Visualising the COVID-19 Pandemic using Reddit posts, Twitter Hashtags and Stock Market fluctuations

Amandeep Singh[1]*, Hitesh Madhukar Patil[2]*
Vikas Kishanrao Thamke[3]*, Vishal Shakya[4]*

Abstract
The ongoing coronavirus pandemic is a crisis unlike any other. Except for a handful of people, no living person has seen a situation unfold on such large proportions. But even with a lock-down or quarantine imposed in most major cities in the world, life is still keeping pace, all thanks to the innovations in technology that have connected more people together taking advantage of the internet. However, this dependence on the internet and social media does not come without its burdens. With unlimited real-time information available all the time, it can become overwhelming, sometimes even overshadowing some important issues. With the world slowly recovering from the pandemic, more and more focus is now being given to bring everything back to normal. But to chart a course for reclaiming the world, the effects of coronavirus have to be studied. This report endeavours to study the impact, and potential correlations, on the economic and social structures of the society, using database management methods & analysis techniques to illustrate the progression of the coronavirus pandemic. While data from Reddit & Twitter is used to study the spread of news and information on the social media, historic stock market prices are used to visualise the response of the financial sector. Finally, an animated dashboard is produced to graphically represent the findings of the project.

Keywords
Coronavirus, COVID-19, Reddit, Twitter, Stock Market, Visualisations, Animations, Correlations

[1] *x19194137@student.ncirl.ie*, [2] *x19147996@student.ncirl.ie*
[3] *x19180080@student.ncirl.ie*, [4] *x19182732@student.ncirl.ie*
***Affiliation**: MSc Data Analytics, Group A (Tutorial 2), National College of Ireland, Dublin, Ireland

Introduction

Thousands have died, millions are infected, and billions have been affected. Just in the first five months of the year 2020. The COVID-19 pandemic has changed the world. The situation is unfolding on such a large scale that it might get difficult to understand the effect of the disease on all the facets of human society. The various organisations are striving to get critical information across to people worldwide. One good way to convey a message is to use a picture, i.e., a visualisation. The World Health Organisation (WHO) has built a very informative dashboard for COVID-19 [1] that updates daily and maps the number of novel coronavirus cases using a world and country-wise maps. Johns Hopkins University implemented a similar dashboard [2] with individual country-wise counts given as a sidebar. Many countries and organisations have made their own dashboards and visualisations to get the masses to understand the scale of things.

All these dashboards convey pivotal information in this current situation. But with the main focus being saving lives, the visualisations showing correlations of this pandemic with other parts and aspects of our society have not been explored. The focal point of this report is to study and correlate the COVID-19 pandemic to the recent changes in the social and economic systems of human society, using understandable visualisations, while also incorporating the informative world maps from the credible organisations mentioned previously.

To investigate the economics angle, one will have to look at the fundamentals, i.e., income statements, balance sheets, cash flow, operational data, and external factors like market position, economic prospects, etc., which requires a deep study and understanding of the financial domain. However, for a neophyte, easiest workaround could be to check the history of stocks of organisation. Increasing values of share price implies financial progress of organisation, whereas declining value suggests the performance is reduced. Thus, to correlate the impact of this pandemic on economy, it was decided to analyse the stock prices of prominent companies from seven different sectors. The industry sectors included for the study involves Airline, Automotive, Pharmaceutical, Construction, Chemical, Entertainment, and Telecommunication.

Social media users usually follow particular subject or interests, like entertainment or game, which helps to commu-

nicate with the people who you have common interests with. This is common in social media platforms like Facebook, YouTube, Quora etc. where users like specific documents of their interest [3]. Reddit is another social media platform which is updated constantly with news, pictures and videos. Since Reddit has handles for various topics, the aims to extract the data from various Reddit sources with the help of programming language python (PRAW library) and web-scraping tool known as Parsehub. The data-set will be generated which will gather all the unstructured data from Reddit and then data will be analysed with the help of sequence query language on PgAdmin software. The PgAdmin provides organised and cleaned structure data which is imported in python software. finally providing the desired results with some visualisations [4].

Twitter is a micro-blogging and social networking service where users post and socialise with messages known as "tweets" and has 330 million monthly active users [5] and continues to upsurge. Twitter's user tweet about any subject within the 140-character boundary and follow others to hold their tweets. Snapchat and Instagram may have hooked up the young people, but Twitter has its own circle to appeal to, and around six of every 10 (63 percent) Twitter users globally are between 35 and 65 years old [5]. In the case of gender, Twitter is more famous among males than females. There were 11.7 million App Store downloads of the Twitter app in the first fraction of 2019 – enrolling a year-over-year increase of 3.6 percent [5]. As we all know, the first cases of COVID-19 were reported in Wuhan, China, in December 2019, and have increased rapidly. Since the initial stages, people have shown their view and shared information, as well as misinformation, about it via social media platforms, such as Twitter. The plan is to investigate a corpus of tweets that correlate to COVID-19 to recognise common responses to the pandemic and how these responses differ across time.

1. Related Works

1.1 Dataset 1 - World COVID-19 Cases

- Since this disease is highly infectious, the number of cases are increasing exponentially. This forced the governments and organisations to make united efforts in tracking the spread of this virus across the world. WHO was on the front-lines reporting on the number of cases and updating them daily. The WHO dashboard [1] guided the authorities globally, which helped in planning and policy-making. The nation-wise list of number of cases clearly showed the spread of the virus, which is why it became the inspiration for this project.

- The extremely informative dashboard developed by researchers at Johns Hopkins University [2, 6] instilled motivation to make the visualisations simple and efficient - to convey the information to the viewer without the need for second glances.

- The data collected by Johns Hopkins University to develop their dashboard was kept in the public-domain. This prompted the development of several API services [7, 8] that extracted and provided data according to requests made by the researchers. This made the data acquisition more fluent.

1.2 Dataset 2 - Stock Market

- In paper [9] by Abdullah M. Al-Awadhi studied Hang Seng Index and Shanghai Stock Exchange Composite Index during the COVID-19 communicable disease outbreak in China along with the visualisations which showed that COVID-19 pandemic associates negatively with stock market return and this return negatively correlated with both regular increase in total confirmed cases and the regular increase in total cases of death caused by COVID-19. Therefore, we got the idea to include the stock market dataset which shows a rapid rise and fall in the stocks.

- In the paper [10] the author explained several reasons causing an impact on the Stock Market including the severity of the Pandemic, comparison to the Spanish Flu, the inter-connectedness of the modern economy, etc. The author also reported that no previous outbreak has powerfully impacted the Stock Market as COVID-19 pandemic which made it easier to choose between different datasets for this project.

- In paper [11] by Nuno Fernandes studied the economic impact on the different industries globally and also elaborated on the economic straits through which economic motion will be affected. It also experiments a rough evaluation of the latent global economic costs of COVID-19 under distinct situations. The author analysed and exhibited the various graphs on the global economy which gave the motivation about the dataset Stock Market.

1.3 Dataset 3 - Reddit

- There are different types of studies analysing the correlation between various social issues (such as depression, hunger, child abuse, gender discrimination etc.) and Reddit posts.

- Dating back to the earliest years of therapy, Michael Tadesse el at. [12] wrote about the presence of depression on the Reddit platform and searched for improvement of depression discovery. He also identified a closer connection between depression and Reddit post usage by implementing Natural Language Processing and text classification techniques.

- T. weninger et al. [13] in his paper, illustrated a positive relationship of karma with subreddits. He explained that the association of karma about closeness and clustering coefficient. He concludes that the nature of Red-

dit's popular content reduces the possibility that it is related to unpopular content.

- According to BG Dastidar [14], The Internet is loaded with publicly accessible information. This data can help solve the problems we have, by examining the theory. According to him, the main advantage of Web Scraping is that it's not restricted to any domain.

- These theories have motivated us to extract information from Reddit and with the increasing popularity in using social media to gather data and finding ways to use that data, it is very useful to study the various aspects that are affected by COVID-19.

1.4 Dataset 4 - Twitter

- In paper [15] the author analysed the number of conversations taking place on Twitter about COVID-19. The author focused on the comparison between location conversations and COVID-19, words being used, etc. The correlation between Twitter conversations and COVID-19 was shown whereas analysis of common myths about COVID-19 which reveals a pattern was also shown. The Twitter conversations and visualisation in this paper gave inspiration to include Twitter in this project.

- In paper [16] the author studied hashtags like Coronavirus using sentiment analysis and topic modelling using Latent Dirichlet Allocation post pre-processing. It was analysed that misinformation spread of Ebola and the Zika outbreak was much higher in comparison to the Coronavirus outbreak. So, this study gave the idea to use distinct hashtags for the study.

- In this research work [17], the author used different machine learning models like Naïve Bayes, SVM, random forest, decision tree, etc., to analyse and visualise the influence of coronavirus with accuracy of 74%. On the other hand, API call was used to collect the data and then use that data for visualisation in Python. The author used sentiment analysis to study the classification of coronavirus effects whereas for this project different hashtags were used to study the increasing tweets during outbreak.

2. Methodology

2.1 Data Acquisition

Dataset 1 - World COVID-19 Cases

- The number of coronavirus cases was a crucial link in this analysis, i.e., getting credible information was paramount. Additionally, the possibility of updating the data in the future had to be incorporated. This made it certain that the route of choosing an API-based system would be suitable.

- The world-map portraying the nation-wise number of coronavirus cases had to be kept in the final visualisation because it provided a 'bird's eye'-view of the pandemic. Several packages were available in Python to aid in the custom world map plotting, namely Matplotlib Basemaps, Folium, Geopandas, etc.

- Development on Matplotlib Basemaps is stagnant and it has not been receiving updates for the past few years, which negated its inclusion in the project considering future-proofing. Folium was good contender, but the package itself was big, and it was slow to run on personal laptops. Finally, Geopandas was chosen.

- Geopandas is a python package used to plot world maps because it is designed to handle geometry shape files. As the name suggests, Geopandas is built on the Pandas package, meaning that it is highly compatible with Pandas dataframes, which is an added benefit in this case.

Dataset 2 - Stock Market It is possible to get Stock Market data in several ways as listed below:

- API: The Motley Fool, Yahoo Finance, Metastock are some of the most trusted sites which provides API service to collect data [18].

- Google Spreadsheets: Data from Google Finance can be fetched to Google Spreadsheets using commands [19].

- Python Libraries: Developers have developed libraries like Stocker, yfinance to get stock data and made open source for everyone [20, 21].

- Web Scraping: It is possible to scrape data of any stock market website and collect data.

The live share price information of any stock could be obtained easily using APIs with limited access. However, all the reputed companies charge some amount of money to provide historical data and unlimited access. Thus, after reading terms and conditions of various companies, it is decided to crawl historical data from Yahoo Finance as it allows to do so.

On the webpage of Yahoo Finance, the historical data is present in the form of table, that comes after inserting name of company in the search box and clicking on Historical Data tab. The data present in the table is in decreasing order of date i.e. most recent to older. Also, the website has used Dynamic Loading of information, which means the further data will be loaded if scrolled down to the bottom. As it could only be done through automation, it is decided to automate the process to reach the data and fetch the HTML page.

The framework used for automation is Selenium with the Chromium Driver to perform task on Chrome Browser. Initially a list is created which has names of all the companies from each sector. The automation driver fetches webpage of each company as explained in the steps below:

1. For each company name, driver navigates to yahoo finance home page, and waits till the page is loaded.
2. Finds the search box on page using provided xpath and enters the name of company.
3. Once the page of specific company is loaded, driver searches for the Historical Data tab using provided xpath, clicks on it, and waits till the table is loaded.
4. Then the page is scrolled down three times to reach historical data of last three months.
5. Page is fetched and stored in a variable.

As the webpage is available, it is parsed using BeautifulSoup, and required information is extracted with the help of tags and classes of HTML page.

Dataset 3 - Reddit To find out the motivation behind the concept of data, extraction and interpretation of information is a necessity. The requirement for extraction is expected to consistently improve the information.

- To extract data about COVID-19, a library called PRAW was used as a Reddit API for python.
- For the extraction of articles about COVID-19, an app was created on Reddit and information was extracted in various configurations in. JSON format.
- But with Reddit API can't access data older than seven days or 250 entries.
- Older data was accessed from the archives available on old.reddit.com
- A web scraping tool, i.e., Parsehub was used.

Dataset 4 - Twitter

- Initially, the aim was to look for the data containing several hashtags using API but it was returning data before 7 days which was not meeting up with the strategy. Social media is an essential part of this project so the idea popped up to get data from some publicly available data with the hashtags.
- All the unstructured data was handled by MongoDB to deal with its quantity and performance. Thousands of records can be inserted in a second by MongoDB. The database called Twitter was created in MongoDB to collect the number of tweets related to each Hashtags.
- PostgreSQL is open-source and highly extensible with numerous characteristics aiming to create applications, administrators to preserve data uprightness and develop fault-tolerant circumstances, which also helps to maintain data no matter the size of the dataset. A table was created in the Twitter database to insert the data into it so that dictionaries can be converted into data frames and finally merge these data frames into a single data frame.
- The data was downloaded to the local system from Kaggle which is the world's largest data science community. The data was modified according to the necessity like taking the cumulative sum by implementing cumsum() function in python. The downloaded data was satisfying as it was containing all the needed data like user ID along with the hashtags and then through these IDs tweets were fetched which showed the distinct hashtags. These hashtags add an extra edge to this project because hashtags exhibit the number of users using that particular hashtag in a day.
- After all the above steps, the central focus was on the visualisations. The number of tweets of that particular hashtag was shown through the bar graph individually. All the datasets were merged to exhibit correlation with the other datasets and the conclusive animation of the distinct datasets collectively with the date counter as the epicentre which portrays the data of that particular date on the screen.

2.2 Database Management

The data obtained would be mostly unstructured in JSON format and required to be stored somewhere. MongoDB is the best database to store unstructured data as it has automatic scaling, high performance, and stores data in JSON format. There are multiple advantages of storing raw data in own database such as it can be used anytime as per requirement, as the data is fetched for only one time from provider, charges would be reduced etc. It is possible that the data obtained has missing values, unnecessary information, duplicate values etc. which is then checked, and data is cleaned. The clean data can be stored in structured databases such as PostgreSQL in the form of tables.

2.3 Process Flow Diagrams

Dataset 1 - World COVID-19 Cases: Figure 1 visualises the process flow for this dataset in the form of a flowchart.

Dataset 2 - Stock Market: Figure 2 visualises the process flow for this dataset in the form of a flowchart.

Dataset 3 - Reddit: Figure 3 visualises the process flow for this dataset in the form of a flowchart.

Dataset 4 - Twitter: Figure 4 visualises the process flow for this dataset in the form of a flowchart.

3. Data Sources & Data Description

3.1 Dataset 1 - World COVID-19 Cases

Data Sources: The API used for this part of the report was chosen to be one that was based on the data provided by Johns Hopkins University [7, 8, 2, 6]. The data contained information regarding the country name, date, number of active cases, number of confirmed cases, number of recovered cases, and number of deaths. This was sufficient for the scope of this project.

Data Description: The data extracted from the API was in JSON format. Figure 5 depicts the structure of the data. The size of this JSON file was $\sim 110MB$.

3.2 Dataset 2 - Stock Market

Historical data of all the companies is available on Yahoo Finance website [22], could be obtained by entering Company name and navigating to Historical Data section, e.g., the historical data of American Airlines is available at [23]. The data has features like Opening Price, Closing Price, Peak Price, Lowest Price, and Adjusted Closing Price of each day when the share market was open.

As the process is automated and driver waits till the page is loaded at each step, it takes around 8-10 minutes to collect the data considering usual internet speed.

Data Description: As the data is parsed using BeautifulSoup, it is initially formatted in the form of JSON, as shown in Figure 9.

3.3 Dataset 3 - Reddit

Data Sources: The COVID-19 outbreak has a huge impact on our social life amid lock-down. There is a massive increase in people and expert engagement in social media platform. The data about COVID-19 with massive data analysis is available on many subreddit posts. We examine engagement and interest in the COVID-19 issue. Reddit provides API services to scrape the data but does not provide data older than 7 days. Parsehub software used as a web scraping tool to access archives. An automated script was run onto old.reddit.com to collect information.

Data Description:

- Information obtained from various subreddits for keywords like 'Covid-19', 'Coronavirus', 'Pandemic', 'Lockdown' & 'World news'.
- Information is collected from Hot posts, Controversial Posts, New Posts and All time Popular posts.
- Data such as date, title, number of comments, and votes are extracted from each post.
- Data was stored in JSON format and uploaded to the MongoDB database.
- Uploaded unstructured data has various entry fields like Date, Title, Id, Comments and Votes. Furthermore, this data was uploaded to PostgreSQL as structure data.
- Figures 17, 19.

3.4 Dataset 4 - Twitter

Data Sources:

- The dataset can be accessed on Kaggle at: LINK [24]. Tweet IDs were made available by Kaggle in agreement with Twitter's terms and conditions. The data of tweets that have not been deleted was fetched using the Twitter API.
- Thee dataset contains 69,465,876 tweets from 22 January, 2020 to 13 April,2020. The tweets with hashtags used are: #virus and #coronavirus since 22 January, 2020 whereas #ncov19 and #ncov2019 since 26 February, 2020, and #covid since 7 March, 2020.
- After data collection, it was uploaded to MongoDB to create a database and in that database a table was created in PostgreSQL to insert the data.
- This dataset which is a social media platform was selected over others because it provided the data related to COVID-19 Hashtags whereas other platforms like Facebook, Snapchat, Instagram, etc., were not providing enough data. Twitter was the first preference as the social aspect of this project because most of the celebrities, musicians, political leaders, etc., post their tweets on Twitter over other platforms.
- Postman was used as a software to fetch the data from the IDs through Twitter API. It was not possible to directly fetch the data as it was returning data 7 days prior to the current date. It took lot of time to understand this process to fetch the data using Postman. The memory taken by the data is 1.07 GB.

Data Description: The data extracted from the API was in JSON format. Figure 23 depicts the structure of the data.

4. Data Cleaning and Transformation

Data cleaning, transformation and feature selection was in Python using the Pandas and NumPy packages.

4.1 Dataset 1 - World COVID-19 Cases

Data Preparation: There were 7 steps involved in preparing the data for visualisations:

1. The dataframe contained the cases for each province/county in China and USA. But this additional bifurcation of the number of cases was not needed. The final world map was to be plotted using nation-wise cases, so for each country, all the cases for each province/county were summed up and added under a single country-name row.
2. The date sequence was extracted from the dataframe and a separate Date dataframe was created.
3. To plot the world map, a shape-file containing all the information regarding the geometry (latitude/longitude) of each country was imported as a Geopandas dataframe. Further, a separate dataframe containing the names of all countries in the world was extracted from the shape-file.

4. The two dataframes containing country names and dates were merged such that each country had all the dates from the date-sequence. Numerically, there were 251 countries in the Country dataframe and 83 dates in the Date dataframe. So the new Country-Date dataframe had $251 \times 83 = 20833$ rows. This was a crucial step because this dataframe will form the base of the final visualisation. Now, the Country-Date dataframe had all the countries in the world, and each country had a date sequence from 22nd January to 13th April 2020. Two pieces of information were now needed for the visualisation - the number of cases for each country on each date, and the geometry (latitude/longitude) of that country.

5. The geometry column from the shape file dataframe, and the columns with all the number of cases from the coronavirus dataframe were merged on top of the Country-Date dataframe such that all the countries that did not have any records of coronavirus would NOT be removed, instead they would contain null values. A single dataframe was created from this merged file.

6. The null values in this merged dataframe were replaced by zero, i.e., if a country does not have records for the number of cases of coronavirus, the entry will be zero cases. Further, a separate column was created for the base-10 logarithm of the number of confirmed cases. Taking logarithm was a conscious decision so that all the number of cases are re-scaled and the plotting is more consistent.

7. Antarctica was removed from the dataframe because it doesn't have any inhabitants.

Cleaned and Transformed Dataset: Figures 6 & 7 show the data dictionary and the info about the final dataset.

4.2 Dataset 2 - Stock Market

Data Preparation: As the share market remains closed on every weekend and on occasional holidays, it is observed that the data stored in MongoDB has those values missing. In 9, data of 18th April and 19th April is missing as the share market was closed on these days because of weekend. Also, observed that Date, Opening Price and Closing Price could be sufficient for the analysis.

The data from all the collections of StockMarket database of MongoDB is fetched for each date. The values, which are found to be unavailable for a date, are replaced with values of previous date. Opening Price and Closing Price of each company is selected from the structured data, and new data is uploaded to PostgreSQL to store in tabular form.

Each company has different share price. The companies like Johnson & Johnson, Disney has high stock prices with peak of around 160 dollars. Whereas American Airlines, General Motors has average price of around 30 dollars. If the data is plotted on different scale, the line plots of lower values would get suppressed, and analysis would not be precise. Thus, the data of each company is re-scaled using Normalisation Technique.

Cleaned and Transformed Dataset: Figure 10 shows the data fetched from PostgreSQL. It can be observed that the values of 18th and 19th April are same as 17th April.

4.3 Dataset 3 - Reddit

Data Preparation: As most of the data was unstructured, multiple values were missing in each dataset stored in MongoDB. Date format was modified and some values were missing from the data also, some data entries were repeated a few times, hence data

Cleaning was necessary. For dates, some values of previous entries were missing, so their value was replaced with the previous date.

Cleaned and Transformed Dataset: Figure 18

4.4 Dataset 4 - Twitter

Data Preparation:

1. The language used for cleaning and transformation was Python and the packages used to do the entire process were NumPy, pandas, urllib, time, BeautifulSoup, pymongo, psycopyg2, DateTime, os, and pylab. The number of tweets was counted for each hashtag. The file in the JSON is shown in figure 23.

2. All the extra unnecessary values were removed to plot the graphs and visualize the data. The data was then converted from JSON to CSV. There were some rows with zero entry as there were no data related to the required hashtags.

Cleaned and Transformed Dataset: The twitter table after transformation in figure 24 was collected from PostgreSQL which was later used for visualisations. The multiple bar graph was the most suited way to represent this data.

The final size of the data after cleaning and transformation was 4.1KB as seen in figure 25.

5. Charts, Plots, & Animations

5.1 Dataset 1 - World COVID-19 Cases

The progression of the virus across the world can be tracked in figure 8

5.2 Dataset 2 - Stock Market

Individual Analysis: Stocks of all the companies with original values are visualised in figure 11, which are affected in mid-March of 2020. Approaching towards the end of March, stocks of Johnson & Johnson, General Motors, Jacobs, started gaining value. Whereas prices kept on lowering till mid-April 2020 for American Airlines.

The figure 12 shows the plot of all companies with re-scaled data for precise comparison.

The most benefited company Johnson & Johnson is compared with the most suffered company American Airlines in figure 13.

The industries General Motors and Jacobs, whose average share value started to rise from the start of April 2020, are compared in figure 14.

Disney and AT&T, the well-established companies, which did not have much variation in stock prices before March 2020, are compared with dynamically growing Albemarle as shown in figure 15. Stock prices of these companies increased from start of April 2020 but started decreasing from mid-April.

Progression of the Virus: The effect of the progression of the virus can be seen in figure 16

5.3 Dataset 3 - Reddit

Individual Analysis: The bar graphs in figures 20, 21, 22 display the distribution of the Reddit Posted divided into each genre after normalisation. A cumulative function used in python to create unbiased data. The average number of posts and comments about COVID-19 are doubled, average votes per post are almost tripled during the rise of COVID-19. To show every aspect of the data use of normalisation function was necessary.

Progression of the Virus: The effect of the progression of the virus can be seen in figure 32

5.4 Dataset 4 - Twitter

Individual Analysis: The packages used for the visualisation and plotting of various graphs are Matplotlib and selenium. The visualisation after the data transformation was used accordingly to represent the tweets by hashtags according to the date.

In figure 26, the graph exhibits the multiple bar graph where each bar represents a distinct hashtag. The data was fetched from 22 January 2020 to 13 April 2020 because that date is starting from 22 January 2020.

In figure 27, the graph exhibits that with the increasing date the number of tweets for hashtag coronavirus keeps increasing with the date.

In figure 28, the graph shows the number of tweets for hashtag covid which came in use after 07 March 2020.

In figure 29, The graph shows the number of tweets for hashtag ncov19 which came in use after 26 February 2020.

In figure 30, it can be seen clearly that the bar represents the number of tweets for hashtag ncov2019.

As can be observed from figure 31 that the use of the hashtag virus started from the very beginning of the pandemic and the number of tweets was increasing continuously with the date.

Progression of the Virus: The effect of the progression of the virus can be seen in figure 33

5.5 Combined Visualisations and Animation

By combining the findings of this project into a single animation, the correlation of the spread of the coronavirus and the effect on stock market prices, Reddit posts, & Twitter hashtags can be observed clearly. Four different dates were selected, each almost a month apart, and the screenshots of the animations for those four dates were made into a collage – figure 34.

6. Conclusions

As the world is expecting to get vaccine soon, the Pharmaceutical Industry which manufactures it has benefited the most. Migration of people is stopped worldwide, which affected severely to Airline Industry in terms of finance. Even though people are spending most of their time watching Movies, TV Series etc., Entertainment Industry does not show much rise in terms of share price.

This report gives an in-depth analysis of the effect of COVID-19 on different social aspects affecting our day to day life. Data was taken from a different platform such as Yahoo finance, Twitter, and Reddit. Key data from these platforms described the spread evaluation structure and the surge in social media post related COVID-19.

This project showed that growth of tweets and hashtags are based on the trend for example COVID-19. The data was collected from API calling and then uploaded to MongoDB and then to PostgreSQL. The final dataset was used for visualisation which showed that hashtags played a major role in the increase of the tweets.

Finally, an animation was made by including the different pieces of analysis done on stock market, Reddit, & Twitter. The animation clearly showed correlations between the spread of the virus and its social & economic impact on the human society.

Acknowledgements

The authors of this report are extremely grateful to the National College of Ireland (NCI) and Prof. (Dr.) Athanasios Staikopoulos for their constant support, guidance & encouragement, without which this project would be impossible to create. The authors tried their best to offer their share of analysis and understanding to promote sharing of accurate and credible information in these bleak times.

The authors wish everyone all the best and sincerely hope that the world overcomes all these hurdles and comes out stronger on the other side.

References

[1] WHO . Who covid-19 dashboard.

[2] JHU . Operations dashboard for arcgis.

[3] Jack Burke and Ben Wagner. Rednet, a different perspective of reddit. In *2015 IEEE Integrated STEM Education Conference*, pages 145–146. IEEE, 2015.

[4] Hoang Nguyen, Rachel Richards, Chien-Chung Chan, and Kathy J Liszka. Redtweet: recommendation engine for reddit. *Journal of Intelligent Information Systems*, 47(2):247–265, 2016.

[5] Ying Lin. 10 twitter statistics every marketer should know in 2019 [infographic], 07 2019.

[6] CSSEGISandData. Cssegisanddata/covid-19, 03 2020.

[7] Saiprasad Balasubramanian. backtrackbaba/covid-api, 05 2020.

[8] Covidapi.

[9] Abdullah M Al-Awadhi, Khaled Al-Saifi, Ahmad Al-Awadhi, and Salah Alhamadi. Death and contagious infectious diseases: Impact of the covid-19 virus on stock market returns. *Journal of Behavioral and Experimental Finance*, page 100326, 2020.

[10] Scott R Baker, Nicholas Bloom, Steven J Davis, Kyle J Kost, Marco C Sammon, and Tasaneeya Viratyosin. The unprecedented stock market impact of covid-19. Technical report, National Bureau of Economic Research, 2020.

[11] Nuno Fernandes. Economic effects of coronavirus outbreak (covid-19) on the world economy. *Available at SSRN 3557504*, 2020.

[12] Michael M Tadesse, Hongfei Lin, Bo Xu, and Liang Yang. Detection of depression-related posts in reddit social media forum. *IEEE Access*, 7:44883–44893, 2019.

[13] T. Weninger, X. A. Zhu, and J. Han. An exploration of discussion threads in social news sites: A case study of the reddit community. In *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*, pages 579–583, 2013.

[14] Bhaskar Ghosh Dastidar, Devanjan Banerjee, and Subhabrata Sengupta. An intelligent survey of personalized information retrieval using web scraper. *International Journal of Education and Management Engineering*, 6(5):24–31, 2016.

[15] Lisa Singh, Shweta Bansal, Leticia Bode, Ceren Budak, Guangqing Chi, Kornraphop Kawintiranon, Colton Padden, Rebecca Vanarsdall, Emily Vraga, and Yanchen Wang. A first look at covid-19 information and misinformation sharing on twitter. *arXiv preprint arXiv:2003.13907*, 2020.

[16] Dr Prabhakar Kaila, Dr AV Prasad, et al. Informational flow on twitter–corona virus outbreak–topic modelling approach. *International Journal of Advanced Research in Engineering and Technology (IJARET)*, 11(3), 2020.

[17] Muthusami Ra, Bharathi Ab, and Saritha Kc. Covid-19 outbreak: Tweet based analysis and visualization towards the influence of coronavirus in the world.

[18] Top 8 best stock market sites to check investments (2020) — rapidapi, 11 2019.

[19] Googlefinance - docs editors help.

[20] Will Koehrsen. Stock analysis in python, 01 2018.

[21] Ritvik Kharkar. How to get stock data using python, 01 2020.

[22] Yahoo finance - business finance, stock market, quotes, news, 2000.

[23] American airlines group, inc. (aal) stock historical prices data - yahoo finance.

[24] Covid-19 tweets dataset.

Appendix: Images

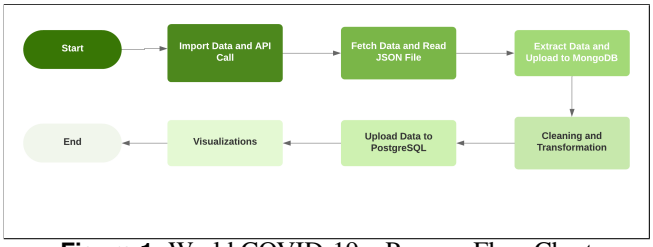

Figure 1. World COVID-19 – Process Flow Chart

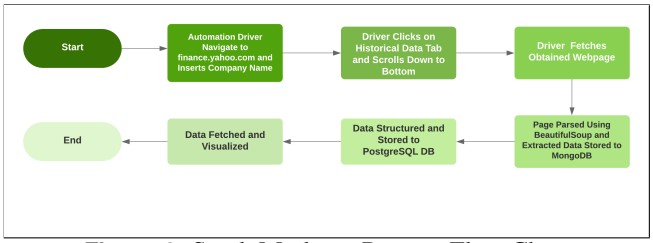

Figure 2. Stock Market – Process Flow Chart

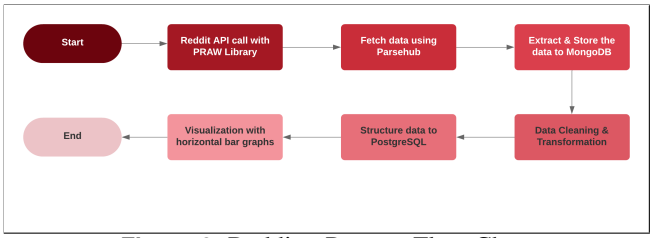

Figure 3. Reddit – Process Flow Chart

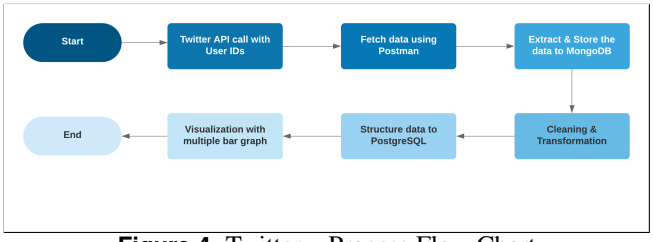

Figure 4. Twitter – Process Flow Chart

```
[
    {
        "Country": "Afghanistan",
        "CountryCode": "AF",
        "Province": "",
        "City": "",
        "CityCode": "",
        "Lat": "33.94",
        "Lon": "67.71",
        "Confirmed": 0,
        "Deaths": 0,
        "Recovered": 0,
        "Active": 0,
        "Date": "2020-01-22T00:00:00Z"
    },
    {
        "Country": "Afghanistan",
```


Figure 5. World COVID-19 – JSON File

final.head()

	Country	Date	Confirmed	Deaths	Recovered	Active	geometry	log_Confirmed	log_Deaths
0	Aruba	2020-01-22	0.0	0.0	0.0	0.0	POLYGON ((-69.88223 12.41111, -69.94695 12.436...	0.0	0.0
1	Aruba	2020-01-23	0.0	0.0	0.0	0.0	POLYGON ((-69.88223 12.41111, -69.94695 12.436...	0.0	0.0
2	Aruba	2020-01-24	0.0	0.0	0.0	0.0	POLYGON ((-69.88223 12.41111, -69.94695 12.436...	0.0	0.0
3	Aruba	2020-01-25	0.0	0.0	0.0	0.0	POLYGON ((-69.88223 12.41111, -69.94695 12.436...	0.0	0.0
4	Aruba	2020-01-26	0.0	0.0	0.0	0.0	POLYGON ((-69.88223 12.41111, -69.94695 12.436...	0.0	0.0

Figure 6. World COVID-19 – Data Dictionary

```
final.info()
<class 'geopandas.geodataframe.GeoDataFrame'>
RangeIndex: 20750 entries, 0 to 20749
Data columns (total 9 columns):
Country          20750 non-null object
Date             20750 non-null datetime64[ns]
Confirmed        20750 non-null float64
Deaths           20750 non-null float64
Recovered        20750 non-null float64
Active           20750 non-null float64
geometry         20750 non-null geometry
log_Confirmed    20750 non-null float64
log_Deaths       20750 non-null float64
dtypes: datetime64[ns](1), float64(6), geometry(1), object(1)
memory usage: 1.4+ MB
```

Figure 7. World COVID-19 – Info


```
Started: General Motors
{'Date': 'Apr 24 2020', 'Values': {'Open': '21.71', 'High': '22.24', 'Low': '21.54', 'Close': '21.95', 'Adj Close': '21.95',
'Volume': '11990500'}}
{'Date': 'Apr 23 2020', 'Values': {'Open': '21.55', 'High': '22.06', 'Low': '21.45', 'Close': '21.52', 'Adj Close': '21.52',
'Volume': '10673000'}}
{'Date': 'Apr 22 2020', 'Values': {'Open': '21.65', 'High': '21.78', 'Low': '21.07', 'Close': '21.30', 'Adj Close': '21.30',
'Volume': '9553600'}}
{'Date': 'Apr 21 2020', 'Values': {'Open': '21.27', 'High': '21.89', 'Low': '20.98', 'Close': '21.24', 'Adj Close': '21.24',
'Volume': '13507300'}}
{'Date': 'Apr 20 2020', 'Values': {'Open': '21.72', 'High': '22.64', 'Low': '21.44', 'Close': '22.38', 'Adj Close': '22.38',
'Volume': '15912400'}}
{'Date': 'Apr 17 2020', 'Values': {'Open': '21.92', 'High': '22.54', 'Low': '21.83', 'Close': '22.48', 'Adj Close': '22.48',
'Volume': '17166300'}}
{'Date': 'Apr 16 2020', 'Values': {'Open': '21.64', 'High': '21.65', 'Low': '20.56', 'Close': '20.87', 'Adj Close': '20.87',
'Volume': '11501700'}}
{'Date': 'Apr 15 2020', 'Values': {'Open': '21.92', 'High': '22.27', 'Low': '21.47', 'Close': '21.66', 'Adj Close': '21.66',
'Volume': '12760900'}}
{'Date': 'Apr 14 2020', 'Values': {'Open': '23.60', 'High': '23.77', 'Low': '22.67', 'Close': '22.98', 'Adj Close': '22.98',
'Volume': '12654700'}}
```


Figure 9. Stock Market – JSON File

	date_time	gm_open	gm_close	johnson_open	johnson_close	jacobs_open	jacobs_close	albemarle_open	albemarle_close	disney_open	disney_close
171	2020-04-16	21.64	20.87	148.31	149.67	79.50	79.05	58.26	58.57	103.53	102.02
172	2020-04-17	21.92	22.48	151.99	152.02	81.22	84.20	60.58	61.47	106.21	106.63
173	2020-04-18	21.92	22.48	151.99	152.02	81.22	84.20	60.58	61.47	106.21	106.63
174	2020-04-19	21.92	22.48	151.99	152.02	81.22	84.20	60.58	61.47	106.21	106.63
175	2020-04-20	21.72	22.38	150.93	151.67	83.70	80.70	60.00	60.00	103.58	102.26
176	2020-04-21	21.27	21.24	150.12	149.68	79.73	78.21	58.50	57.86	100.01	100.54
177	2020-04-22	21.65	21.30	152.81	152.99	79.70	79.51	58.96	59.45	101.80	100.99
178	2020-04-23	21.55	21.52	154.25	155.51	80.56	79.75	60.67	60.66	100.65	101.00

Figure 10. Stock Market – Data Dictionary

Figure 11. Stock Market – Line Chart

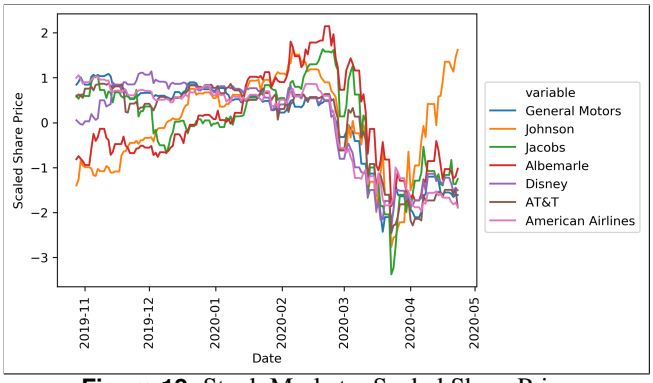

Figure 12. Stock Market – Scaled Share Price

Figure 13. Stock Market – American Airlines vs Johnson & Johnson

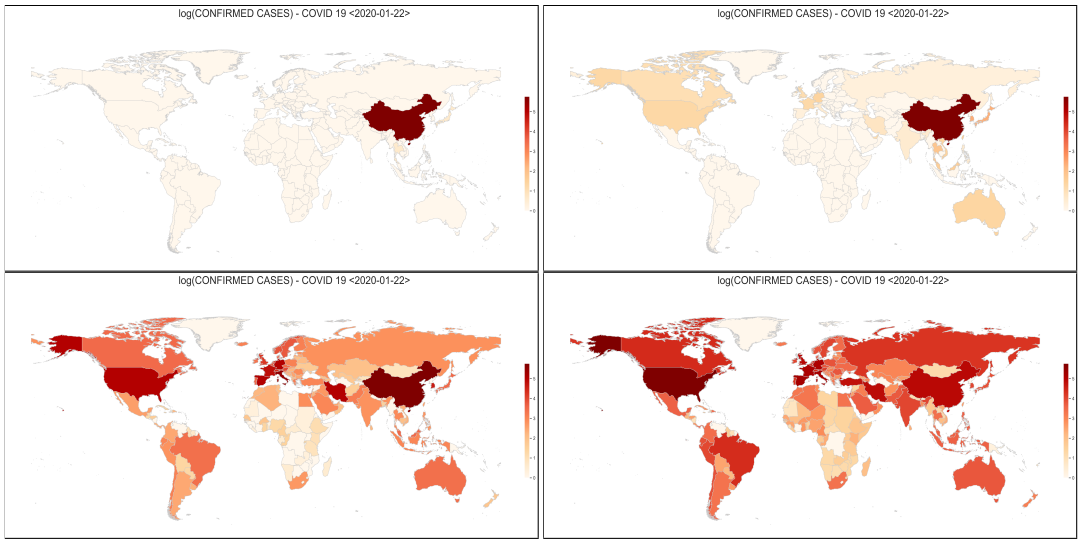

Figure 8. World COVID-19 – Progression of the Virus

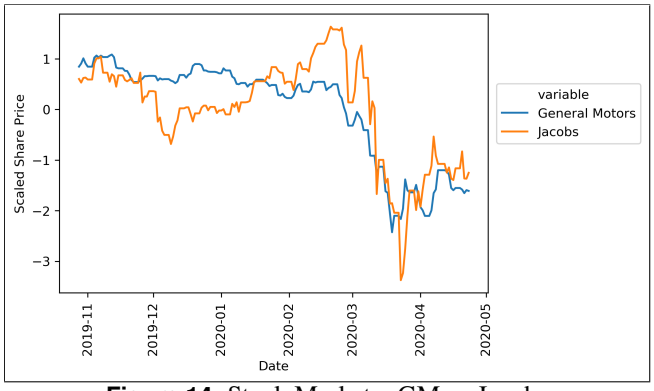

Figure 14. Stock Market – GM vs Jacobs

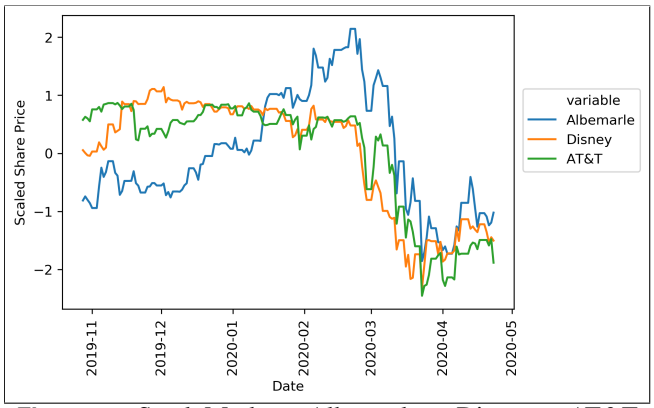

Figure 15. Stock Market – Albemarle vs Disney vs AT&T


```
{'_id': ObjectId('5eb1f61f9d2010355e5f2441'), 'votes': '52787'}
{'_id': ObjectId('5eb1f61f9d2010355e5f2442'), 'votes': '52564'}
{'_id': ObjectId('5eb1f61f9d2010355e5f2443'), 'votes': '52162'}
{'_id': ObjectId('5eb1f61f9d2010355e5f2444'), 'votes': '52140'}
{'_id': ObjectId('5eb1f61f9d2010355e5f2445'), 'votes': '52035'}
{'_id': ObjectId('5eb1f61f9d2010355e5f2446'), 'votes': '51962'}
{'_id': ObjectId('5eb1f61f9d2010355e5f2447'), 'votes': '51102'}
{'_id': ObjectId('5eb1f61f9d2010355e5f2448'), 'votes': '50968'}
{'_id': ObjectId('5eb1f61f9d2010355e5f2449'), 'votes': '50979'}
{'_id': ObjectId('5eb1f61f9d2010355e5f244a'), 'votes': '72109'}
{'_id': ObjectId('5eb1f61f9d2010355e5f244b'), 'votes': '71403'}
{'_id': ObjectId('5eb1f61f9d2010355e5f244c'), 'votes': '70764'}
{'_id': ObjectId('5eb1f61f9d2010355e5f244d'), 'votes': '1'}
{'_id': ObjectId('5eb1f61f9d2010355e5f244e'), 'votes': '68438'}
{'_id': ObjectId('5eb1f61f9d2010355e5f244f'), 'votes': '67936'}
{'_id': ObjectId('5eb1f6209d2010355e5f2450'), 'votes': '67007'}
{'_id': ObjectId('5eb1f6209d2010355e5f2451'), 'votes': '66802'}
{'_id': ObjectId('5eb1f6209d2010355e5f2452'), 'votes': '65839'}
{'_id': ObjectId('5eb1f6209d2010355e5f2453'), 'votes': '65065'}
```


Figure 17. Reddit – JSON File

	Votes_name	Date_Time	Comments_name	c_countstr	c_count	v_count
0	128721.0	2020-03-10	4416 comments	4416	4416	128721
1	118808.0	2020-03-22	1855 comments	1855	1855	118808
2	105662.0	2020-03-22	2219 comments	2219	2219	105662
3	6.0	2020-04-08	2334 comments	2334	2334	6
4	101981.0	2020-03-16	1422 comments	1422	1422	101981
...	...	...	...	...	...	...
2695	6569.0	2020-04-22	401 comments	401	401	6569
2696	6572.0	2020-04-17	396 comments	396	396	6572
2697	6559.0	2020-04-14	393 comments	393	393	6559
2698	6554.0	2020-03-27	492 comments	492	492	6554
2699	6857.0	2020-04-20	209 comments	209	209	6857

2574 rows × 6 columns

Figure 18. Reddit – Data Dictionary

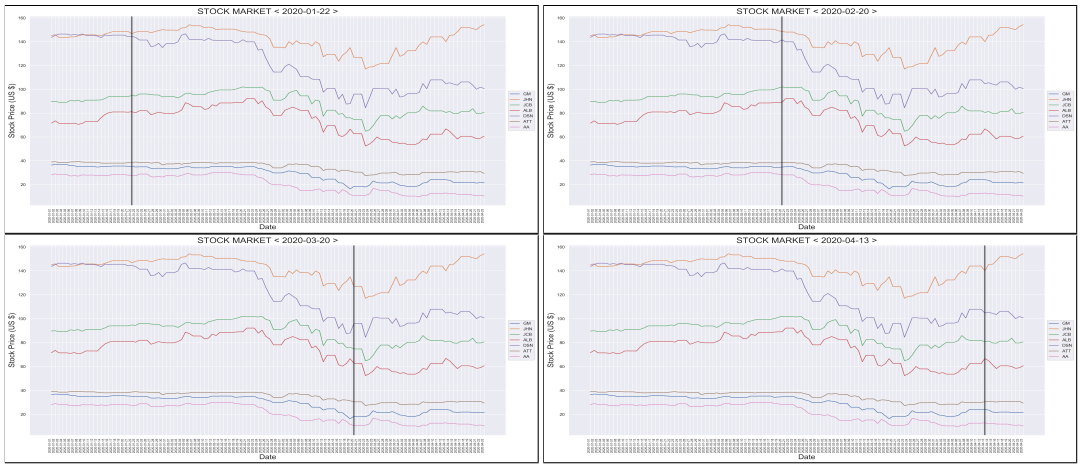

Figure 16. STOCK MARKET – Progression of the Virus

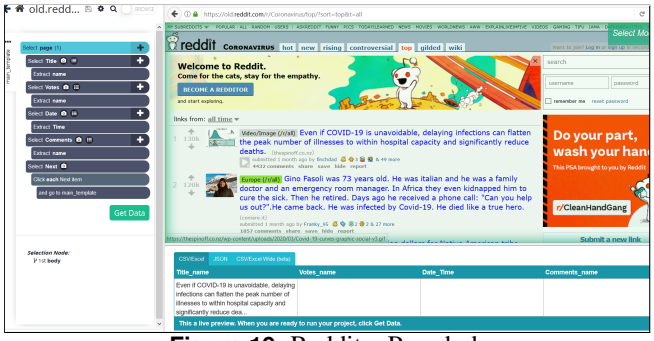

Figure 19. Reddit – Parsehub

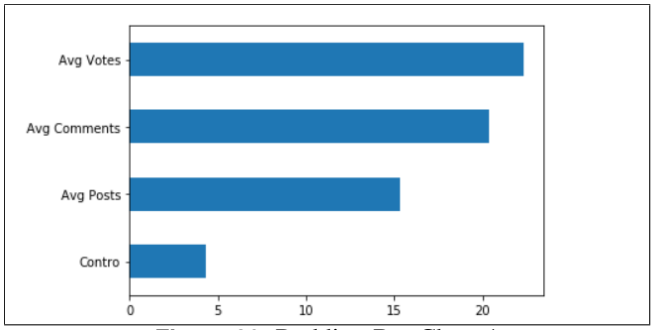

Figure 20. Reddit – Bar Chart-1

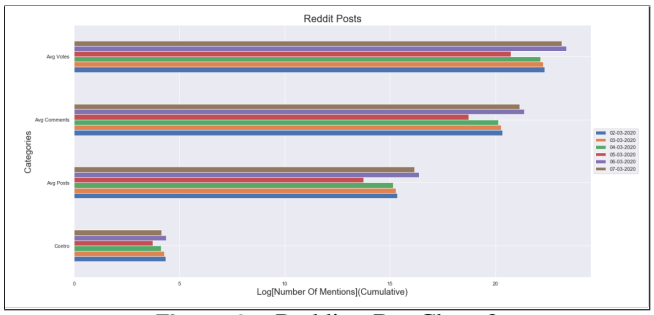

Figure 21. Reddit – Bar Chart-2

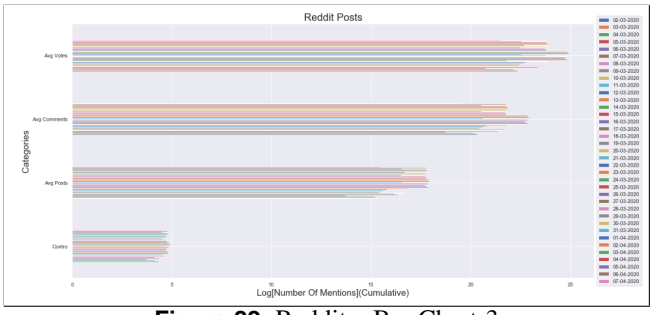

Figure 22. Reddit – Bar Chart-3


```
[{'Date': '2020-02-26',
  'ncov2019': 6,
  '_id': ObjectId('5eb2d39e205306fbeee86e8b')},
 {'Date': '2020-02-27',
  'ncov2019': 1419,
  '_id': ObjectId('5eb2d39e205306fbeee86e8c')},
 {'Date': '2020-02-28',
  'ncov2019': 2119,
  '_id': ObjectId('5eb2d39e205306fbeee86e8d')},
 {'Date': '2020-02-29',
  'ncov2019': 1489,
  '_id': ObjectId('5eb2d39e205306fbeee86e8e')},
 {'Date': '2020-03-01',
  'ncov2019': 1208,
  '_id': ObjectId('5eb2d39e205306fbeee86e8f')},
 {'Date': '2020-03-02',
  'ncov2019': 1994,
  '_id': ObjectId('5eb2d39e205306fbeee86e90')},
 {'Date': '2020-03-03',
```


Figure 23. Twitter – JSON File

	date_time	coronavirus	covid	ncovn	ncontn	virus
0	2020-01-22	7468.0	0.0	0.0	0.0	5662.0
1	2020-01-23	34028.0	0.0	0.0	0.0	23635.0
2	2020-01-24	40687.0	0.0	0.0	0.0	27243.0
3	2020-01-25	40742.0	0.0	0.0	0.0	36094.0
4	2020-01-26	42585.0	0.0	0.0	0.0	42071.0

Figure 24. Twitter – Data Dictionary

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 84 entries, 0 to 83
Data columns (total 6 columns):
date_time      84 non-null datetime64[ns]
coronavirus    84 non-null float64
covid          84 non-null float64
ncovn          84 non-null float64
ncontn         84 non-null float64
virus          84 non-null float64
dtypes: datetime64[ns](1), float64(5)
memory usage: 4.1 KB
```

Figure 25. Twitter – Info

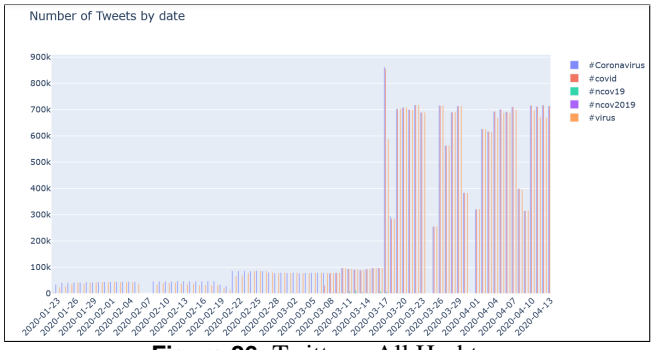

Figure 26. Twitter – All Hashtag

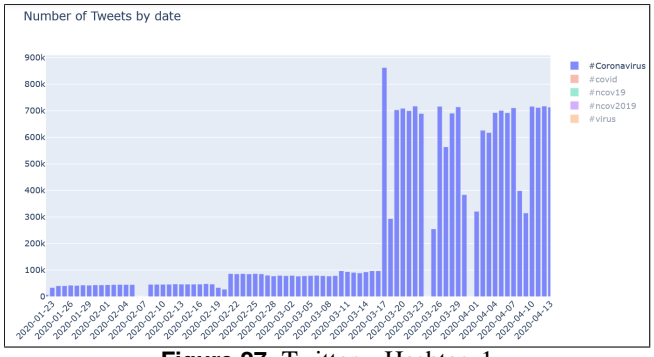

Figure 27. Twitter – Hashtag-1

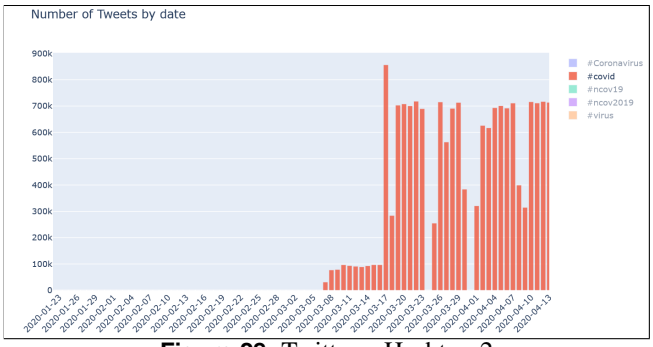

Figure 28. Twitter – Hashtag-2

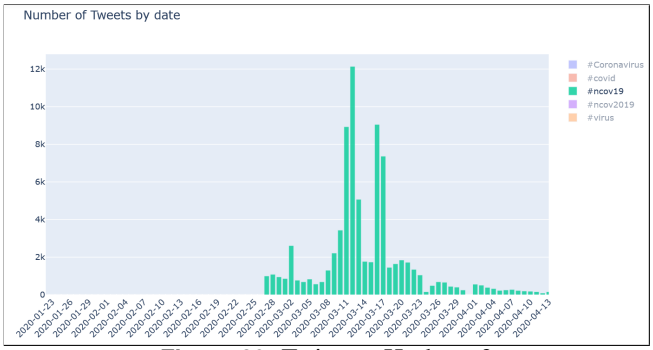

Figure 29. Twitter – Hashtag-3

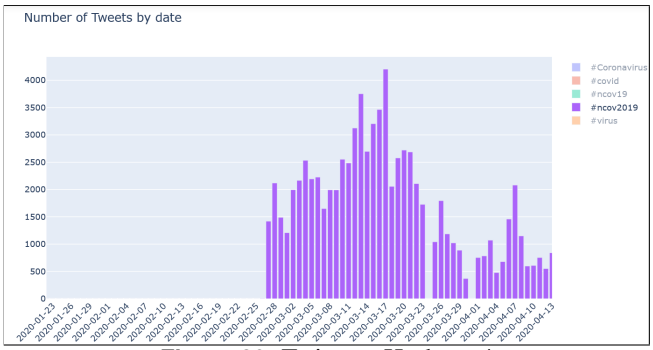

Figure 30. Twitter – Hashtag-4

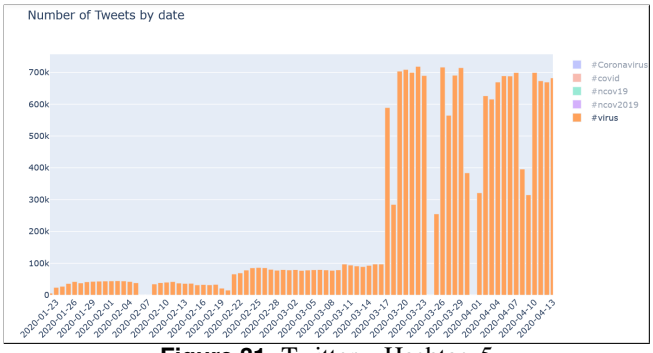

Figure 31. Twitter – Hashtag-5

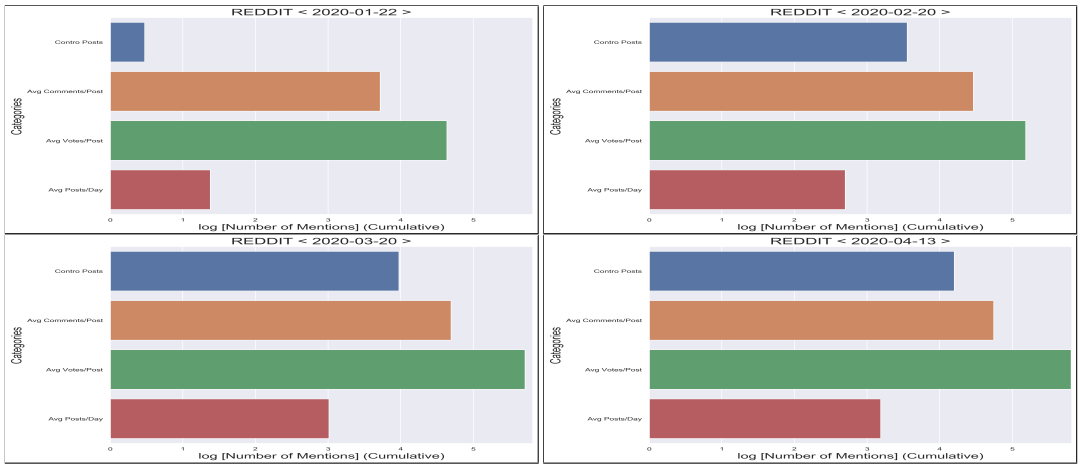

Figure 32. REDDIT – Progression of the Virus

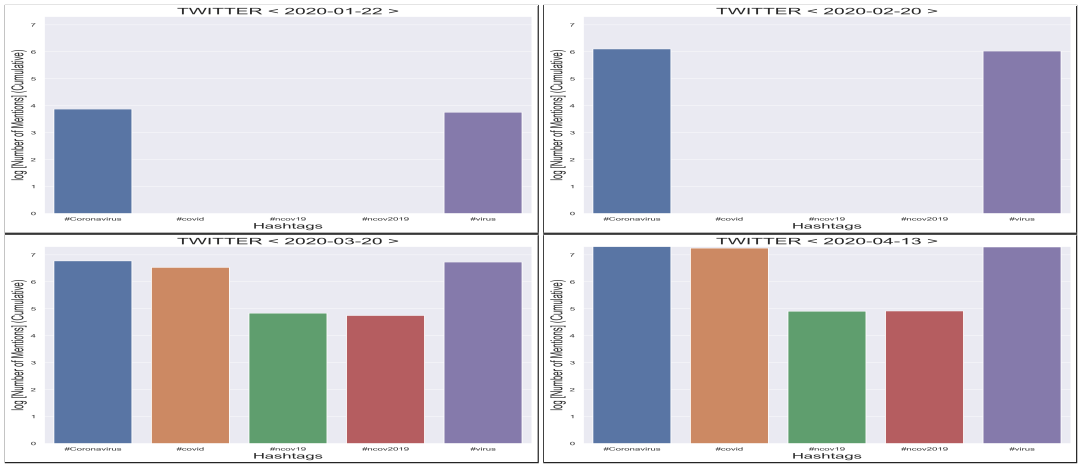

Figure 33. TWITTER – Progression of the Virus

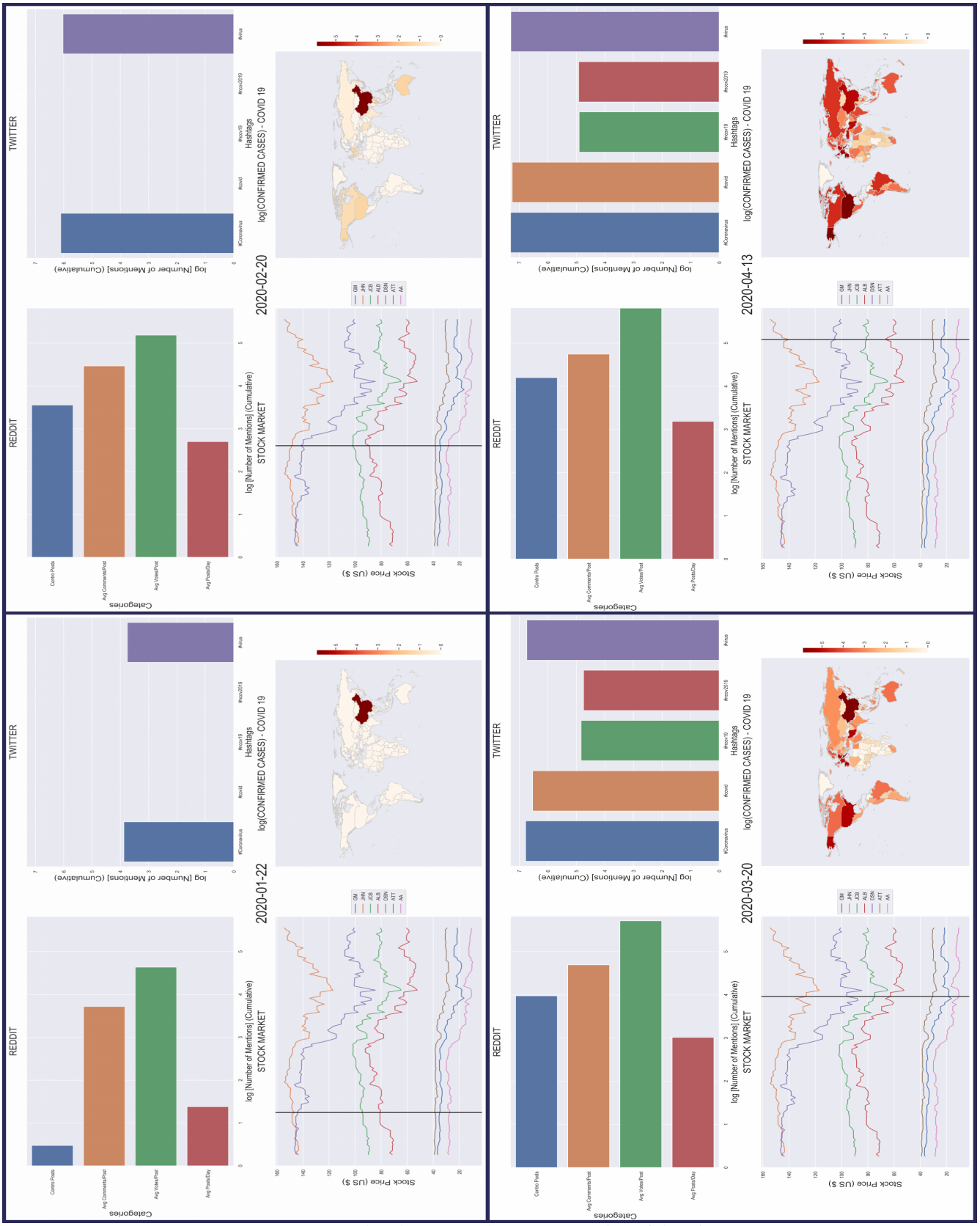

Figure 34. Animation screenshots – Progression of the Virus