Visualization of outbreaks of mental disorders around the world using twitter Metadata: Covid-19 Sentiment Analysis

Student
Mavridi Anastasia (mtp197)

Abstract

Due to the nature of the data and public interaction,twitter is becoming more and more useful to understand and model various events. The goal of this work is to use tweets as the information shared by the people to visualize topic modeling,study subjectivity and to model the human emotions duringthe COVID-19 pandemic. The main objective is to explore the psychology and behavior of the societies at large which can assist in managing the economic and social crisis during the ongoing pandemic as well as the after-effects of it. The novel coronavirus(COVID-19) pandemic forced people to stay at home to reduce the spread of the virus by maintaining the social distancing. However, social media is keeping people connected both locally and globally. People are sharing information on social media platforms which can be helpful to understand the various public behavior such as emotions, sentiments, and mobility during the ongoing pandemic.In this work, we develop a live application to observe the tweets on COVID-19 generated all over the world. In this paper, we have generated various data analytics over a period of time to study the changes in topics, subjectivity, and human emotions.

Introduction

At the time of writing this document, there were more than 54 million confirmed cases of novel corona virus cases all over the world. The number of total fatal cases exceeded 1 million globally. The number of infected people, active cases, and fatality keeprising every day. Every country is taking preventive measurements to fight against the COVID-19 pandemic. Social distancing or stay-at-home became the most widely used directive all over the world. Social distancing is forcing people to stay at home, and as a result, it is impacting the public event, business, education,and almost every other activity associated with the human life.People are also losing their jobs, and getting infected from corona and thus, stress is rising at the personal and at the community levels. Studies of behavioral economics show that emotions (Joy, Anger, Worry, Disgust, Fear, etc.) can deeply affect the individual behavior and decision-making.

Social networks have the hidden potential to reveal valuable outcomes on human emotions at the personal and community level. Monitoring tweets could be valuable particularly during and after COVID-19 pandemic as the situation and peoplere action both are changing every moment during this unpredictable time. Therefore, the analysis of twitter data might play acrucial role to understand the people behavior and responseduring the COVID-19 pandemic.

To find out the useful insights from public opinions and shared posts in social media, and to model the public emotions,we have started collecting tweets from March 2020. That’s the reason why we developed a web application in order to track, collect and analyze tweets related to COVID-19.

Related Work

At the time of this writing, the coronavirus disease (COVID-19) pandemic outbreak has already put tremendous strain on many countries' citizens, resources, and economies around the world.

One related paper to my work is “Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set” (29 May 2020) [11]. This specific work describes a multilingual COVID-19 Twitter data set that is available to the research community via our COVID-19-TweetIDs GitHub repository. Since the inception of this paper’s collection, maintained and updated the GitHub repository on a weekly basis. Over 123 million tweets were published, with over 60% of the tweets in English. This paper also presents basic statistics that show that Twitter activity responds and reacts to COVID-19-related events.

Another interesting work is “A Content Analysis of Coronavirus Tweets in the United States Just Prior to the Pandemic Declaration” (14 Dec 2020) [12]. This study examined public comments on Twitter about coronavirus in the weeks after news stories across the globe on the coronavirus outbreak. A total of 600 tweets were assessed ($N = 600$) for sentiment, risks presented, attribution of blame, and outrage. A sample of 300 tweets was taken from Sunday, February 9, through Wednesday, February 19, 2020, and an additional 300 tweets from Sunday, March 1, through Wednesday, March 11, 2020, to assess how public communication changed over time. Results show that risk, blame, and outrage differed significantly between February and March in a variety of ways. Specifically, more significant risks were noted in February than in March, with the majority of March tweets not noting any specific risks. Although most tweets did not present any specific blame, more blame was presented in February than in March, although more tweets in March attributed blame to countries and governments. Finally, outrage varied in several ways, with hazards generally being higher in February and outrage being higher in March.

Another related work is “The COVID-19 Infodemic: Twitter versus Facebook” (17 Dec 2020) [10]. The global spread of the novel coronavirus is affected by the spread of related misinformation -- the so-called COVID-19 Infodemic -- that makes populations more vulnerable to the disease through resistance to mitigation efforts. Here has been analyzed the prevalence and diffusion of links to low-credibility content about the pandemic across two major social media platforms, Twitter and Facebook. This work characterizes cross-platform similarities and differences in popular sources, diffusion patterns, influencers, coordination, and automation. Comparing the two platforms, it has been found divergence among the prevalence of popular low-credibility sources and suspicious videos. A minority of accounts and pages exert a strong influence on each platform. These misinformation "superspreaders" are often associated with the low-credibility sources and tend to be verified by the platforms. On both platforms, there is evidence of coordinated sharing of Infodemic content. The overt nature of this manipulation points to the need for societal-level rather than in-house mitigation strategies. However, limits imposed have been highlighted by inconsistent data-access policies on our capability to study harmful manipulations of information ecosystems.

In this paper, in contrast with the previous works, we are lead to statistics based on the Sentiment Analysis of twits gathered from February 2020 until today. In addition, we observe the tendency of cryptocurrencies such as Bitcoin and Etherum. All of these measurements are visualized via web components on a website.

Chapter 1 - Sentiment Analysis

With the aim to understand the people reaction during the COVID-19 pandemic, we performed extensive analysis on the sentiment of the shared tweets and the users. We used specific hashtags related to our topic in order to collect and manipulate our results.

This study has done sentiment analysis on different samples of COVID-19 specific Twitter data. We ave also analyzed COVID-19 specific tweets collected between February and December 2020, and performed sentiment analysis and topic modeling to identify Twitter users' interaction rate per topic.

Data collection

Twitter provides two API types: search API [8] and streaming API [9]. The Standard version of search API can be used to search against the sample of tweets created in the last seven days, while the Premium and Enterprise versions allow developers to access tweets posted in the previous 30 days (30-day endpoint) or from as early as 2006 (Full-archive endpoint) [8]. The streaming API is used for accessing tweets from the real-time Twitter feed [9]. For this study, the streaming API is being used since February, 2020.

Keywords

Specific keywords are being tracked for streaming the tweets.The number of keywords can be evolved continuously since the inception of this study.As the pandemic grew, a lot of new keywords can be added.

The sentiment scores

The sentiment scores are defined in the range [-1,+ 1]. If a score falls between (0,+ 1], the tweet is considered to have a Positive sentiment. Similarly, a score in the range [-1, 0) represents a Negative sentiment. And the score "0" denotes a Neutral sentiment. Scores in the extremes of the range [-1,+ 1] represent strongly Negative sentiment and strongly Positive sentiment, respectively.

Chapter 2 - Tech Stack (MEVN)

A web server is required to implement this web application. In our case one of the most modern tools was used, NodeJS. NodeJS is an open-source, cross-platform Javascript environment that executes Javascript code outside of a web browser.
It is considered one of the top tools for creating REST APIs because it uses non-blocking, event driven I/O to remain light and efficient against real time applications with high data requirements. How it works in terms of design is quite interesting. Compared to traditional web hosting techniques where each connection creates a new thread, takes up system RAM, and finally maximizes available RAM, Node.js works on a single thread, using non-blocking I/O calls. , allowing it to support tens of thousands of simultaneous connections (made in the event loop).

A quick calculation: assuming each thread has potentially 2 MB of bundled memory, running on a system with 8 GB of RAM brings us to a theoretical maximum of 4000 simultaneous connections (calculations taken from Michael Abernethy's article "Exactly what a node is"), published on IBM developerWorks in 2011; unfortunately, the article is no longer available), plus the cost of switching threads. This is the scenario you usually encounter with traditional web service techniques. Avoiding all of this, Node.js achieves scaling levels for more than 1 million concurrent connections and over 600,000 concurrent web connectors. There is, of course, the issue of sharing a thread between all customer requests and it is a potential pitfall of Node.js. First, heavy computing could choke Node's single thread and cause problems for all customers (more on that later) as incoming requests will be blocked until that computation is complete. Second, developers must be very careful not to allow an exception to leak into the core of the Node.js event loop, which will cause the Node.js instance to terminate (effectively interrupting the program).

The technique used to avoid surface-to-surface exceptions is to return errors to the caller as redirect parameters (instead of dropping them, as in other environments). Even if an untapped exception manages to go up, tools have been developed to monitor the Node.js process and perform the necessary retrieval of an interrupted presence (although you will probably not be able to retrieve the current state of the user session), or using a different approach with external boot and monit tools, or even just booting.

ExpressJS

Express is a minimal and flexible Node.js web application framework that provides a powerful set of capabilities for web and mobile applications. With countless HTTP utility methods and middleware, creating a powerful API is quick and easy. Through Express we manage the routes of our API, the authentication that is needed where it is needed as well as the interface with our database. Express is responsible through nodejs to call requests to external services for the provision of data.

MongoDB

We use MongoDB to enter the data we collect. Mongo is a document-oriented / NoSQL database and uses JSON type documents with optional formats. MongoDB was developed by MongoDB Inc. and is licensed under the Public Server License (SSPL). Each relational database has a standard schema design that shows the number of tables and the relationship between those tables. In MongoDB, there is no sense of relationship. The advantages of MongoDB over an RDBMS base are:

- Schema less: MongoDB is a document database in which a collection contains different documents. The number of fields, the content and the size of the document may differ from document to document.
- The structure of an object is clear.
- There are no complicated joins.
- Deep query-ability: MongoDB supports dynamic queries in documents using a document-based query language that is almost as powerful as SQL.
- No conversion / mapping of application objects to database objects is required.
- Uses internal memory to store the (windowed) work set, allowing faster access to data.

Why use MongoDB?

- Document Oriented Storage - Data is stored as JSON documents.
- Index on any attribute
- Replication and high availability
- Auto-Sharding
- Rich queries
- Fast in-place updates
- Professional support by MongoDB

Where can I use MongoDB?

- In Big Data
- Content Management and Delivery
- Mobile and Social Infrastructure
- User Data Management
- Data Hubs

So the reason we used Mongo, in addition to its speed performance due to the available indexing and dynamic queries, is that the data we collect and store from twitter is in JSON format. As a result, we do not need to keep any format for our database and update it along with all database records every time we need a new twitter field. Also our built-in indexing is very useful due to the large volume of data as well as the formatting that must be done by the server to be returned in the appropriate format to the client.

VueJS, VueX, VueCLI

The front-end library VueJS was used for the visual representation of the application. Vue.js is an open source model model – view – viewmodel Javascript framework for creating user interfaces and one-page applications. Created by Evan You and maintained by him and other active members of the team from various companies such as Netlify and Netguru. It is one of the most popular frontend frameworks in the period we are going through and this is because it is very small in size and simple to learn, without lacking features. We used the Vuex and VueCLI libraries to create the Med Covid-19 Single Page Application. In this way vue provides us through vuex a state management pattern through which data management in the frontend is facilitated (Data Manipulation on Frontend). VueCLI is a tool that helps us to create and manage complex SPA projects with several libraries and special setup. We used the webpack to bundle the files. The modules that were bundled are: Vuejs, Vuex, VueCLI, linter (eslint), css preprocessor (scss), library chartjs, bootstrap, element-ui, font-awesome.

Element.io, ChartJS

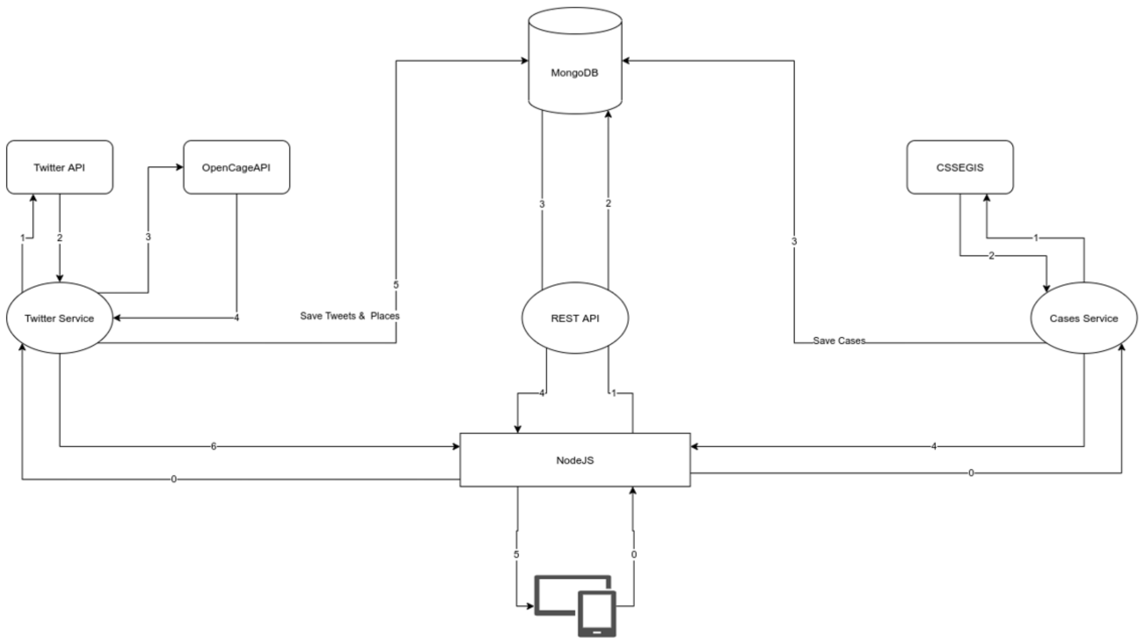
In the User Interface (UI) part we used the element.io library as well as some libraries to manage the data through vuejs. These libraries are:

- Vue-chartjs [1] for graph representation with the ChartJS library
- Bootstrap-vue [2] for grid management
- Vu2-google-maps [3] to create a responsive map
- axios library [4] for the implementation of http request.

Data Fetching & Manipulation
The data collection used the git repository [5] of the Johns Hopkins Whiting School of Engineering which was responsible from the beginning of the pandemic for recording case data. For the analysis and collection of data from the markets (stock data) we used the API of the official site of crypto comparison [6]. Also due to the fact that the twitter api does not provide the country in which the user is located but simply its coordinates, this information had to be mapped by country. To do this we used the opencage geocoder API [7]. This way before saving the data in our database we have added the country information to each twitt.

Chapter 3 - Web Services/Code Analysis

The MED COVID-19 application relies heavily on Web Services and their implementation. These services are divided into three parts. The first part concerns the collection of data from twitter, their processing and storage. The second part concerns the collection of data from CSSEGIS, their processing and storage. The third part concerns the REST API that has been implemented to serve the data in each client application. (In our case the vue single page application).



But let's look at each piece in a little more detail to understand its role and function.

Part A - Twitter Service

The Twitter Service is responsible for collecting data from twitter and storing it in the database. It is not limited to just making a get request and a save as the tweets as mentioned above do not return an area but lat and long. As a result in order to be able to group them by region and do analysis we need a geoIP service like OpenCageAPI. Through this we can request the area based on the lat and long data we have from twitter for each twit. GeoIP services are not free but this one has the most calls available in the free package so it covers us. In case the limit is exceeded, a mechanism has been implemented for automatic change to a 2nd API KEY (if available), in order to collect even more data. When this process is completed we can now save the twit to our database as well as the correlation of lat / long with the area. That way if a twit is done again from the same coordinates, we will not refer to the OpenCage API to request it, but we will receive it directly from our database

Part B - CSSEGIS

The CSSEGIS service is responsible for the collection of Covid-19 cases. To do this, because we did not find any immediately available APIs, we collect the data through the official git repository of Johns Hopkins University. A cron job has been created to download the data daily and then call the import service of the application. Once the process is complete the data is entered into the database and is available through the web application.

Part C - REST API

The third part of the application concerns the connection of the database with the frontend single page application. In order for the web app to receive the data, it is necessary to have a REST API which will undertake to take the request, will validate the check, will understand what data is requested (routing) and after receiving it from the database will configure it properly so that returned together with the response to the client (web app - single page application) At this point it would be right to emphasize the important role of data handling done by MongoDB since the volume of data is quite large. In correlation with other sizes (eg Stocks, Bitcoin, Etherum) we understand that the performance and especially the indexing provided by mongo is of particular importance and is consistent with the best user experience (User Experience). However, mechanisms and special techniques have not been used only in the backend. The front-end part of the project contains the most up-to-date reusable code techniques, with component and slot creation.

More specifically, vuejs is a library that provides us with all the necessary tools for the production of clean and reusable code. It essentially pushes us through the Single File Components to create as much reusable code as possible and call it where we need it. we created different components for each different type of card. The diagrams that had a common type use the same component. In addition to SFC, vue also provides us with the technique of slots. , footer) and the inside of these elements can be "filled" by the respective father who calls the component. This technique is particularly useful in creating the layout of the page as well as in other places (cards, etc.). Vuex we have the ability for direct and easy data manipulation through the state.

Vuex uses a single state tree - that is, this single object contains the entire state of your application level and serves as a "single source of truth". This also means that you will usually only have one store for each application. A single state tree makes it easy to locate a specific part of the state and allows us to easily take snapshots of the current state of the application for debugging purposes. The data you store in Vuex follows the same rules as the data in a Vue presence, ie the status object must be simple.

Part D - SENTIMENT ANALYSIS

Sentiment analysis (also known as sentiment mining or emotion AI) is the systematic detection, retrieval, quantification, and study of affective states and subjective knowledge using natural language processing, text analysis, computational linguistics, and biometrics.

Sentiment research is commonly used in customer-facing materials such as ratings and survey results, as well as web and social media.

For applications ranging from marketing and customer service to clinical medicine, sentiment analysis is widely used on voice of the consumer materials including feedback and survey responses, web and social media, and healthcare materials.

Classifying the polarity of a given text at the document, sentence, or feature/aspect level—whether the expressed opinion in a document, a sentence, or an entity feature/aspect is positive, negative, or neutral—is a basic activity in sentiment analysis. For example, advanced "beyond polarity" sentiment classification explores emotional states like pleasure, anger, disgust, sadness, fear, and surprise.

The General Inquirer [provided clues toward quantifying trends in text] and, separately, psychological studies that analyzed a person's psychological state based on examination of their verbal actions were forerunners to sentimental analysis.

Following that, Volcani and Fogel outlined a system in a patent that looked explicitly at sentiment and defined individual terms and phrases in text in relation to different emotional scales. EffectCheck, a current method based on their work, offers synonyms for increasing or decreasing the amount of evoked emotion in each scale.

Many additional attempts that were less sophisticated, they relied on a simplistic polar view of opinion, from positive to negative, such as Turney and Pang's work detecting the polarity of product and movie ratings, respectively.

This is a document-level project. Pang and Snyder, among others, tried to define the polarity of a text on a multi-way scale.

Pang and Lee extended the simple task of categorizing a movie review as positive or negative to predict star ratings on a 3- or 4-star scale, while Snyder examined restaurant reviews in detail, predicting ratings for different aspects of the restaurant, such as the food and atmosphere (on a five-star scale).

The 2004 AAAI Spring Symposium was the first step toward putting together different approaches—learning, lexical, knowledge-based, and so on—where linguists, computer scientists, and other interested researchers aligned interests and suggested mutual tasks and benchmark data sets for systematic computational research on affect, appeal, subjectivity, and sentiment in text.

Even though the neutral class is typically ignored in statistical classification methods since neutral texts are presumed to be near the binary classifier's boundary, some researchers argue that, as with any polarity query, three categories must be defined.

Furthermore, it has been shown that some classifiers, such as the Max Entropy and SVMs, benefit from the inclusion of a neutral class and increase overall classification accuracy. In theory, there are two ways to work with a neutral class.

Either the algorithm detects neutral language first, filters it out, and then analyses the remainder in terms of positive and negative emotions, or it produces a three-way classification in one step.

This subsequent methodology frequently includes assessing a likelihood dissemination over all classes (for example gullible Bayes classifiers as actualized by the NLTK). Regardless of whether and how to utilize an unbiased class relies upon the idea of the information: if the information is unmistakably grouped into nonpartisan, negative and positive language, it bodes well to sift the impartial language through and center around the extremity among positive and negative assumptions. On the off chance that, conversely, the information are for the most part unbiased with little deviations towards positive and negative effect, this methodology would make it harder to plainly recognize the two posts.

An alternate strategy for deciding slant is the utilization of a scaling framework whereby words usually connected with having a negative, unbiased, or positive conclusion with them are given a related number on a −10 to +10 scale (generally negative up to best) or essentially from 0 to a positive furthest breaking point, for example, +4. This makes it conceivable to change the supposition of a given term comparative with its current circumstance (for the most part fair and square of the sentence). At the point when a piece of unstructured content is examined utilizing common language preparing, every idea in the predefined climate is given a score dependent on the manner in which opinion words identify with the idea and its related score. This permits development to a more refined comprehension of notion, since it is currently conceivable to change the opinion estimation of an idea comparative with alterations that may encompass it. Words, for instance, that heighten, unwind or refute the assumption communicated by the idea can influence its score. Then again, writings can be invigorated a positive and negative supposition score if the objective is to decide the slant in a content instead of the general extremity and strength of the content.

There are different kinds of notion examination like-Aspect Based supposition investigation, Grading slant investigation (positive,negative,neutral), Multilingual slant investigation and discovery of feelings.

Sentiment of Emojis

There is another age of emojis, called emoticons, that is progressively being utilized in versatile interchanges and web-based media. In the previous two years, more than ten billion emoticons were utilized on Twitter. Emoticons are Unicode realistic images, utilized as a shorthand to communicate ideas and thoughts. Rather than the modest number of notable emojis that convey clear enthusiastic substance, there are many emoticons. However, what are their enthusiastic substance? We give the primary emoticon feeling dictionary, called the Emoji Sentiment Ranking, and draw an assumption guide of the 751 most regularly utilized emoticons.

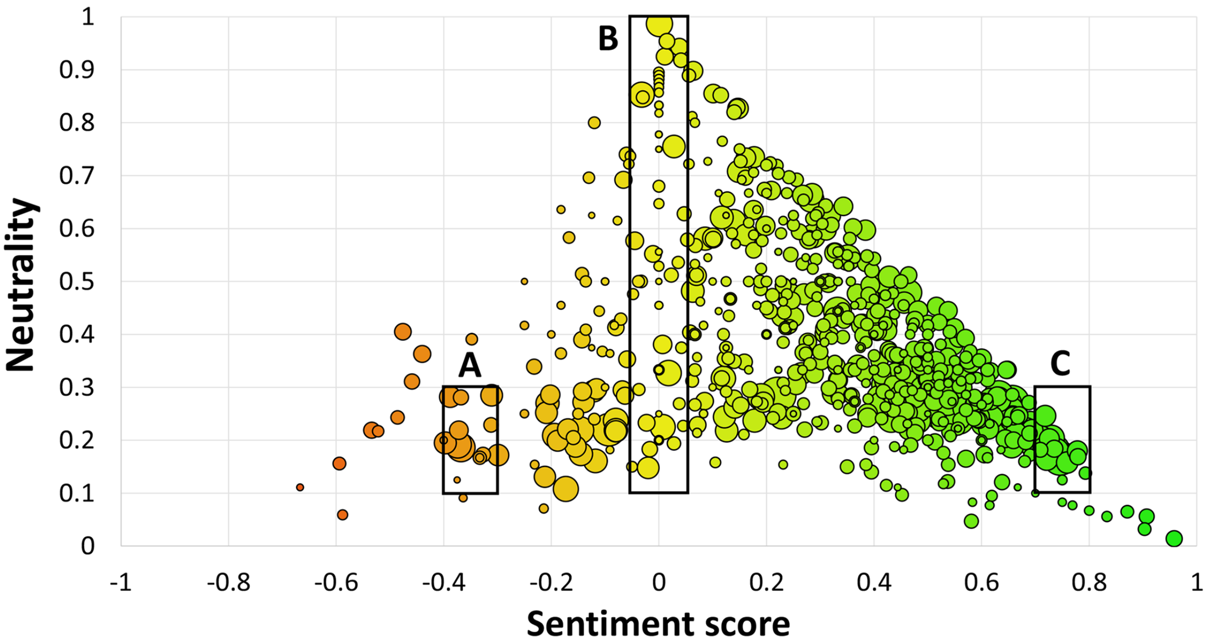
The assessment of the emoticons is processed from the notion of the tweets in which they happen. We connected with 83 human annotators to mark over 1.6 million tweets in 13 European dialects by the opinion extremity (negative, impartial, or positive). About 4% of the explained tweets contain emoticons. The assessment examination of the emoticons permits us to make a few intriguing inferences. Incidentally, the vast majority of the emoticons are positive, particularly the most well known ones. The assumption dissemination of the tweets with and without emoticons is altogether extraordinary. The between annotator concurrence on the tweets with emoticons is higher. Emoticons will in general happen toward the finish of the tweets, and their conclusion extremity increments with the distance. We notice no critical contrasts in the emoticon rankings between the 13 dialects and the Emoji Sentiment Ranking.

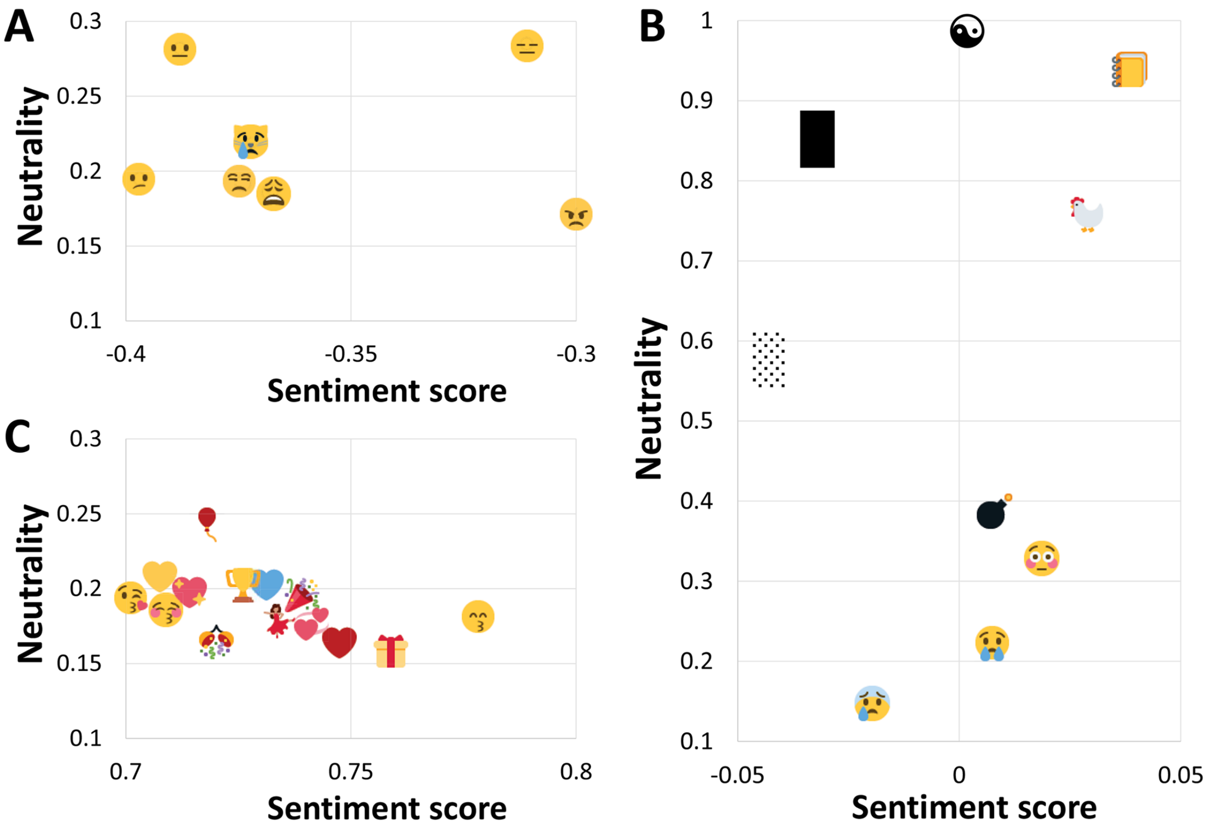
Thus, we propose our Emoji Sentiment Ranking as an European language-free asset for robotized slant examination. At long last, the paper gives a formalization of feeling and a novel perception as a conclusion bar.

Emoji	N	Position	p_-	p_0	p_+	$\bar{s}$	Name
😂	14,622	0.80	0.25	0.29	0.47	0.22	face with tears of joy
❤	8,050	0.74	0.04	0.17	0.79	0.75	heavy black heart
♥	7,144	0.75	0.04	0.27	0.69	0.66	black heart suit
😍	6,359	0.76	0.05	0.22	0.73	0.68	smiling face with heart-shaped eyes
😭	5,526	0.80	0.44	0.22	0.34	-0.09	loudly crying face
😘	3,648	0.85	0.05	0.19	0.75	0.70	face throwing a kiss
😊	3,186	0.81	0.06	0.24	0.70	0.64	smiling face with smiling eyes
👌	2,925	0.80	0.09	0.25	0.66	0.56	ok hand sign
💕	2,400	0.76	0.04	0.29	0.67	0.63	two hearts
👏	2,336	0.78	0.10	0.27	0.62	0.52	clapping hands sign

Emojitracker	Tweets with emojis ~4 billion	Different emojis used 845	Pearson correlation /	Spearman rank correlation /
Emoji Sent. Rank.				
$N \geq 1$	69,673	969 (721)	0.945*	0.897*
Emoji Sent. Rank.				
$N \geq 5$	69,546	**751** (608)	0.944*	0.898*

doi:10.1371/journal.pone.0144296.t001





In our case we used the sentiment library for nodejs.

As the library states:

> “Sentiment is a Node.js module that uses the AFINN-165 wordlist and Emoji Sentiment Ranking to perform sentiment analysis on arbitrary blocks of input text.”

Sentiment provides several things:

- Performance
- The ability to append and overwrite word / value pairs from the AFINN wordlist
- The ability to easily add support for new languages
- The ability to easily define custom strategies for negation, emphasis, etc. on a per-language basis

The library upholds AFINN which is a rundown of words appraised for valence with a whole number between less five (negative) and in addition to five (positive). Assessment examination is performed by cross-checking the string tokens (words, emoticons) with the AFINN list and getting their individual scores. The relative score is basically: amount of every token/number of tokens. So for instance we should take the accompanying:

I love cats, but I am allergic to them.

That string results in the following:

```
{
    score: 1,
    comparative: 0.1111111111111111,
    calculation: [ { allergic: -2 }, { love: 3 } ],
    tokens: [
        'i',
        'love',
        'cats',
        'but',
        'i',
        'am',
        'allergic',
        'to',
        'them'
    ],
    words: [
        'allergic',
        'love'
    ],
    positive: [
        'love'
    ],
    negative: [
        'allergic'
    ]
}
```

- Returned Objects
    - Score: Score calculated by adding the sentiment values of recognized words.
    - Comparative: Comparative score of the input string.
    - Calculation: An array of words that have a negative or positive valence with their respective AFINN score.
    - Token: All the tokens like words or emojis found in the input string.
    - Words: List of words from input string that were found in AFINN list.
    - Positive: List of positive words in input string that were found in AFINN list.
    - Negative: List of negative words in input string that were found in AFINN list.

For this situation, love has an estimation of 3, hypersensitive has an estimation of -2, and the leftover tokens are nonpartisan with an estimation of 0. Since the string has 9 tokens the subsequent similar score resembles: (3 + - 2)/9 = 0.111111111

This methodology leaves you with a mid-purpose of 0 and the upper and lower limits are obliged to positive and negative 5 separately (equivalent to every token!). For instance, we should envision a unimaginably "positive" string with 200 tokens and where every token has an AFINN score of 5. Our subsequent similar score would resemble this:

```
(max positive score * number of tokens) / number of tokens
(5 * 200) / 200 = 5
```

Tokenization

Tokenization works by parting the lines of info string, at that point eliminating the uncommon characters, lastly parting it utilizing spaces. This is utilized to get rundown of words in the string.

Benchmarks

An essential inspiration for planning estimation was execution. Thusly, it incorporates a benchmark content inside the test index that thinks about it against the Sentimental module which gives an almost comparable interface and approach. In view of these benchmarks, running on a MacBook Pro with Node v6.9.1, estimation is almost twice as quick as elective usage:

```
sentiment (Latest) x 861,312 ops/sec ±0.87% (89 runs sampled)
Sentimental (1.0.1) x 451,066 ops/sec ±0.99% (92 runs sampled)
```

Validation

While the precision given by AFINN is very acceptable thinking of it as' computational presentation (see above) there is consistently opportunity to get better. In this manner the slant module is available to tolerating PRs which change or alter the AFINN/Emoji datasets or execution given that they improve precision and keep up comparable execution attributes. To set up this, we test the assessment module against three named datasets given by UCI.

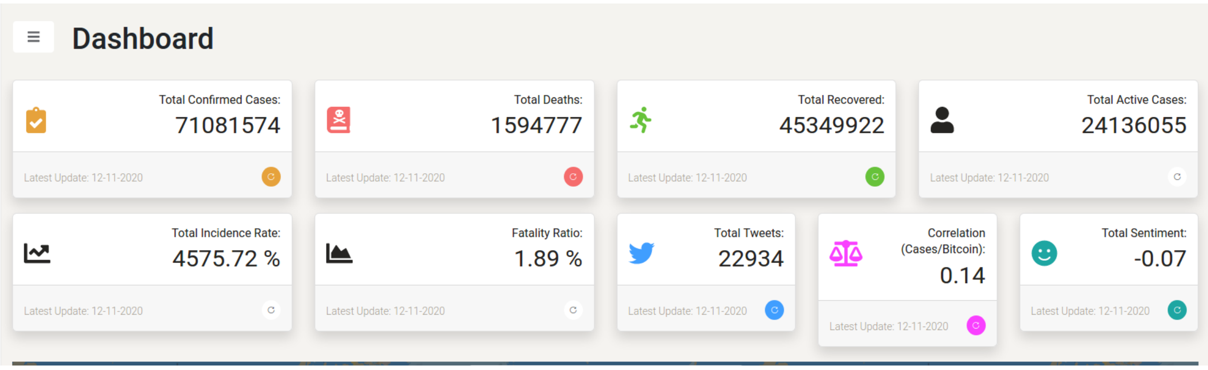
Rand Accuracy

```
Amazon:   0.726
IMDB:     0.765
Yelp:     0.696
```

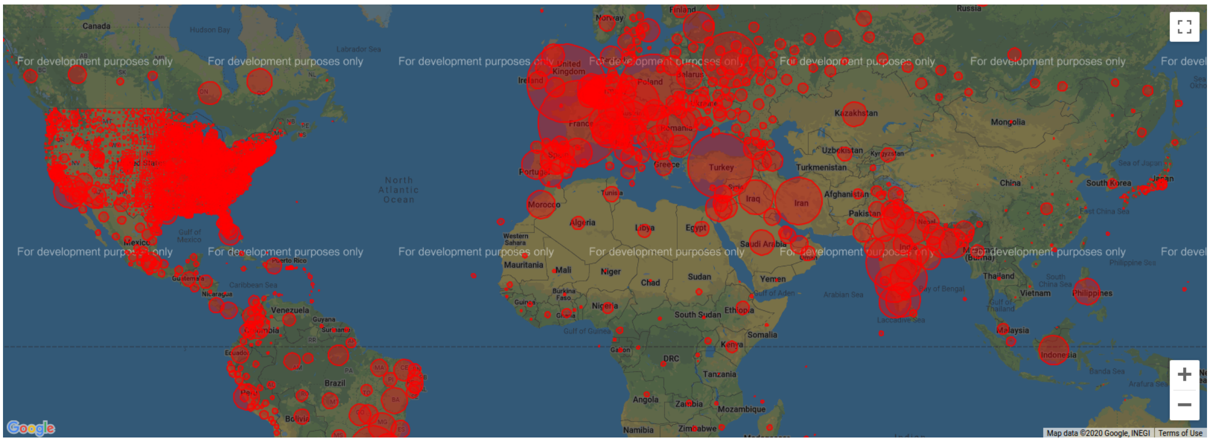
Chapter 4 - Graph Analysis / Components

Dashboard

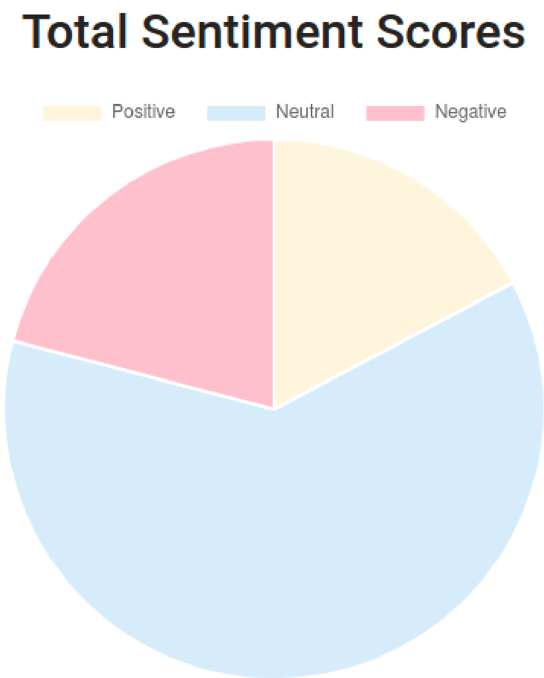
In the dashboard as we see from the pictures below are presented some general data about the course of COVID-19. (Total cases worldwide, total deaths, total treated patients, active cases). We also see some statistics such as the total growth rate of cases since the beginning of the pandemic at a rate of 14001.71% (total incidence rate), the death rate worldwide 2.33%. We also have the total number of tweets mentioned in COVID-19 as well as the citizen sentiment resulting from these tweets. Through this number we can understand the negative psychology of the public. We also have a value that shows the correlation between COVID-19 and bitcoin which we see is not great but certainly not negligible 0.06.

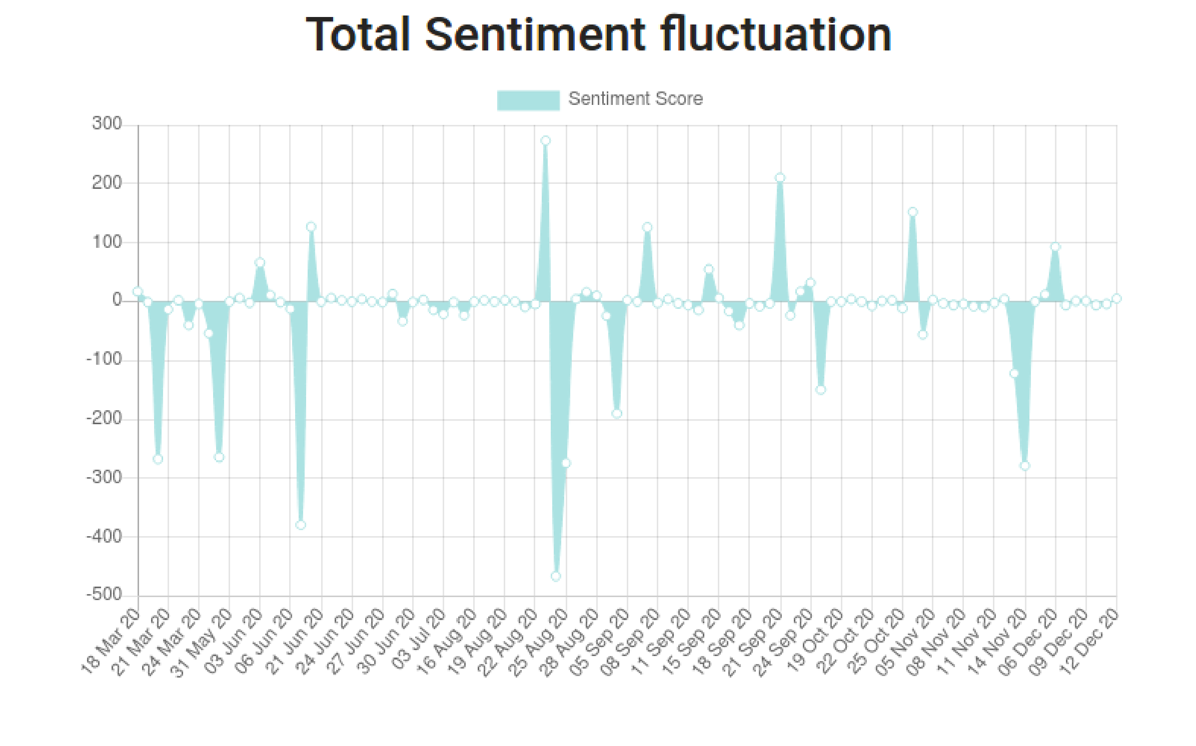


Then we have a map that represents cases across the globe. More specifically, we see the amount of cases per region and we can click on the country to see more information.



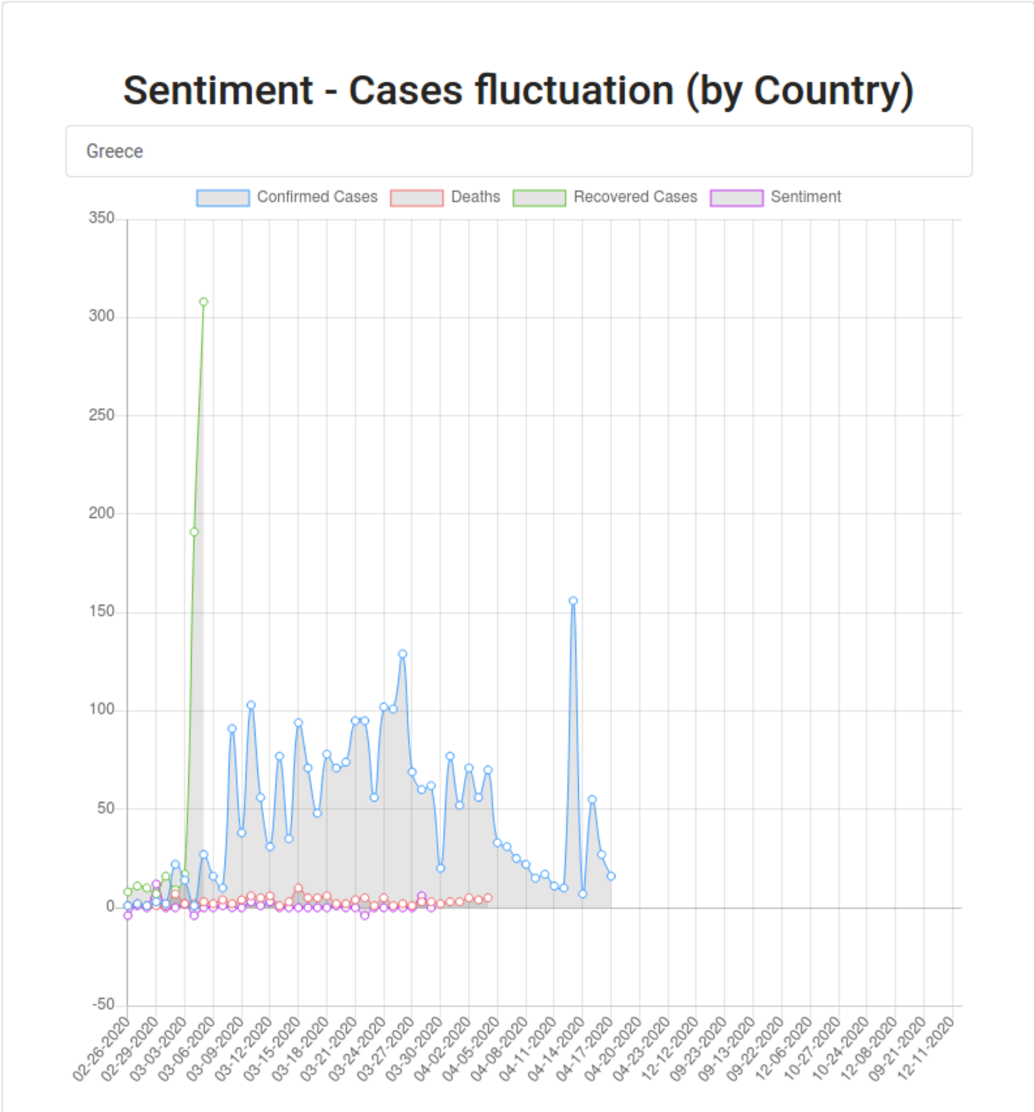
At the bottom of the page we see how the sentiment scores are distributed. That is, how many positives we have, how many negatives and how many neutrals, as well as their distribution over time.





Continuing on the Charts page we see some more graphs that give us a broader picture of the data.

In the first diagram we see the distribution of cases over time, divided into confirmed cases, deaths, cases that have recovered. In the second diagram we have a comparison of COVID-19 data with crypto coins such as etherum and bitcoin. In other words, we observe the evolution of the pandemic every day and whether it affects the markets on a daily basis. In the third diagram we observe the gradation of the emotions and the psychology of the world through the tweets that we collected. This chart may provide useful information on when and why COVID-19 affects the world's psychology the most. Then as we see the countries with the most cases per category are represented (eg confirmed, deaths, recovered). Finally we have a diagram that represents / compares the variation of sentiment in relation to the COVID-19 data by country.



Continuing on the Tables page we see some more tables with detailed data. First we see a table with the tweets that we collected and concern COVID-19. From here we can see the date, the place, the text of the tweet as well as the sentiment analysis score. Finally, there is a table with the total data of the cases per country. The user can search through the search the country he wants.

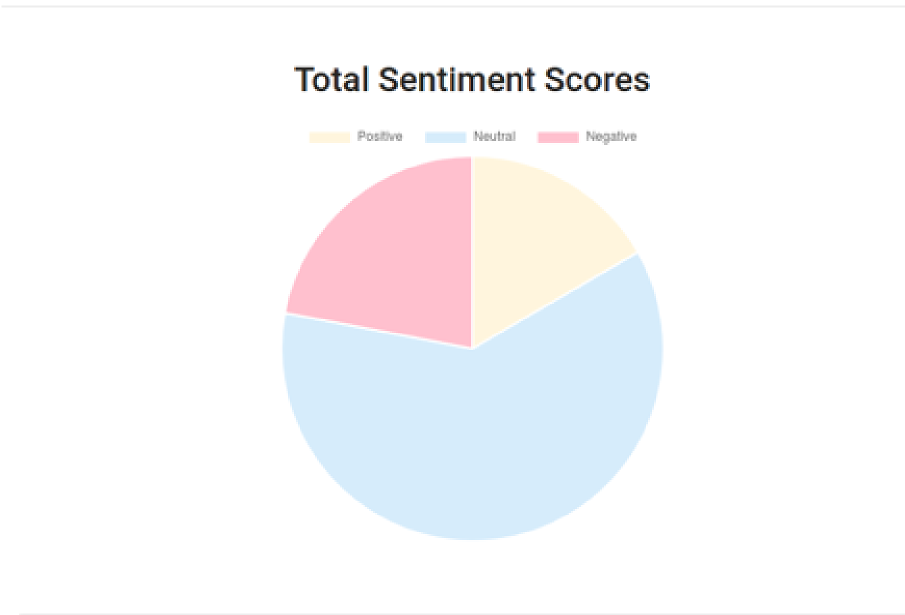
Chapter 5 - Results

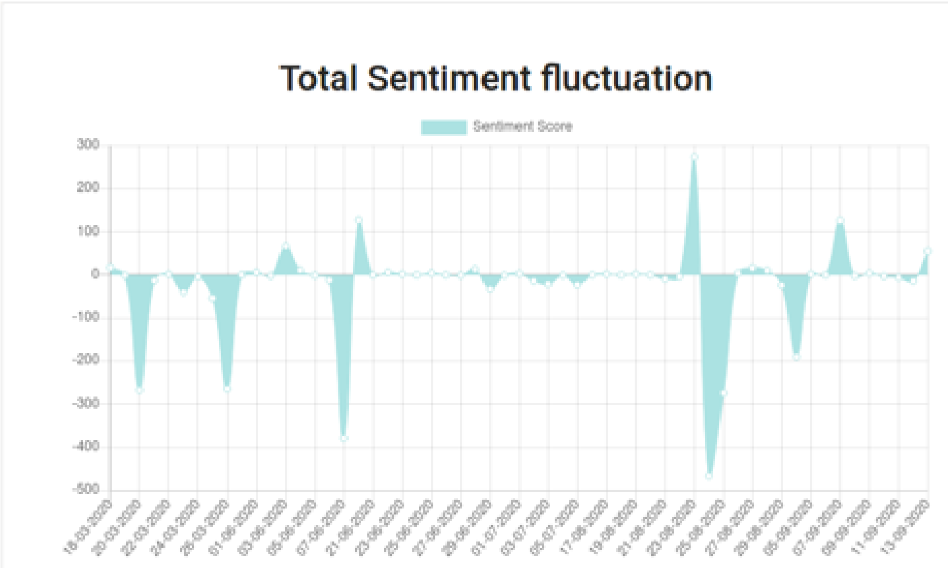
In this project, we developed an interactive web application for tweets tracking on COVID-19 and producing insights dynamically. We performed sentiment analysis and related that with trending topics to find out the reason behind a sentiment for better understanding of the human emotions.

The conclusions that emerged from the implementation of this mechanism are evident from the diagrams and we can group them into two categories which we will study / analyze later in the chapter.

5.1 Covid Cases - Sentiment Analysis

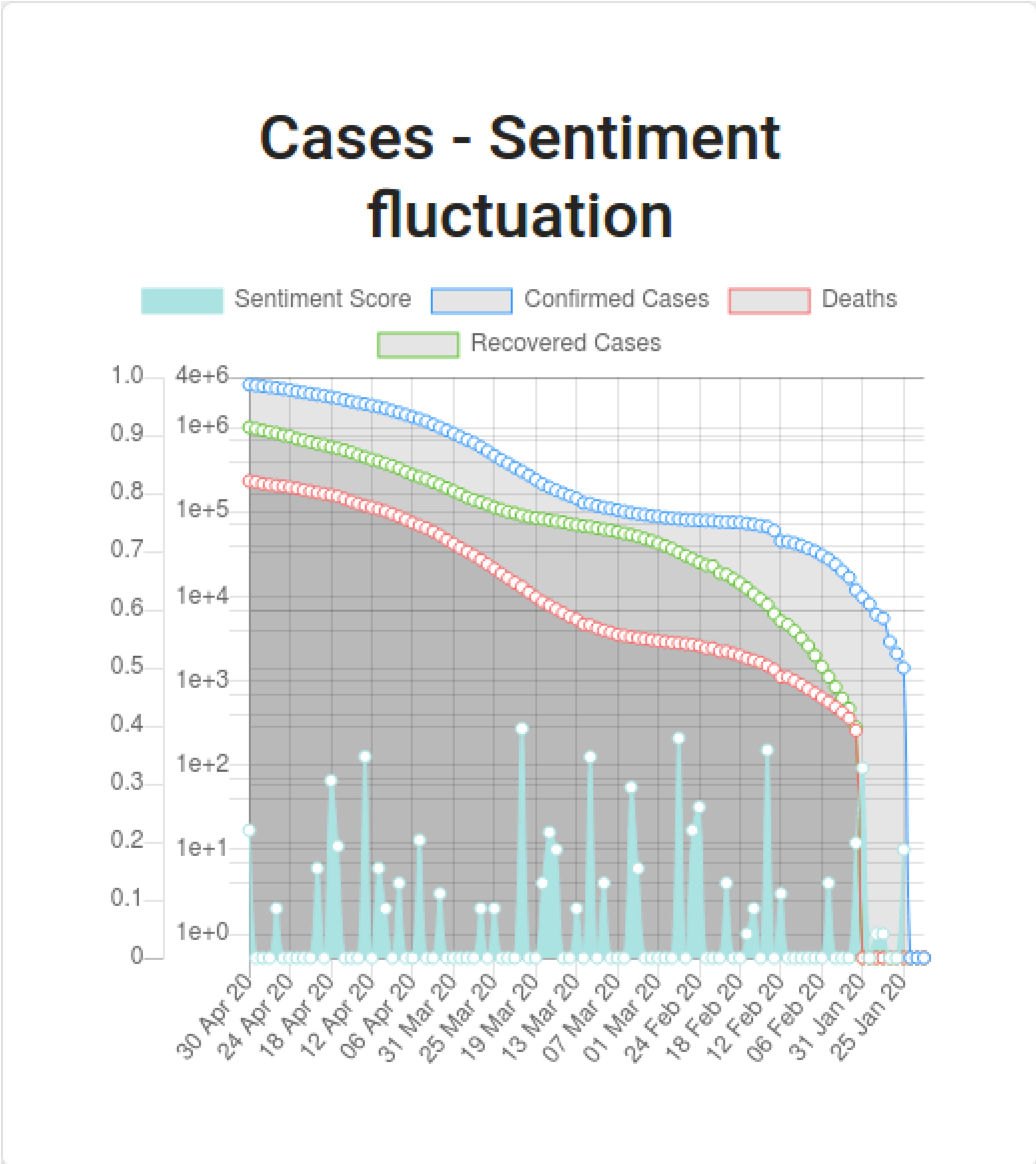
Based on the total sample of tweets we collected and applying the algorithm of sentiment analysis we visualized the results ending in the following diagrams.



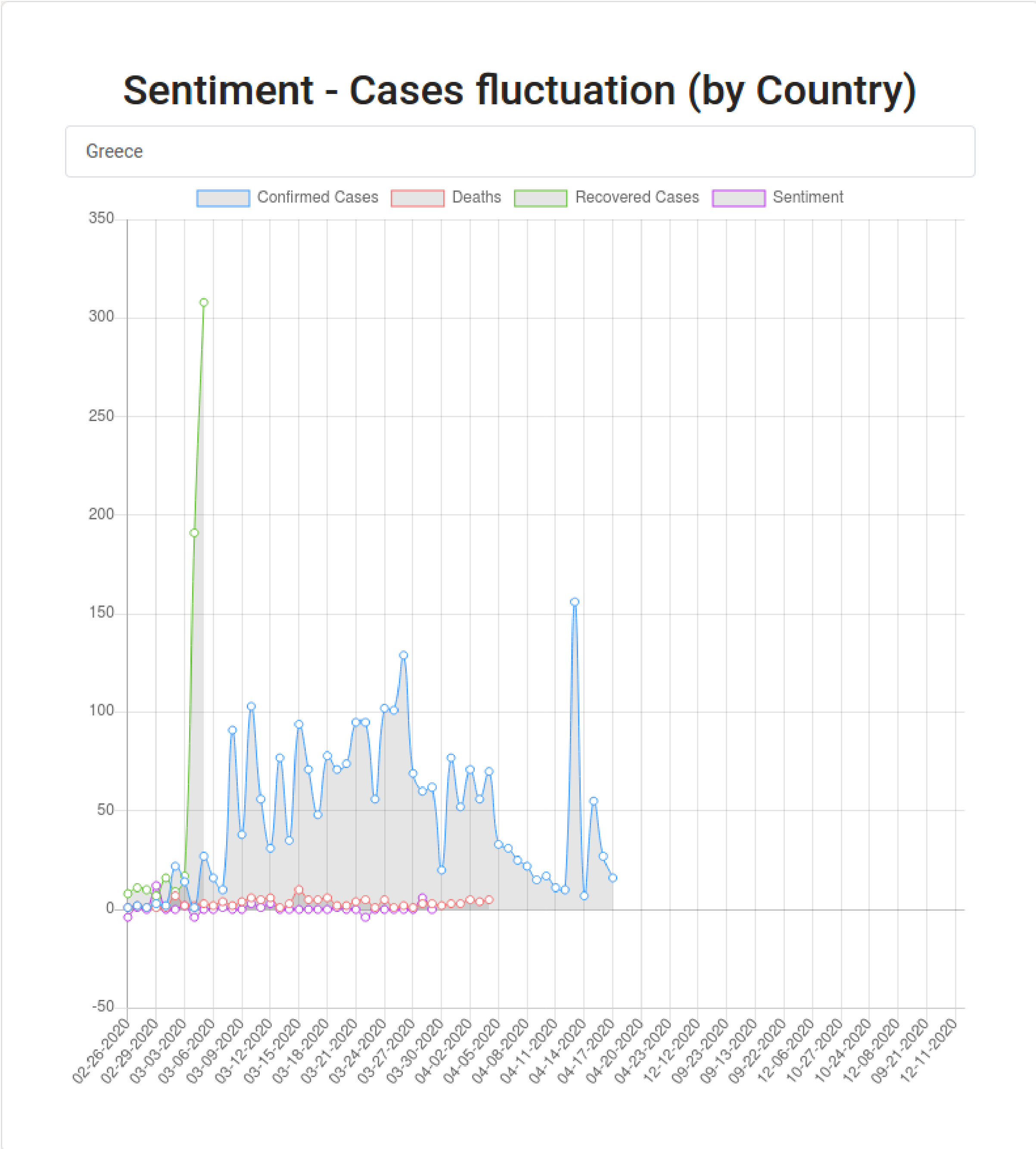


* Data on tweets have been collected since February 2020.

From the above diagrams we can draw useful conclusions about the influence of Covid-19 on human psychology both overall from the beginning of the pandemic (Chart 1) and by time period.



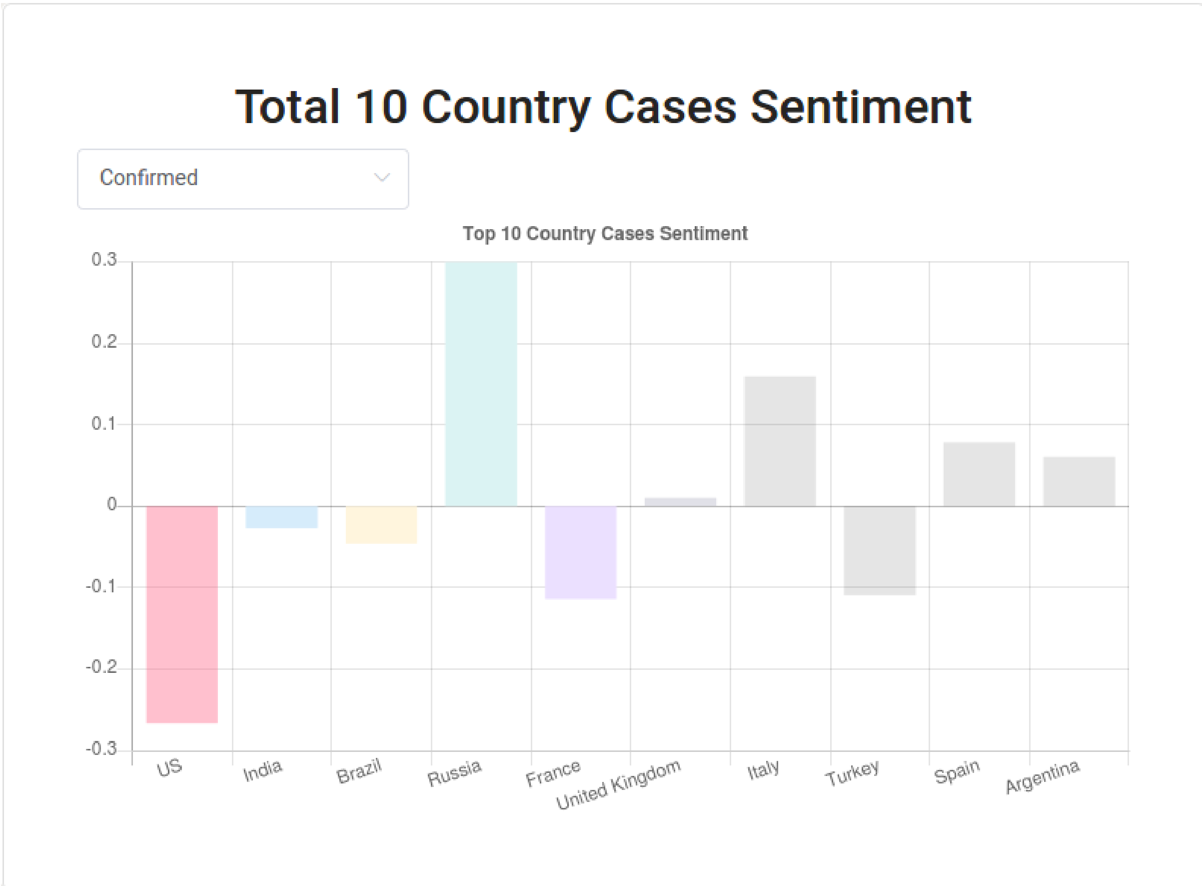
Even through the diagram Sentiment - Cases fluctuation (by Country), we can draw conclusions about how much and when each country was psychologically affected by the pandemic.



So we see that countries like the United States and the United Kingdom that have the most cases show a big change in psychology at times of rising cases and especially at the beginning of the pandemic (eg US 26/02).

However, this conclusion can be further analyzed as the sample of tweets we have is not very large and corresponds to the number of cases in the countries.

Finally, by studying the diagrams Total 10 Country Cases and Total 10 Country Cases Sentiment we can draw useful conclusions for the correlation of the number of cases, deaths or recovered with the psychology of their citizens.

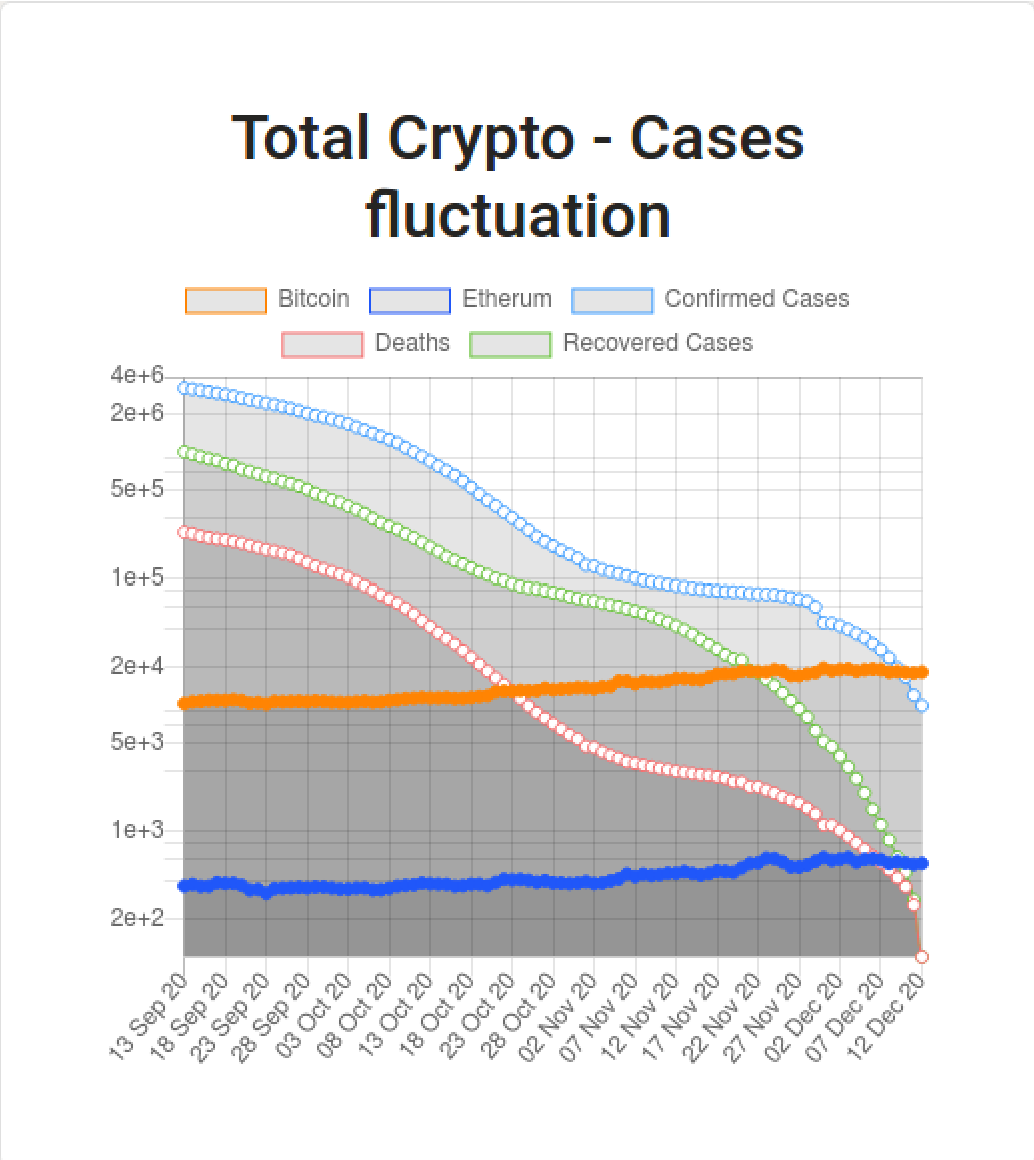


For example, the conclusions we can draw are that the country with the lowest sentiment score (psychology index) is America, which seems logical since it has the most cases in the world.

Even countries like Russia, which is the first to start vaccinations, it seems that the psychology of its inhabitants has returned to exceptional levels.

5.2 Covid Cases - Economical Analysis

In this section we will study possible correlations between Etherum, Bitcoin and Covid cryptocurrencies.
As we can see from the diagram, cryptocurrencies do not seem to be negatively affected by the pandemic.



We observe a particular increase in both etherum and bitcoin and any declines observed last only a short distance in relation to their overall course in the time we are studying. Trying to cross-reference the information we came up with corresponding results.

Finally, this analysis indicates that there is initially a negative relationship between the number of reported cases, deaths, Bitcoin; however, the relationship becomes positive in the later period. The findings for Ethereum are also similar to the Bitcoin evidence, however, the interactions are weaker compared to Bitcoin. This shows the hedging role of cryptocurrencies against the uncertainty raised by COVID-19.

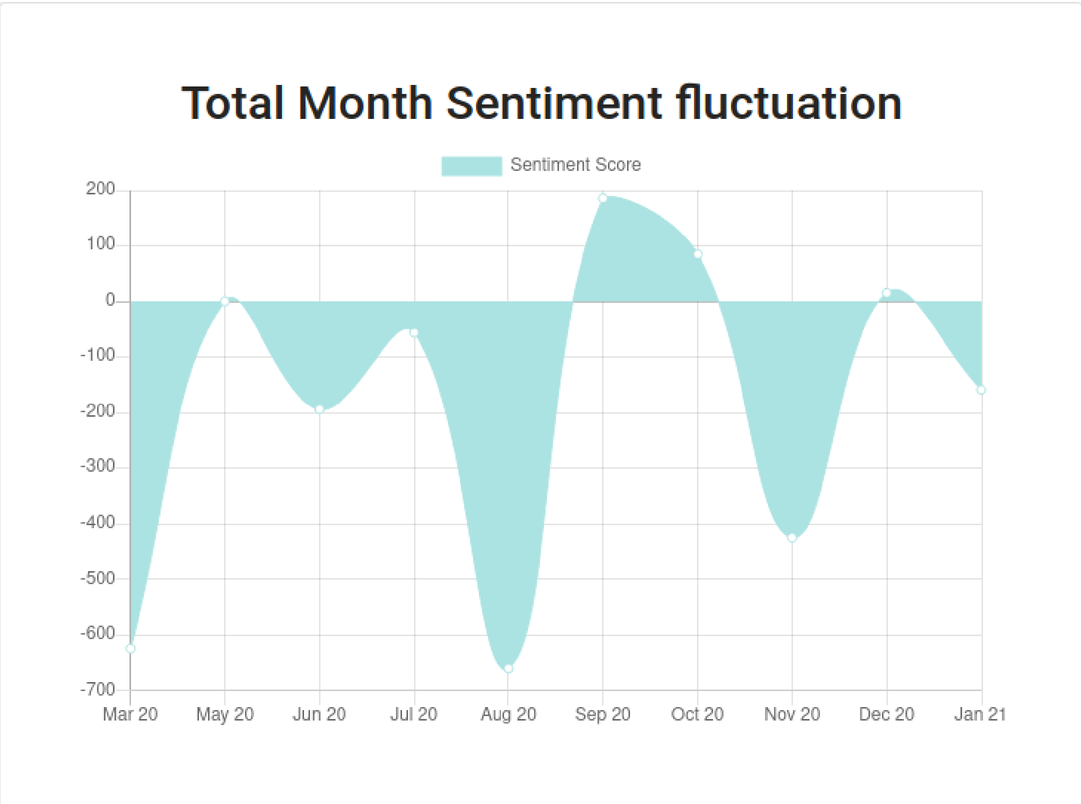
5.3 Conclusions

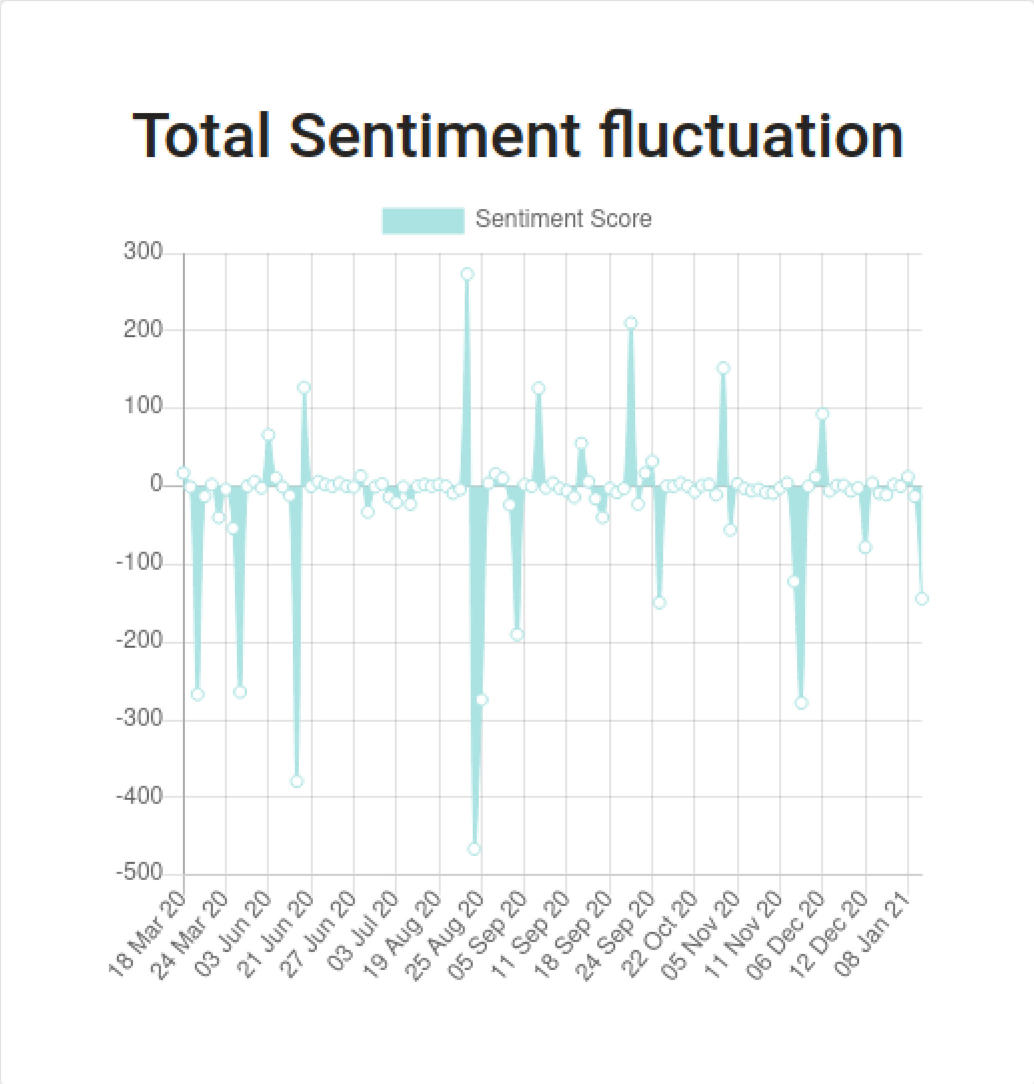
As a result the conclusions of the project can be grouped in two categories. The first one is the data analysis conclusion and the second one is the technological-dev conclusion.

Data Analysis Conclusion

The data analysis conclusion refers to the understanding of the diagrams.

We can see from the visuals that as the pandemic goes on the sentiment of the population has been affected. In some cases we have ripples given the fact that we have some extreme changes on the daily covid cases or deaths.

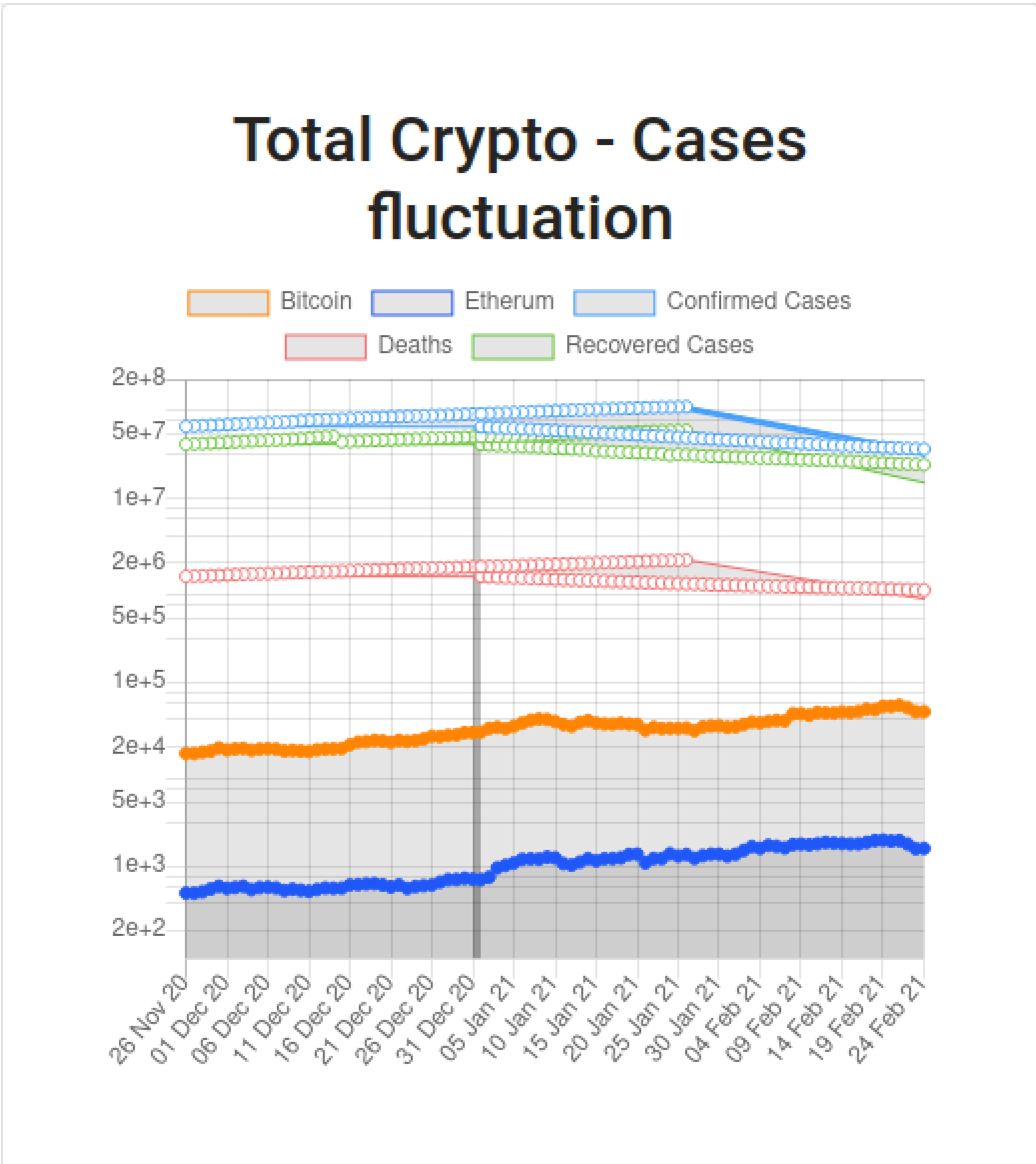




This reasonably affects the psychology of the population.
We can easily analyze some of the data we receive from the previous chart.
At first we see that on the beginning of the pandemic the sentiment score is pretty low. This fact can be characterized as logical due to the new situation the world came in.
This curve lasts until the end of May where the lockdowns were started to come to an end.
All this time the psychology of the people was under 0 which indicates the new situation we were all introduced.
Another interesting observation is that from late July to the End of August the psychology started to drop again.
The reason behind that is the beginning of the second covid-19 wave which lasted until September.
The 9nth month the psychology was up the markets started to open and we thought that the pandemic was behind us.

Again we were wrong, and this situation lasted only for a month.
On October the cases were increased and so the sentiment went bellow zero.
Since December though with the announcement of the Vaccines the psychology started to stabilized near zero with some ups and downs.
The sentiment analysis can be called pretty accurate and it reflects the history of the events.

Additionally we can extract some conclusions from the cryptocurrency charts.



We can see that the price is constantly increased during the pandemic with small daily drawbacks. The reason behind that are the changes in the global economy. During the pandemic the global economy has suffered a lot, the law of supply and demand affects millions of products and its prices. The factories in China have been closed for quite some time during the last time and now with the healthy protocols that have been applied the production chain has been taken a serious hit.

As a result one of the most necessary tools for cryptocurrency and miners, which is the graphic card (GPU) has received a large price increase.
Graphic cards are responsible for the most heavy computations on the mining process and as a result is a crucial part.
While the pandemic continues the production of GPUs is low and the prices have increased. Also the stocks are empty and the supply chain has received a big hit.
As a result the prices on the cryptocurrencies have received a big increase. Bitcoin reached 33.000$/bitcoin in the first month of 2021 which is a record and it continues to break record after record.

Another explanation for the rise in bitcoin prices may be that large-scale institutions such as pension plans, university endowment funds and investment trusts have had a large inflow of investors.

In 2017, the cryptocurrency system was ruled by person retail speculators, numerous of whom were pulled in to bitcoin’s paucity and the truth that it stood exterior the worldwide money related framework. Many huge names have contributed intensely. This all makes a difference to extend believe within the cryptocurrency and shows that it is getting to be more common.

In addition, Bitcoin has been supported by some expansive consumer-facing installment names. PayPal presently permits clients to purchase, hold and offer bitcoin specifically from their PayPal accounts. The number of merchants tolerating bitcoin as a shape of installment is developing quickly.

In October it reported a modest bunch of bitcoin-related credit and charge cards with driving crypto trade Coinbase. With increasingly ways of utilizing bitcoin, it ought to cruel that more individuals will need to hold it. Also, Bitcoin has gotten to be much more develop since the days when it was utilized basically as a strategy to buy drugs on the dim web on Silk Street.

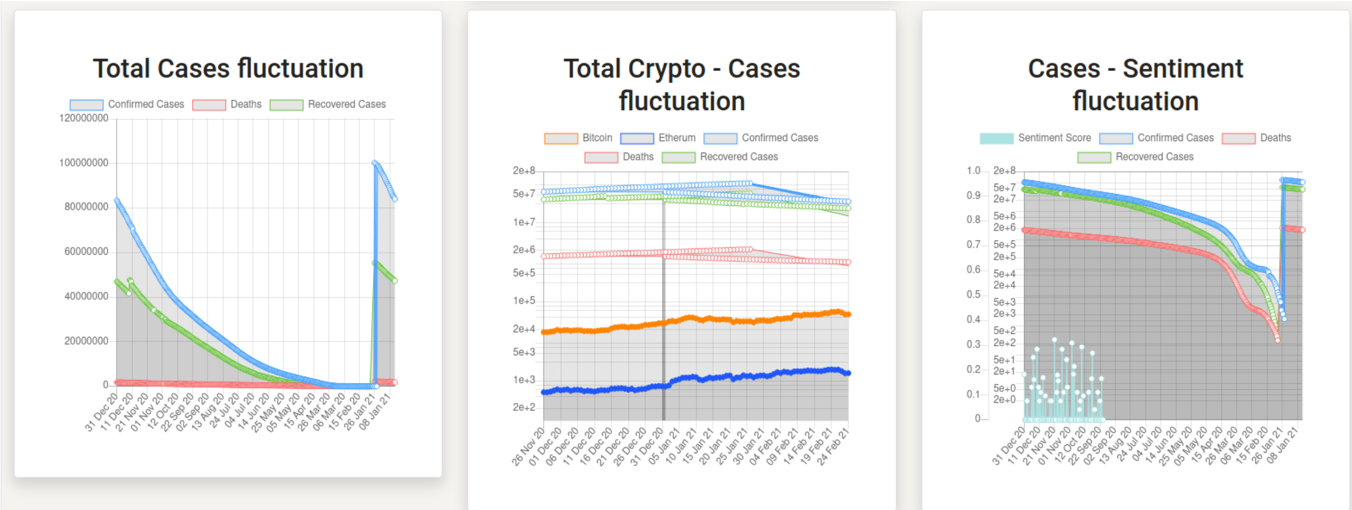
Other than all this standard entusiasm, the massacre brought by COVID-19 has driven to colossal boost bundles from governments around the globe and numerous central banks printing more cash. This may drive up swelling, which in turn brings down people’s acquiring control. Within the confront of this risk, speculations like bitcoin are being consider a store of esteem.

Additional conclusion that we can extract from the pie chart is that most of the people and their psychology stays neutral and that the positive and negative values they share two pieces of the pie.

Based on the pandemic curves we saw on the previous charts this is reasonably explained by the points and the curves on the map.

Except some extreme diversions the sentiment value tends to 0.

From the next row of charts we can find data analysis and representation for the cases fluctuation on par with cryptocurrencies and sentiment once again.



We can have an overview of the generic course of the pandemic and create comparisons on the results.

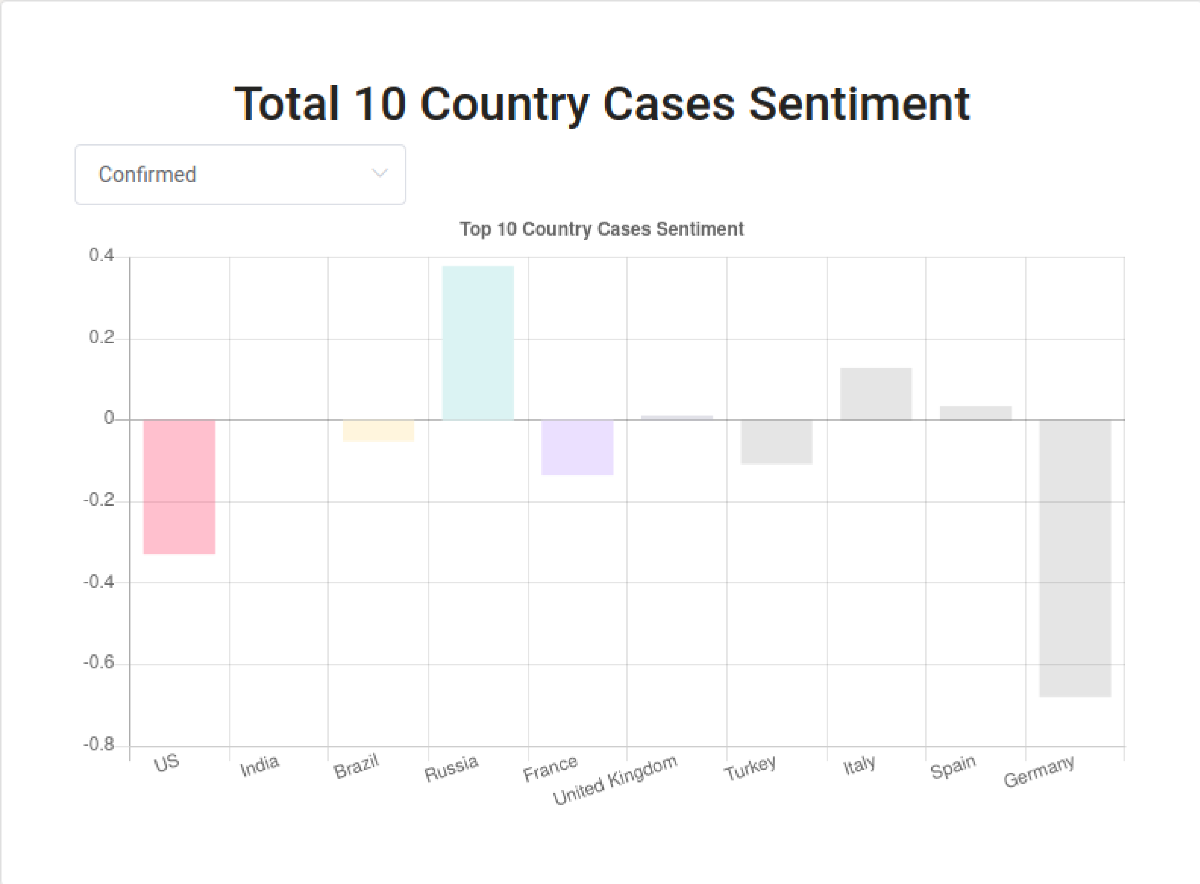
Another useful tool here is the option to filter further the results and see only the deaths or the recovered cases. With the current data were the deaths follow the cases and also the recovered people this comparison can not be that useful but in future with the appearance of the vaccines that probably will change.

On the next and on the last rows we have very interesting sections.

We are able to analyze the top 10 countries based on their cases and also analyze their sentiment.

In the most cases as soon as the country is higher on the list the opposite is true for the sentiment.

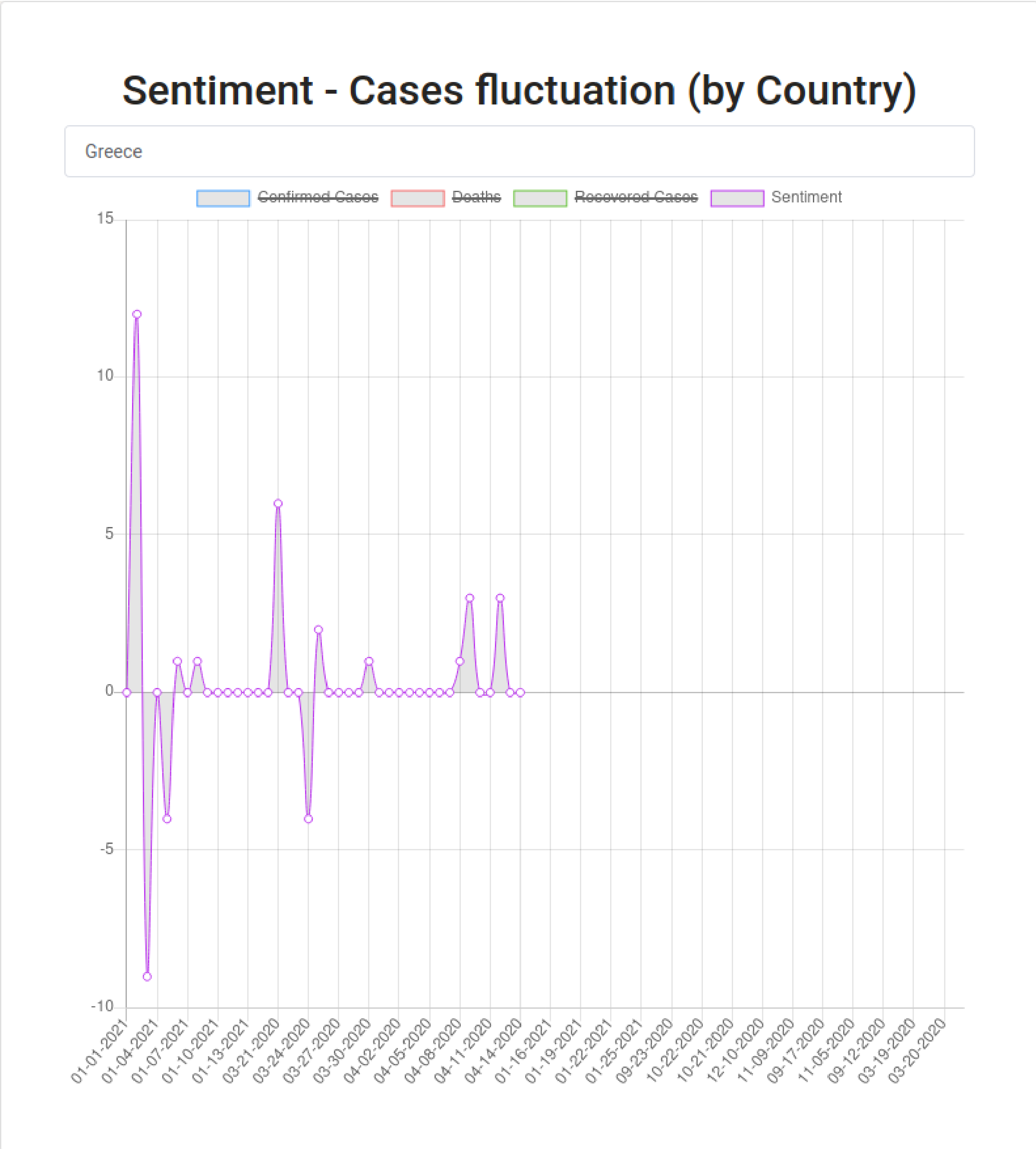
Russia is a strange case since the vaccine was introduced since last summer and the sentiment is pretty high even though the cases are still up.



Germany on the other hand which has so many deaths and cases daily event until now the sentiment is pretty low. One of the reasons for that could be the continued lockdowns which they use to control the pandemic.

For Greece we can not extract a result from the data and relate it with the current status on the country.

Greece such as Germany and many other European countries use global lockdowns as precautionary measure. This measures create multiple fluctuation on the psychology of the civilians and this is also represented on the chart.


Sentiment - Cases fluctuation (by Country)
Greece
Confirmed Cases
Deaths
Recovered Cases
Sentiment
15
10
5
0
-5
-10
01-01-2021
01-04-2021
01-07-2021
01-10-2021
01-13-2021
03-21-2020
03-24-2020
03-27-2020
03-30-2020
04-02-2020
04-05-2020
04-08-2020
04-11-2020
04-14-2020
01-16-2021
01-19-2021
01-22-2021
01-25-2021
09-23-2020
10-22-2020
10-21-2020
12-10-2020
11-09-2020
09-17-2020
11-05-2020
09-12-2020
03-19-2020
03-20-2020

Development Analysis Conclusion

Based on the information and the description we provided in the previous chapters we have created a mechanism for storing tweets ordered by date and a second one for covid-19 cases retrieval.

Due to the large scale of data, manipulation of results for each query is a major issue. The chart.js library we used and any other frontend library is not able to handle and represent easily charts for millions of data mixed.
In order to avoid this issue we created computational services for the “heavy” queries.

The purpose of these services is to handle and format the data of each query based on the information the client requests. With that procedure we are able to serve data to the user easily without any additional delay.

Twitter’s Standard API that we used for tweets retrieval has some limitations.
We were able to get tweets only for the last 7 days. With that limitation we were not able to select randomized data for all the hashtags that we wanted.

This limitation can easily be bypassed or skipped by selecting another plan but that comes with an additional cost. It is also possible for universities and other organizations to get licenses for academic purposes.

In any of these cases we just have to change the API keys and the application will collect the additional tweets.

Another issue that we faced was the amount of data that we had to store for tweets.
We did some testing for a completed week and we found out that we needed to store about 200.000 tweets for only a week only for one hashtag. Those tweets occupy ~60Mb. That means 240Mb/month , 2.8GB/year only for one covid-19 related hashtag.

We couldn’t find a plan to store our data for free in a MongoDB hosting service (Atlas).
This problem can be addressed by upgrading to a more expensive plan on the hosting service.

Another issue that is also generated from the big amount of tweets we were receiving is the geolocation service.

This is a free service which comes with a limit of 2500 requests / day.
Even if we split our fetching mechanism and add a limit of 2500 tweets per day we are going to ignore some tweets from our sample.

A simple solution here is to upgrade the license and add more requests per day in order to get at least 170.000 per week or to create a second key and switch to that one once the limit has been reached. The latest solution is not that simple and we have to restart the server for that but it should work.

Other than that we didn't find any other mentionable issues with the implementation. I believe that all of these issues are able to be addressed with proper licenses and pricing plans.

Chapter 6 - Conclusions and discussion

This thesis aimed to identify and monitor the relationship between the COVID-19 pandemic and the human psychology and bitcoin market.

Based on the tweets we collected and the sentiment analysis we applied on those tweets, it can be concluded that social media and as a result the human psychology is indeed affected by the course of the pandemic.

The results, which are represented by some charts on the application, indicate that during the pandemic the human psychology, as it was reason enough, is affected. This affection is clearly seen on social media and in our case on twitter.

With the additional charts we created it is now easier to search, find and monitor the most affected places on earth and also the change on their psychology.

A possible and very interesting additional comparison that can be added to the project is the comparison of the sentiment of the users before and afte the pandemic. In that case we could easily see and identify by how much the psychology of the human race was affected.

Due to the limitations we discussed in previous chapters we were not able to collect an amount of tweets of that size.

Nevertheless we believe that with the appropriate licenses and some small modifications on the application the results would be very interesting for future analysis.

Another aspect of this thesis was to monitor and identify the influence of the pandemic to the cryptocurrencies. This analysis can be characterised as more complex due to the economical and political factors that have an impact on the price of the cryptocurrencies.

Based on the diagrams we created with actual and live data during the pandemic we can clearly see that affection but we are not absolutely sure if this huge increase is caused by the pandemic.

To have a more clear view on that, we have to wait and continue to monitor the course of cryptocurrencies. After the end of the pandemic as soon as the global markets and the global economy reach a stable point we will then re-evaluate the course of cryptocurrencies analyzing if and by how they were affected.

With the data we have right now it seems insecure to arrive at a specific conclusion.

As additional topics for study and further development in future could be study of other social media for specific covid-related topics.

Another aspect could be the automation of the monitoring process on social media.
In this case the user will be able to select a theme/hastag/topic and receive statistics based on social media.

References

[1] https://www.chartjs.org/
[2] https://bootstrap-vue.org/
[3] https://github.com/xkjyeah/vue-google-maps
[4] https://github.com/axios/axios
[5] https://github.com/CSSEGISandData/COVID-19
[6] https://www.cryptocompare.com/
[7] https://opencagedata.com/
[8] Twitter: Standard search api (2020). https://developer.twitter.com/en/docs/tweets/search/overview
[9] Twitter: Filter realtime tweets (2020). https://developer.twitter.com/en/docs/tweets/filter-realtime/overview
[10] The COVID-19 Infodemic: Twitter versus FacebookKai-Cheng Yang, Francesco Pierri, Pik-Mai Hui, David Axelrod, Christopher Torres-Lugo, John Bryden, Filippo Menczer
[11] Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set, reviewed by Adam Jonas, Edson Da Silva, and Muhammad Saud
[12] A Content Analysis of Coronavirus Tweets in the United States Just Prior to the Pandemic Declaration, Amanda D. Damiano and Jennifer R. Allen Catellier