Sentiment Analysis on COVID-19 Twitter Data: A Sentiment Timeline

Makrina Karagkiozidou, Paraskevas Koukaras, Christos Tjortjis

Sentiment Analysis on COVID-19 Twitter Data: A Sentiment Timeline

Makrina Karagkiozidou[1], Paraskevas Koukaras[1][0000−0002−1183−9878], and Christos Tjortjis[1][0000−0001−8263−9024]

The Data Mining and Research Analytics Group, School of Science and Technology, International Hellenic University, 57001, Thessaloniki, Greece
{mkaragkiozidou, p.koukaras, c.tjortjis}@ihu.edu.gr

Abstract. COVID-19 has been one of the most dominant discussion topics on Twitter since 2019. Users express their opinions representing public sentiment on the topic. This paper presents a sentiment timeline of Twitter users, regarding COVID-19 vaccines. This work raises concerns about the extracted information with regards to sentiment analysis, the dominance of each sentiment and its influential power. During the implementation of the analysis, several datasets were examined for the creation of the model. Various algorithms were employed with Random Forest performing best and therefore selected for training the model, achieving an accuracy of 91.5%. Our findings indicate that the majority of Twitter users are positive regarding COVID-19 vaccines and support WHO's recommendations. Negative tweets comprising the minority of the tweets, appear to have a higher influential power with their retweet rates, outperforming positive and neutral sentiments.

Keywords: Sentiment analysis · Text mining · Social media · Vaccination · COVID-19.

1 Introduction

Since the creation and wide spread of Social Media (SM) their users have been offered a platform to post and publish opinions in public. That has led companies and organizations to gather and benefit from the information published. Therefore, SM have been widely used for a variety of analytics [12] also including predictions [23] that may integrate sentiment as an extra feature [14]. In microblogs users can share their opinion on any matter, even serious social and political affairs. Such is the COVID-19 virus, a coronavirus that first appeared in December of 2019 in Wuhan, China [24].

Aiming to stop the spread of the virus, each country has taken various measures, including mandatory mask use, local and nationwide lockdowns and more. The citizens of each community have developed increased anxiety levels and insecurity about their future [2]. Even with the creation and production of vaccines, it occurred that a part of each country's community was unwilling or against it, expressing their negative opinion on various types of SM platforms [13]. In

this study, sentiment analysis on Twitter data has been conducted, regarding COVID-19 vaccines. The aim is to identify patterns and changes in users' public opinion throughout time (during the past year, 2021) and based on milestone events and important announcements of the World Health Organization (WHO) but also actions from authorities. The remaining of the paper is structured as follows. Section 2 reviews related work. Section 3 elaborates on the methodology, while Section 4 presents the results of the experimentation we conducted. The paper concludes with Section 5 discussing research accomplishments and future work.

2 Related work

This section presents and analyses recent related work. The aim is to provide a concrete background knowledge regarding sentiment analysis using SM data.

Nemes and Kiss [20] conducted COVID-19 related sentiment analysis on data gathered from Twitter. They focused on comparing the utilized methods. In their training model they verified their findings with an external open-source dataset. Overall, a positive sentiment was identified in the posts, that was maintained over time. An increase in negative sentiment was also observed, leading to a diversity in sentiment polarisation.

Manguri et al. [16] conducted sentiment analysis based on posts published on Twitter during the week of 09-04-2020 to 15-04-2020 related to the keywords 'Covid19' and 'coronavirus'. The findings showed a high percentage of polarisation among users. Furthermore, they related daily opinion changes, depending on government and media actions, broadcasts and new guidelines. To further support their findings they identified higher sample quality on Twitter compared to other SM.

Kruspe et al. [15], conducted cross-language sentiment analysis. The research time frame expanded from December 2019 to April 2020, including countries such as the UK, Spain, Germany, Italy, France and Netherlands. Their findings show that until February 2020, there was little reference to COVID-19 related keywords and topics. Additionally, with the announcement of a lockdown, the sentiments were mostly negative, but improved over time.

Garcia and Berton [10] focused on Brazil and the USA. In their research, they detected and ranked 10 topics, ranging from economic impacts, politics and case reports to anti-racism protests, online events and sports. The dataset includes tweets within a four-month time span, between April and August 2021. Their analysis showed that negative emotions were dominant, especially for the topics of case studies, 'proliferation care' and 'statistics'.

Boon-Itt and Skunkan [5] completed a sentiment analysis study aiming to provide insights on the public perception of COVID-19 using Twitter data. Data were collected during the period from the 13th of December 2019 up to the 9th of March 2020, establishing this research as one of the first on the topic. They created a timeline of the frequency of each symptom mentioned in the posts. Meanwhile, the results indicated that there was in general a negative sentiment for COVID-19.

The analysed literature focuses on specific time periods, topics and Twitter sentiment without considering the total sentiment changes throughout the last year. Therefore, in this study, open-source Twitter datasets are combined with twitter data collected during the period expanding from 15-09-2021 up to 10-12-2021 to identify the sentiment of users after public announcements of governments, WHO, and various FDA vaccination approvals.

3 Methodology

The main purpose of this research is to identify and elaborate the sentiment of Twitter users with regards to the topic of COVID-19 vaccines. Therefore, a clear and well-defined methodology was structured. To identify the response of Twitter users on important COVID-19 vaccine events and announcements, published tweets on the topic were collected in combination with external datasets. RapidMiner was employed for partial data gathering (for some periods) but mainly for data processing. The research model was developed with the use of the Python programming language. The environment in which the code was developed, was Jupyter Notebook. The external dataset was trained with the use of the Random Forest classifier. The main dataset was pre-processed, leading to the prediction of each tweet's sentiment. Finally, the results of the analysis were extracted and visualized using word clouds.

3.1 Dataset

The utilized data in the analysis were collected using Twitter API v2. For the collection of tweets, a workflow on RapidMiner was implemented. Then 10 'Twitter Search' nodes were implemented for each one of the researched keywords. These were: 'covid19 vaccines', 'coronavirus', 'pandemic', 'Pfizer Vaccine', 'Delta Variant', 'Vaccine Certificate', 'Covidiots', 'Covidscam', 'PCR test', 'Rapid test'. For each search, the most recent English tweets were selected.

For integrating more data in the sentiment analysis approach we also employed an existing coronavirus dataset [7]. It contains more than 165,000 tweets with their sentiment annotation. The tweets were notated as positive, neutral and negative. 55% of the tweets were labelled as neutral, 23% as positive and 22% as negative. The downside of this dataset is that there is no data description, therefore, we could only assume the process followed for the annotation of each tweet. To establish the validity of that work, we conducted manual data evaluation.

Dataset structure
With the use of Twitter API v2 and RapidMiner for the collection of the tweets, there was a specific number of features extracted for every tweet. The main tweet information that is required for the analysis contains the text of the tweet, the date and time of its creation, its retweet count, and the name of the author. Therefore, since there was a combination of self-gathered and external datasets, they both needed to contain the same information.

Pre-processing
The pre-processing of the dataset was implemented using the Python programming language and was divided into two parts. Firstly, the dataset was pre-processed, cleaning the 'text' field from elements that reduce the accuracy of the model based on the research of Beleveslis et al. [4]. The implemented pre-processing function aimed at the removal of the retweet feature, the '@usernames' mentioned in the main part, the '#' symbol, the possible link that might accompany the text and the numbers. Since the final data used in the analysis originated from various sources, their interoperability characteristics were important. Therefore, actions to eliminate duplicate records were implemented. Then, we executed a function for cleansing every tweet text.

3.2 Sentiment Model development

The libraries that were used throughout the analysis were 'Pandas', 'NumPy', 'Re', 'Seaborn' and 'Sklearn'. It occurred from the examination of the dataset, that tweets characterised as neutral were dominant totalling 98,844 records while positive and negative followed with counts of 40,693 and 40,322, respectively. It appeared that the selected dataset was imbalanced in favour of neutral tweets. For that reason, the overall dataset was divided into three subsets using under-sampling and according to their sentiment. Under-sampling is a method for balancing unequal datasets that involves maintaining all of the data in the minority class while reducing the size of the majority class. This operation caused the most dominant sentiments, neutral and positive to become equal to the number of negative tweets. Positive tweets were represented by the numerical annotation '3', neutral by '2' and negative by '1'.

The next action was to split data into the training and test sets. 80% for training set and 20% for testing was chosen after running multiple variations of these percentage thresholds and checking the output results. For the vectorization of the tweets we used TF-IDF (Term Frequency - Inverse Document Frequency) vectorizer [11].

In the final part of the model, Random Forest Classifier was employed. The model was trained to fit the training set and finally, it predicted the label of the test set, indicating the final accuracy of the model. The accuracy percentage achieved was 91.58%. Other classifiers including Support Vector Machine (SVM), Naïve Bayes and kNN were also tested, but they achieved lower accuracy rates.

3.3 Sentiment analysis

After successfully training the model, the prediction of the gathered dataset's sentiment took place along with the data pre-processing. Only English tweets and unique tweets were maintained. Next, feature selection was conducted using the 'Optimize Selection' operator[1] from RapidMiner, eliminating features that did not include valuable to the research information. The second external dataset was imported and pre-processed in the same way as the first one. Furthermore, the sentiment of each tweet was calculated and stored using the TF-IDF vectorizer.

[1] https://docs.rapidminer.com/latest/studio/operators/

Tweets were split into tokens of words for the generation of the general, negative and positive word clouds. The findings and results of this analysis are presented and discussed in the next section.

4 Results

4.1 Overview

The gathered tweets, cover a one-year time frame, from 12 Dec 2020 to 10 Dec 2021. During this year, the first COVID-19 vaccine was produced and shared but also mass vaccination became available. At the same time, objections against the vaccines by part of the public were also made. The word clouds presented are indicators of the discussion topics of Twitter.

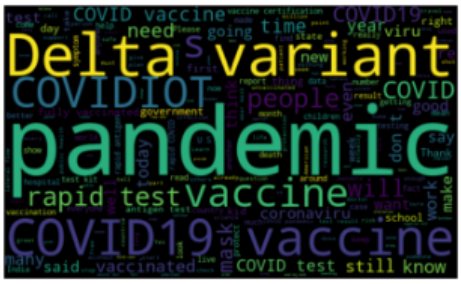

Fig. 1. General word cloud.

In the general word cloud (Fig. 1) the most dominant topic is 'pandemic', showing that most people were discussing the pandemic in general, not only 'vaccines' which follow, and 'Delta variant', which is one of the topics that gained attention in the last five months. Other words that appear to dominate in the word cloud are 'COVID19' and 'COVIDIOT', indicating polarity among the sentiments.

In the positive word cloud (Fig. 2), users tweet about 'vaccines', the new variant, but apart from the topic words, there are noted some smaller very indicative words. Namely, 'still', 'need', 'good', 'work', 'well', 'right', are proof that supporters of the vaccine, try to convince the rest of users of the vaccine's necessity. It is also noticed, that the term 'COVIDIOT' appears in this word cloud indicating polarity among the sentiments.

On the contrary, the negative word cloud (Fig. 3) contains fewer terms, indicating that most users were focusing on a smaller variety of topics. Respectively, for this sentiment, the most dominant terms are 'pandemic', 'vaccine'. Twitter users of this sentiment, appear to use the term 'COVIDIOT'. Other less dominant terms include 'mask', 'work', 'sick','which indicate their focus on more practical aspects of the topic. Researchers in [6] claim that most negatively driven

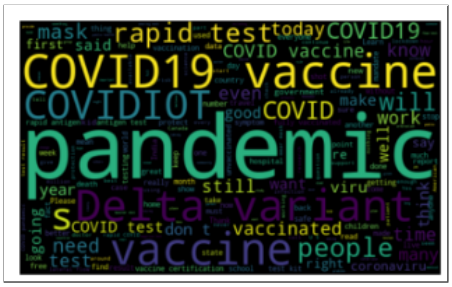

Fig. 2. Positive word cloud.

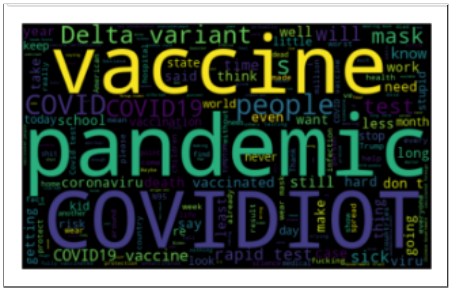

Fig. 3. Negative word cloud.

users, use the 'mask' topic to express their opposition to political or ideological beliefs rather than to the actual effects of masks on health and COVID-19 spread prevention.

4.2 Twitter Timeline

Based on the information of authorised announcements about the virus and WHO's response in their COVID-19 pandemic website [25], a timeline of the expansion and WHO's response is presented starting in December 2020 (Table 1).

Isolation and strict measurements has led to shifts in sentiment on Twitter, such as increments in observed anger for every new wave of the pandemic [1]. Therefore, this work is implemented to monitor Twitter user sentiment on the announcement of each event from WHO and the authorities.

4.3 Sentiment Overview

The dataset used in the analysis initially consisted of 611,022 Tweets in total. After data pre-processing and cleansing, the total number of records was decreased to 388,220. With the prediction of the label per tweet, it appears that the majority of tweets are positive. This observation shows that Twitter users are positive and support COVID-19 vaccines. In detail, positive tweets that are labelled with '3' comprise 67.8% of the records, showing the support and belief of users for the vaccines released. 23% were labelled as Neutral, while only 9.2% of the tweets were labelled as Negative, indicating disbelief, anger and disapproval of the vaccines.

4.4 Sentiment Timeline

We found that the average sentiment during the past year, ranges from 2.25 to 2.8 indicating a neutral to positive sentiment (Fig. 4). The lowest sentiment was noted in the week between 17 Jun 2021 and 24 Jun 2021, while the highest was almost a month later between 07 Aug 2021 and 14 Aug 2021. Based on the plot presented, it is obvious that sentiment has shifted daily. One of the most intense decreases was noted at the end of Jan 2021, where the sentiment average reached approximately 2.42. Considering the timeline with the most important events, on

Table 1. Coronavirus timeline.

Date	Event
14 Dec 2020	The first vaccination shot took place in the United States [22]. United Kingdom reported a SARS-CoV-2 variant to WHO.
31 Dec 2020	WHO issued the emergency use validation for a COVID-19 vaccine, focusing on equitable global access.
29 Jan 2021	WHO publishes their recommended COVID-19 tests (PCR and Antigen).
17 Mar 2021	A statement was made by WHO regarding the AstraZeneca vaccine safety. The reason was reports of rare blood coagulation disorders in people that had recently received the vaccine.
14 Jun 2021	Lockdown extended in England by four weeks due to Delta Variant [18].
18 Aug 2021	US government announces the initiation of booster doses from September [21].
23 Aug 2021	The Pfizer 2-dose COVID-19 vaccine receives full FDA approval [8].
19 Nov 2021	FDA Authorizes boosters of the Pfizer and Moderna COVID-19 vaccines for adults [9].
22 Nov 2021	Austria is the first country in Europe to impose a 'lockdown' both for vaccinated and not [3].
26 Nov 2021	The latest variant called Omicron is detected in South Africa and Botswana [26, 27].

25 Jan WHO released recommendations for use of the Moderna vaccine, while on 29 Jan they publish a list of recommended COVID-19 tests. The most intense general drop in sentiment average over the past year was at the beginning of summer, especially in the week between 12 Jun 2021 and 19 Jun 2021.

Further researching on the information published about COVID-19, it appears that on 14 Jun 2021, the UK announced the extension of the lockdown for four more weeks, based on the threat of the Delta Variant. Since the selected tweets were written in English, it is expected that many users were located in the UK and the US. Therefore, the sentiment could partly reflect their discomfort. The higher sentiment average was noted in the middle of August reaching almost 2.8 with 3 representing a purely positive dataset. Moreover, booster doses were announced in the US on 18 Aug 2021 and the FDA approval of the Pfizer vaccine, on 28 Aug 2021 allowing people to feel more certain that the vaccine is safe and effective. Until 10 Dec 2021, the sentiment was maintained at high levels, with a small drop on 12 Nov 2021. At that period, once again cases started to increase, restrictions to non-vaccinated people were applied in Europe.

4.5 Influential Power

The influence of a tweet is described by several metrics, including 'replies', 'if exists', 'link clicks', 'mentions' and 'retweets' [19]. Based on the findings of a

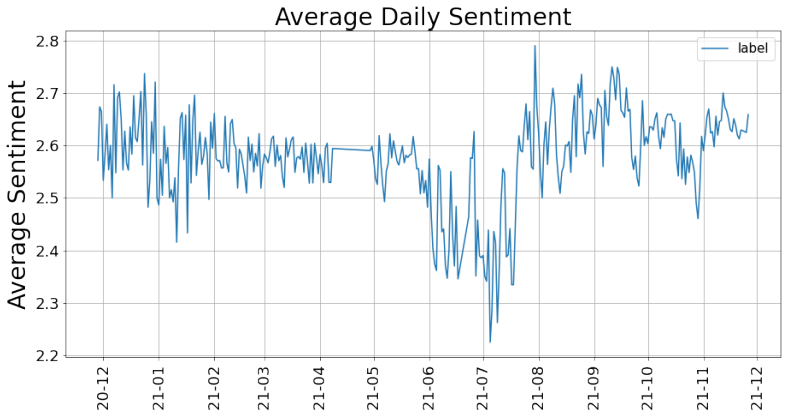

Fig. 4. Average daily sentiment.

similar approach from Medford et al. [17], the tweets with the most retweets were those with a negative context. In this research, their finding can be also be verified. Regarding the retweets per sentiment, some controversial findings occurred. It was found that the number of retweets, compared to the actual number of each sentiment tweets show a different dynamic at each label. It shows that negative tweets tend to have a higher influence and retweet rate per tweet reaching 47.1% with positive having 31.8% and neutral 21.1% respectively. Daily analysis of the sentiments is difficult to be implemented as retweets can happen over several days.

5 Conclusion

5.1 Discussion

During COVID-19, SM were a means for people to express their opinion publicly. It was found that the majority of the users were positive for the course of the pandemic, supporting the work of scientists, governments and WHO. On the contrary, the negative sentiment tends to have a higher engagement rate. The novelty of this work can be attributed to the fact that it investigates the trajectory of the COVID-19 Twitter sentiment over a year and associates it with important events. It also integrates a machine learning approach to evaluate sentiment data from SM. A sentiment timeline is generated to validate certain social behaviors, based on the same features, yet from multiple data sources, i.e. different Twitter datasets. Such an multi-source approach may be employed by officials or researchers, leading to useful knowledge extraction, considering switches in sentiment polarity in SMinitiated by or associated with real-world events.

Moreover, according to our findings, it was observed that people being against the vaccination prefer to share already published thoughts. It can also be concluded that researchers, that base their findings on the work of external datasets

are prone to bias potentially introduced by the original dataset curators. This may happen due to missing dataset descriptions and small volume of daily tweets from a complementary dataset.

All in all, the sentiment of every geographical domain is a multidimensional feature, for which more specialized analysis could be conducted. Each country reacted in a different way during the battle with COVID-19. For that reason, a customised analysis per country could be more effective, while rendering governmental responses to the pandemic waves more easily manageable.

5.2 Future work

This work can be used as a basis and inspiration for future work. Researchers can focus either on a specific geographic location or extend it targeting additional vaccine related keywords. Since the pandemic and its relationship with people is affected by various aspects, such as social, health, political and financial factors, researchers could attempt to extract knowledge based on this research and also considering such parameters. Furthermore, the analyzed dataset might be expanded by incorporating other COVID-19 related search results from Twitter, such as 'Moderna', 'AstraZeneca' vaccines and others. Scientists may also use this work as a point of reference to enhance or doubt their findings on similar attempts. Finally, this work can be imitated or utilized for the extraction of information in any domain of interest, apart from the case of coronavirus.

References

1. Aiello, L.M., Quercia, D., Zhou, K., Constantinides, M., Šćepanović, S., Joglekar, S.: How epidemic psychology works on twitter: Evolution of responses to the covid-19 pandemic in the us. Humanities and Social Sciences Communications **8**(1), 1–15 (2021)
2. Atalan, A.: Is the lockdown important to prevent the covid-19 pandemic? effects on psychology, environment and economy-perspective. Annals of medicine and surgery **56**, 38–42 (2020)
3. BBC: Austria to go into full lockdown as Covid surges (2021), https://www.bbc.com/news/world-europe-59343650
4. Beleveslis, D., Tjortjis, C., Psaradelis, D., Nikoglou, D.: A hybrid method for sentiment analysis of election related tweets. In: 2019 4th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM). pp. 1–6. IEEE (2019)
5. Boon-Itt, S., Skunkan, Y., et al.: Public perception of the covid-19 pandemic on twitter: sentiment analysis and topic modeling study. JMIR Public Health and Surveillance **6**(4), e21978 (2020)
6. Cerbin, L., DeJesus, J., Warnken, J., Gokhale, S.S.: Understanding the anti-mask debate on social media using machine learning techniques. International Journal for Computers & Their Applications **28**(3) (2021)
7. Dhawan: Sentimental analysis of covid-19 tweets | Kaggle, https://www.kaggle.com/dhruvdhawan/sentimental-analysis-of-covid19-tweets/version/1

8. FDA: FDA Approves First COVID-19 Vaccine | FDA, https://www.fda.gov/news-events/press-announcements/fda-approves-first-covid-19-vaccine
9. FDA: Coronavirus (COVID-19) Update: FDA Expands Eligibility for COVID-19 Vaccine Boosters (2021), https://www.fda.gov/news-events/press-announcements/coronavirus-covid-19-update-fda-expands-eligibility-covid-19-vaccine-boosters
10. Garcia, K., Berton, L.: Topic detection and sentiment analysis in twitter content related to covid-19 from brazil and the usa. Applied soft computing **101**, 107057 (2021)
11. Gupta, D.K., Ekbal, A.: Iitp: Supervised machine learning for aspect based sentiment analysis. In: SemEval@ COLING. pp. 319–323 (2014)
12. Koukaras, P., Tjortjis, C.: Social media analytics, types and methodology. In: Machine Learning Paradigms, pp. 401–427. Springer (2019)
13. Koukaras, P., Tjortjis, C., Rousidis, D.: Social media types: introducing a data driven taxonomy. Computing **102**(1), 295–340 (2020)
14. Koukaras, P., Tsichli, V., Tjortjis, C.: Predicting stock market movements with social media and machine learning. In: Proceedings of the 17th International Conference on Web Information Systems and Technologies - Volume 1: WEBIST,. pp. 436–443. INSTICC, SciTePress (2021). https://doi.org/10.5220/0010712600003058
15. Kruspe, A., Häberle, M., Kuhn, I., Zhu, X.X.: Cross-language sentiment analysis of european twitter messages duringthe covid-19 pandemic. arXiv preprint arXiv:2008.12172 (2020)
16. Manguri, K.H., Ramadhan, R.N., Amin, P.R.M.: Twitter sentiment analysis on worldwide covid-19 outbreaks. Kurdistan Journal of Applied Research pp. 54–65 (2020)
17. Medford, R.J., Saleh, S.N., Sumarsono, A., Perl, T.M., Lehmann, C.U.: An "infodemic": leveraging high-volume twitter data to understand early public sentiment for the coronavirus disease 2019 outbreak. In: Open forum infectious diseases. vol. 7, p. ofaa258. Oxford University Press US (2020)
18. Morton, B., Lee, J.: Covid: Lockdown easing in England to be delayed by four weeks (2021), https://www.bbc.com/news/uk-57464097
19. Muñoz-Expósito, M., Oviedo-García, M.Á., Castellanos-Verdugo, M.: How to measure engagement in twitter: advancing a metric. Internet Research (2017)
20. Nemes, L., Kiss, A.: Social media sentiment analysis based on covid-19. Journal of Information and Telecommunication **5**(1), 1–15 (2021)
21. O'donnell, C., Aboulenein, A.: U.S. to begin offering COVID-19 vaccine booster shots in September (2021), https://www.reuters.com/world/us/us-start-offering-covid-19-vaccine-booster-doses-september-2021-08-18/
22. Ritchie, H., Mathieu, E., Rodés-Guirao, L., Appel, C., Giattino, C., Ortiz-Ospina, E., Hasell, J., Macdonald, B., Beltekian, D., Roser, M.: Coronavirus pandemic (covid-19). Our world in data (2020)
23. Rousidis, D., Koukaras, P., Tjortjis, C.: Social media prediction: a literature review. Multimedia Tools and Applications **79**(9), 6279–6311 (2020)
24. Singhal, T.: A review of coronavirus disease-2019 (covid-19). The indian journal of pediatrics **87**(4), 281–286 (2020)
25. World Health Orgainisation: Timeline of WHO's response to COVID-19, https://www.who.int/news-room/detail/29-06-2020-covidtimeline
26. World Health Orgainisation: Tracking SARS-CoV-2 variants (2021), https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/
27. World Health Orgainisation: Update on Omicron (2021), https://www.who.int/news/item/28-11-2021-update-on-omicron