DHBW
Duale Hochschule
Baden-Württemberg
Mannheim

Baden-Wuerttemberg Cooperative State University

Data Exploration Report

CovaxAnalytica – Data and Sentiment Analysis in Vaccine Discourse on Twitter

Authors:	Ayman Madhour, Lukas Bach
Course:	WWI-19-DSA
Time frame:	11.05.2021 – 13.07.2021

Contents

List of Figures

1 Motivation

COVAX is a worldwide initiative aiming at a fair distribution of COVID-19 vaccines, especially in developing countries. It is directed by the global alliance for Vaccines and Immunisation (GAVI) and the world health organisation (WHO).

> "With a fast-moving pandemic, no one is safe, unless everyone is safe."

Under this slogan, the initiative is striving to distribute WHO approved Vaccines, namely: Oxford-AstraZeneca, Pfizer-BioNTech, Moderna, Sinopharm, Sinovac and Johnson & Johnson. However, when it comes to vaccines, there are vastly polarising opinions, ranging from absolute vaccine-acceptance all the way to FUD and conspiracy theories.

Identifying triggers of negative sentiment and vaccine-hesitancy might be the first step towards a united society, pulling together against COVID-19. CovaxAnalytica aims to analyze the overall sentiment towards each vaccine, which allows for an in-depth analysis of opinions and their causes.

2 Related Work

Cotfas et al. conducted an analysis on COVID-19 Vaccination Tweets at the beginning of 2021. The analyzed time period was from November 2020 to January 2021.[1] Furthermore, DeVerna et al. provide insights into vaccine-hesitancy through their analysis of Twitter posts.[2] The results of both papers were relevant for our work, however vaccination campaigns skyrocketed over the course of the last few weeks. Thus further research capturing the most recent changes is needed.

3 COVID-19 Tweets

The first dataset consisted of COVID-19 related tweets.footnote[3] Due to twitter policies, only tweet IDs can be published, which then have to be aggregated in the hydration process. An aggregated dataset contains more than 110.000 Tweets that were parsed by specific keywords (e. g "covid", "covid_19", "pandemic", etc.). Over the course of the data protection, keywords were continuously enhanced. The sentiment of the tweets was determined by TextBlob, a Library for processing textual data, and included in the Dataset. It provides a simple API for Sentiment Analysis. Sentiment values range from -1 for very negative to 1 for very positive. For the analysis only the time

[1] Cotfas et al. 2021.
[2] DeVerna et al. 2021.
[3] Lamsal 2021.

period from 1st of November 2020 till 12th of June 2021 was considered.

3.1 Findings

The dataset mainly includes tweets from the United States (44.824 Tweets), the United Kingdom (21.110 Tweets), Canada (15.610 Tweets) and India (7.687 Tweets). What can be clearly observed is that some countries like Russia and China for example are not represented in the dataset because of the language barrier and the fact that they mostly use their own social networks (VK, WeChat).

Average Sentiment and Tweet Amount

Overall a very neutral Sentiment can be observed, however some countries are under represented in the dataset. This leads to the sentiment of some countries being set by only a few tweets. Thus no finite claims on the overall sentiment can be made. Nonetheless, a significant changeover can be observed in India.

In April India celebrated Kumbh Mela, the biggest festival of the world, with 30 Million attendants. [4] The weeks following this had a 1800 percent increase of COVID-19 cases, this translates to more than 169.000 new cases on 12th of April alone. [5] Which resulted in an increase of Tweets and a slightly more negative Sentiment. The United Kingdom has two significant sentiment developments. The day Prime Minister Boris Johnson announced the second Lockdown in England lead to a downwards trend in sentiment whereas the new year's lead to a positive spike in sentiment.[6]

4 Vaccine Tweets

The basis of the vaccine sentiment analysis is a tweet dataset, containing more than 100.000 vaccine-related tweets. Out of the whole dataset, only tweets about WHO-approved vaccines were extracted, as well as russian SputnikV and BharatBiotechs Covaxin. For the sentiment prediction, pythons Natural Language Toolkit (NLTK) library for sentiment analysis was used. The values then got smoothed through a savitzky-golay filter so that peaks appear more clear. The peak dates were then used to identify notable trends that could have potentially lead to that sentiment. Trends were obtained through PyTrends, an unofficial api for the Google Trends service.

4.1 Findings

The overall mean sentiment of each vaccine is neutral with a positive tendency. Out

[4](Redaktion and Wawatschek 14.04.2021)
[5](Manoj Chaurasia 30.05.2021)
[6]GOV.UK 12.07.2021.

of all analyzed vaccines, SputnikV has the lowest and Johnson & Johnson the highest sentiment. But because the tweet volume of these two tweets wasn't particularily high, caution is advised when drawing conclusions.

PfizerBiontech, the first approved COVID-19 vaccine, was the most tweeted about in the beginning of the year. However it got quickly overrun by Moderna and Covaxin. The tweet volume is of importance because the more a vaccine is talked about the less the risk of a few opionion leaders influencing the absolute sentiment.

Most positive tweets tend to talk about a successful vaccination experience or are longing for positive things like traveling again. Negative tweets tend to be about conspiracy theories or bad vaccine experiences (side-effects, deaths, etc.). Neutral tweets are mostly articles about cases and available doses.

As with all Vaccines predominantly used in the U.S., there is an observable dip in Vaccination on valentines day and memorial day. Other than that, it seems like more people get vaccinated on weekends than during the week.

The first negative peak in sentiment seems to be influenced by news about pfizer biontech causing 23 deaths in norway. The news had a strong impact overseas and caused an increase in vaccine-hesitancy, as observable in the following twee

> 'U MAY die if u become covid positive. But #PfizerBioNTech shot will DEFINITELY kill you. Stuck between devil and sea.' [7]

Sinovac and Sinopharm have a notable peak in positive sentiment in the beginning of june. This was most likely a reaction to the fact that the vaccine got approved by the WHO.

On 12. April, Johnson & Johnson (JandJ) experienced a sudden shift from very positive sentiment to neutral with negative tendency. That is roughly the time during which the news broke that johnson and johnson causes blood clots. Trending search terms were "johnson and johnson vaccine blood clot symptoms" which hints that people who are vaccinated feel afraid. JandJ was largely used in the U.S. which during that time was vaccinating its population at full speed.

A more in depth look reveals that over the course of 4 days the amount of positive tweets fell and neutral/negative tweets gained traction.

4.2 Text Analysis

As of june/july 2021, the delta variant of the COVID-19 virus is posing serious threats to progress made in the fight against the virus. The discourse about the delta variant shows that Covaxin, an indian vaccine, is the most mentioned. This doesn't come as a surprise as the delta variant appeared first in India, which means that this region is likely the

[7] Index No. 29798 in geo_vaccine_tweets_with_sentiment.csv

most affected.

Apart from this, the whole corpus ran through a natural language processing pipeline with several iterations of word and document vectorization. The main aim was to cluster tweets depending on their content but because tweets are short in text, no clear cluster could be identified. To achieve this, further data engineering and hyperparameter tuning is needed. The word vectorization showed nonetheless a few relevant insights that can be used for further research. A Word2Vec model trained on the whole corpus of 100.000+ tweets yields the following:

> the most similar word to common conspiracy theory terms like "plandemic" is: great reset, bill gates, genocide, drfauci, merck.

5 Conclusion

The results appear promising, however we can't make definite conclusions as there are a few limiting factors:

The language barrier can't be neglected. Some countries weren't considered as only english tweets were analyzed. Some contries have their own social media platforms and don't partake in the discourse on twitter. The largest demographic group of twitter users are betwen ages 18 and 29, that means that other age groups aren't represented in the discourse.

Nonetheless, fact checkers are widely needed and remain necessary to reduce misinformation. The way social media is designed is that people stay in their own "bubble" and only consume tweets/news that validate their views and thoughts. This confirmation bias is dangerous as it can lead to extreme views. A mixed and civil discourse would be the most beneficial to unite society as a whole.

References

Cotfas, Liviu-Adrian et al. (2021). "The Longest Month: Analyzing COVID-19 Vaccination Opinions Dynamics From Tweets in the Month Following the First Vaccine Announcement". In: *IEEE Access* 9, pp. 33203–33223. DOI: 10.1109/ACCESS.2021.3059821.

DeVerna, Matthew R. et al. (2021). *CoVaxxy: A Collection of English-language Twitter Posts About COVID-19 Vaccines*. arXiv: 2101.07694 [cs.SI].

GOV.UK (12.07.2021). *Prime Minister's statement on coronavirus (COVID-19): 5 November 2020*. URL: https://www.gov.uk/government/speeches/prime-ministers-statement-on-coronavirus-covid-19-5-november-2020.

Lamsal, Rabindra (2021). "Design and analysis of a large-scale COVID-19 tweets dataset". In: *Applied Intelligence* 51.5, pp. 2790–2804.

Manoj Chaurasia, The Guardian (30.05.2021). "Kumbh Mela: how a superspreader festival seeded Covid across India". In: *The Guardian*. URL: https://www.theguardian.com/world/2021/may/30/kumbh-mela-how-a-superspreader-festival-seeded-covid-across-india.

Redaktion, Br24 and Veronika Wawatschek (14.04.2021). "Hindu-Fest Kumbh Mela findet ohne Corona-Beschränkungen statt". In: *BR24*. URL: https://www.br.de/nachrichten/deutschland-welt/hindu-fest-kumbh-mela-findet-ohne-corona-beschraenkungen-statt,SUYYgVM.

6 Appendix - Visualizations

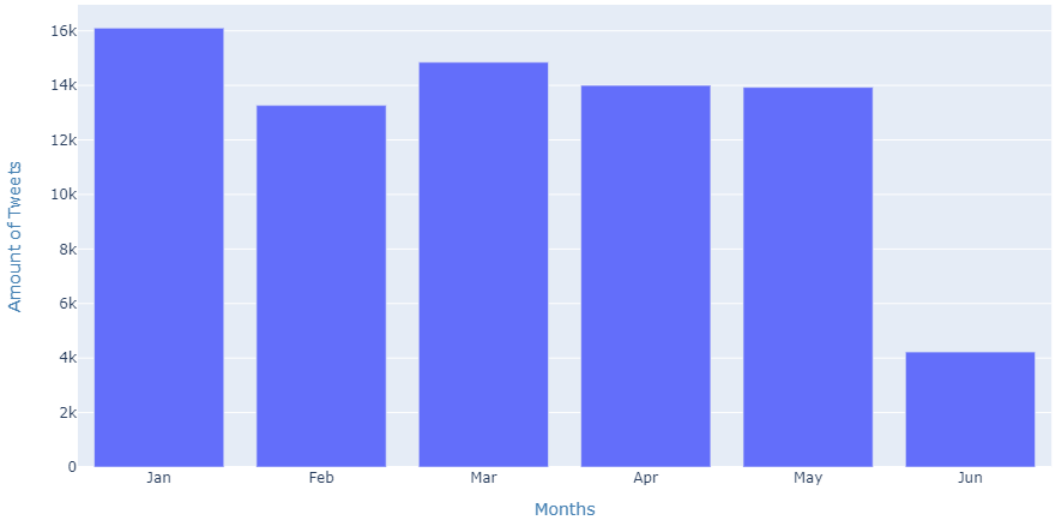

Figure 6.1: Tweet Amount Worldwide

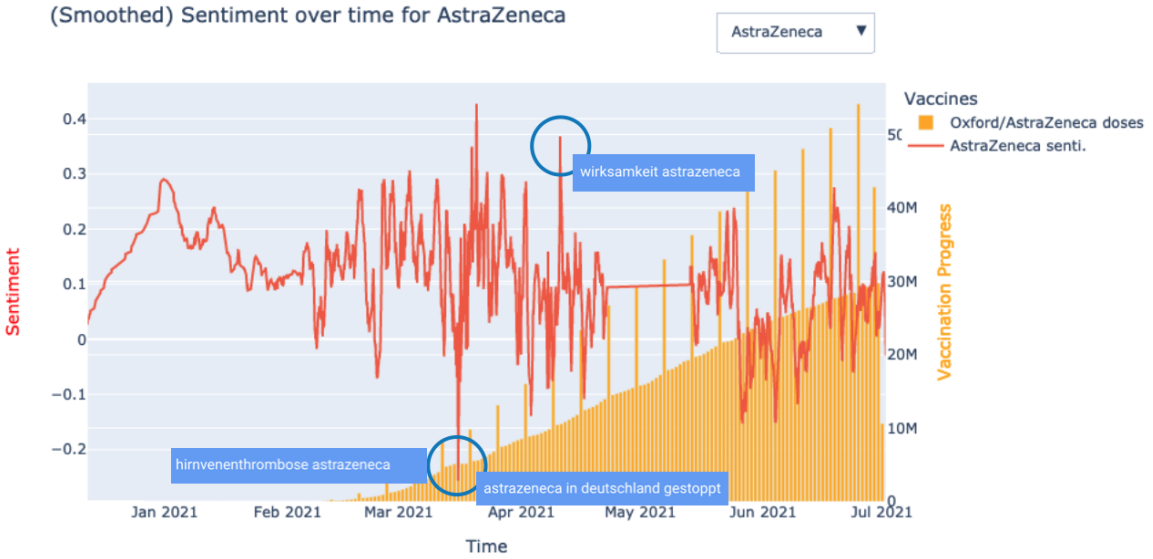

Figure 6.2: AstraZeneca Sentiment

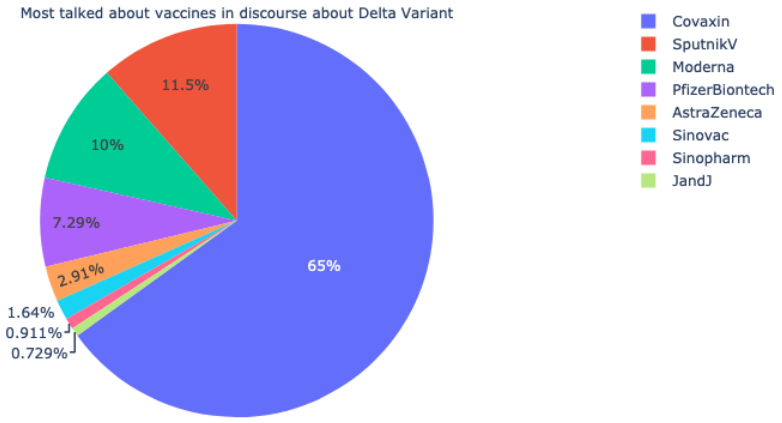

Figure 6.3: Delta Variant

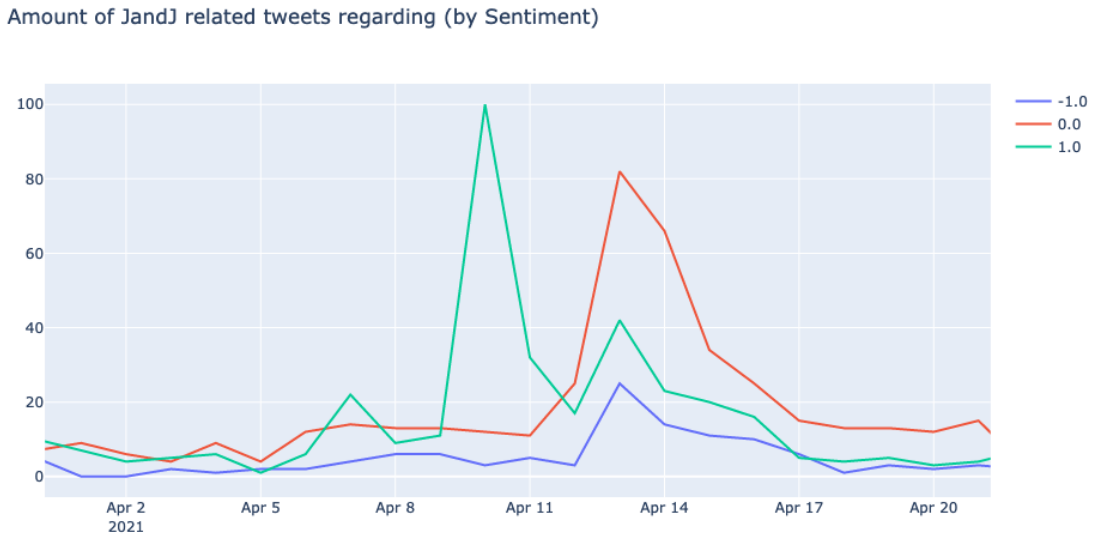

Figure 6.4: Johnsson & Johnsson Sentiment broken down

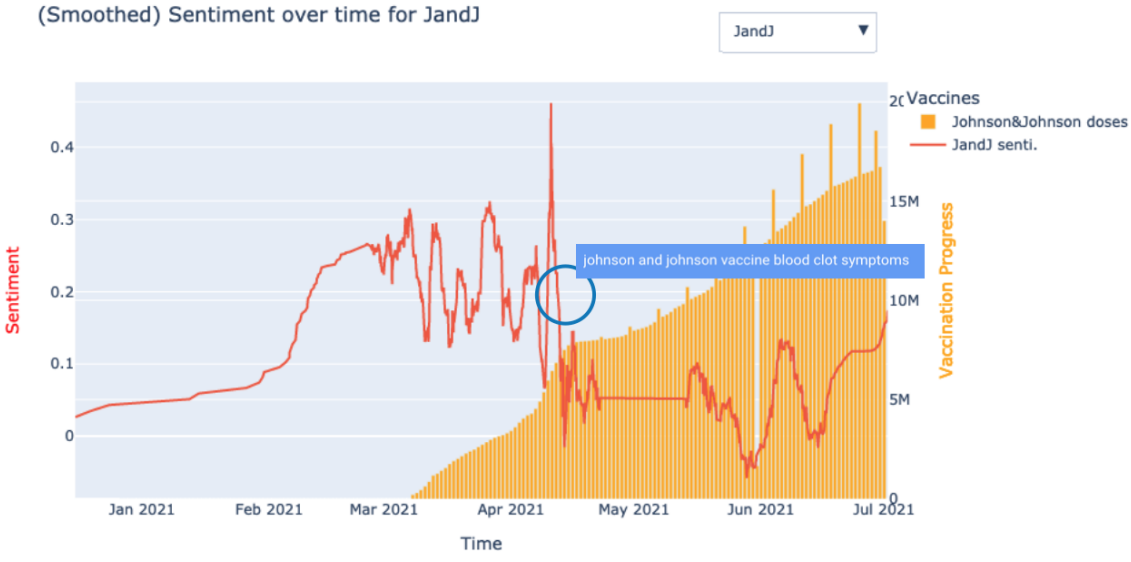

Figure 6.5: Johnsson & Johnsson Sentiment

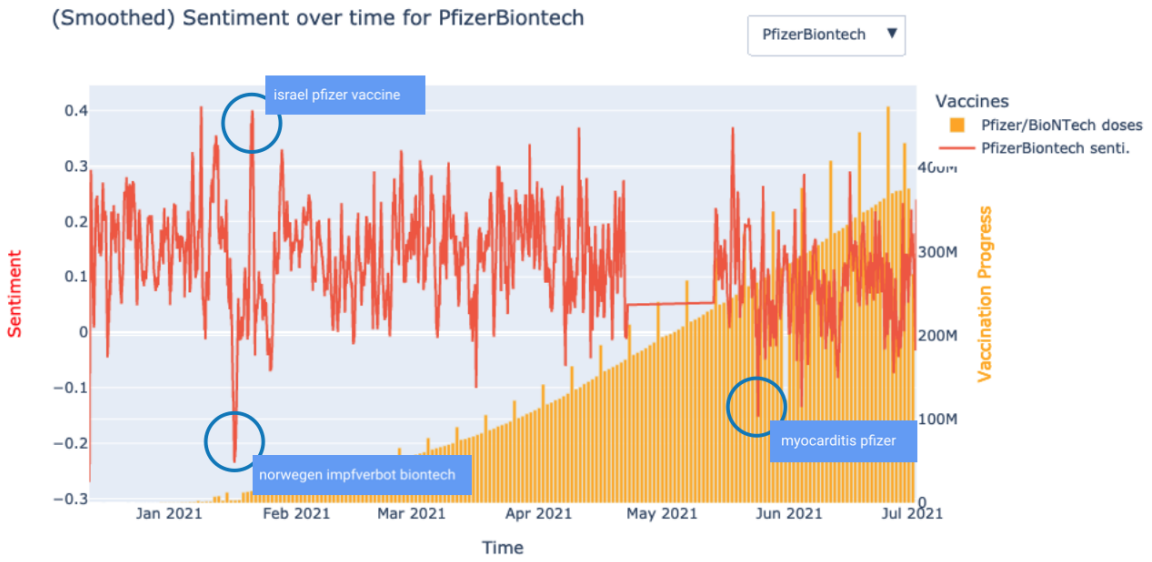

Figure 6.6: Pfizer/Biontech Sentiment

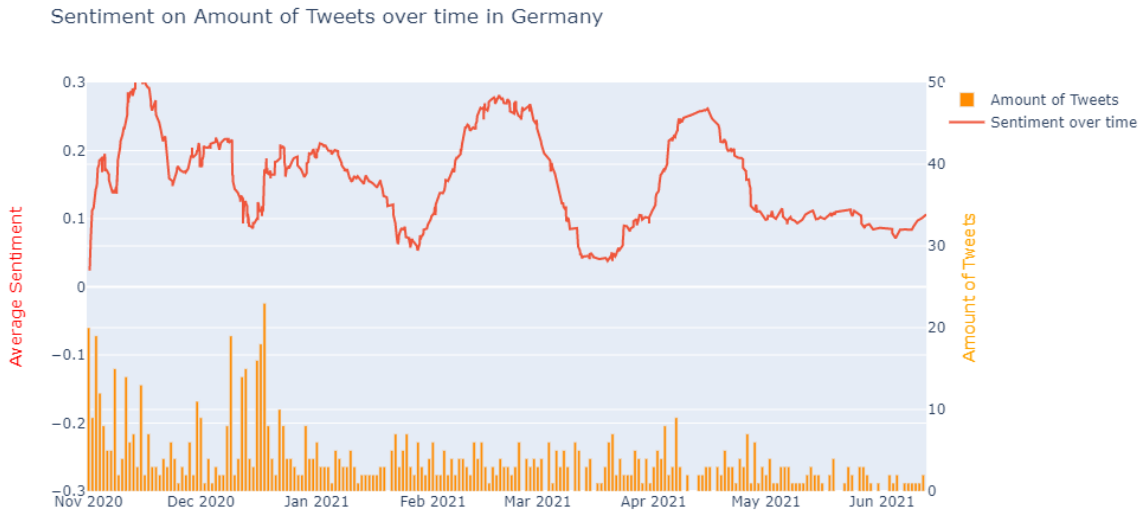

Figure 6.7: Sentiment Germany

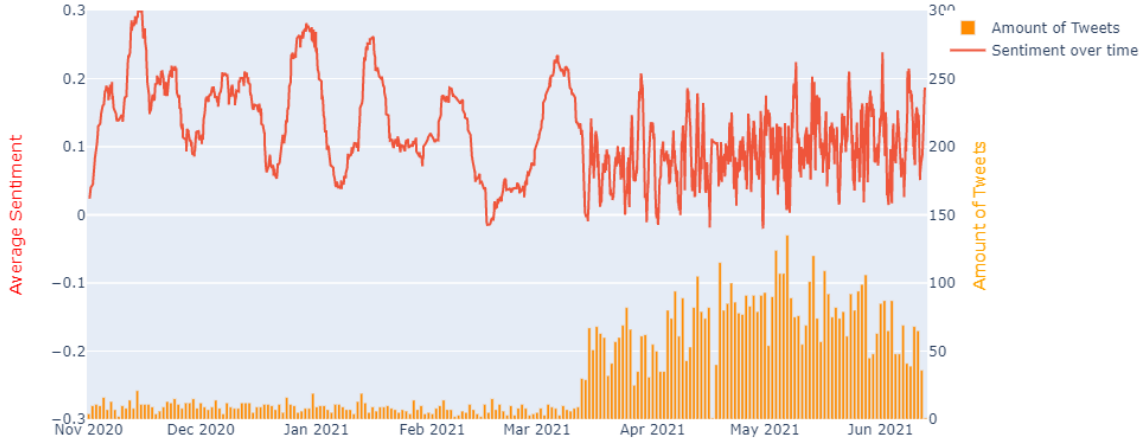

Figure 6.8: Sentiment India

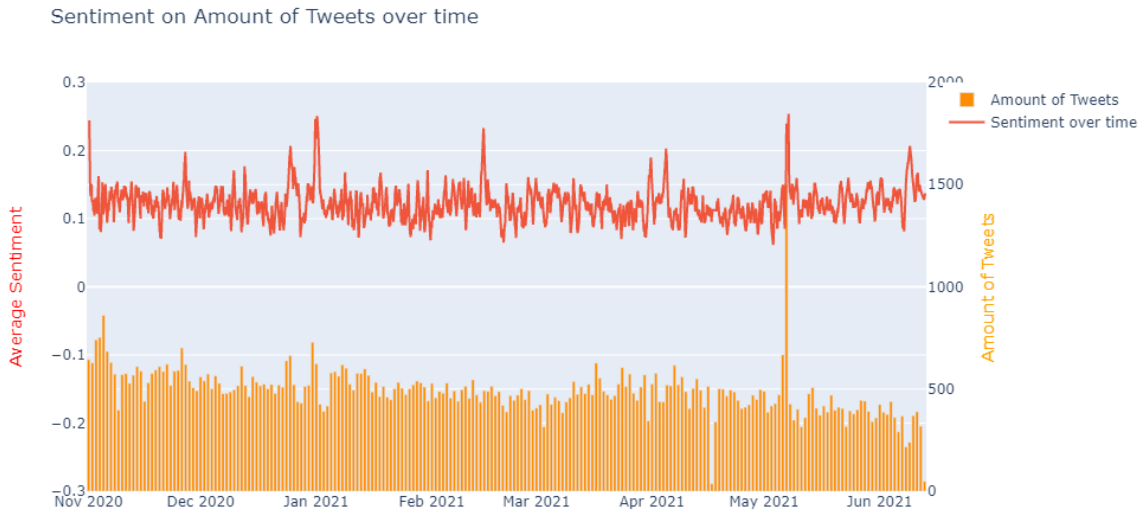

Figure 6.9: Sentiment Overall

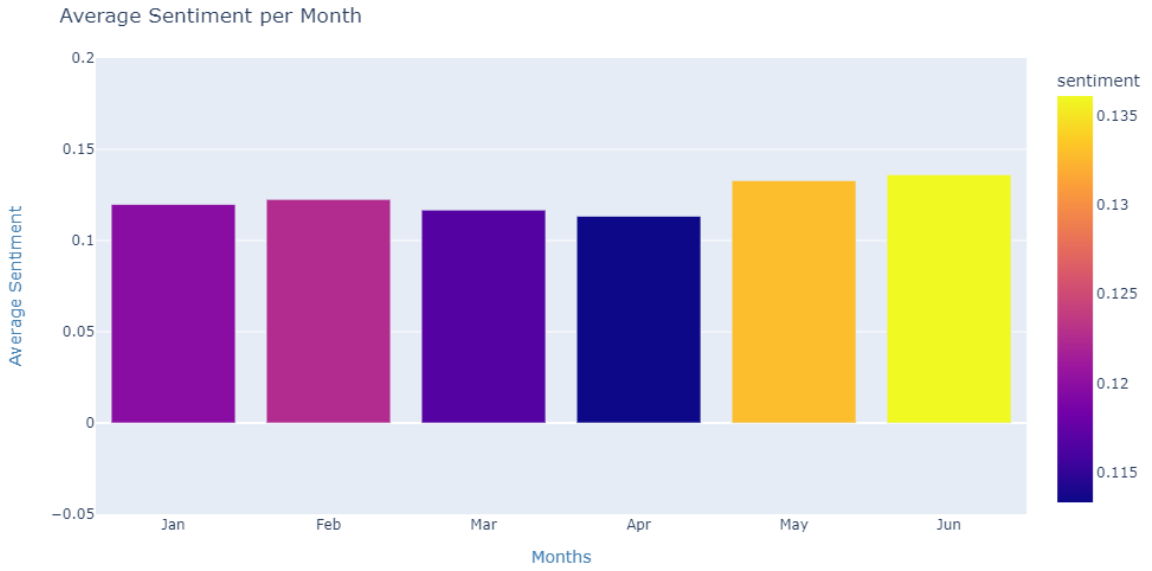

Figure 6.10: Sentiment per Month

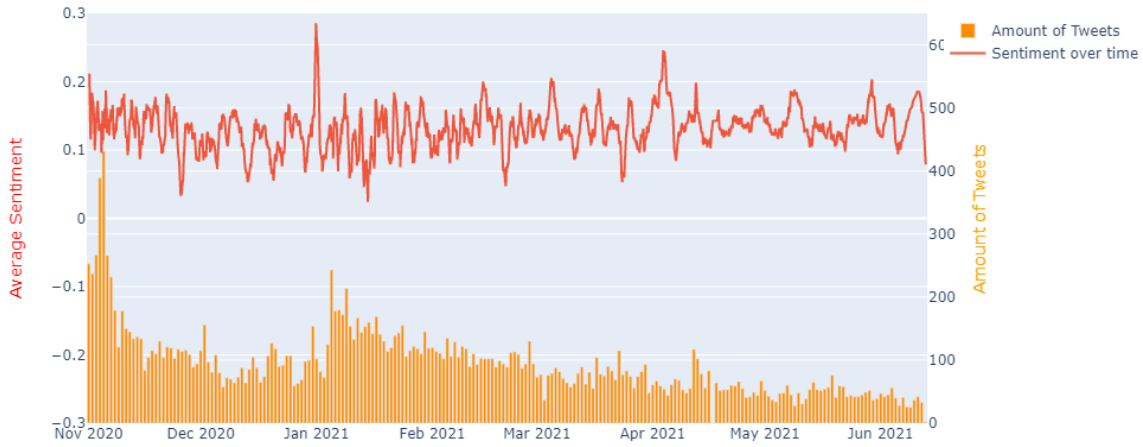

Figure 6.11: Sentiment United Kingdom

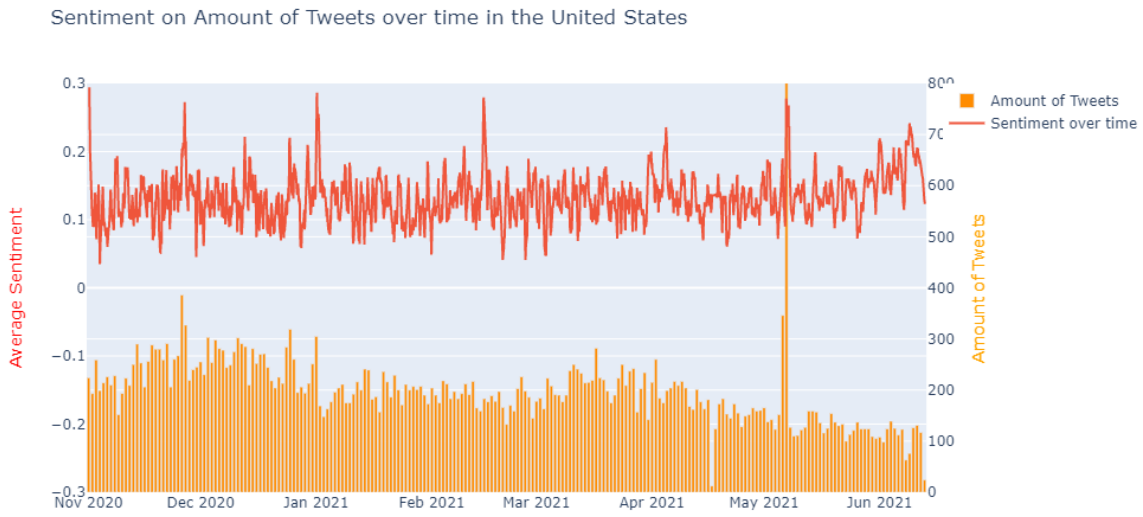

Figure 6.12: Sentiment United States

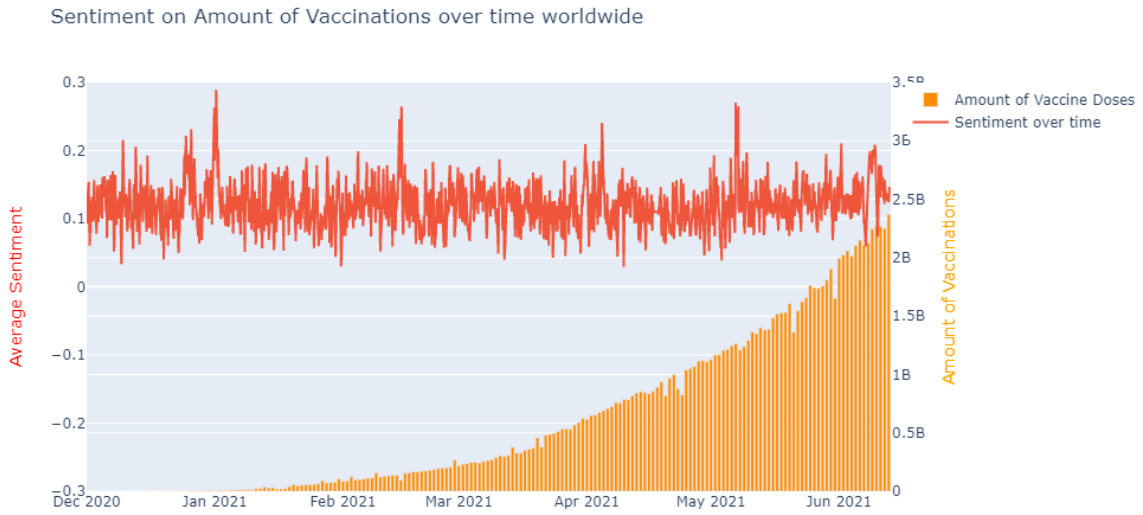

Figure 6.13: Sentiment Vaccinations

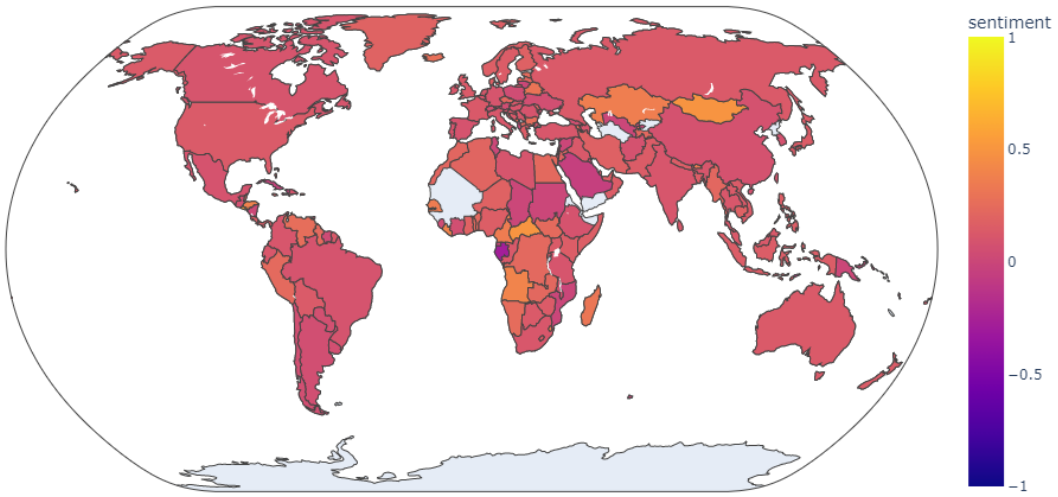

Figure 6.14: Sentiment Worldwide

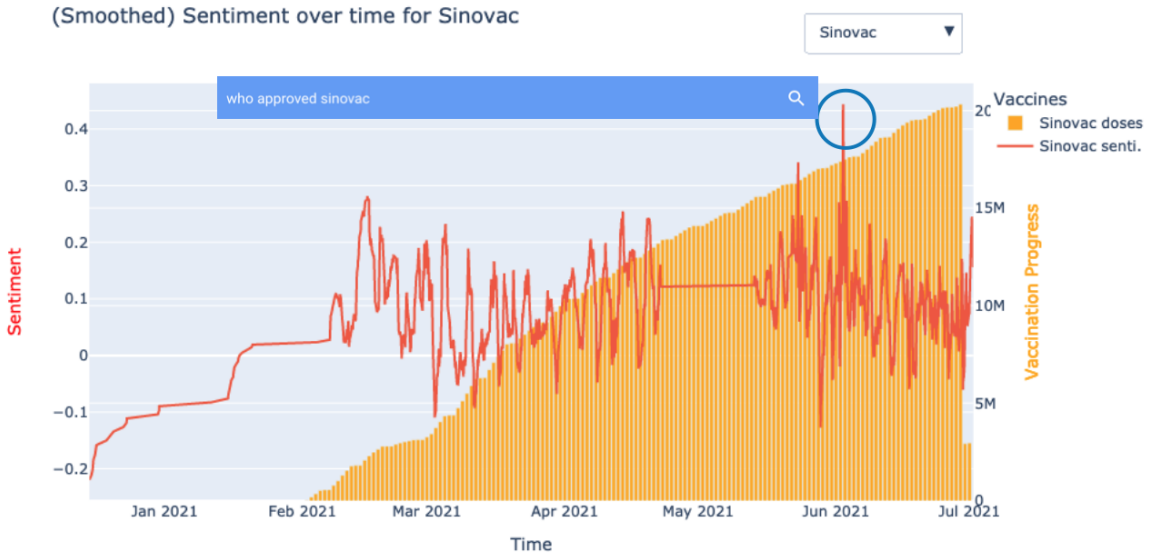

Figure 6.15: Sentiment

Figure 6.16: Tweets Worldwide

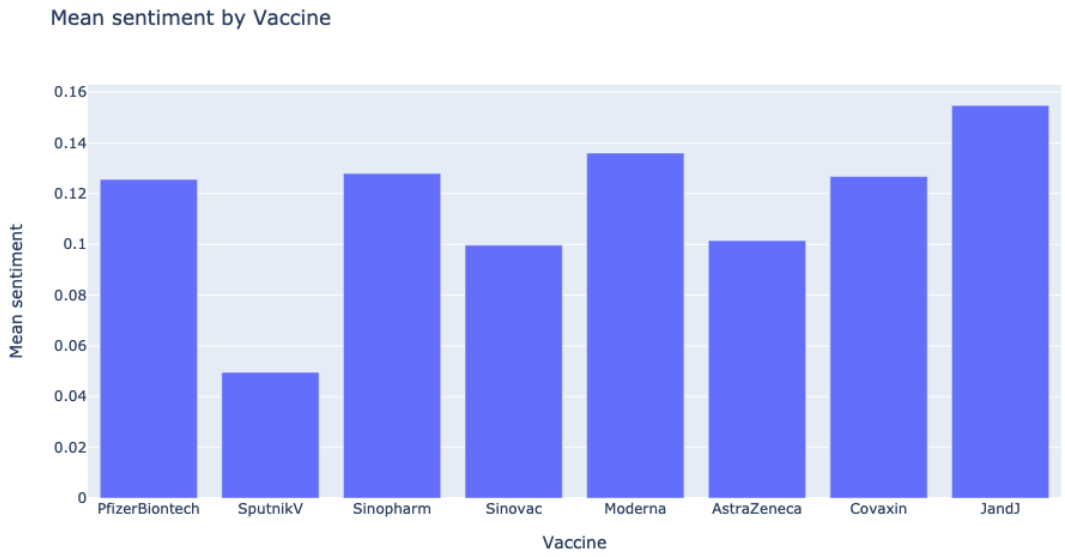

Figure 6.17: Vaccine Sentiment

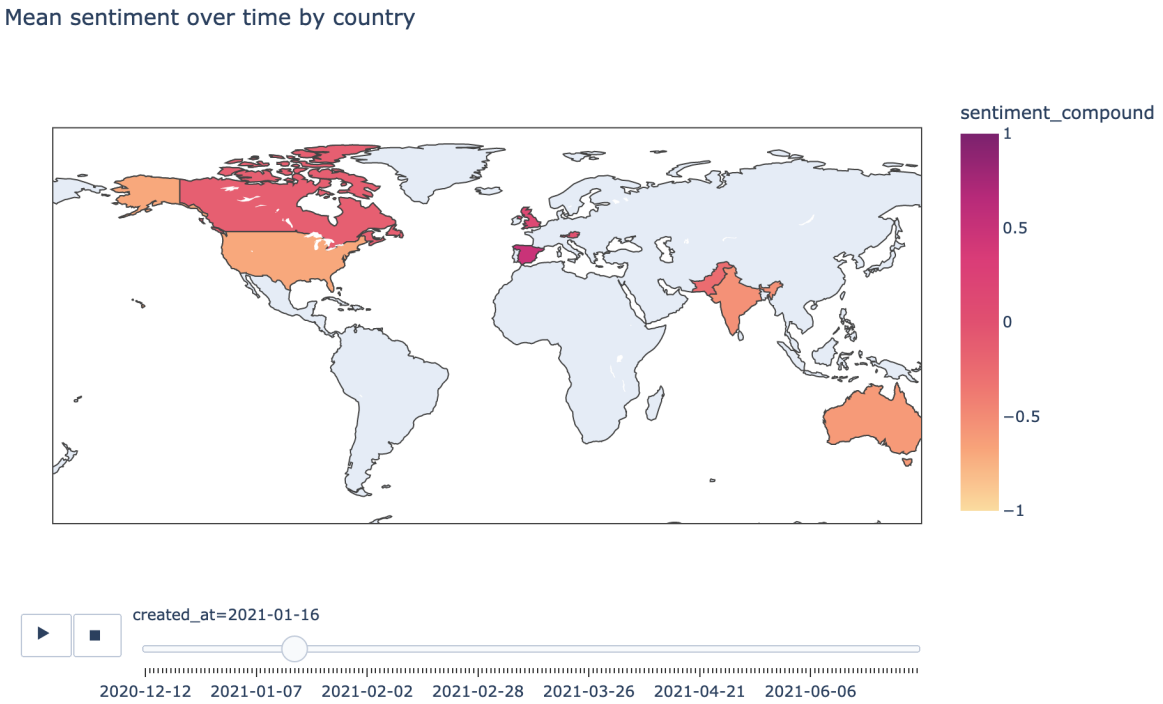

Figure 6.18: World Sentiment