CART 451: Project Proposal

December 9th, 2022

Amanda Clement

ID 40136259

PROTOTYPE SUMMARY

This prototype is an interactive data visualization that illustrates the emotional impact of the COVID-19 pandemic. Each dot represents a Tweet posted on March 11th, 2020, which is the day the WHO declared the COVID-19 outbreak a global pandemic. COVID-19 related hashtag usage spiked on that day.

The initial data set used for this visualization was a list of 1,145,287 COVID-19-related Tweet IDs from March 11th, 2020 was collected from the following GitHub repository: https://github.com/echen102/COVID-19-TweetIDs. These IDs were passed through the Hydrator app which uses Twitter's API to return the corresponding Tweet information. 701,747 Tweets were successfully hydrated. All non-English Tweets were discarded, leaving 380,505 Tweets to be processed. The Tweets were then put through the language processing algorithms and the visualization was created using the extracted information.

STEPS / TECHNOLOGIES

Step 1: Gathering the data

- A list of 1,145,287 COVID-19-related Tweet IDs from March 11th, 2020 was collected from the following GitHub repository: https://github.com/echen102/COVID-19-TweetIDs. This date was chosen as it was the day the WHO declared the COVID-19 outbreak a global pandemic so COVID-19 related hashtag usage spiked on that day. These IDs were passed through the Hydrator app which uses Twitter's API to return the corresponding data from Twitter as a CSV file. 701,747 Tweets were successfully hydrated.
- Text files containing lists of English words and their associations with the emotions of interest (anger, fear, trust, joy and sadness) from the NRC Emotion Lexicon were gathered from: https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm. I modified these files by removing certain words that, in the context of my project, seemed misclassified.

Step 2: Processing the data

- The CSV file of hydrated Tweets was imported into MongoDB collection, where the following fields were retained from the set of data: text, language, location, number of retweets.
- Mongoose queries, complimentary to NodeJS, were used to fetch the data. Only English Tweets were fetched as I could not properly analyze or process other languages. 380,505 Tweets remained after this step.
- The following language processing steps were executed on the text from each Tweet:
    - An array of normalized versions of terms was used to convert parts of the text to simpler form to prepare it for the following steps (e.g. aren't ⇒ are not).
    - All characters were converted to lowercase.
    - All special characters (e.g. emojis) were removed from the text as the emotion analysis only works on textual data.
    - Stopwords were removed from the text. The n-stopwords package was used for this step, but modified it so that it does not remove negation words.
    - The text was tokenized using the Natural package.
    - The text was stemmed using the Natural package.
    - Negation words were searched for in the text and if found, the tokenized word that came directly after it was removed.
    - Emotions from the emotion text files (e.g. anger.txt) were searched for in the remaining tokens. These results were stored.
- The following language processing steps were executed on the location from each Tweet:
    - For Tweets that were associated with a location, the location text was converted to lowercase.
    - Country names and codes were searched for in the location text. States and provinces were also searched for. If found, the country code was stored.
- The stored results, which included the Tweet texts, country codes, and retweet counts were outputted to a JSON file. This JSON file was then imported into MongoDB to create a new collection.

Step 3: Creating the visualization

- The files were set up using NodeJS and ExpressJS for server-side scripting.
- The data from the form the user is presented with when they launch the website is passed via POST request, which is then used to decide how large a sample size to fetch from the MongoDB collection. Once the sample data has been fetched via a Mongoose query, the results are passed along via POST request to the Client.
- Once the response is received, the data is used to create the visualization, which uses p5JS for canvas drawing.

FEATURES IMPLEMENTED / FEATURES REJECTED / FEATURES TO IMPLEMENT

The main feature changed between the prototype and this later version is that I used emotion analysis instead of sentiment analysis. This change involved many steps, from finding a new data source, to hydrating the Tweets, to modifying the language processing algorithms. Since the submission of the prototype, the following features were also added: introduction screen with HTML form that is used to determine the sample size used, separation of language processing and visualization to improve speed performance, overall improvement of language processing algorithms, addition of location filtering option, adaptations to visualization (e.g. ellipse sizes are loosely based on number of retweets), and addition of content to the About page.

The following features were proposed but not implemented: date filtering and keyword filtering options, and addition of information from government and news databases alongside the visualization. The date filtering was excluded as I ended up only using Tweets dated to March 11th because the volume of data to process would otherwise be too large. The keyword filtering was excluded to avoid querying Mongoose once the visualization loads for better speed performance. The addition of information from government and news databases was not implemented due to a lack of time, so that feature will be added in future developments of the project. Additionally, I plan on refining the location filtering option to be more visually interesting, detailed and informative.

PROJECT ANALYSIS

Overall, I consider the development of my visualization to be a successful research-creation project. However, the focus somewhat drifted from the original goal of analyzing emotional trends

related to the COVID-19 pandemic towards studying the efficiency and capacity of natural language processing algorithms in general. During the development of the first version of my prototype which used sentiment analysis, I realized how problematic and misleading some of the language processing algorithms are. Part of the issue relates to my lack of understanding on how to use them correctly and to understand the depths of how they work, but the process nonetheless brought other issues to light. What I found more interesting and informative was how much data gets rejected by these natural language processing algorithms, and this can completely change the extracted meaning. For instance, before I made any modifications to it, the n-stopwords package removed almost all negation words. Hence, the sentence "I am not happy" would be converted to "happy" after tokenization and stopword removal, which completely flips the original meaning.

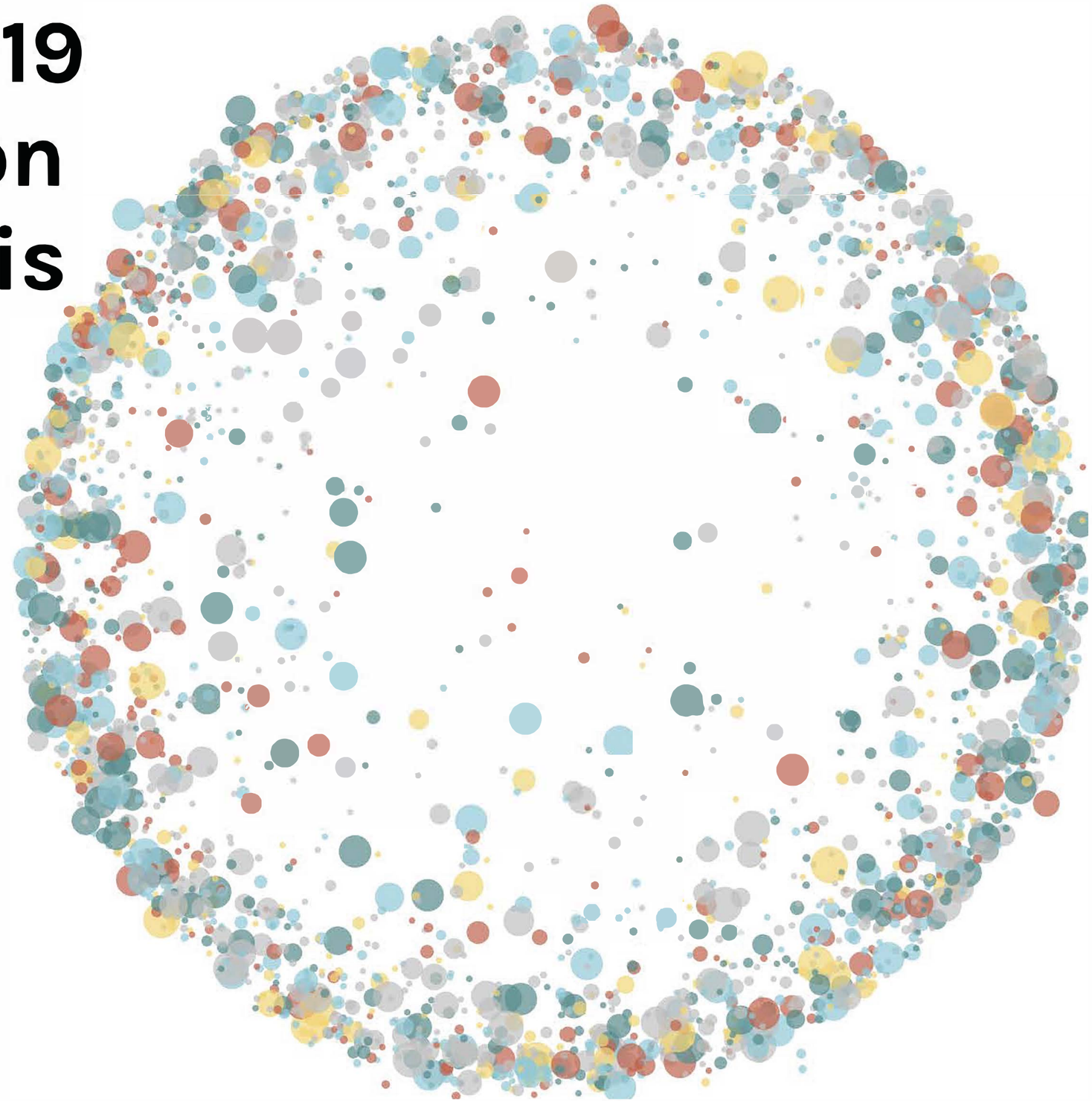
Furthermore, the visualization was less telling than expected. This is largely due to many Tweets being misclassified, mainly those involving sarcasm or more cryptic messages. For example, the following Tweet gets assigned joy as its emotion: *At some point, we have to talk about the fact that, in what is consistently touted as a thriving economy, the key concern w/ closing schools b/c of COVID-19 is "children who rely on school food won't get to eat." If the economy is so great, why aren't babies eating?* Evidently, joy is not an appropriate label for this Tweet but words like *thriving* and *great* lead to this misclassification. However, due to the large volume of data included in the visualization and the refinements made to the language processing algorithms, I consider the visualization as a whole to be mostly accurate and informative. I also believe that my original goal of arranging and compressing the data in a visualization to make it more approachable and digestible was met.

INSPIRATIONS

Besides the inspirations discussed in my proposal, another source of inspiration for the development of my prototype was the natural language processing workshop done in class. It encouraged me to improve my Tweet processing algorithm by adding additional steps to help extract emotions from the text more efficiently and accurately. Although I did not get the chance to refine the algorithm nearly as much as I would have liked, I plan on exploring new ways of improving it in future developments.

Intro screen: user is prompted to enter a number of Tweets

COVID-19 Emotion Analysis



anger

fear

joy

sadness

neutral

country cluster

reset visualization

Sample size: 3000 Tweets

Anger: 226

Fear: 309

Joy: 192

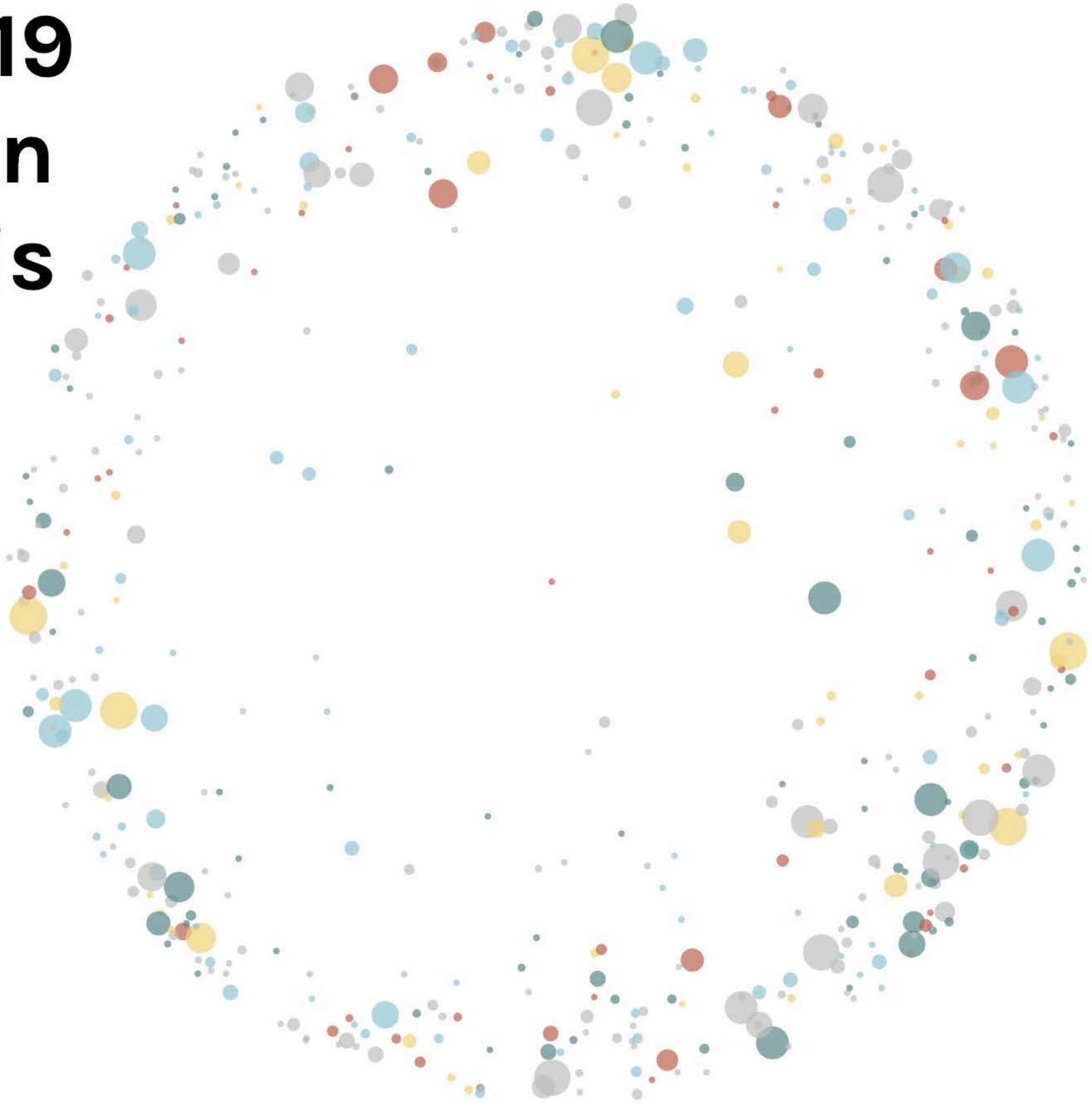
Sadness: 369

Neutral: 294

Since the user entered 3000 on the intro form, a visualization featuring the data from 3000 Tweets is displayed. These 3000 Tweets are chosen at random from the dataset.

COVID-19 Emotion Analysis

About the Project



anger

fear

joy

sadness

neutral

country cluster

reset visualization

Sample size: 500 Tweets

Anger: 34

Fear: 61

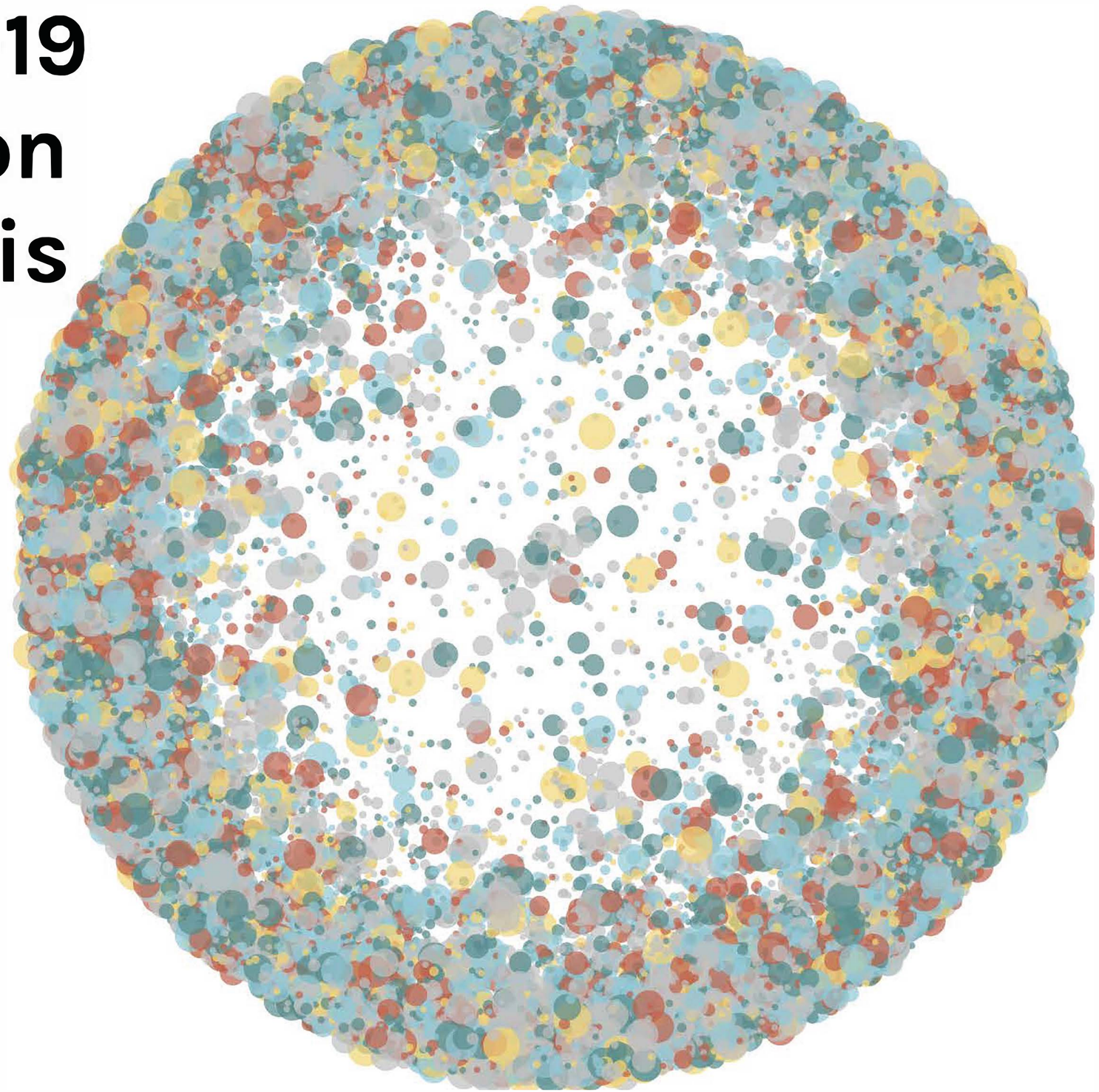
Joy: 33

Sadness: 60

Neutral: 53

A visualization featuring 500 Tweets chosen at random from the dataset.

COVID-19 Emotion Analysis



anger

fear

joy

sadness

neutral

country cluster

reset visualization

A visualization featuring 30,000 Tweets chosen at random from the dataset.

Sample size: 30000 Tweets

Anger: 2260

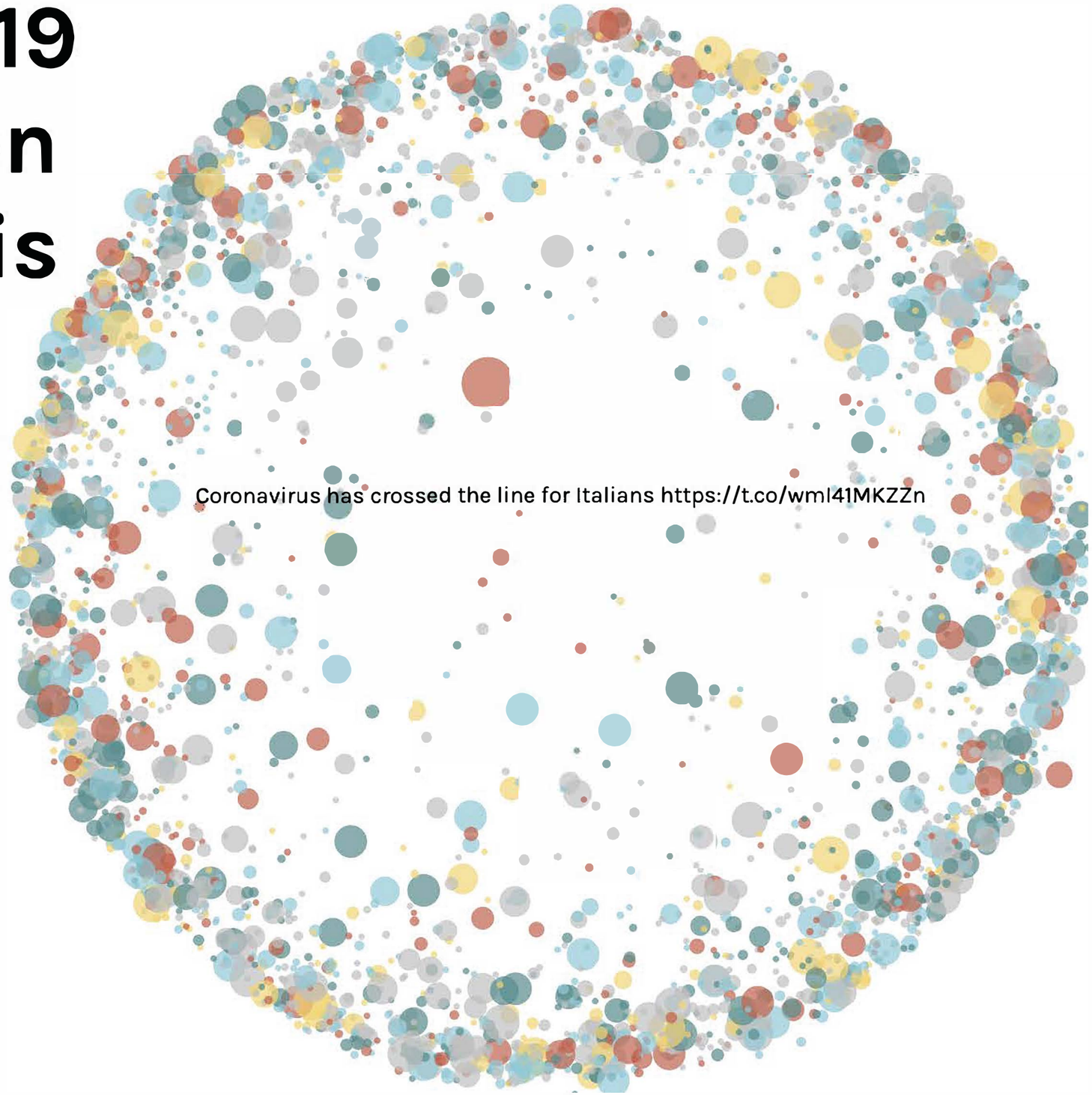
Fear: 3254

Joy: 1896

Sadness: 3665

Neutral: 3105

COVID-19 Emotion Analysis



anger

fear

joy

sadness

neutral

country cluster

reset visualization

Sample size: 3000 Tweets

Anger: 226

Fear: 309

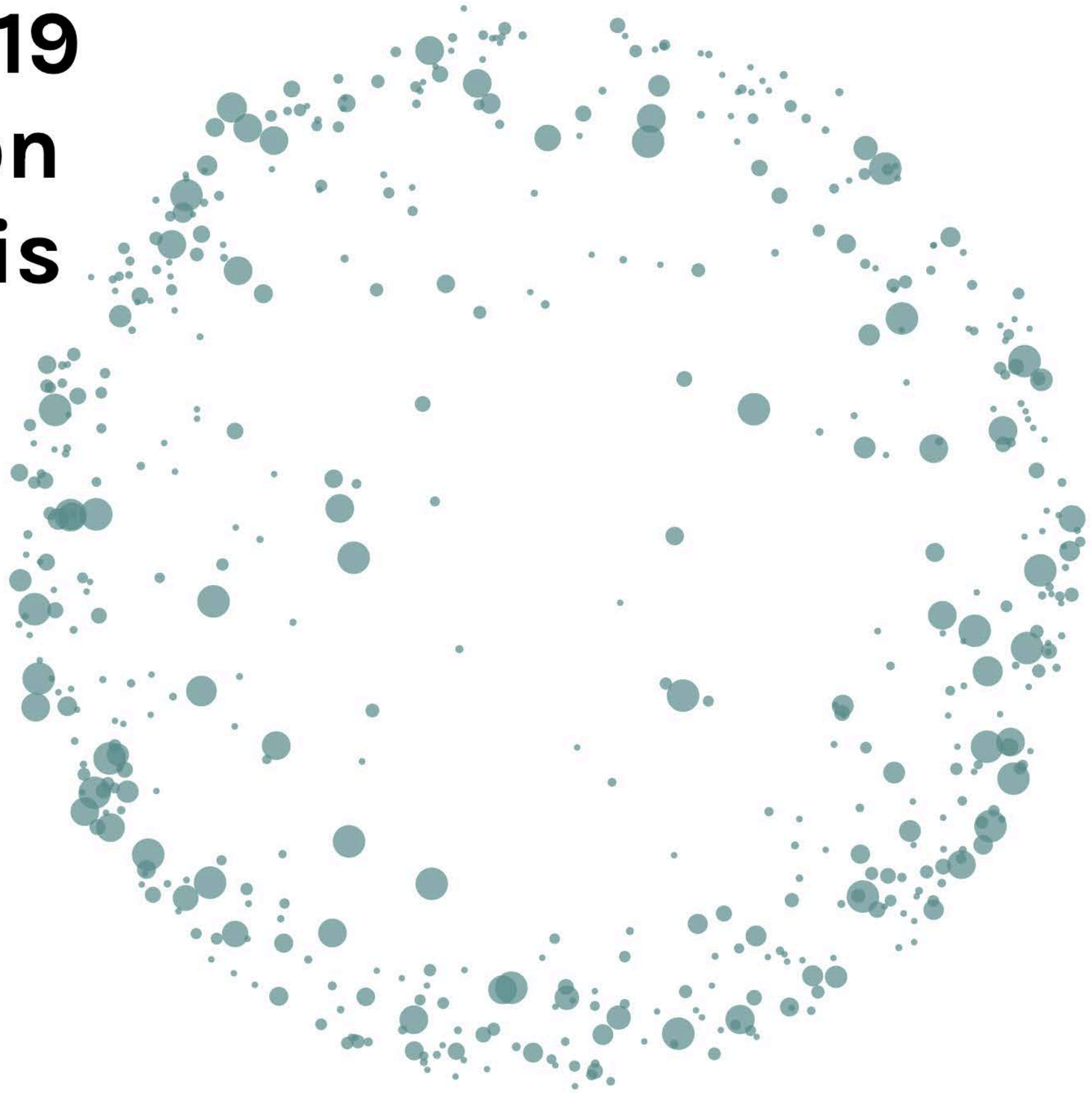
Joy: 192

Sadness: 369

Neutral: 294

Demonstrates what happens when the users presses down on an ellipse.

COVID-19 Emotion Analysis



anger

fear

joy

sadness

neutral

country cluster

reset visualization

Sample size: 3000 Tweets

Anger: 226

Fear: 309

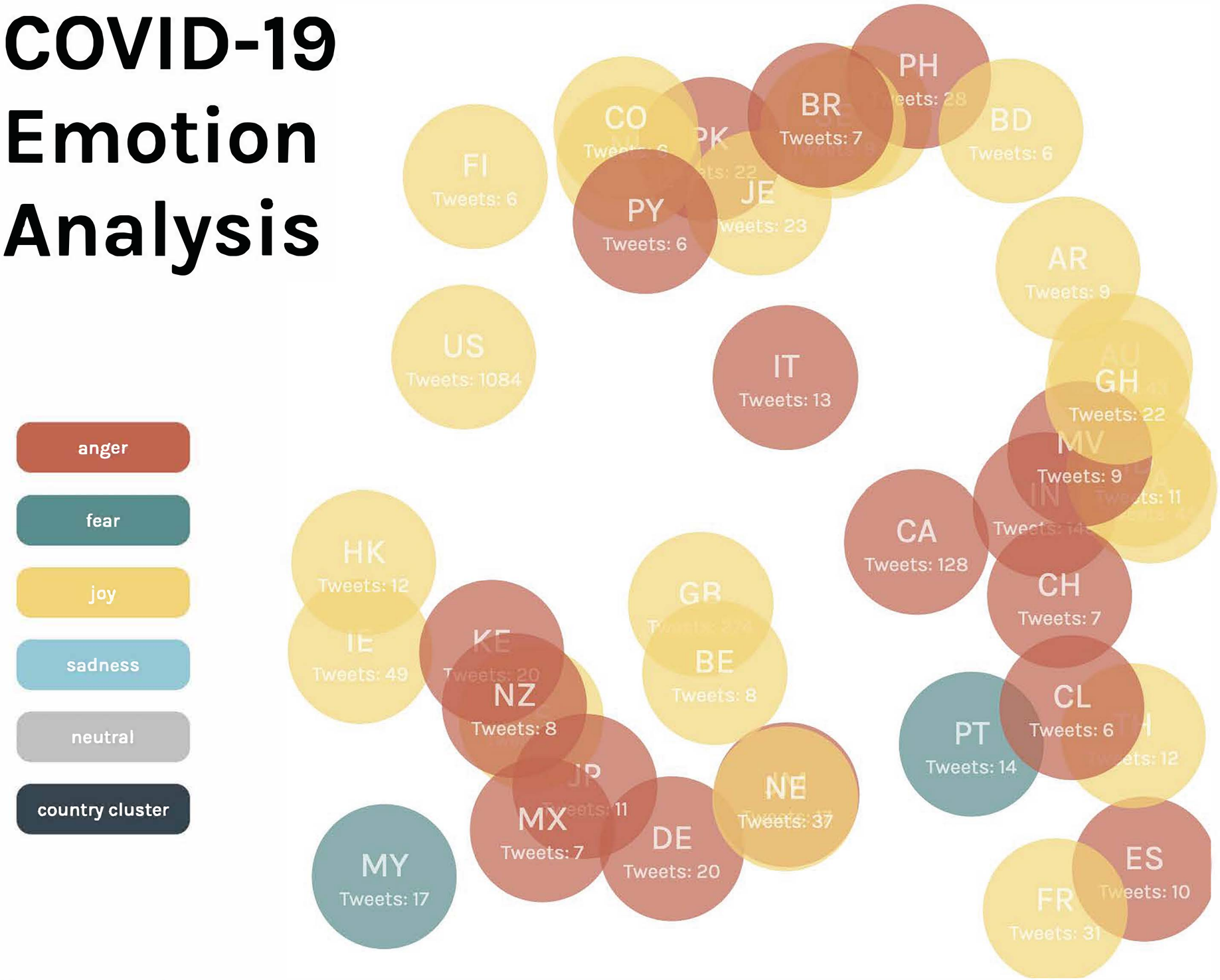
Joy: 192

Sadness: 369

Neutral: 294

Demonstrates what happens when the users clicks the button labeled "fear".
Only Tweets associated with that emotion are shown.

COVID-19 Emotion Analysis



Sample size: 8000 Tweets

Anger: 631

Fear: 889

Joy: 482

Sadness: 956

Neutral: 826

Demonstrates what happens when the users clicks the button labeled "country cluster".
The ellipses are coloured based on the prevailing emotion found in the collection of Tweets tagged with each country.

About the Project

x

This prototype is an interactive data visualization that illustrates the emotional impact of the COVID-19 pandemic. Each dot represents a Tweet posted on March 11th, 2020, which is the day the WHO declared the COVID-19 outbreak a global pandemic. COVID-19 related hashtag usage spiked on that day. The initial data set used for this visualization was a list of 1,145,287 COVID-19-related Tweet IDs from March 11th, 2020 was collected from the following GitHub repository: https://github.com/echen102/COVID-19-TweetIDs. These IDs were passed through the Hydrator app which uses Twitter's API to return the corresponding Tweet information. 701,747 Tweets were successfully hydrated. All non-English Tweets were discarded, leaving 380,505 Tweets to be processed.

The following language processing steps were executed on each Tweet: normalizing the text, removing special characters, removing stopwords, tokenization, stemming, and handling negations. Then, emotions were searched for in the remaining tokens using lists of English words and their associations with the emotions of interest (anger, fear, joy, and sadness) from the NRC Emotion Lexicon: https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon. These lists were slightly modified to remove certain words that, in the context of this project, felt misclassified. Finally, country names and codes, as well as states and provinces, were searched for in the Tweet's location text.

The visualization was then created using these findings, with each dot representing a Tweet that's coloured based on the emotion it's associated with and sized based on the number of retweets it received. For Tweets linked to multiple emotions, the dot's colour is randomly chosen from the set of colours of its associated emotions. Neutral (grey) dots represent Tweets that aren't linked to any of the four target emotions.

The corresponding post is revealed when a dot is pressed down on. The buttons to the left of the visualization can be used for filtering by emotion or to arrange the data based on the Tweet locations. When arranged by location, only countries that have at least 10 associated Tweets are considered. These dots are coloured based on the prevailing emotion found in the collection of Tweets tagged with that country.

Next update: language processing algorithm will be refined and duplicate Tweets will be removed.

The About page.