
retweets	*int*	Number of retweets of the tweet.
favorites	*int*	Number of favorties (equivalent to likes) of the tweet.
is_retweet	*bool*	Whether the tweet is a retweet (True/False)
sentiment	*str*	The sentiment category of the tweet (Positive/Neutral/Negative)

17172 tweets about the Pfizer/BioNtech vaccine.

12/12/2020 – 22/08/2021.

Sentiment Analysis – VADER

What is VADER?

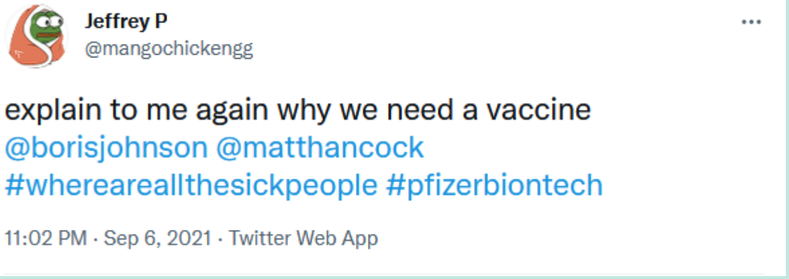
A lexicon and rule-based sentiment analysis tool, specifically attuned to sentiments expressed in social media.

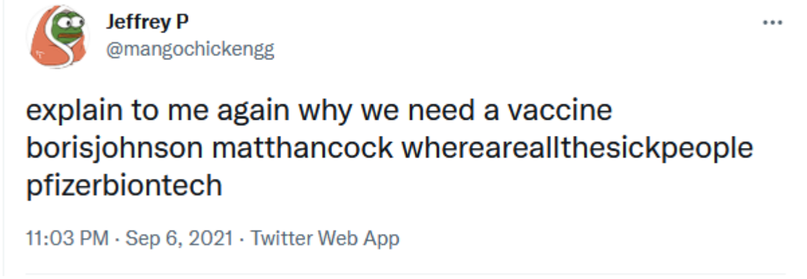
Text Preprocessing?

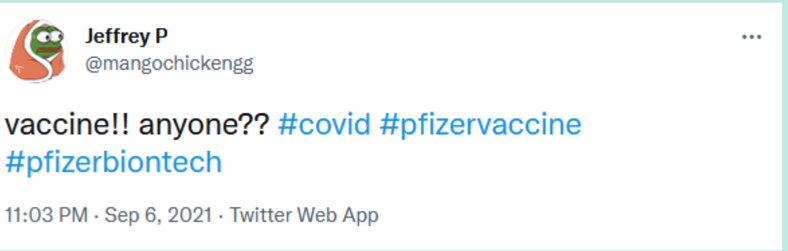
Very minimal, VADER is smart enough to understand the valence of non-conventional text: including emojis 🙃, CAPITALIZATION, and extended punctuation!!!!!!!

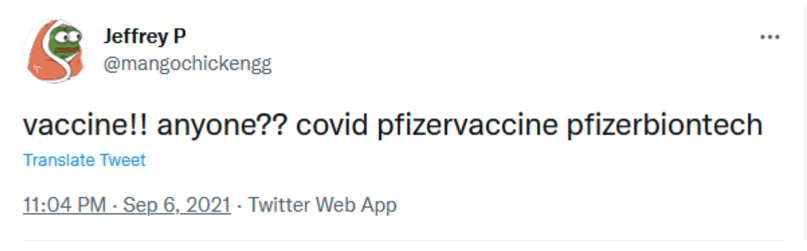
Sentiment Scoring?

- Positive: compound score >= 0.05
- Neutral: -0.05 < compound score < 0.05
- Negative: compound score <= -0.05


Jeffrey P
@mangochickengg
explain to me again why we need a vaccine @borisjohnson @matthancock #whereareallthesickpeople #pfizerbiontech
11:02 PM · Sep 6, 2021 · Twitter Web App


Jeffrey P
@mangochickengg
explain to me again why we need a vaccine borisjohnson matthancock whereareallthesickpeople pfizerbiontech
11:03 PM · Sep 6, 2021 · Twitter Web App


Jeffrey P
@mangochickengg
vaccine!! anyone?? #covid #pfizervaccine #pfizerbiontech
11:03 PM · Sep 6, 2021 · Twitter Web App

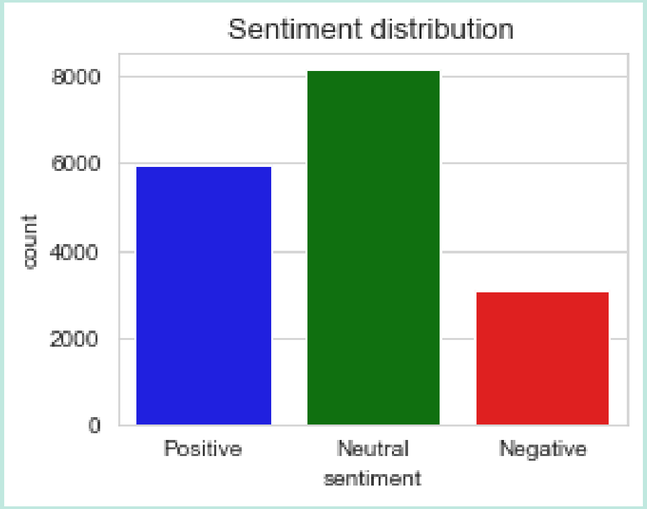

Jeffrey P
@mangochickengg
vaccine!! anyone?? covid pfizervaccine pfizerbiontech
Translate Tweet
11:04 PM · Sep 6, 2021 · Twitter Web App

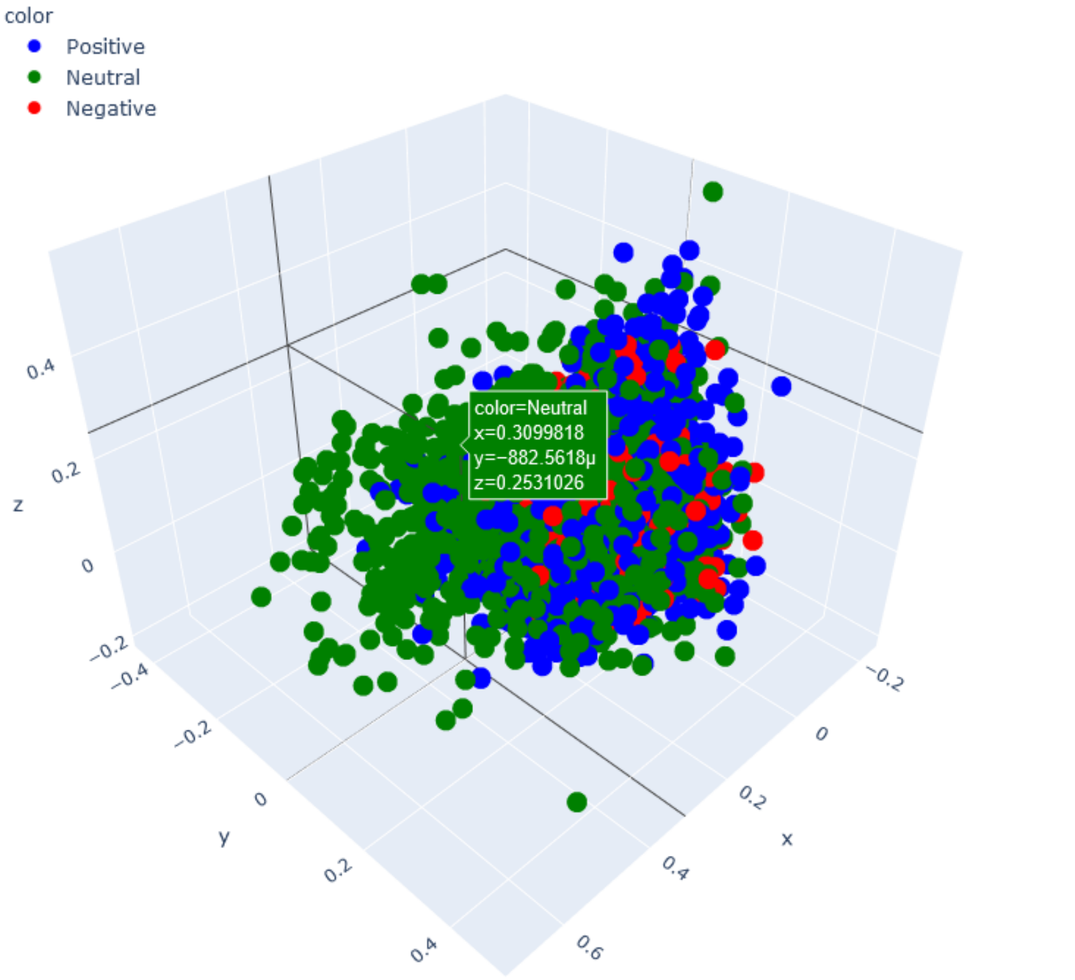
Sentiment Clusters

The clusters (generated with a PCA scatterplot) do not seem very distinct from one another.

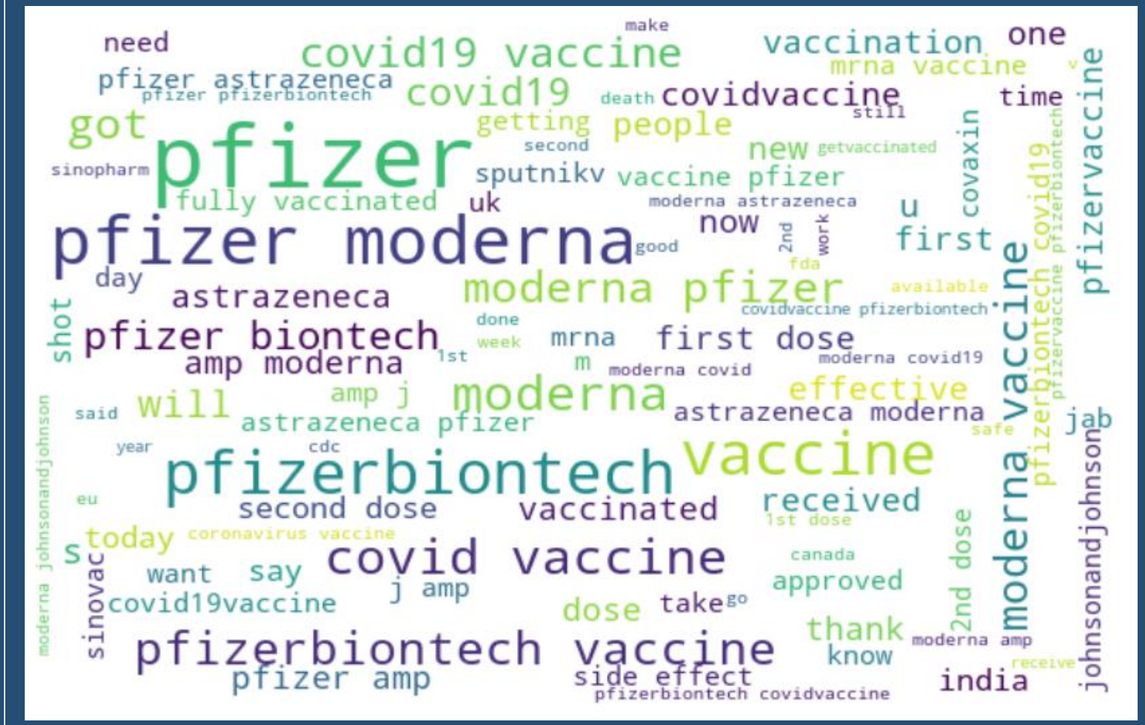
This is rather concerning, but also understandable because the distinctions between sentiments are typically not straightforward.

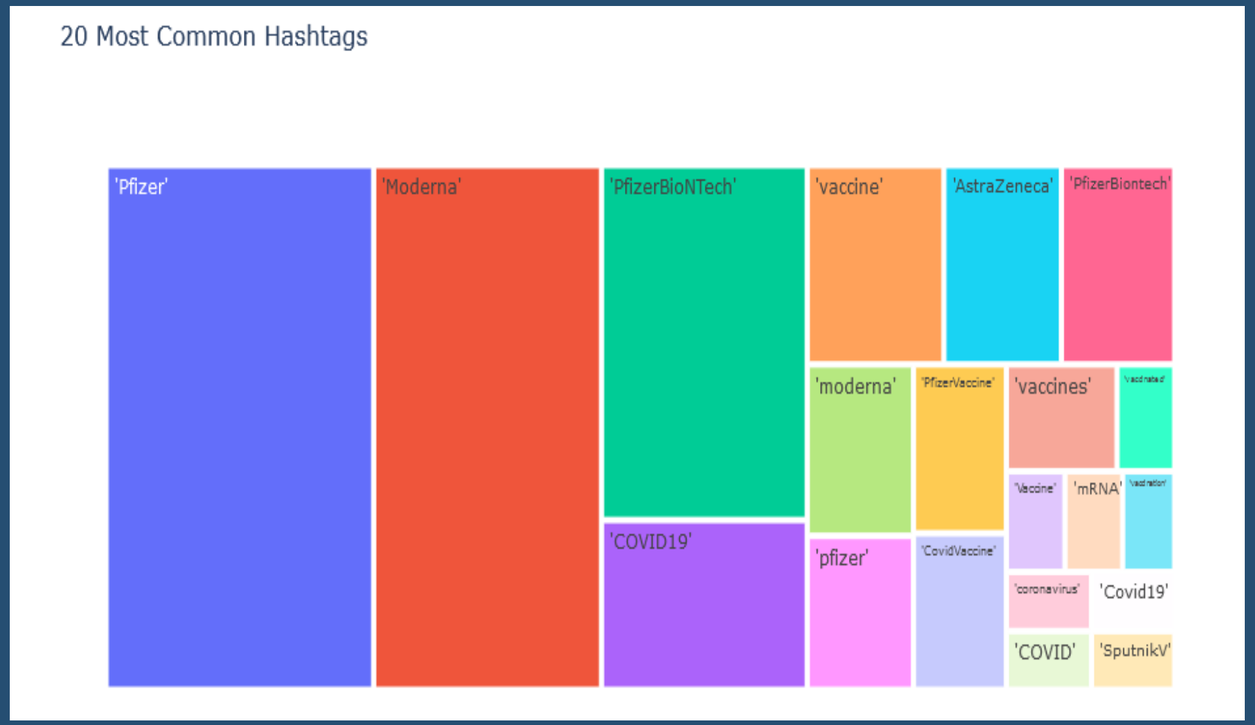
Sentiment classification is a highly subjective task.



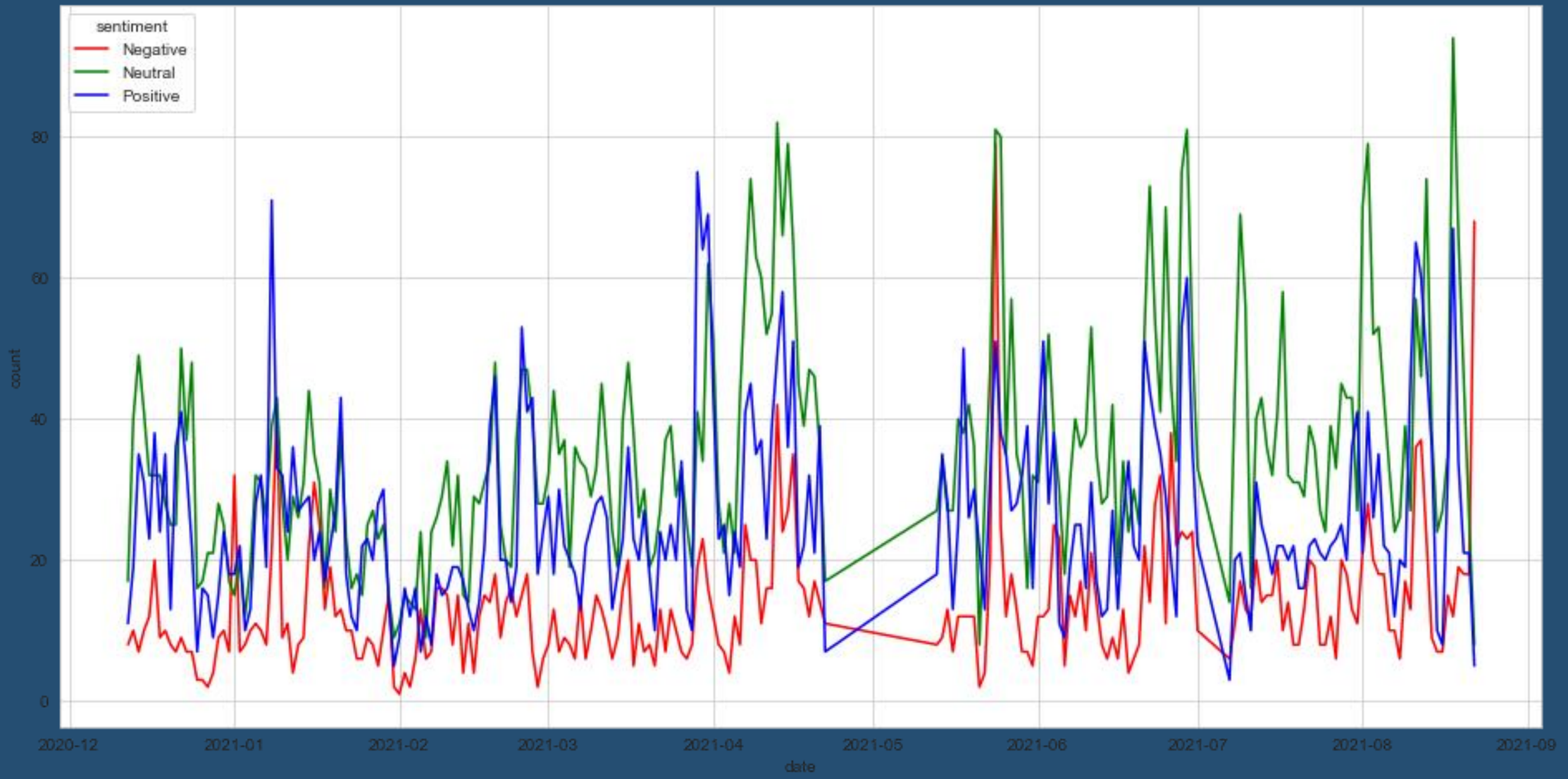


Insights – Word Cloud

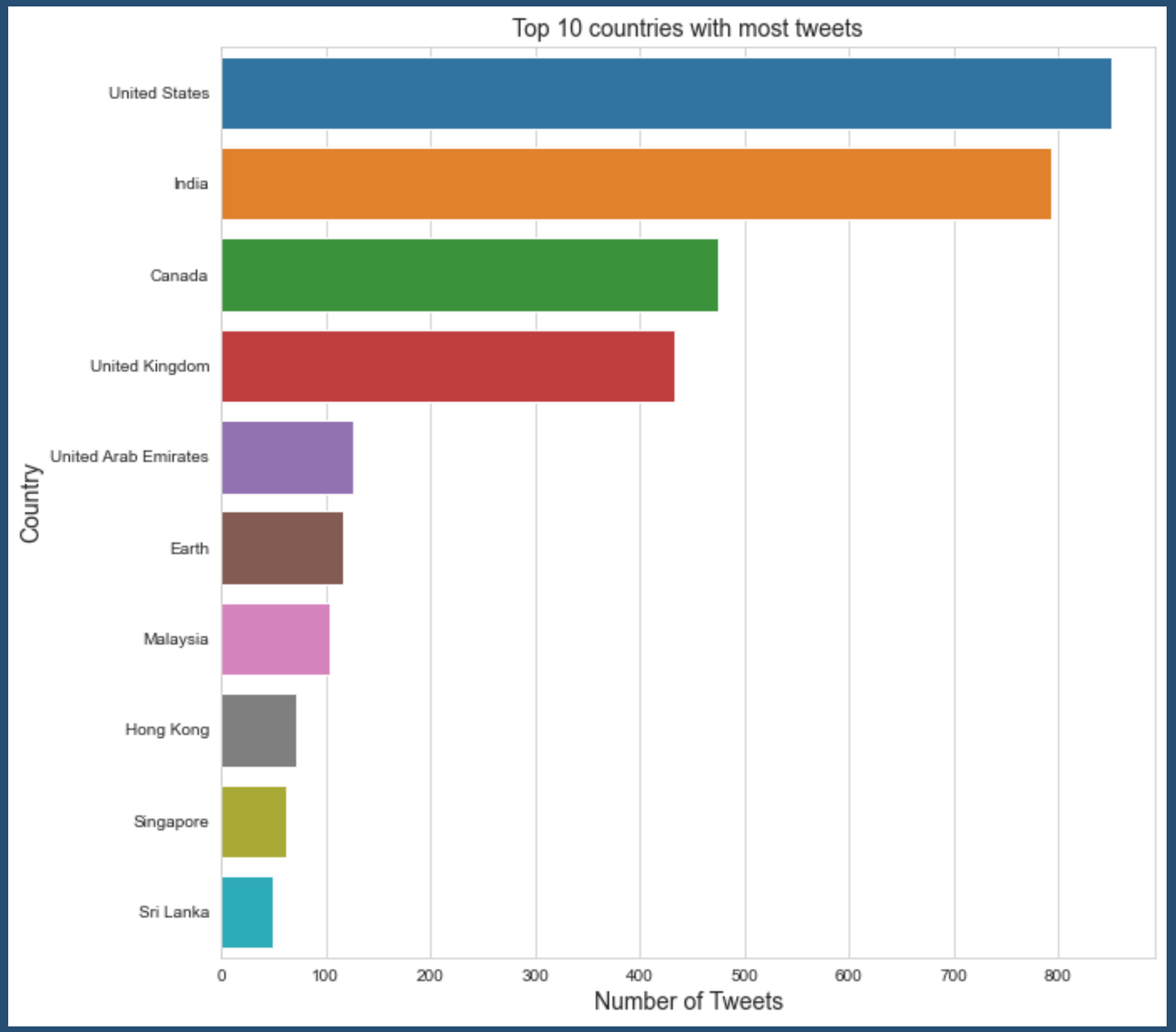




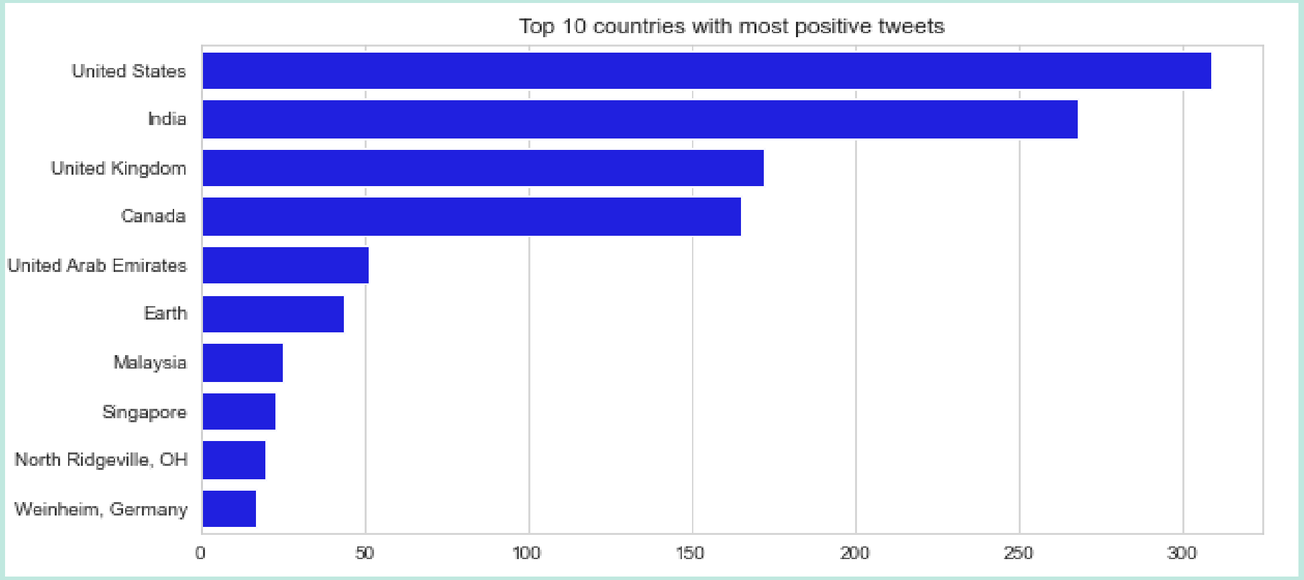
Insights – Time Trends

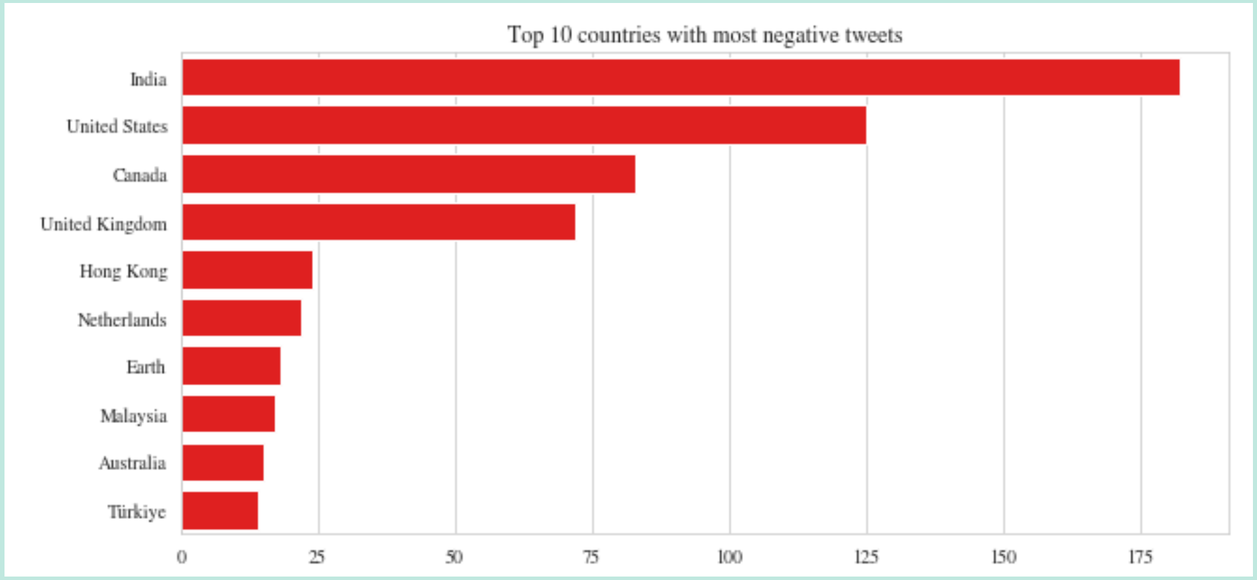


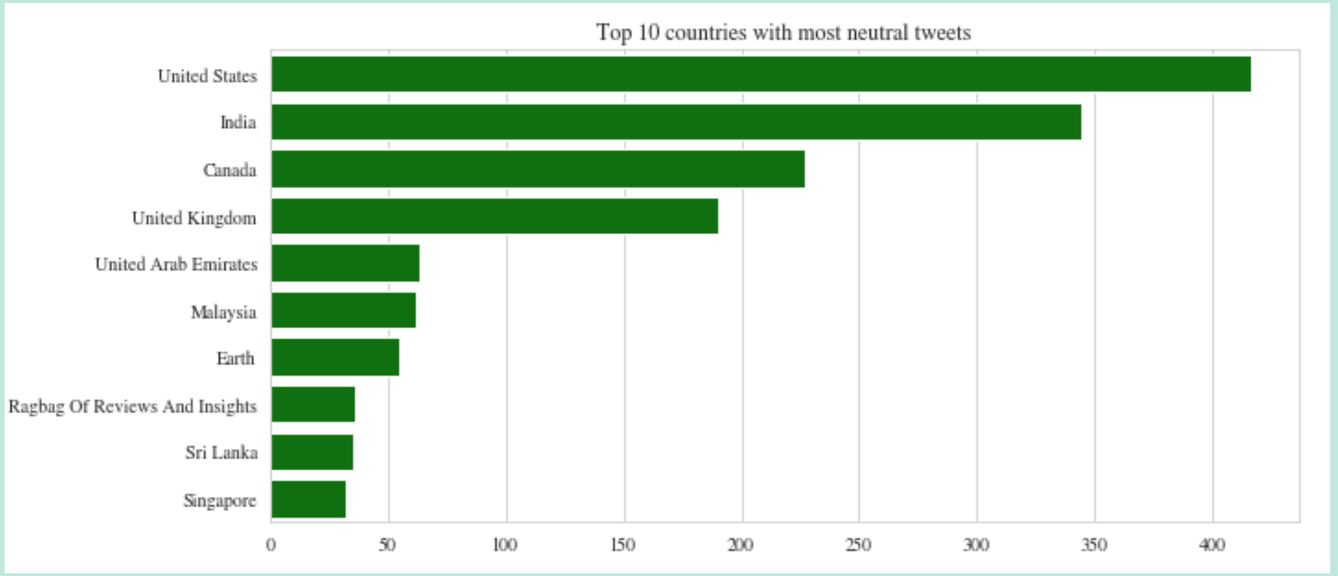
Insights – Top 10 Countries



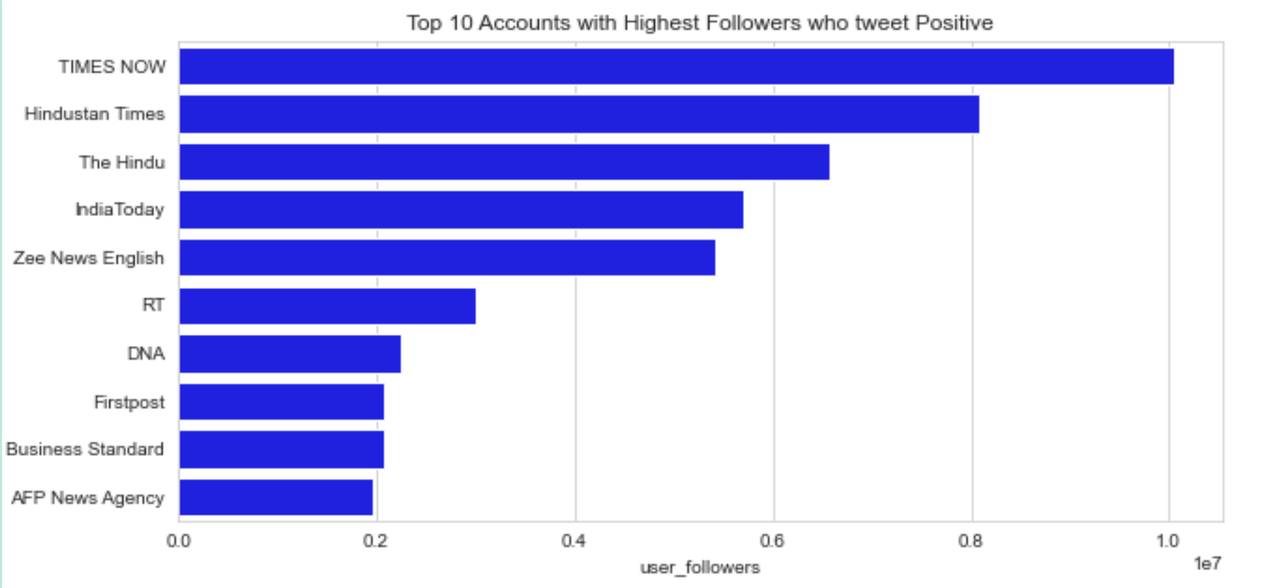
Insights – Top 10 Countries

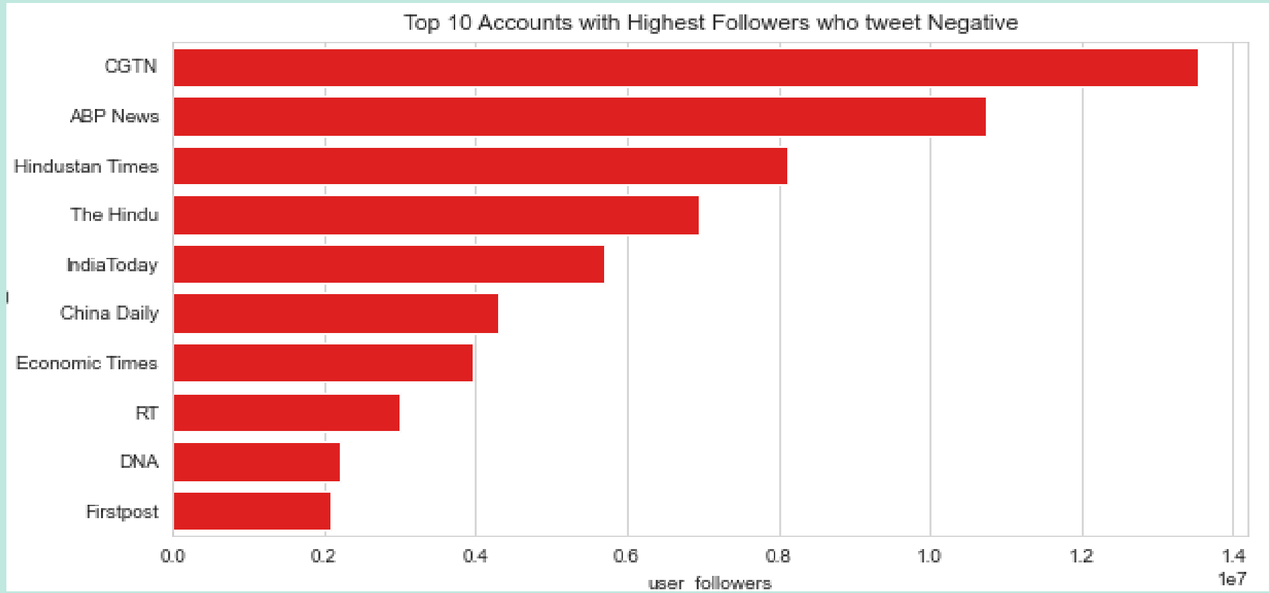


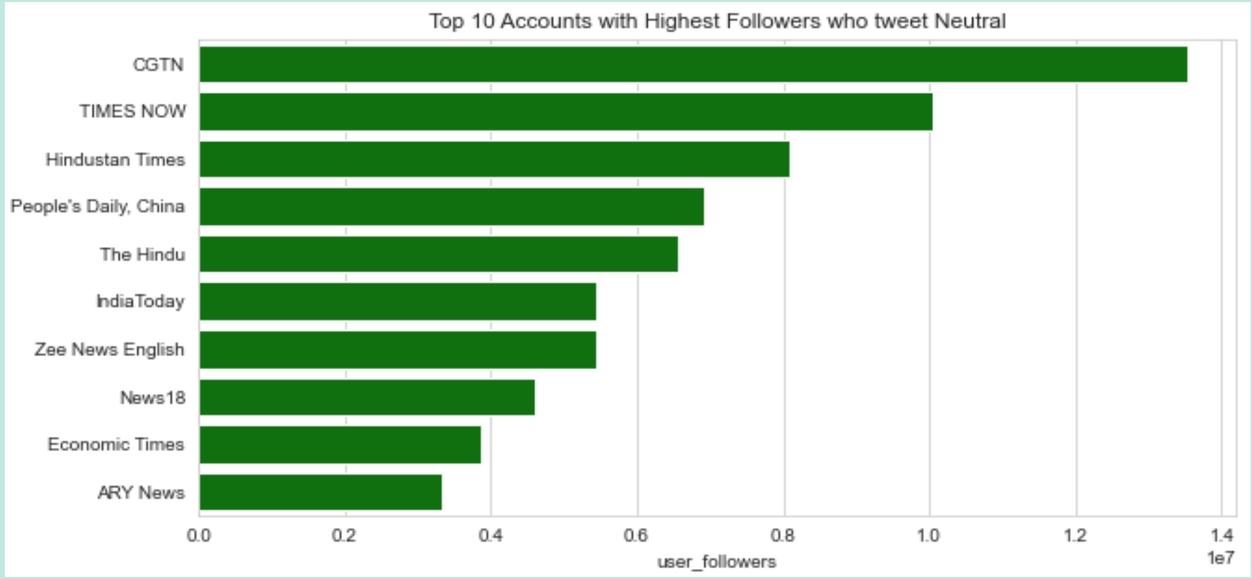




Insights – Accounts w/ most followers







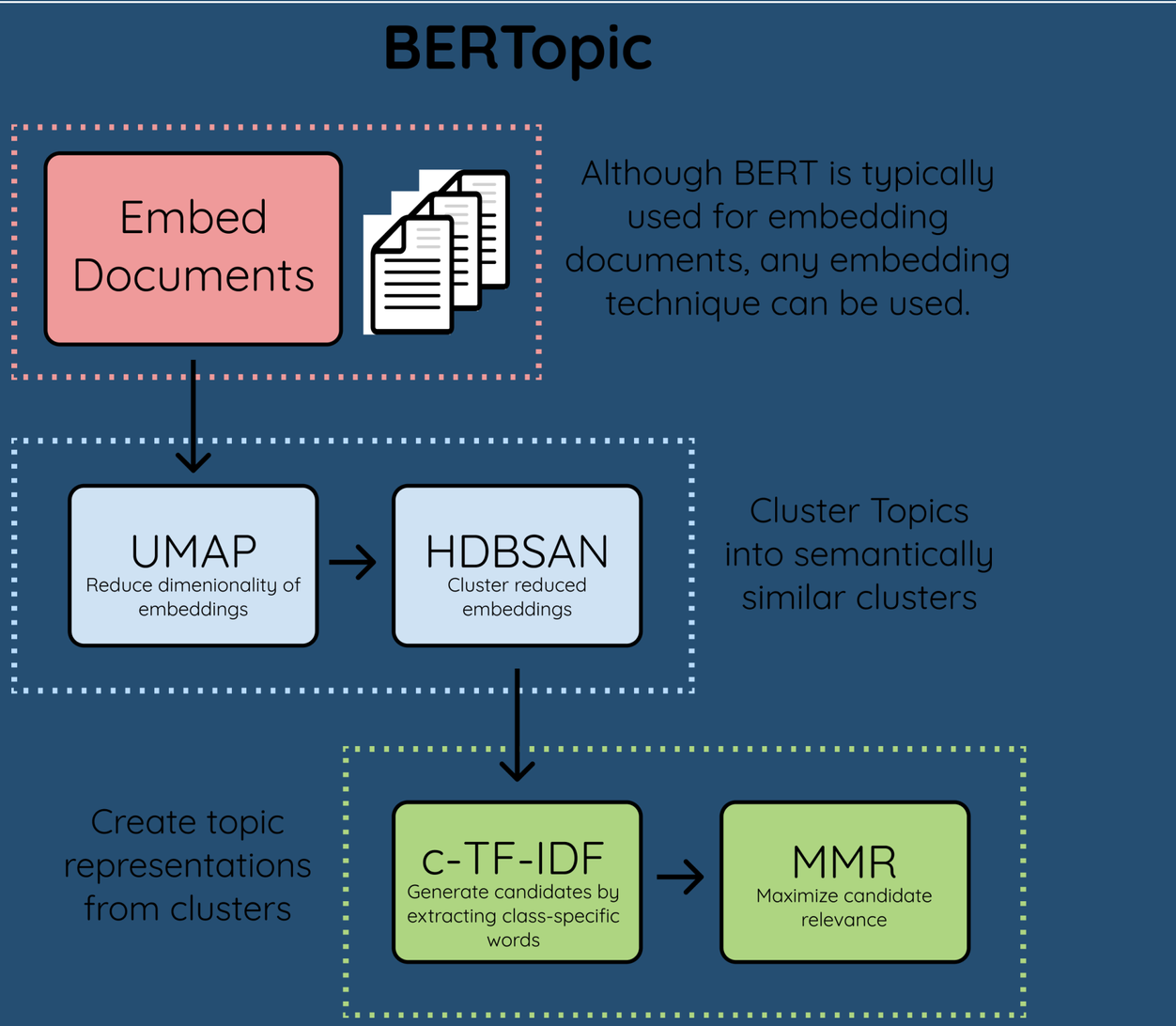
Topic Modelling

What is BERTopic?

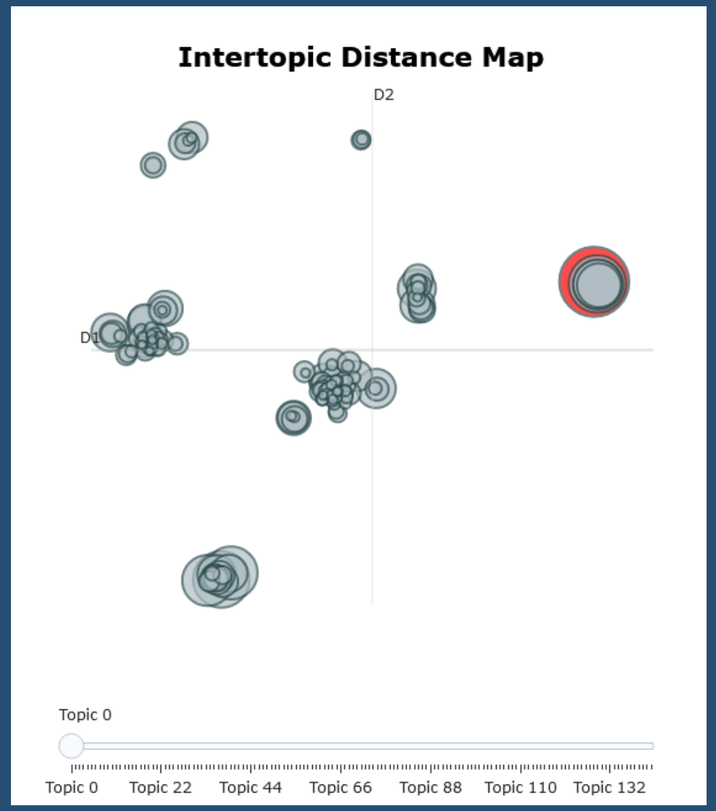
BERTopic is a topic modeling technique that leverages transformers and c-TF-IDF to create dense clusters allowing for easily interpretable topics.

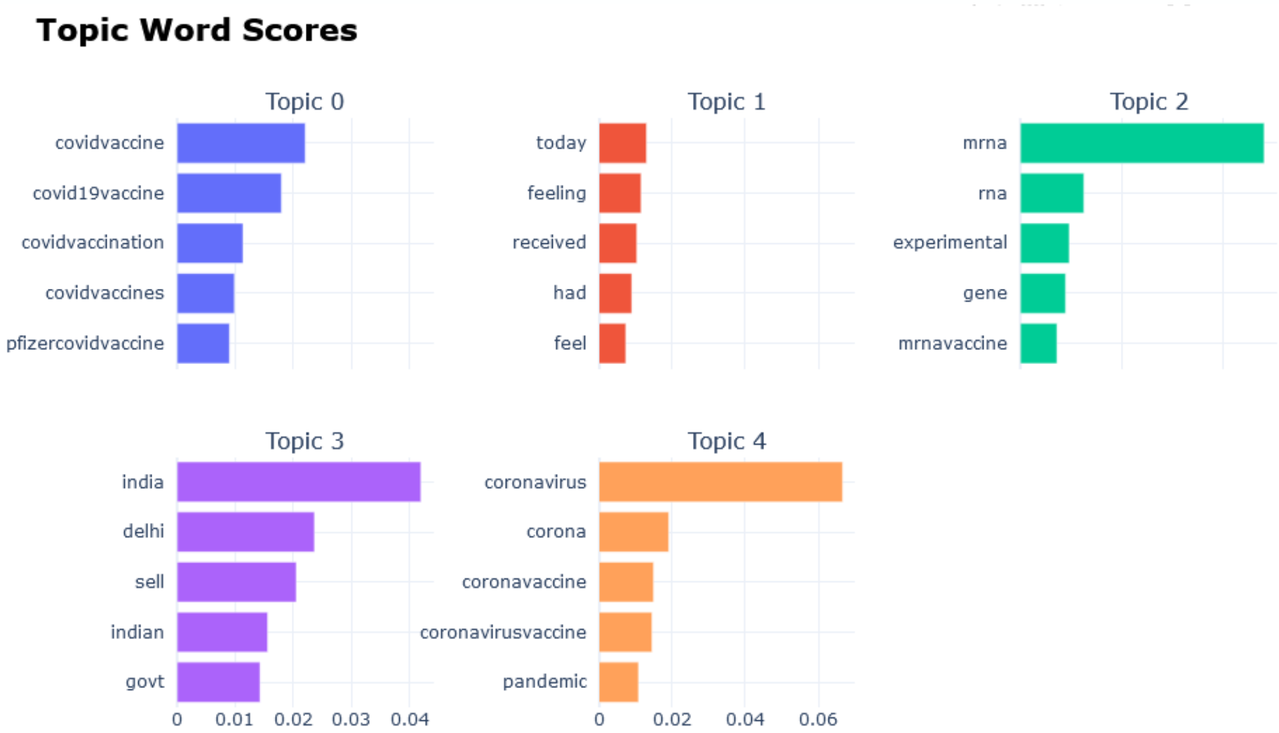
Embedding Model?

paraphrase-mpnet-base-v2 (a model that provides a higher quality, but takes more compute time)

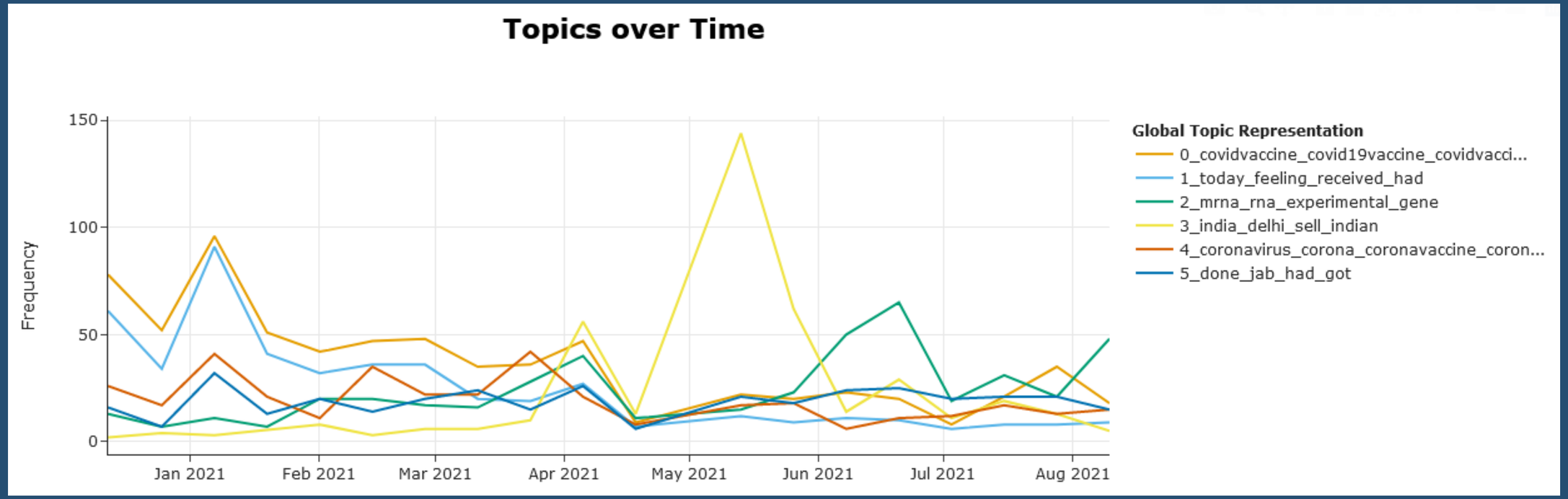


Topic Clusters

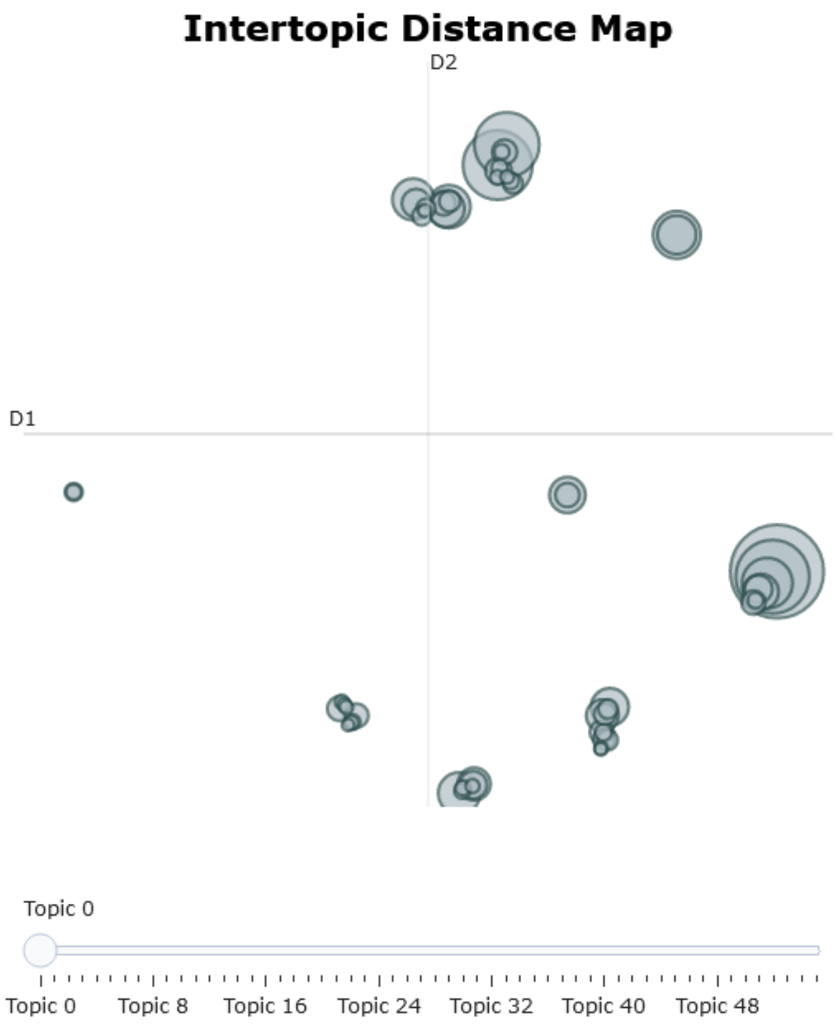




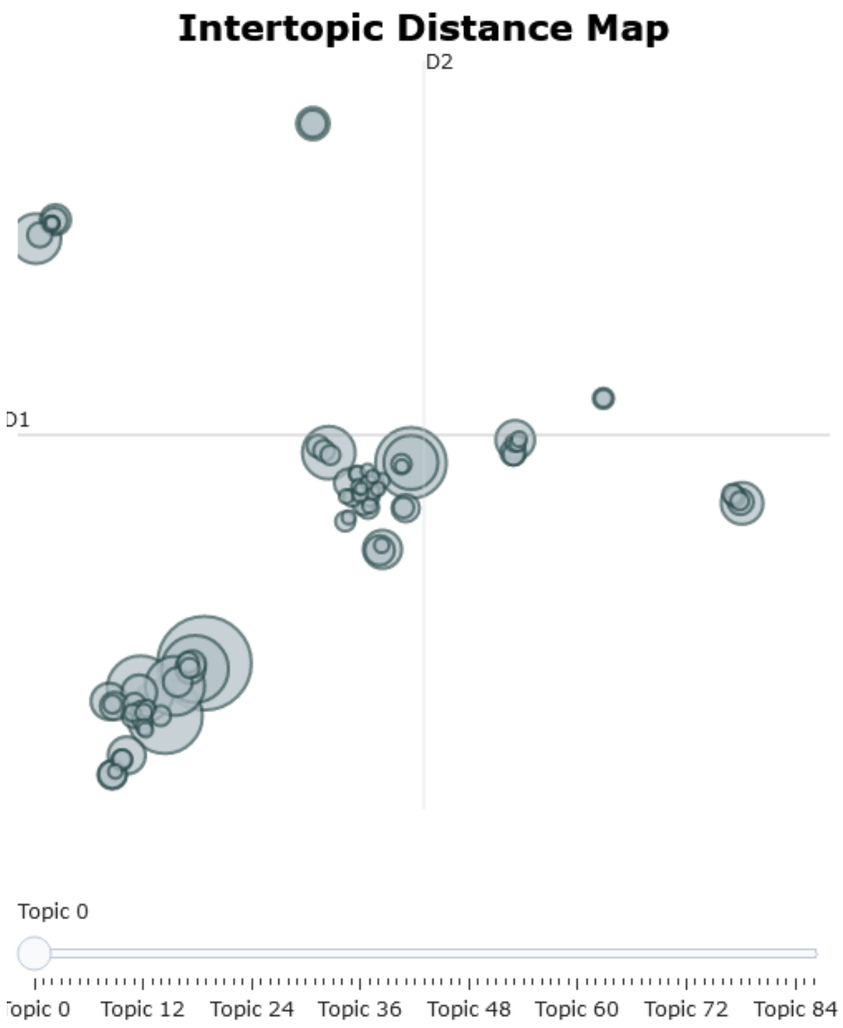
Topics over Time



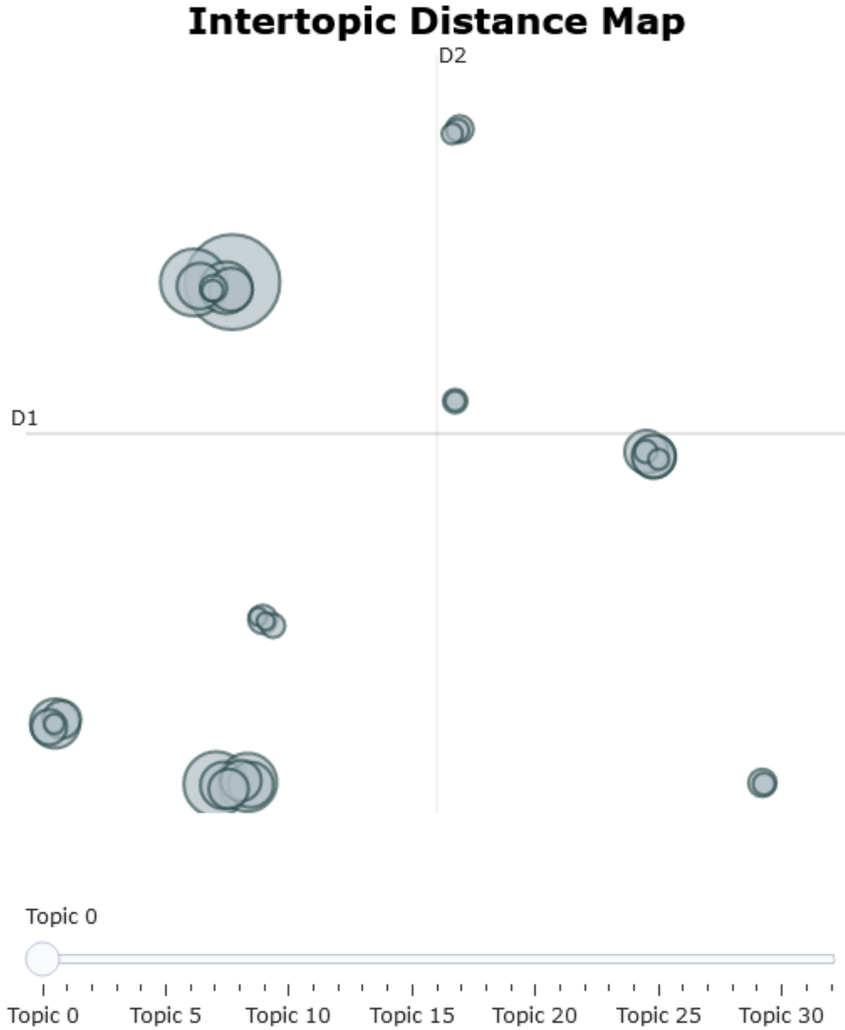
Topic Clusters



Positive



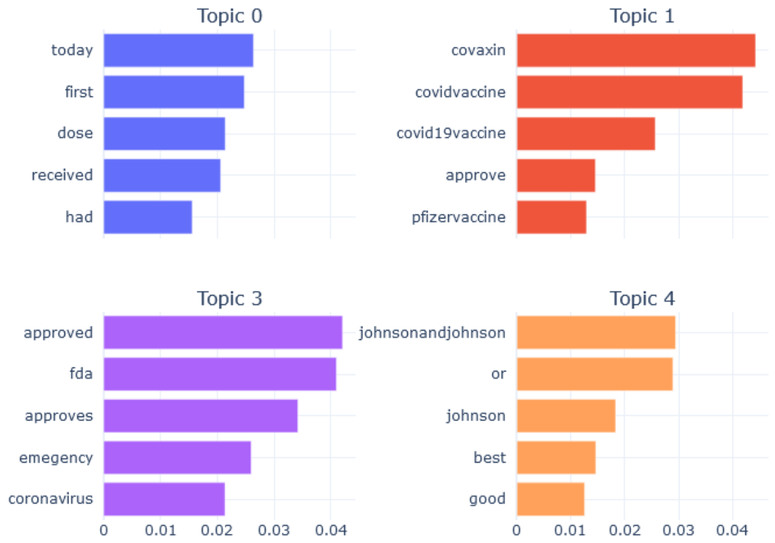
Neutral



Negative

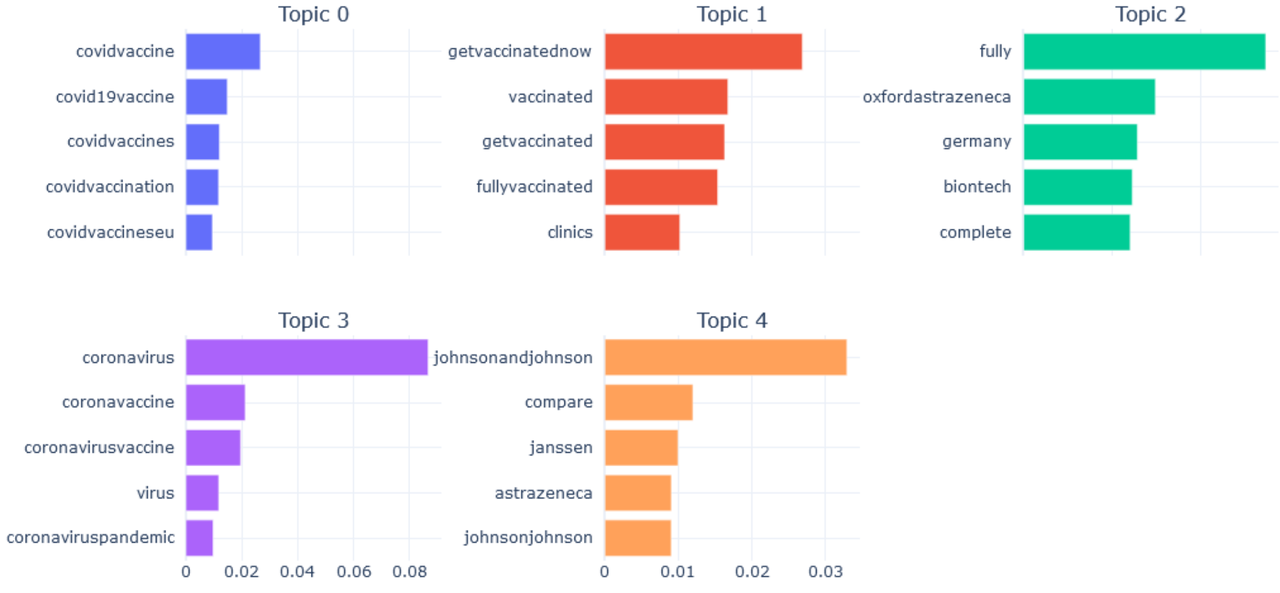
Topic Word Scores

Positive



Negative

Neutral



Conclusions & Recommendations

1. Although the Pfizer/BioNtech vaccine received the FDA approval on 23 Aug 2021, it was preceded with negative sentiment about how the vaccine is less effective against the Delta variant (Israel outbreak). In order to leverage on the FDA approval status, the company will have to address this controversy - if it's true, perhaps more R&D will have to be invested into a 'vaccine upgrade'. From the sentiment analysis, we saw that the highest negative spike was due to the Delta variant outbreak in India.

2. As seen from the sentiment distribution within each country, there is no obvious country with tremendously positive sentiment towards the Pfizer/BioNtech vaccine. As such, when it comes to location targeting, the company should look towards countries with the large populations and lowest vaccination rates.

3. The Johnson & Johnson vaccine was received positively due to its effectiveness against the Delta variant. In return, the Astra Zeneca vaccine has had a negative history of problems. The company should study other vaccine manufacturers and learn directly from their competitors.

THE END