Sentiment Analysis on COVID - 19 Twitter Data

Ahindrila Saha, Ahmed Rabbani, Jonathan Chow,
Li-Hsing Huang, Saurabh Aggarwal, Siddharth Sen

Dashboard: `https://reurl.cc/529QD6`

1 Introduction - Motivation

The COVID–19 pandemic has disrupted the world tremendously. The core problem that allowed the spread of Covid-19 lies in the mixed reception and responses from different people and regions toward government policies. To address this, our group believes that using sentiment analysis on a large scale to better understand state-level emotions regarding certain policies or information will allow for more successful execution of policies.

2 Problem Definition

To understand the complex emotions behind the tweets, we will process the data through a model that accounts for these complications with text embedding and natural language processing and classify Twitter data into positive, neutral, and negative categories. Once we have created a library containing state-level sentiment of all the Twitter data, we plan to examine sentiment data around key milestone dates to verify if emotions are accurately captured within our analysis.

3 Survey

To date, numerous studies have been conducted related to sentiment analysis of tweets relevant to COVID–19. For instance, in [10], [14], [11], [18], [21], [8], [6] and [5], sentiment analysis was performed to analyze public opinion regarding general events related to pandemic whereas in [20] and [13] specific themes related to pandemic such as social distancing and potential implications of popularity on social media accuracy, were explored. Data collection and feature generation from relevant tweets is an important aspect of our project and multiple papers have adopted a variety of approaches for performing this task. For instance, in [20] ,[18],[21] and [5] DOC NOW hydrator and Python libraries namely Tweepy, SNSCRAPE and GetOldTweets API were used to scrap relevant tweets. [2],[7], [11] and [18] employed critical methods such as Bag of Words, TF-IDF scores and Word2Vec methods for meaningful data cleaning tasks such as removing stop words, URL, retweets, and hashtags. Feature extraction is followed by sentiment tagging and the most common approach is to employ lexicon-based applications such as NRC Lexicon, VADER Lexicon and SentiStrength ([14], [11], [19], [20], [18] and [21]). Variety of modelling techniques are generally used to train models for sentiment analysis. PCA, SVM classifiers, DICE and Deep Learning models were used for model training in [10], [9],[15] and [16]. Several techniques such as LSTM, GRU RNN, CART and J48 were compared for model training in [17] and [3]. The sentiments are visualized using packed bubbles, stacked bar charts, word clouds, sentiment rivers and some these were implemented in [1],[13] and [12]. Whereas an entire methodology regarding presenting data as paintings was detailed in [4].

4 Proposed Method - Intuition - why should it be better than the state of the art?

In most of these research papers, the analyses methodology used was implemented mainly on small sets of data, which can possibly lead to statistically insignificant results. Further, existing analyses tend to classify tweets data using hashtags, limiting the possibilities of further personal or regional analysis. As a part of our project, we intend to improve on these approaches by performing our analysis on a large data set and classifying our data into more dimensions, including emotions, locations, and time. By doing so, we can expand the sentiment analysis to a more in-depth level. Most importantly, we target to visualize our analysis using an interactive dashboard which is also a clearer and innovative way to demonstrate the observations. The dataset we are planning to tackle has 145 million tweets all related to COVID-19.

5 Proposed Method

5.1 Data Extraction

The initial data set used had over 145 million Tweets included in it, so memory issues were a constant problem throughout the project. To circumvent that, the group divided the data into 12 months with each member handling 2 months respectively from March 2020 to February 2021. Since only US data was required for our analysis, a good deal of the data was eliminated. Post-filtering, each member worked with a dataset of roughly 5 million each, for each month.

A random sample of 500k tweet id's was taken for each month and then their corresponding tweets was extracted. One of the earliest problems encountered was Twitter's limitation on data extraction. To protect their internal servers from overload, the Twitter API enforces a limit of 900 Tweet extractions per 15 minutes. Factoring the magnitude of several million Tweets for each member, the cap was a clear bottleneck. The group ultimately adopted the use of Hydrator, a 3rd party software, which was able to somewhat bypass this limit block to extract data. By running Hydrator in the background during downtimes, the relevant data was successfully extracted.

5.2 Data Cleaning

Since the data we were primarily interested in was the text content in Tweets, we had to process the data prior to placing it into our NLP models. We primarily used Regex to clean mentions, hashtags, links, retweets, and non-alpha numeric characters in each tweet. We repeated this process for each month of data ranging from March 2020 to Febuary 2021. The location was cleaned to extract the state corresponding to each tweet's location.

5.3 Data Modelling

There were two models that the group considered using. The first model was the VADER model, while the second model focused on using unsupervised learning through word-embeddings. The key difference between the two is that VADER sentimental analysis relies on a dictionary that maps lexical features to emotion intensities known as sentiment scores, the sentiment score of a text can be obtained by summing up the intensity of each word in the text, while the embedding model focuses on assigning vectors to words. In the embedding model, similar words will have similar vectors, which can then be grouped together to determine an overall sentiment of a tweet.

Upon completion of our embedding model, we found its sentiment results to be diluted. The embedding model approach was done by first getting individual word embeddings from GloVe, then averaging the returned values of all words in a sentence to get the overall embedding of a sentence. However, this approach ran into the issue of each word being weighted equally, resulting in diluted sentiment because many of the words required for grammatical syntax (and, the, a etc) were all classified as neutral, which resulted that all tweets had an average value close to neutral values. Even if we remove the stopwords from the model, giving different weights to different words would still be an issue for the embedding model. Thus, we elected to use the Vader model going forward because of its better performance.

5.4 Data Aggregation

After applying sentiment analysis on our tweets, we first grouped our data on date, location, and sentiment. After doing this step, we were able to get the aggregated sentiment score of each type of sentiment for each state on a given day along with the corresponding number of tweets on that day. This data was primarily used to make the choropleth map and tree diagram in our final dashboard.

After a series of aggregation steps we were able to structure our data to display the count of hashtags with their corresponding location and sentiment for each date. A similar approach was applied on the tokenized attribute of our resulting data set from the previous section. This resulted in a data set that displayed the count of most common words with their corresponding state and sentiment for each day. For both hashtag and word count These data sets were then used to make word clouds and bar charts in our final dashboard.

5.5 User Interface

The main components of our user interface can be split into two parts: a choropleth map and word clouds. The choropleth map is divided on a state level, and color divided, with red indicating negative, green positive, and grey neutral sentiments. The intensity of the color also indicates the sentiment strength that was obtained from the Vader Lexicon model. The dropdown on the top right-hand corner of the choropleth can be used to select dates to view how sentiments are changing over time, and also simulate the change or flow of sentiments over time across all the states in USA. Below the choropleth graph, there is trend chart that shows the aggregated average sentiment of the US over time. This chart can be drilled down to view trend at a year, quarter, month, week and date level, so as to capture bith macro and micro trends and changes. Additionally, there is a tree-map at the very bottom that can be used to view the visualize the volume of tweets and overall sentiment by each state for the given day.

The right-hand side of the dashboard displays two word clouds: hashtags and tokens (words or unigrams), across the entire time frame of our model. These can be used to visualize the most trending hashtags and words used by users in their tweets. Hovering upon the desired hashtag or token also reveals the date on which it was most trending or used on, which can be tha=en used to draw correlations between the sentiment seen on the coropleth map. As expected, Covid and its variants are consistently the largest words. The words clouds also have added functionality to allow the end user to filter out words based on the frequency of the words as desired.

6 Experiments/ Evaluation - Description of your testbed; list of questions your experiments are designed to answer

The top priority of this experiment is to determine whether the Vader model accurately captures sentiments about Covid 19 Twitter data. We will validate this by comparing sentiment levels for time periods following key milestone events. Since we are working on historic data, the outcomes of these events are known, in that whether they will have a positive, negative, or neutral impact and thus can be used for accurate verification.

7 Experiments/ Evaluation - Details of the experiments, observations

To analyze the sentiment geographically, a choropleth map of the United States visualizing the result of our VADER model has been created. The graph uses shades of colors to demonstrate emotions in each state with a date selection drop-down menu so that we can observe the sentiment change on a day-to-day level.

In order to verify the accuracy of our VADER model, we have selected 4 major events in 2020 and checked the sentiment scores corresponding to the specific time periods.

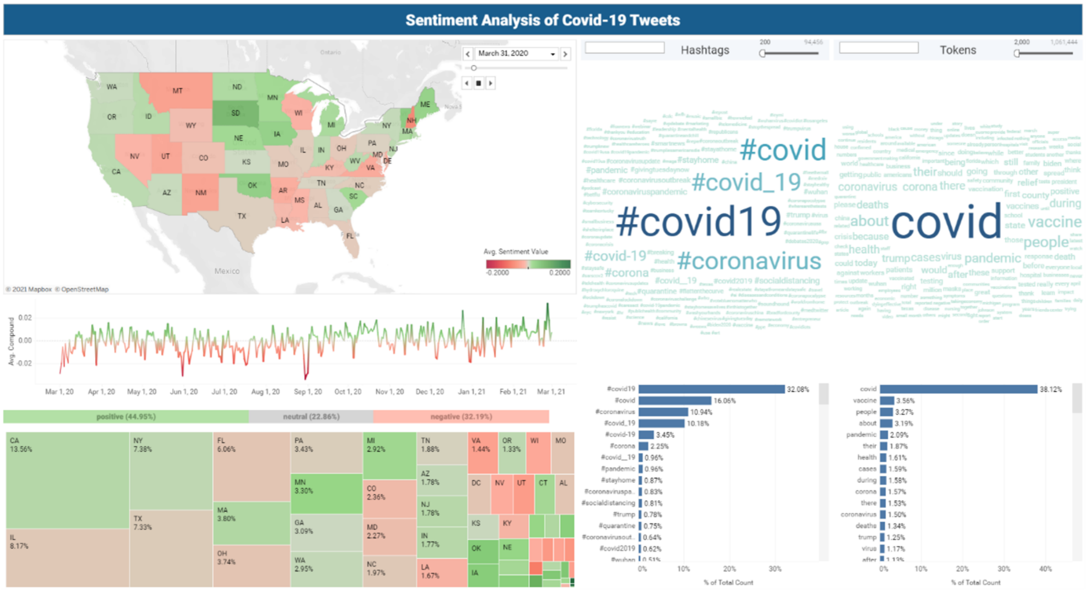

Figure 1: Dashboard - Sentiment Analysis of Covid-19 Tweets

The 4 major events are as follows:

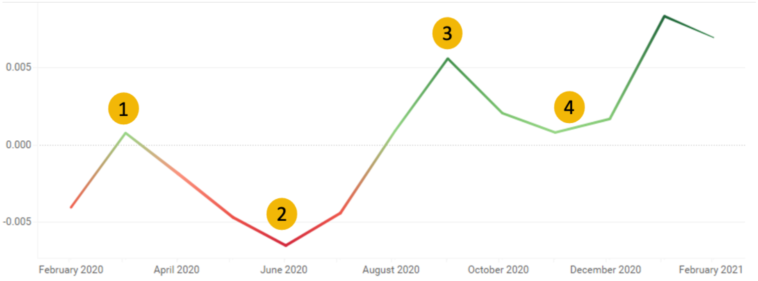

Figure 2: trumphascovid hashtag

- Stay Home Order: The announcement was issued in March 30th 2020, which ignited people's concern and anger. The Stayhome trended on that day. Our model correctly captures the negative emotions and therefore we can see a significant drop in sentiment scores starting from April 2020. Also, states such as Florida and Ilinios contanstly have negative sentiments as compared to states like California

- Positive Announcement from Moderna and AstraZeneca: Starting from July 2020, both Moderna and AstraZeneca had shown promising results and record-breaking progress in the trials, enhancing people's belief that effective Covid vaccines were about to come out and therefore leading to a steady and significant increase in sentiments.

- The United States Hitting a Record High Number of Confirmed Daily Covid Cases: On Oct 24th, 2020, the United States had hit its record-high single day confirmed daily Covid cases which once again aroused people's fear of this seemingly-never-ending disease. This was also coupled with Donald Trump, getting infected with coronavirus. The hashtag trumphascovid was also trending on Twitter at that time and can also be seen on the dashboard. All these effects combined could have results in the observed drop in sentiments.

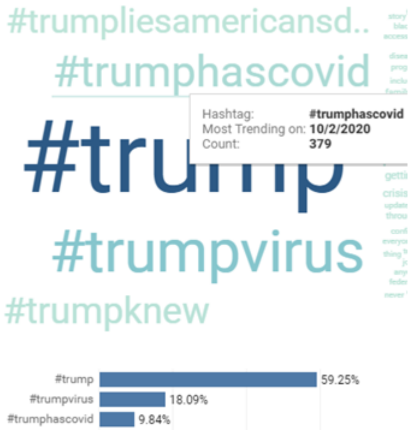

Figure 3: trumphascovid hashtag

- Execution of Covid Vaccinations in the United States: The United States started the distribution of vaccines on 14th Dec 2020, marking major progress of people against the Covid-19. With the expected protection from the vaccines, people strongly believed that the whole situation had been starting to move in the positive direction, leading to a significant and steady increase in the sentiments.

An interesting result from our sentiment analysis was that most of the averaged sentiments were along an extremely small scale, roughly between -0.4 to 0.4. Examination of the data led us to believe that this is a result of inertia from too many irrelevant Tweets because we did not further separate the Tweets into smaller groups. As such, many "irrelevant" tweets would result in very small numerical sentiment changes.

8 Conclusions and discussion

Overall we have very positive results. Our chosen major events and the expected sentiment change regarding them have all been matched very well. This shows that the spatio-temopral analysis conducted constructed is a suitable tool for capturing the sentiments and analysing their trends. We believe that this model can be further extended to more practical purposes. For instance, policymakers can use this information both to decide what policies to pursue, as well as the most optimal method with which to execute their policies.

As our future work, we plan to enhance the accuracy and efficacy of our model by doing further data preprocessing upstream. As noted previously, the overly broad nature of our dataset made it difficult for us to analyze specific topics. Hence, the logical next step would be for us to set up an accurate sorting pipeline for the Tweets gathered post cleaning for even better results. Based on our successful implementation with Twitter data, we anticipate that the improved sentiment analysis tool can also be applied to other social media platforms as well as other topics (besides COVID) to gather sentiment data on an ever-larger level to optimize policy choices and execution.

9 Distribution of team member effort

All team members have contributed a similar amount of effort.

References

[1] Aljoharah Almjawel, Sahar Bayoumi, Dalal Alshehri, Soroor Alzahrani, and Munirah Alotaibi. Sentiment analysis and visualization of amazon books' reviews. In *2019 2nd International Conference on Computer Applications & Information Security (ICCAIS)*, pages 1–6. IEEE, 2019.

[2] Arpita, Pardeep Kumar, and Kanwal Garg. Data cleaning of raw tweets for sentiment analysis. In *2020 Indo – Taiwan 2nd International Conference on Computing, Analytics and Networks (Indo-Taiwan ICAN)*, pages 273–276, 2020.

[3] Felipe Bravo-Marquez, Marcelo Mendoza, and Barbara Poblete. Combining strengths, emotions and polarities for boosting twitter sentiment analysis. In *Proceedings of the second international workshop on issues of sentiment discovery and opinion mining*, pages 1–9, 2013.

[4] Andreas Buja, John Alan McDonald, John Michalak, and Werner Stuetzle. Interactive data visualization using focusing and linking. In *Proceedings of the 2nd conference on Visualization'91*, pages 156–163, 1991.

[5] Koyel Chakraborty, Surbhi Bhatia, Siddhartha Bhattacharyya, Jan Platos, Rajib Bag, and Aboul Ella Hassanien. Sentiment analysis of covid-19 tweets by deep learning classifiers—a study to show how popularity is affecting accuracy in social media. *Applied Soft Computing*, 97:106754, 2020.

[6] Emily Chen, Kristina Lerman, Emilio Ferrara, et al. Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. *JMIR public health and surveillance*, 6(2):e19273, 2020.

[7] Stephen Hill and Rebecca A Scott. An approach to harvesting, cleaning, and analyzing data from twitter using r. *Information Systems Education Journal*, 15(3):42, 2017.

[8] Zhiyuan Hou, Fanxing Du, Hao Jiang, Xinyu Zhou, and Leesa Lin. Assessment of public attention, risk perception, emotional and behavioural responses to the covid-19 outbreak: social media surveillance in china. *Risk perception, emotional and behavioural responses to the COVID-19 outbreak: Social media surveillance in China (3/6/2020)*, 2020.

[9] Ramandeep Singh Kathuria, Siddharth Gautam, Arjan Singh, Smarth Khatri, and Nishant Yadav. Real time sentiment analysis on twitter data using deep learning (keras). In *2019 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, pages 69–73. IEEE, 2019.

[10] Mohammad Abu Kausar, Arockiasamy Soosaimanickam, and Mohammad Nasar. Public sentiment analysis on twitter data during covid-19 outbreak.

[11] Jolin Shaynn-Ly Kwan and Kwan Hui Lim. Understanding public sentiments, opinions and topics about covid-19 using twitter. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 623–626. IEEE, 2020.

[12] Yafeng Lu, Xia Hu, Feng Wang, Shamanth Kumar, Huan Liu, and Ross Maciejewski. Visualizing social media sentiment in disaster scenarios. In *Proceedings of the 24th international conference on world wide web*, pages 1211–1215, 2015.

[13] Kamaran H Manguri, Rebaz N Ramadhan, and Pshko R Mohammed Amin. Twitter sentiment analysis on worldwide covid-19 outbreaks. *Kurdistan Journal of Applied Research*, pages 54–65, 2020.

[14] Amrita Mathur, Purnima Kubde, and Sonali Vaidya. Emotional analysis using twitter data during pandemic situation: Covid-19. In *2020 5th International Conference on Communication and Electronics Systems (ICCES)*, pages 845–848. IEEE, 2020.

[15] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113, 2014.

[16] Usman Naseem, Imran Razzak, Katarzyna Musial, and Muhammad Imran. Transformer based deep intelligent contextual embedding for twitter sentiment analysis. *Future Generation Computer Systems*, 113:58–69, 2020.

[17] Ru Ni and Huan Cao. Sentiment analysis based on glove and lstm-gru. In *2020 39th Chinese Control Conference (CCC)*, pages 7492–7497. IEEE, 2020.

[18] Khairiyah Mohamed Ridhwan and Carol Anne Hargreaves. Leveraging twitter data to understand public sentiment for the covid-19 outbreak in singapore. *International Journal of Information Management Data Insights*, page 100021, 2021.

[19] Amrita Shelar and Ching-Yu Huang. Sentiment analysis of twitter data. In *2018 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 1301–1302. IEEE, 2018.

[20] Carol Shofiya and Samina Abidi. Sentiment analysis on covid-19-related social distancing in canada using twitter data. *International Journal of Environmental Research and Public Health*, 18(11):5993, 2021.

[21] Tanmay Vijay, Ayan Chawla, Balan Dhanka, and Purnendu Karmakar. Sentiment analysis on covid-19 twitter data. In *2020 5th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE)*, pages 1–7. IEEE, 2020.