Sentiment analysis of the COVID-19 pandemic using English tweets

Sam Kuilboer[1,2], Roman Peerboom[1,2], Gabor Banyai[1], and Enrico Menarini[1,2]

[1] Vrije Universiteit Amsterdam, Amsterdam, Netherlands
[2] Universiteit van Amsterdam, Amsterdam, Netherlands
sam.kuilboer@student.uva.nl, g.banyai@student.vu.nl,
enrico.menarini@student.uva.nl, roman.peerboom@student.uva.nl
VU.nl

Abstract. The Coronavirus outbreak quickly spread around the globe as well as on social media. Moreover, since the beginning of the pandemic, people tend to spend more time on social media then they ever did before. Twitter partially reflects the social media agenda and provides a great amount of text to be analysed. This paper relies on the the existing Twitter data from the GeoCOV19Tweets Dataset [15] and the US Vaccination data retrieved from the Centers of Disease Control and Prevention [2]. With such data, the current study aims to explore the relation between sentiment score for US states and respective vaccination rate. Exploratory data analyses methods have been used and the results suggest that no clear correlation can be found. Still, when zooming into the top five states with respect to number of geo-located tweets, different online sentiments reflecting their current real-life experience can be found.

Keywords: Covid19 · Sentiment analysis · Twitter · Geo-located tweets · Vaccinationrate.

1 Introduction

Up to today, the coronavirus (COVID-19) pandemic can be considered one of the greatest crises worldwide due to the deaths registered and the psychological and economical damage it caused [11]. According to World Health Organization (WHO), more than 283 million individuals have contracted the disease and more than 5434 thousand individuals have died (WHO — World Health Organization, 2021/12/29). When a crisis occurs, people are keen to spend more time on social media platforms than they would do under usual circumstances [13]. Indeed, this phenomenon could be explained both from a physiological response that people have when they face such crises, as well as the role that social media acquired during the years that made them also active sources of information [14]. Events of everyday life are discussed on social media, and everybody is free to discuss and share their opinions about them and nevertheless, the COVID-19 has been one of the most discussed as well as spreading diseases worldwide [1].

The increased usage of social platforms offers scientists a large amount of data that can be analyzed to understand the social aspect of the virus. Numerous research has shown that tweets related to a crisis can help to understand the social tendencies surrounding the event [8], and the data on the platform is easily accessible for researchers using Twitter's API.

On the other hand, the year 2021 has been characterized by the massive amount of vaccinations that took place all over the world in response to the pandemic. The COVID-19 vaccine is currently the best option available to fight the virus but current hesitancy in undertaking vaccination shots are threatening its effectiveness [17].

This paper aims to visualize an in-depth understanding of the sentiment is displayed in tweets regarding the coronavirus. The purpose of this paper is to analyze the possible difference in sentiment between countries, as the sentiment displayed over time. Finally, the sentiments expressed on Twitter within the United States (US) will be compared with the US people's willingness to take the vaccine shot. To do so, an existing data set of Coronavirus related tweets [15] and official vaccination rates published by the official health agency of the United States, the Centers of Disease Control and Prevention (CDC)[2], will be used.

Hence, the following research question has been formulated: 'Is there a correlation between the sentiment score for Twitter data and vaccination rate in the US from March 2020 to December 2021?'. To answer this question, multiple sub-questions were formulated:

- How is the data-set distributed over the globe?
- How is the data-set distributed within the US?
- How does the sentiment analysis change over time?
- What are the vaccination rates in every state?
- Is there a correlation between the sentiment score and the vaccination rate in a state?

The project is structured as follows: section 2 elaborates on the current literature in the research field. Section 3 shines light on both the data collection, the preprocessing approach and the software used for the project. Section 4 consists of the results which answer the subquestions. Lastly, section 5 discusses the results and compare them with the current literature.

2 Current literature

There is a large amount of studies that investigated Twitter data related to the COVID-19 pandemic. Several studies have been gathering and sharing large-scale data-sets in order to permit further study into the public debate on COVID-19 [15]; the following articles are all relying on the just mentioned data-set.

Some studies, such as Qazi, Imran and Ofli [21], collected Twitter data from all around the world and thus in different languages. With the goal of understanding public sentiment toward COVID-19 vaccines, Monselise et al. [19] analyzed

with a combination of topic detection and sentiment analysis the discussions about vaccines on social media vaccines when the vaccines were started to be administered in the United States. By identifying both positive and negative emotions, their findings imply that administration and vaccination access were among the public's top worries, and that fear was the most common emotion expressed in tweets, followed by joy.

Similarly, Cotfas et al. [10] studied the dynamics of public opinion on COVID-19 vaccination over the course of a month after the first vaccine announcement. Their method consisted in machine learning and deep learning algorithms and their findings indicate that most of the tweets had a neutral stance, while the number of in favor tweets were in greater amount than the number of against tweets.

Lastly, Chopra et al. [9] by using community detection algorithms, analyzed the temporal evolution of different emotion categories (e.g. hesitation) with influencing factors (e.g. vaccine rollout) as lexical categories created from Tweets. Their findings indicate that hesitancy has the highest mentions when talking about vaccines followed by contentment. By observing changes over time, they also found that negative emotions like rage and sorrow gained high importance.

3 Methods

This section elaborates on the methods used to investigate the Lamsal et al. [15] data-set, as well as the gathering of the vaccination rates data.

3.1 Data collection

The data-set consists of tweet ids instead of tweets as may seem unconvincing at first sight. However, this is due to Twitter regulations [5]. Due to the Twitter developer policies, no data-set consisting of actual tweets can be legally found on the web. Every time one wants to make use of Twitter data, new data has to be downloaded. To still be able to work with the same data-set over multiple studies, it is legal to share the IDs of tweets [7]. The same applies to the data-set used in this research [15]. During the research of Lamsal et al., two data-sets were collected over a 5 month period and analyzed. However, after the paper was published, they have kept collecting tweets [15]. As of 8 December 2021, a database of almost 2 years was collected. The first database consists of all tweets within the period of time mentioning one of the words shown in table 3, in the appendix. The second data-set is a subset of the first data-set and consists of only the tweets which have a geo-location hidden in the tweet-object [6] [7]. For this research, only the second is considered. One disadvantage of the second data-set, that it is remarkable smaller than the full data-set. Our research question specific mention the location of the social media, so the use of geo-located tweets is mandatory. This is a fraction of the original data-set.

To be able to make use of the data-set, the tweets had to be so-called hydrated. Hydration is a method that makes use of the Twitter API [4] to gather

all data available of the tweet using its ID as a search term. To do so efficiently the Twarc package was used [3]. Using this package the full tweet and geo object could be collected. By using the hydration method, deleted tweets can not be retrieved, and thus all twitter guidelines are followed.

Once the hydration was fulfilled, the tweets were inspected. One disadvantage of the Twitter set was discovered. The tweet object comes with multiple location statements. However, all locations other than the coordinates within the tweet-object are unreliable. This unreliability is a consequence of the Twitter regulations letting everyone freely decide where they are located, e.g. person X can state in his profile that he lives in California, but actually lives in New York. To overcome this, all coordinates were reverse engineered to a state and a county using the *open street map* library. This library is compared to the google reverse engineer package unreliable, but it is open-source, which the google reverse engineer package is not.

Table 1: The remaining features for every tweet after preprocessing. The example is non-realistic, as this data is not open-source due to twitter regulations.

Feature	Type	Example
Valid_id	string	1249953360375480320
date	date	2021-08-04 18:12:33
country	string	US
lon	obj	-117.566
lat	obj	33.8753
followers	int	301
retweets	int	6
replies	int	3
likes	int	5
quotes	int	0
score	float	0.6
States	String	California
Counties	String	Riverside County

In addition, all meaningless data features were dropped and the sentiment score was added. The final data product left of every tweet is given in table 1.

Two other data-sets were used for the analysis of a possible correlation between the sentiment scores and other variables. One is the data-set on vaccinations in the United States retrieved from the Centers of Disease Control and Prevention [2]. This dataset contains tens of columns of information on vaccinations of every day in the last year for all the American states. We decided to use the column that contained the amount of vaccinations per 100k citizens in the state. The reason for this is that it is a relative number, so this makes it easily comparable. Administered means that a person is vaccinated, so that was the

most relevant to our research. An additional data-set is a Harvard data-set on elections. We used this to get the percentage of democratic voters in each state.

3.2 Sentiment Score

This paper relies on a sentiment score provided by Lamsal et al. [15]. In the articles, the authors explain how they implemented TextBlob's Sentiment Analysis module for the computation of the sentiment score and the preprocessing phase. TextBlob is a Python library for processing textual data and has been chosen since its sentiment analysis model computes the sentiment polarity as a continuous value rather than a category. Indeed, the scores fall in a range [-1,+1], where a score above "0" is considered positive sentiment and the closer is to "+1", the more positive the tweet is considered. On the opposite, a score below "0" is considered negative sentiment and the closer is to "-1", the more negative the tweet is considered. A score of "0" is treated as a neutral sentiment.

3.3 Software

All computations are performed within the python 3.8 environment and were mostly done using the IPython notebook interface. The code is open-source and can be found on GitHub [3]. The maps are generated with geopandas.

4 Results

4.1 Global sentiment analysis

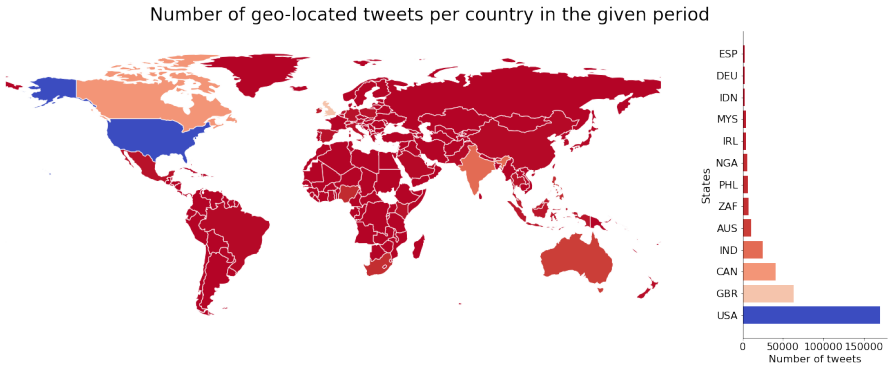


Fig. 1: The number of tweets for every country. The barchart shows only the countries which have more than 2500 tweets. The colors of both graphs match.

[3] https://github.com/Boy0211/SocialWebFinal

Figure 1 shows the number of tweets per country over the total time period. As can be expected for an data-set based on English keywords, most English-speaking countries are over-represented. In Europe only Great-Britain, as was expected, has a high number of tweets. One exception on this is Australia, which has a relative low number of tweets.

To analyze the sentiment of the tweets, the sentiment scores over time and location were visualized. Figure 2 depicts the change of mean sentiment scores over the inspected time frame for the top four countries and globally, while the right y-axis shows the number of new registered Covid cases globally [12]. Some of the 'landmarks' of the virus are clearly reflected on the graph with a sudden increase or decline of the mean sentiment score value. For example the second wave around September-November 2020, the third wave around January-February 2021 and the fourth wave around September-October 2021 are all indicated with a downward change of mean sentiment score.

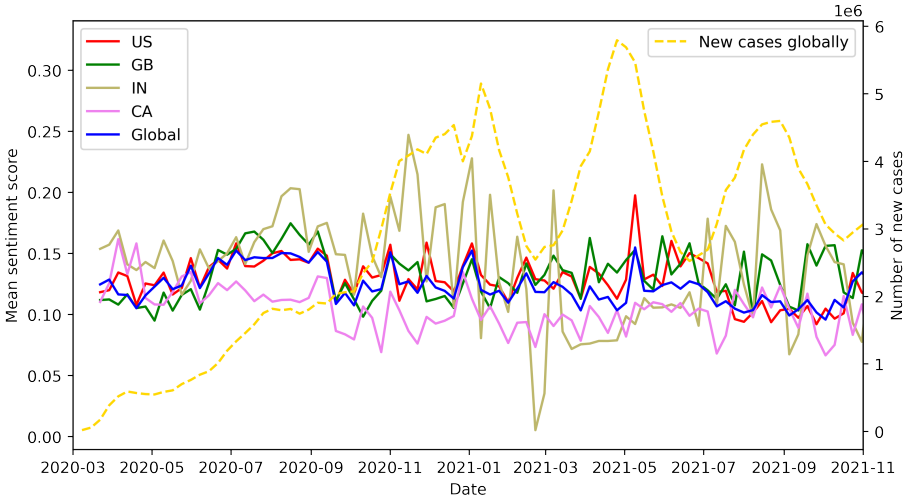


Fig. 2: Mean sentiment scores globally and for the countries United States, United Kingdom, Canada and India. The right y-axis shows the number of new registered Covid cases globally.

Similarly, when the first COVID vaccines have been approved around December 2020, the sentiment of tweets turned more positive. The changes of mean sentiment scores in the US are roughly following the global patterns, with the difference that the US tweets tend to display more extreme values, especially when there is an increase of the mean sentiment score. The most extreme sentiment values are displayed in the tweets originated from India; both on the negative and positive side.

The mean sentiment score globally was 0.130 in the given time frame as can be seen in table 2, while the standard deviation of the sentiment score was 0.253. For the US tweets, the mean sentiment score was 0.134 and the standard deviation was 0.260, which means they do not significantly differ.

Table 2: The mean and standard deviation of the sentiment scores for the countries which had the largest number of tweets.

Country	Mean	St. dev
Global	0.130	0.253
United States	0.134	0.260
United Kingdom	0.136	0.252
Canada	0.106	0.219
India	0.137	0.255

4.2 Sentiment analysis United States

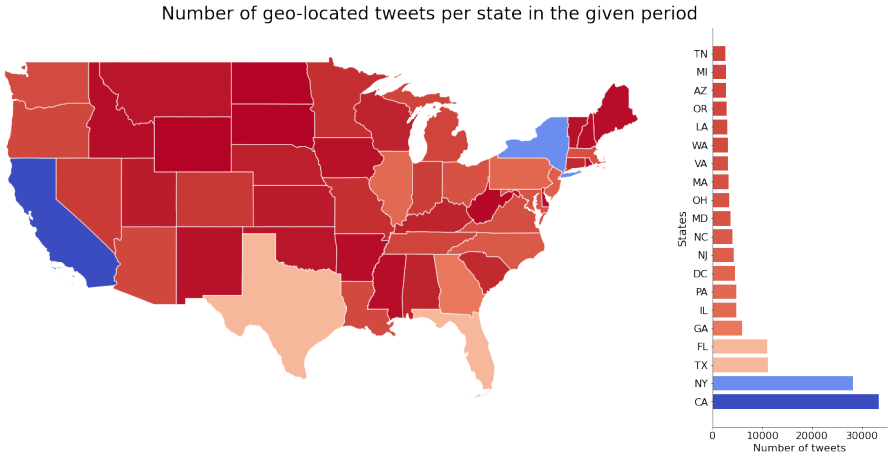


Fig. 3: The number of tweets for every state. The barchart shows only the states which have more than 2500 tweets. The colors of both graphs match.

Figure 3 shows huge differences when it comes to the number of geo-located tweets per state. The four states with the most geo-located tweets are the states which have the most inhabitants, respectively California, New York, Texas and Florida.

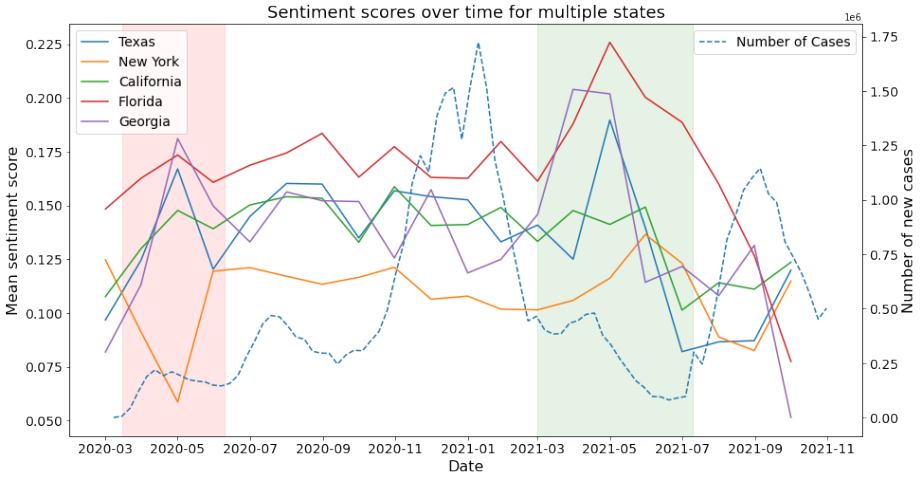


Fig. 4: Sentiment scores for the states Texas, New York, California, Florida and Georgia. The right y-axis shows the number of new registered Covid cases in the United States.

Figure 4 shows the sentiment scores for the five states which have the most geo-located tweets, as shown in figure 3. The mean sentiment scores were calculated per month to prevent lack of clarity. The number of new-registered cases was added to a second y-axis, and shown with a dashed line. When inspecting the figure, two things stand out, shown with the red and green background. From March to June 2020, the red part, all sentiment scores go up except the sentiment score of New York, while the total number of new-registered cases is relative low. To understand this, one has to read the news of March 2020 with respect to New York. New York city back then was tormented with huge numbers of new cases every day, and the hospitals could barely keep up with the incoming cases [18]. The low sentiment score seems to be the consequence of this. The high covid peak from the winter of 2020-2021 does not seem to have an high impact on the sentiment-scores online.

The green part stands out due to its relatively high sentiment scores, followed by a huge drop in sentiment scores. This can be the consequence of the drop of new-registered Covid cases early 2021 and a spring and summer with relatively low numbers of new-registered cases. This is then followed by an huge increase of Covid cases as the result of the new delta-variant, which results in a drop of sentiment score in all states.

To further enhance the insight with respect to the sentiment anlysis the two states with the highest number of tweets were further investigated. Figure 5 shows the states of New York and California, colorized by their sentiment analysis. The figure depicts a clear lower sentiment in the center of New York city which can be explained by the heavy lockdown measures and catastrophic numbers of Covid cases early in the pandemic. However, California does not

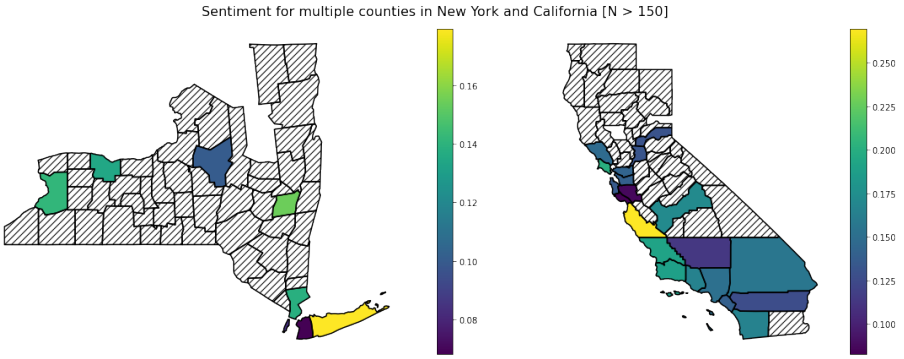


Fig. 5: States of New York and California. The better the sentiment, the more yellow the state colours. The state which are not colourized did not meet the threshold of 150 tweets within the given time period.

show a same pattern. Unfortunately, due the lack of tweets of the more remote areas, a clear comparison is difficult.

Figure 6 shows the development scores for the states with the 10 highest and lowest average sentiment scores. It is visible that states with a high average sentiment score generally have had a high sentiment score during the whole observed period, which makes sense. However, there was a strong dip in sentiment score for these states during the last summer. This dip, that was earlier discussed and was seen in all states, had a stronger profound effect on the states with a high average sentiment effect than states with lower average sentiment.

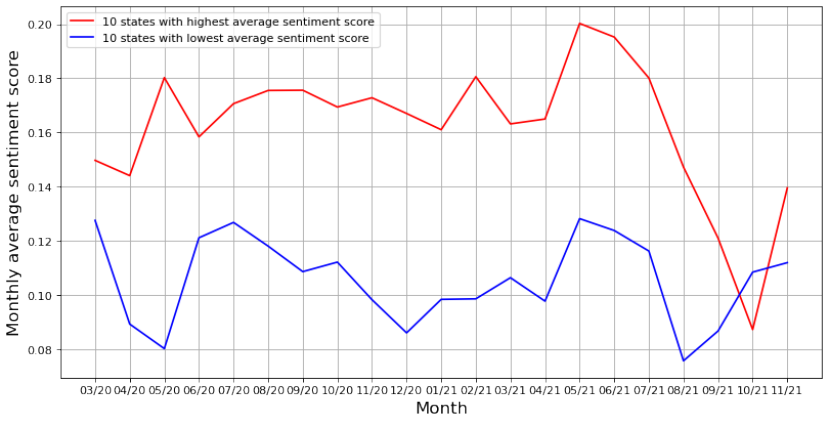
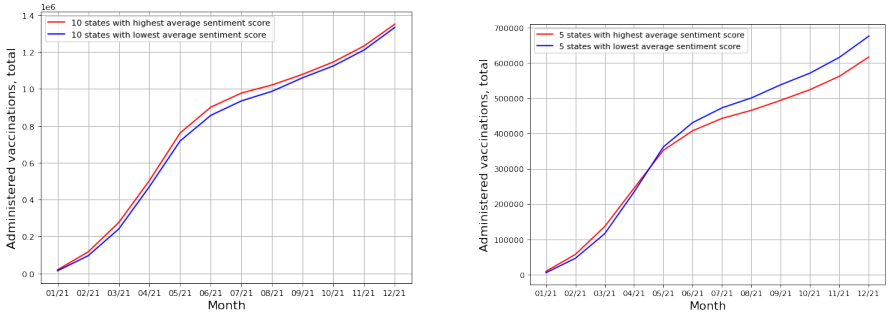


Fig. 6: Development sentiment scores of the 10 states with the highest and the 10 states with the lowest sentiment scores

4.3 Sentiment analysis compared to the vaccination rates

Comparing the sentiment analysis data-set with the CDC data-set of the vaccination rates gave some interesting results. Our expectation was that a low sentiment score in a US state would correlate with a low vaccination rate. The reasoning for this was that a low sentiment score would mean that people generally have negative opinions on vaccinations and thus are less likely to take it. However, this was not the outcome of our analysis. Figure 7 shows that there is barely any difference in the development of the vaccination rate between states with the highest and lowest average sentiment scores. The 10 states with the highest sentiment score had a higher vaccination rate for a couple of months, but the 10 lowest states have caught up in recent months. When comparing the states with the 5 highest and lowest sentiment scores, the reverse of what was expected is visible. The states with the highest sentiment score have a significantly lower vaccination rate. This gap starts to arise in June and July of 2021.



(a) The top and lowest 10 States (b) The top and lowest 5 States

Fig. 7: Vaccination rates over time of multiple states with the lowest and highest average sentiment scores

Republicans are more likely to spread anti - vaccine misinformation, according to [20]. Also, Republican states generally show a greater hesitancy to vaccinations than Democratic ones [16]. Because of this, it was interesting to see if the data shows that states with a lower sentiment score are more Democratic or Republican. Figure 8 shows that contrary to expectation, the states with a lower sentiment score generally have more Democratic voters. This is based on the 2020 Biden/Trump presidential election.

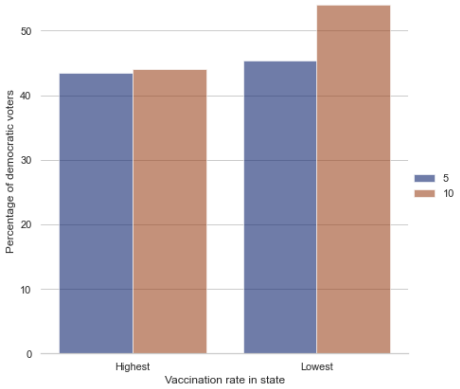


Fig. 8: Percentage of democratic voters in the states with the highest and lowest vaccination rates. The blue bars show the results of the 5 highest or lowest states, the orange bars of the 10 highest or lowest.

5 Discussion

Just four years ago the online policies about social web data were a lot less strict. The most famous example of this, are the elections project done by Cambridge Analytica. By exploiting online information of people in the United States they were able influence the elections. As a consequence, Facebook, but also Twitter, made stricter policies about downloading their social data, which was also a step back for the science in this field as using huge numbers of social web data became more difficult. This horde was also found in this research. Finding the proper data-set which corresponds with the purpose of a research like this was difficult. The research makes again clear that only a small amount of tweets is geo-located and using it as representative for the total sentiment with respect to Covid on twitter is hard. This corresponds with the current literature stating that geo-located tweets do not represent the total population of tweets [22].

However, still some interesting results were found. When zoomed in into the top five states with respect to number of geo-located tweets, different online sentiments reflecting their current real-life experience were found. Especially, figure 4 shows the effect of the huge Covid peak in one of the biggest cities of the United States, New York. The sentiment in March-April 2020 was significantly lower than the other states, which were not fighting such a enormous peak as the state of New York.

Furthermore, the analysis shows a dependency of the sentiment scores on the current covid situation in a certain area. However, due to lack of number of twees its unclear how strong this dependency is. Further research should focus on finding more tweets which can be made geo-located. For example, using network analysis to find close contacts of the users which are already geo-located.

When comparing the sentiment scores of covid to the vaccination rates, no clear correlation can be found. As described before, the sentiment scores are more dependent on the direct covid-situation in a certain area. As vaccination is more of long-term influence, we debate that the influence on the sentiment is relatively low.

Future research should focus on a more specific data-set for a more smaller region to create solid conclusion about the online sentiments with respect to Covid. By doing so, one can conclude more about the current sentiment with respect to Covid in a certain region. This could, for example, say something about the willingness of the people to follow lockdown measurements and make governments able to counter negative sentiments online. However, by zooming in, the privacy of the online-users is at stake. Whether it is ethical to at large scale influence people online is up to debate and not for researchers to decide. It is the task of the researchers bring forth the data to discuss it and ignite this debate.

References

1. Twitter Sentiment Analysis on Worldwide COVID-19 Outbreaks. Kurdistan Journal of Applied Research pp. 54–65 (2020). https://doi.org/10.24017/covid.8
2. Center of disease control and prevention (2021), https://www.cdc.gov/
3. Twarc (2021), https://twarc-project.readthedocs.io/en/latest/
4. Twitter: Api (2021), https://developer.twitter.com/en/docs/tweets/search/overview
5. Twitter: Developer agreement and policies (2021), https://developer.twitter.com/en/developer-terms/agreement-and-policy
6. Twitter: Geo objects (2021), https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/geo-objects
7. Twitter: Object (2021), https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/tweet-object
8. Carley KM, Malik M, L.P.P.J.K.M.: Crowd sourcing disaster management: The complex nature of twitter usage in padang indonesia. Safety Science **90**(1), 48–61 (2016). https://doi.org/https://doi.org/10.1016/j.ssci.2016.04.002, https://doi.org/10.1016/j.ssci.2016.04.002
9. Chopra, H., Vashishtha, A., Pal, R., Ashima, Tyagi, A., Sethi, T.: Mining Trends of COVID-19 Vaccine Beliefs on Twitter with Lexical Embeddings (2021), http://arxiv.org/abs/2104.01131
10. Cotfas, L.A., Delcea, C., Roxin, I., Ioanăş, C., Gherai, D.S., Tajariol, F.: The Longest Month: Analyzing COVID-19 Vaccination Opinions Dynamics from Tweets in the Month following the First Vaccine Announcement. IEEE Access **9**, 33203–33223 (2021). https://doi.org/10.1109/ACCESS.2021.3059821
11. Fenwick, M., McCahery, J.A., Vermeulen, E.P.M.: Will the World Ever Be the Same After COVID-19? Two Lessons from the First Global Crisis of a Digital Age. European Business Organization Law Review **22**(1), 125–145 (2021). https://doi.org/10.1007/s40804-020-00194-9, https://doi.org/10.1007/s40804-020-00194-9
12. Hannah Ritchie, Edouard Mathieu, L.R.G.C.A.C.G.E.O.O.J.H.B.M.D.B., Roser, M.: Coronavirus pandemic (covid-19). Our World in Data (2020), https://ourworldindata.org/coronavirus

13. Imran, M., Castillo, C., Diaz, F., Vieweg, S.: Processing social media messages in mass emergency: A survey. ACM Comput. Surv. **47**(4) (jun 2015). https://doi.org/10.1145/2771588, https://doi.org/10.1145/2771588

14. Imran, M., Ofli, F., Caragea, D., Torralba, A.: Using AI and Social Media Multimodal Content for Disaster Response and Management: Opportunities, Challenges, and Future Directions. Information Processing Management **57**(5), 102261 (2020). https://doi.org/https://doi.org/10.1016/j.ipm.2020.102261, https://www.sciencedirect.com/science/article/pii/S0306457320306002

15. Lamsal, R.: Design and analysis of a large-scale COVID-19 tweets dataset. Applied Intelligence **51**(5), 2790–2804 (2021). https://doi.org/10.1007/s10489-020-02029-z

16. Liu, R., Li, G.M.: Hesitancy in the time of coronavirus: Temporal, spatial, and sociodemographic variations in covid-19 vaccine hesitancy. SSM-population health **15**, 100896 (2021)

17. Machingaidze, S., Wiysonge, C.S.: Understanding COVID-19 vaccine hesitancy. Nature Medicine **27**(8), 1338–1339 (2021). https://doi.org/10.1038/s41591-021-01459-7, https://doi.org/10.1038/s41591-021-01459-7

18. McKingley, J.: New york city region is now an epicenter of the coronavirus pandemic. The New York Times https://www.nytimes.com/2020/03/22/nyregion/Coronavirus-new-York-epicenter.html

19. Monselise, M., Chang, C.H., Ferreira, G., Yang, R., Yang, C.C.: Topics and Sentiments of Public Concerns Regarding COVID-19 Vaccines: Social Media Trend Analysis. Journal of medical Internet research **23**(10), e30765 (oct 2021). https://doi.org/10.2196/30765, https://europepmc.org/articles/PMC8534488

20. Motta, M.: Republicans, not democrats, are more likely to endorse anti-vaccine misinformation. American Politics Research **49**, 1532673X2110226 (06 2021). https://doi.org/10.1177/1532673X211022639

21. Qazi, U., Imran, M., Ofli, F.: Geocov19: A dataset of hundreds of millions of multilingual covid-19 tweets with location information. SIGSPATIAL Special **12**(1), 6–15 (jun 2020). https://doi.org/10.1145/3404820.3404823, https://doi.org/10.1145/3404820.3404823

22. Sloan, L., Morgan, J.: Who tweets with their location? understanding the relationship between demographic characteristics and the use of geoservices and geotagging on twitter. PloS one **10**(11), e0142209 (2015)

Appendix

Table 3: The selected keywords, decided by Lamsal et al.[15], which were used to gather the data-set.

Keywords				
#corona	coronavirus	#coronavirus	covid	#covid
#covid19	covid-19	#covid-19	sarscov2	#sarscov2
sars cov 2	covid_19	#covid_19	2019ncov	ncov
ncov2019	2019-ncov	#2019-ncov	pandemic	#pandemic
#flatteningthecurve	quarantine	#quarantine	flatten the curve	flattening the curve
#flattenthecurve	hand sanitizer	#handsanitizer	#lockdown	lockdown
covid19	sars cov2	#ncov2019	#2019ncov	#ncov
social distancing	#socialdistancing	work from home	#workfromhome	working from home
#workingfromhome	ppe	n95	#ppe	#n95
#covidiots	covidiots	herd immunity	#herdimmunity	pneumonia
#pneumonia	chinese virus	#chinesevirus	wuhan virus	#wuhanvirus
kung flu	#kungflu	wearamask	#wearamask	wear a mask
vaccine	vaccines	#vaccine	#vaccines	corona vaccine
corona vaccines	#coronavaccine	#coronavaccines	face shield	#faceshield
face shields	#faceshields	health worker	#healthworker	health workers
#healthworkers	#stayhomestaysafe	#coronaupdate	#frontlineheroes	#coronawarriors
#homeschool	#homeschooling	#hometasking	#masks4all	#wfh
wash ur hands	wash your hands	#washurhands	#washyourhands	#stayathome
#stayhome	#selfisolating	self isolating		

6 Contributions

Table 4: Contributions to the assignments throughout the course.

What	Sam	Roman	Enrico	Gabor
Notebook 1	x		x	x
Notebook 2		x	x	x
Notebook 3	x	x		x
Notebook 4		x		
Paper presentation	x		x	x

6.1 Contribution Sam

The notebooks were done with the guys available on Thursday. For the first two, I was unavailable. The others did not find a solution of one of the assignments in

Table 5: Contributions to the final assignment.

What	Sam	Roman	Enrico	Gabor
Discussing approach	x	x	x	x
Finding the proper dataset	x	x		
Hydrating tweet dataset	x			
Reverse geocode dataset	x			
Analysis countries	x		x	x
Analysis states & counties	x			
Analysis vaccination & votes		x		
Pitch presentation			x	
Final presentation		x	x	
Writing introduction	x	x	x	x
Writing current literature			x	
Writing methods	x	x	x	
Writing results	x	x		x
Writing disucssion	x			
Cleaning up code and repo	x			
Finalizing	x			x

the first notebook so I finished it afterwards. Together, with Gabor and Roman I finished the third on campus.

When it comes to the final assignment, the focus was on comparing covid vaccinations with fake news. I primarily focused on finding the proper dataset, which was difficult. I finally had to switch to sentiment analysis, because fake-news sets were impossible to find.

I wrote the approach for the paper. Writing down the exact structure and using template figures for the sections 'Global sentiment analysis' and 'Sentiment analysis United States' sections, to help the team. Next to this, I wrote part of the introduction and most of the method. This was all checked by Enrico. Subsequently made figures 1, 3, 4 and wrote the accompanying text. Finally I wrote the discussion and made templates for the contributions part. To finalize, I reread the whole paper and added comments to the part of which I thought had still to be improved. The exact code contributions can be found at github [4]. Be aware, this is a bit outbalanced due the fact that I uploaded part of the datasets. Afterwards I cleaned up the code and repo and submitted the assignment.

6.2 Contribution Roman

I was busy for another course so I did not work on the first notebook. The next two were group efforts in which I was involved. The last one I did by myself because the other guys were working on the paper presentation or our own research.

[4] https://github.com/Boy0211/SocialWebFinal/graphs/contributors

For the research paper, I mainly worked on the analysis of the correlation between vaccination rates and the sentiment scores in states of the US, the last part of our results section. For this, I also did some data preprocessing to link the sentiment dataset with the vaccination rate and election datasets. I also wrote about this in the paper, in addition to some stuff about the preprocessing, writing about figure 6, and an alinea in the Discussion.

6.3 Contribution Gabor

I took part in the work for the first three notebooks, while for the last notebook we decided to split the course work due to time limitations. Roman did the last notebook by himself, and me and Enrico focused on the research paper that our group was assigned to and we put together a presentation about it. For the Covid research paper, I took part in the brainstorming sessions where we clarified the scope of our research, our approach and helped in defining our research questions. I wrote the 'Global sentiment analysis' section and next to this, I made figure 2 which depicts the distribution of mean sentiment scores over time globally and for the countries with the most tweets. I proofread those parts that my teammates wrote and made sure that the document is consistent.

6.4 Contribution Enrico

When it comes to the notebooks, I managed to work on the first two. I was not able to work on the third due to personal circumstances and for the last notebook, we decided to split the work in such a way that I would have dealt with the paper presentation with Gabor. Indeed, Sam, Gabor and I worked and presented the assigned paper. Still, all the works have been shared among the group in order to learn what was done by the others.

On the other hand, for the research paper I worked partially on the analysis. My coding skills were not at the level of my peers and thus I was only able to provide basic descriptive data from the GeoCOV19Tweets Dataset (both at global and US levels) with the help of Gabor. However, I focused on the initial part of cleaning the datasets for the analyses that took most of my time.

When it comes to the writing, I wrote myself the abstract and the current literature. With the help of Sam, I also wrote the introduction. For the method part, my only contribution was to write the sentiment score section to explain the scores and reference the paper from which we used that data. Lastly, I took care of the video presentation. Indeed, I decided the main topics to include but I used a good help from all of the team members to finalize it. Together with Roman we presented the current paper.