Visual Analytics Report

Nicolò Palmiero, Gabriele Marcozzi

September 2021

Contents

1 Introduction

In this report we will describe various visualization techniques realized in order to study the trend of coronavirus tweets all over the world. We decided to focus on Covid-19 tweets trend because an analysis of this data can be useful to different kinds of users that want to understand the trend of tweets grouped by nation and related to their time of publication. Moreover, it could be a good starting point in order check whether there exists a correlation between number of covid_tweets and number of covid_infections. In fact, data analysts could use this project in order to compare our visualization with the Covid-19 epidemic curves.
In order to do that, we needed to find a way to represent information such that the Covid-19 tweet trend could be easily plotted, analyzed and understood, according to the nation and the selected time period.

Therefore, we realized five types of visualization:

- A geographical heat map of the word is used to visualize, through a colorscale, the number of tweets made on each country in a specific time period. The map is interactive: it allows the user to select countries by clicking on the countries themselves in order to show their details.
- A plot showing the tweets trend of specific countries (both the ones selected from the map and the non selected ones) in a specific period. The plot will be interactive: it will be possible to show/hide the trend of each country. This will affect the selected countries in the heatmap. From this view it will be also possible to visualize the plot resulting from the monthly sum and/or average of the tweets of the selected countries.
- A word cloud visualization which shows the most common words used by people in their Covid-19 tweets, in a selected period. In this way the user can easily recognize what the top trending terms in the selected period were and can check if, according to the period these words change.
- A scatter plot representing 6-dimensional data gathered by nation and projected in two dimensions through MDS. Each point in the scatter plot represents a different nation, and its coordinates are the result of 6 features: average tweet length on the number of tweets, number of tweets, average number of retweets on the number of tweets, average number of friends of the author of the tweet on the number of tweets, country population and country cities. These 6 features can represent, almost in part, the engagement of Covid-19 topic during this difficult period. Through this plot the user can have an idea of how similar the behaviour of twitter users from different nations was.
- A plot showing the world tweets trend from March 2020 to December 2020. Here it is possible to perform brushing to select a specific time interval. The brushing operation not only selects the desired interval of time on this

plot, but also for other visualizations, that in fact show only data from the selected period.

2 Related work

Throughout the development of the project, to have an idea of which kinds of visualizations to use, we analyzed the literature, studying papers on Covid-19 tweets domain.
For what concerns some of the employed views, we took inspiration from **Visualizing the Covid-19 Twitter chatter dataset for scientific use** [1].
Here, the authors capture the trend of the tweets related to Covid-19 by showing different views: An heat map showing the different number of tweets according to the country (the darker the color of the country, the higher the number of tweets) and a bar-chart showing the comparison of number of tweets by country in a bar view. We took their idea of showing the number of tweets related to a single country, to make comparison among these numbers and extract useful knowledge from this process.
So, we thought about visualizing our Dataset with a similar heat map, and we decided to implement a trend graph which shows the plot of each selected nation according to the number of tweets and the time interval.
So, our approach is slightly different from the one of the related work. We decided to replace the bar chart (which shows for each country the total number of tweets), with the trend bar, in which you can compare countries according to a selected time period. So, with our representation, the graph is interactive and it changes according to the time interval.
In a first moment we tried to implement a day-by-day representation of the data, but what came out from this experiment was that data on our dataset was too fine grained and the final result could appear very confusing to the final user. For this reason we decided to show the data through a moth-by-month representation, that allows to aggregate more data and show more regular trends than a day-by-day representation, at the same time this choice guarantees also a good understanding of the changes involved in the process that we are analyzing.

Design and analysis of a large-scale COVID-19 tweets dataset [2] was of great inspiration for representing the total number of tweets in a month-by-month representation. The paper shows the world-wide trend of Covid-19 tweets from March to July. Starting from this paper, we decided to implement an interactive trend-graph which shows the total number of tweets in each month (from March 2020 to January 2021). Moreover, we decided to implement the brushing operation, in which we can focus to a particular time period.

For what concerns the visualization of the tweets most common words, the word cloud, we took inspiration from **Topics, Trends, and Sentiments of Tweets About the COVID-19 Pandemic: Temporal Infoveillance Study** [3], which examines key themes and topics of English-language COVID-19–related

tweets posted by individuals and explores the trends and variations in how the COVID-19–related tweets. In our project, we decided to show the most common words in tweets according to the selected time period, and try to see if these trending words change over time or remain nearly the same.

After reading this article on **Public risk perception and emotion on Twitter during the Covid-19 pandemic** [4], we thought about adding another visualization technique. In particular we added a MDS to show the proximity of countries based on more than two parameters like the time of publication and the number of tweets, and see if these new parameters could add more information to what we discover using only the number of tweets or not.

3 Dataset

The **Coronavirus (Covid-19) geo-tagged tweets dataset** [5] from IEEE-DataPort contains information from March 2020 to January 2021 about covid-19 related tweets from all over the world.

3.1 Data Preprocessing

The original dataset contains 278425 tweets with the following attributes: *coordinates*, *created_at*, *retweet_count*, *text*, *user_friends_count*, *id*, *id_str*, *truncated*, *display_text_range*, *entities*, *source*, *in_reply_to_status_id*, *in_reply_to_status_id_str*, *in_reply_to_user_id*, *in_reply_to_user_id_str*, *in_reply_to_screen_name*, *user*, *geo*, *place*, *contributors*, *is_quote_status*, *favorite_count*, *favorited*, *retweeted*, *possibly_sensitive*, *lang*.

We decided to preprocess the dataset to shrink the dimensions and use only the useful data for our work. Initially we decided to drop the following columns: *id*, *id_str*, *truncated*, *display_text_range*, *entities*, *source*, *in_reply_to_status_id*, *in_reply_to_status_id_str*, *in_reply_to_user_id*, *in_reply_to_user_id_str*, *in_reply_to_screen_name*, *user*, *geo*, *place*, *contributors*, *is_quote_status*, *favorite_count*, *favorited*, *retweeted*, *possibly_sensitive*, *lang*, because we thought they were not so meaningful for our intended analysis.

Moreover, a lot of them contained sparse data (i.e. not every data entry had a value for those attributes). In particular, we decided to remove all those entries which had no *created_at* value and no *coordinates* value. We decided to do that because we cannot know where and when that specific tweet has been written.

Finally, we decided to create another column called *country_id*, in which, starting from the coordinates, we calculated the country in which the coordinates fall within.

. We added two more columns from two different datasets that are *popula-*

tion, that is the number of people that lives in each country, adn *cities* that is the number of cities of each country, to have more contextual informations to conduct an analysis. At the end of this preprocessing final dataset is composed by 7 columns and 274242 rows. So respecting the AngeliniSantucci index with AS = 1,919,694.

4 Visualization Techniques

4.1 World Map

The world map boundaries are implemented parsing a geojson through D3js and drawing these boundaries scaling them to the dimensions of the monitor. The world map allows to data through a color scale. Each color represents the number of tweets of that specific country: the darker the color, the higher the number of tweets. the entire map is interactive: when the user points with the mouse to a specific nation a tooltip, that contains the name of the nation and the precise number of tweets, is shown. By clicking on a country, this is selected, and more details are shown in the different views of the page.

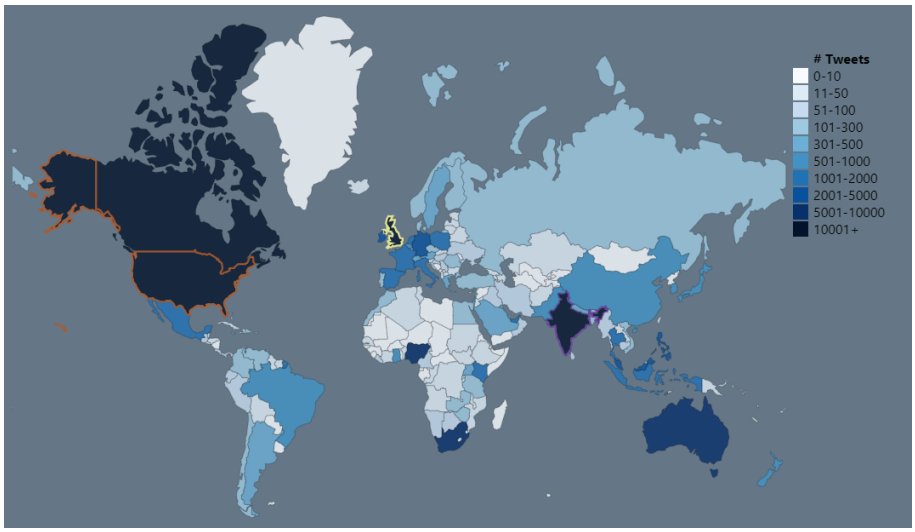

Figure 1: A figure of the world map representing the tweets

4.2 Trend Nation Plot

The trend nation plot shows the tweets trend of specific countries (both the ones selected from the map and the non selected ones) in a specific period (month-by-month).

The X-axis of the cartesian plane represents the month in which a particular number of tweets has been written.

The Y-axis shows the number of tweets.

This view is interactive: we can select/un-select the plot of a specific country by clicking on the plot itself. At the same time, the country in the world map will be highlighted/un-selected.

In each plot, for each month, a dot is shown. If the user goes with the mouse over one of these dots, a tooltip appears and it shows the number of tweets made in the dot's corresponding month in the trends country.

The aim of this view is to compare the number of tweets of different countries in a particular time interval.

It is also possible to compute at runtime the plot of the monthly sum and/or average of the selected nations. In this scenario the user can compare those calculated plot (sum and average) with the plots of other nations.

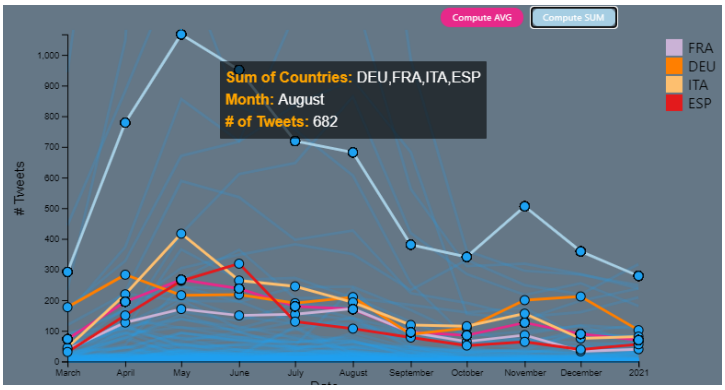

Figure 2: A figure of trend nation plot, showing the trend of the tweets made in Italy, France and Spain from march 2020 to January 2021. The average plot is shown in red. The sum plot is shown in dark blue. All the other countries are shown in light blue.

4.3 General Trend Plot with brushing features

The general trend plot allows to visualize the general trend of Covid-19 tweets from March 2020 to January 2021. For each month, a dot is displayed in order to obtain through a tooltip the total number of twets of that particular month.

From this graph it is possible to select specific time intervals through the brushing technique. After having performed the brush, all the other views will be updated showing just the data which belong to that specific period.

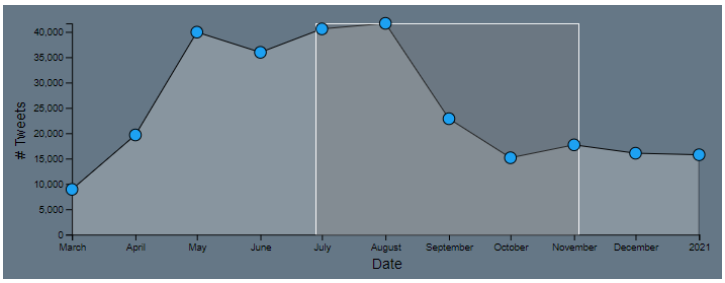

Figure 3: General Covid-19 trend plot with brush from July to November

4.4 Word Cloud

The word cloud is a visualization that permits to show the 10 most used keywords in the selected period of time between the tweets dataset. The most used words are bigger in the visualization, while the other words become smaller as the occurrencies of the target word decrease. To underline the importance of more frequent words a colorscale is also used, that goes from a deep blue for the least occurring word, while goes to a dark red for the most appearing word during the selected period. This colorscale was found using *Colorbrewer* (https://colorbrewer2.org/) which is a tool online that suggests different colorscales depending on how many classes you need to handle and many other factors, like colorblindness-friendly. We decided to implement also a button that activates the computation of this wordcloud, instead of computing directly when the brushing on the time trend plot is made. This choice was made because the computing time of the word cloud is a bit slow and a user could not be interested to visualize the top trending words for the selected period right after he made brushing. When the button is clicked it recomputes the wordcloud plot according to the last interval of time that was selected by brushing on the time trend plot. Through this visualization what a user can immediately notice is that the top trending words are nearly always the same, regardless from the selected period, and this is an important information for our analysis. The top trending words are obviously "Covid-19", "Coronavirus", "Covid", that for the purposes of our analysis are meaningless, but right after thees three words that

are simply the name of the topic that we are treating some words like "lockdown" and "quarantine" appear. This could mean that most of these tweets are not directly related to health issues caused by Covid-19, but to the difficult situation that people was forced to live not going to work and staying at home.

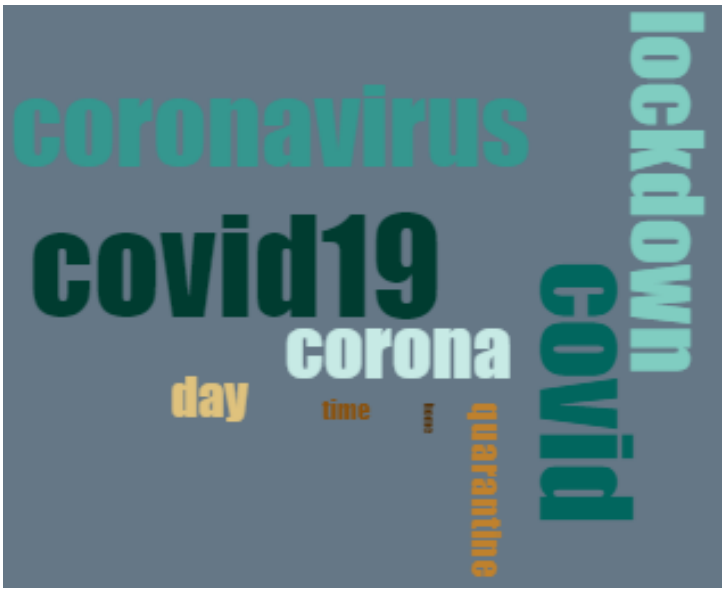

Figure 4: WordCloud plot, showing the most used words during the whole period of our analysis

4.5 Multidimensional scaling

Multidimensional scaling plot is the one on the left of the wordcloud visualization on the dashboard that we have created. This plot can be very meaningful to the goals of our project, because while the other plots, in particular the ones on the upper part of the screen analyze the tweets only from a quantitative point of view, this plot carries more kinds of information to the sight of the user and analyze the correlation between each nation for what concerns the Covid-19 tweets. This plot is a simple two-dimesional scatterplot that, through Multidimensional scaling, achieves the goal of showing six-dimensional data in a user friendly way. This six-dimensional data is obviously gathered for nations, in fact the points in the scatterplot will be always 173 (The number of nations that we analyzed) regardless of the selected time period. Each point represents a nation, and each nation has these four features: $Number\ of\ tweets$, $\frac{Number\ of\ retweets}{Number\ of\ tweets}$, $\frac{Number\ of\ followers}{Number\ of\ tweets}$, $\frac{Cumulative\ tweet\ length}{Number\ of\ tweets}$,

$\frac{Number\ of\ tweets}{countrypopulation}$, $\frac{Numberoftweets}{countrynumberofcities}$. This data is meaningful because the number of retweets and the number of followers of a user are indices of how much engagement the tweet could do, while the cumulative tweet length is a parameter that if compared with the number of tweets can measure, almost in part how much time a user spent on typing his tweet, understanding how much a nation is involved in Covid-19 topic.

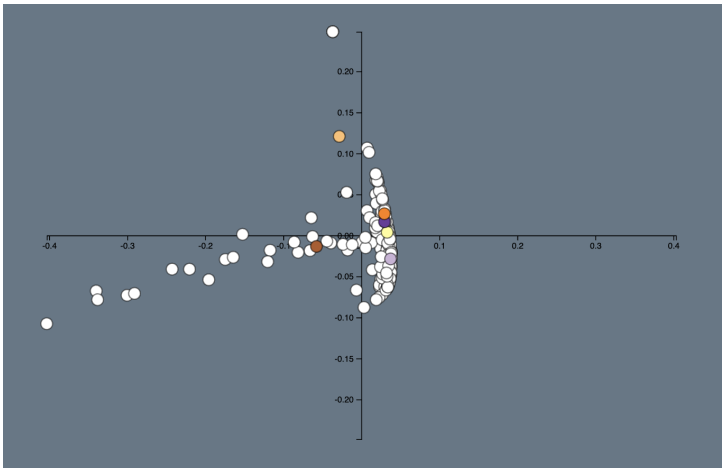

Figure 5: MDS plot, showing the correlation between each nation using six dimensional data, projected in two dimensions

5 Case studies

5.1 Analysis of number of tweets correlated to the population

One important insight that it is possible to carry out from this project is that the number of tweets has to be always related to the number of inhabitants of each nation. This was the role of the MDS plot. Let's take a look on the case of the USA: one could think that USA has by distance the highest number of tweets in the world, so the covid pandemic hit them hard also as a trending topic. But this conclusion is not completely true, although they are the first nation for number of infections we have always to compare these data with the number of inhabitants. If we do that as we have done in the MDS we can see that the USA are not an outlier, but are near other nations that have twitted less. From this analysis we can conclude that we have always to confront raw data with number of inhabitants of a nation before arriving to a conclusion.

5.2 Temporal correlation between tweets and Covid-19 cases

Another case study that one user could attempt to resolve using these views is the temporal correlation between Covid cases and number of tweets. From this analysis a user could verify if there is an effective correlation between people's perception of the pandemic and the effective curve of infection. Let's start by selecting our whole period of study, from March 2020 to January 2021.

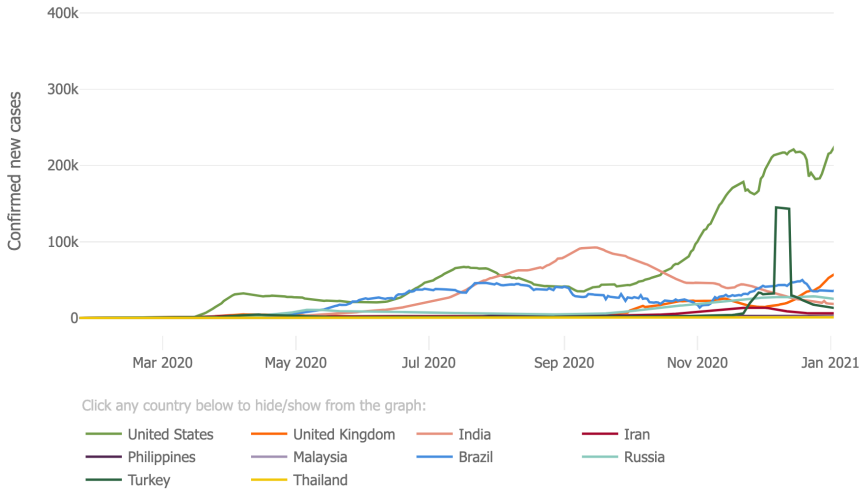

Figure 6: Covid cases curve from March 2020 to January 2021

From these two curves it stands out how much people at the beginning of the pandemic were scared by this virus although the number of the infections was relatively low, while the number of tweets increased fast until touching its maximum in August 2020. The maximum of this tweets function is probably due to the incoming of the second wave of the pandemic, that was pre-announced right in August 2020, in fact we can see that from this date the curve of infections starts increasing rapidly, while the curve of the number of tweets decreases. From this kind of analysis we evince that people was worried about the next few months rather than the present, this was probably due to the situation of uncertainty that this new pandemic introduced.

5.3 Spatial correlation between tweets and Covid-19 cases

An interesting case study is to study how the spread of Covid-19 in different countries is reflected in the amount of tweets posted in each country. For this study, we take the ranking of total Covid-19 cases as a reference (https://www.

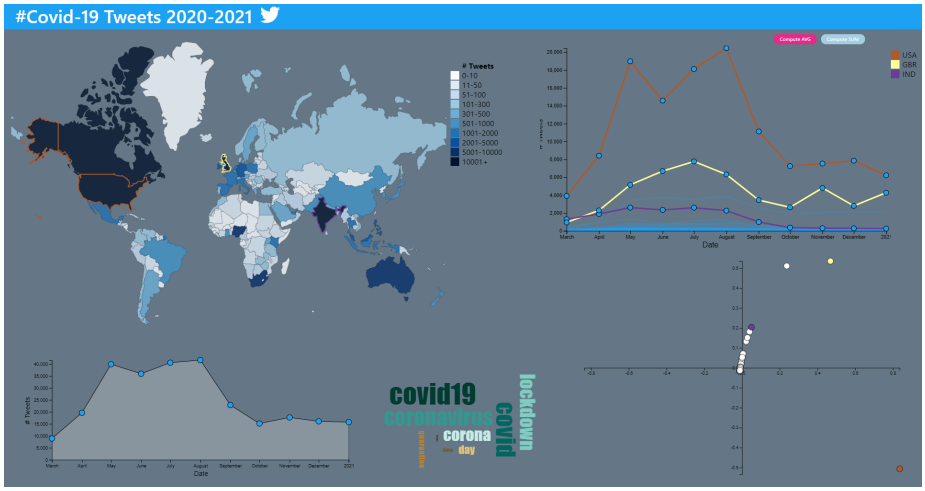

Figure 7: Tweets curves from March 2020 to January 2021

worldometers.info/coronavirus/). Comparing the above ranking with the world map, we can notice the following:

- First, what we can notice is that the country with the highest number of tweets, the USA (124,414 tweets), is also the country with the highest number of total cases of Covid-19.
- Obviously, there is not a 1:1 ratio between the world map and the Covid-19 total case rankings. A discordant example may be Brazil: despite being third in the world for covid-19 cases, the total number of tweets barely touches 600. This can be justified by the fact that technological progress in Brazil is certainly slower than in other countries.
- England is the european country with the most covid-19 cases, and it is also the european country with the highest number of tweets. Also, what we can notice from our project, is that the sum of all tweets from all European countries is less than the number of tweets made in England. This is because, as we know, the English variant of covid-19 made a lot of headlines as soon as it was discovered.
- In addition, we can also see a correlation with India. It stands second in the ranking of total covid-19 cases, while it is the fourth nation with the highest number of tweets. Again, the reason may lie in the fact that the Indian variant of covid-19 has caused many deaths and caused a lot of fear among the population.

So, in summary, the countries with the highest number of tweets are more the ones where new variants have been discovered:

- The 'English variant' or 'Alpha' was first detected in November 2020. Analyzing the trend of tweets in England month by month, we can in fact see an increase in the number of tweets precisely in the month of November.
- The 'South African variant' or 'Beta' was first detected in December 2020. Analyzing the trend of tweets in South Africa month by month, we can in fact see an increase in the number of tweets precisely in the month of November/December.
- The 'Brazilian variant' or 'Gamma' was first detected in January 2021. As previously mentioned, the number of tweets in Brazil does not reflect the true extent of damage caused by Covid-19. Nevertheless, it is also true that the Brazilian variant started to spread in 2021, while our reference Dataset stops at the first month of 2021.
- The 'Brazilian variant' or 'Gamma' was first detected in December 2020.

Covid-19 variants information taken from *Istuto Superiore di Sanità* (https://www.iss.it/cov19-faq-varianti)

6 Conclusion

We introduced this visual analytics tool to support the analysis of Covid-19 tweets during the pandemic.
This tool offers to a user a global view of all Covid-19 tweets trend from all over the world divided by countries from March 2020 to January 2021. He can investigate how Covid-19 trend is related to the tweets during that period, checking both the temporal and spatial correlations.

The work could be expanded including also an animated bar chart showing how the tweets number for each country is evolving month by month. Furthermore in a future work we could include a dataset composed of tweets published in 2021.

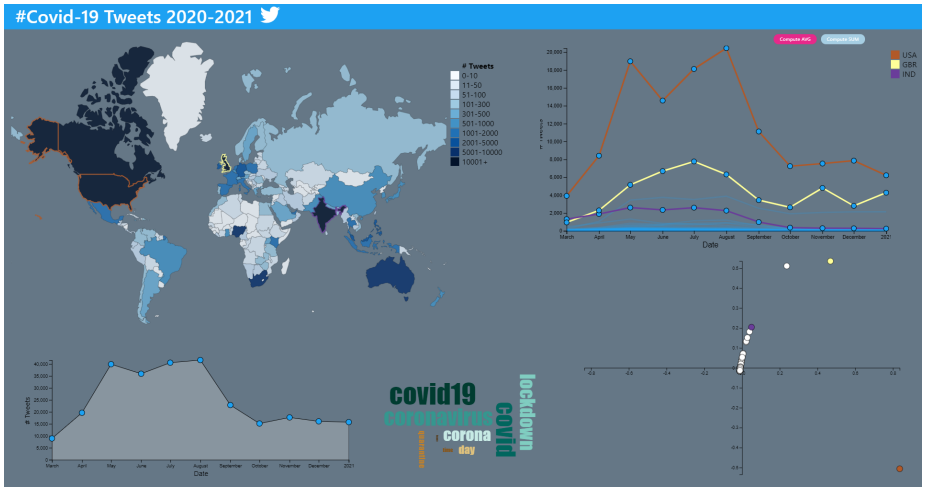

Figure 8: A picture of the whole system

References

[1] Ramya Tekumalla Juan M. Banda and Gerardo Chowell-Puente. "Covid-19 Twitter chatter dataset for scientific use". In: (2021). DOI: https://doi.org/10.5281/zenodo.3723939.

[2] Rabindra Lamsal. "Design and analysis of a large-scale COVID-19 tweets dataset". In: *Applied Intelligence* 2790.2804 (2021), p. 51. DOI: https://doi.org/10.1007/s10489-020-02029-z.

[3] Ranganathan Chandrasekaran. "Topics, Trends, and Sentiments of Tweets About the COVID-19 Pandemic: Temporal Infoveillance Study". In: *PubMed* 22.1438-8871 (2021), pp. 1439–4456. DOI: https://pubmed.ncbi.nlm.nih.gov/33006937.

[4] Joel Dyer. "Public risk perception and emotion on Twitter during the Covid-19 pandemic". In: *Applied Network Science* 99.5 (2020), p. 1. DOI: https://doi.org/10.1007/s41109-020-00334-7.

[5] Rabindra Lamsal. *Coronavirus (COVID-19) Geo-tagged Tweets Dataset*. 2020. DOI: 10.21227/fpsb-jz61. URL: https://dx.doi.org/10.21227/fpsb-jz61.