Analysis of COVID Vaccine Sentiment Using Twitter

Anasse Bari
New York University

Ryan Cohen
New York University

Levente Szabo
New York University

***Abstract*—The development of a vaccine that could end the global COVID-19 pandemic would require a large-scale program to vaccinate a majority of the population. With multiple vaccines in rapid development, we anticipate a variety of barriers regarding widespread vaccination. In order to gauge public perception regarding vaccination we construct a sentiment analysis system that analyzes Twitter data in real time. To facilitate healthcare professionals in overcoming vaccine hesitancy this system would give insight into the reasoning behind specific attitudes as well as highlighting spatiotemporal differences.**

***Index Terms*—sentiment analysis, AWS, COVID-19, vaccine perception**

I. Introduction

As the COVID-19 pandemic has progressed, differing opinions surrounding the virus and the recommended health guidelines have emerged [7]. For example, within countries such as the U.S., while surveys have found that the majority of citizens adhere to mask-wearing recommendations to control the spread of the virus [8], a considerable amount of resistance to these recommendations within the country has also emerged [1]. Now, with discussions about the development and availability of vaccine candidates for COVID-19 occurring, a similar diversity of opinion has become apparent through sources such as a September 2020 Pew Research Center survey indicating that 49

A vaccine for COVID-19 has the potential to decrease the threat of the virus through the achievement of herd immunity. Herd immunity refers to a phenomenon that occurs when the proportion of a population susceptible to a virus is small enough to prevent future outbreaks [3]. One study has estimated that in order for COVID-19 herd immunity to be achieved, 82

The existence of a growing vaccine hesitancy movement predates the COVID-19 pandemic, with vaccine hesitancy even being identified as a top ten global health threat by the World Health Organization (WHO) in 2019 [4]. Reasons for opposition to vaccination have included safety concerns, conspiratorial beliefs, feelings that vaccine requirements infringe on individuals' health freedom, and more. The current presence of this opposition is emphasized by the fact that as COVID-19 pandemic shutdowns occurred, anti-vaccination groups experienced content and engagement growth. Additionally, in October 2020, it was found that Facebook anti-vaccination groups and YouTube anti-vaccination channels have followings of 31 million people and 17 million people, respectively [10]. Gaining a more thorough understanding of the reasons for anti-vaccination sentiment and hesitancy has the potential to support efforts to encourage the public to receive a COVID-19 vaccination.

Twitter has previously been identified as a tool for representing sentiments and moods of the public. For example, Bollen et al. used collections of daily tweets as expressions of public sentiment in studying the effect of the public mood on the stock market [6]. Related to vaccination specifically, Raghupathi et al. applied sentiment analysis to 9,581 vaccine-related tweets from 2019 as a means of generally understanding the public's perception of vaccination [13]. Additionally, Du et al. previously used Twitter as a source from which public opinions of human papillomavirus vaccination could be extracted. They developed a sentiment analysis system to extract these opinions from tweets, acknowledging the potential of their approach to support public health professionals in monitoring the response of the public [9].

Given the potential of Twitter to serve as a tool for monitoring the opinions of the public, we aim to use it as a data source for understanding feelings towards vaccination in the age of the COVID-19 pandemic. Specifically, we aim to develop a method for continually collecting and analyzing the contents and sentiments of tweets, with the goal of developing a regularly-updating dashboard reflecting the results of our analyses. This dashboard has the potential to serve as a valuable resource to medical and public health professionals aiming to monitor public responses as conversations about COVID-19 vaccination continue.

II. Methodology

A. Overview

To gauge public sentiment regarding vaccination in real time our social media analytics pipeline (Figure 1) consists of a variety of platforms and primarily utilizes the Amazon Web Services (AWS) cloud infrastructure. In order to generate valuable insights from newly published Tweets our system takes data through 3 phases; an Ingestion, Analysis and Visualization phase. In addition to these aspects of our pipeline we also performed a data modelling phase by selecting key features, designing a database schema and bringing in external data sources.

1) **Ingestion:** Java application using Twitter4j deployed on an EC2 virtual machine. Tweets mentioning "vaccine"

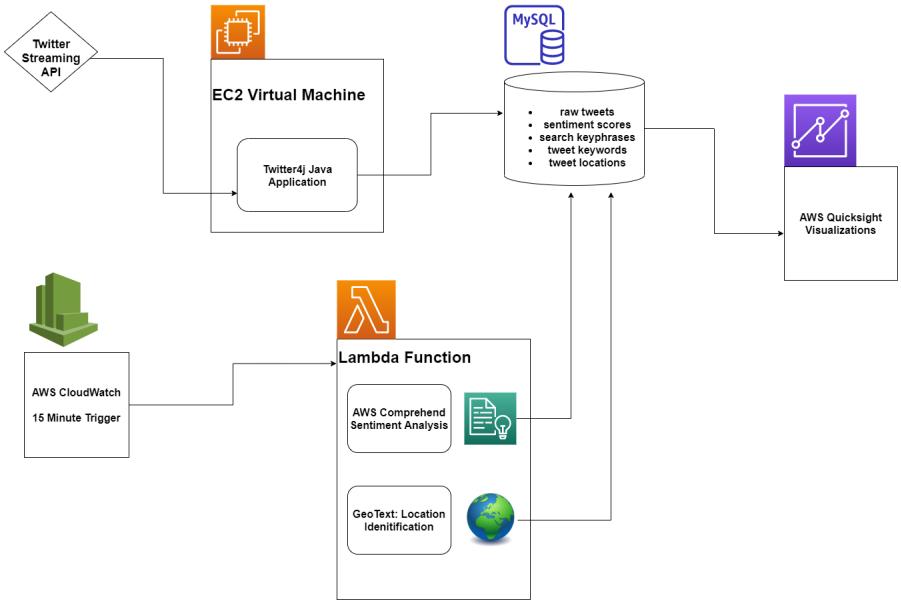

Fig. 1. AWS System Architecture

and other predetermined keywords are preprocessed and stored into a MySQL relational database.

2) **Analysis:** On regular time intervals a CloudWatch service initiates a serverless Lambda function to read unprocessed tweets from the MySQL database. Sentiment analysis and Keyword extraction are performed with Comprehend and location matching using the Python geotext package.
3) **Visualization:** Analytics regarding the breakdown of tweet sentiments, most common keywords and user locations is displayed as a dashboard using QuickSight. Data is constantly refreshed to provide a real-time user experience.

B. Data Modelling

Tweet Data: Our central data schema is described in Figure 2. Every collected tweet is placed into the **raw-tweets** table which contains the primary key status-id which is tied as a primary foreign key to several other tables. The **sentiment-scores** table serves to store the results of sentiment analysis for each tweet, it is the output of an AWS Comprehend workload and gives specific details regarding the level of each sentiment label (positive, negative, mixed, neutral). The **tweet-keywords** table serves as an alternative to **search-keyphrases**, our curated set of keywords and stores additional outputs of AWS Comprehend. Using external data in conjunction with the **country-coordinates** table we construct the **tweet-locations** data which transforms user defined location strings into specific coordinates for mapping. For each collected tweet, sentiment and keyword detection are run on the content of tweets detected as being written in English, while location extraction is run on the user-provided location information (as further described in 2.3). Our MySQL data schema attempts to normalize and facilitate the storage and analytics of sentiment detection, keyword detection, and location extraction results for the collected tweets.

Keywords: To give context to the problem of vaccine hesitancy a curated list of keywords was constructed regarding a variety of different attitudes and topics. For each of these keywords, a category (negative, positive, or undecided) and a subcategory (e.g. conspiracy, safety, key names, hesitancy, etc.) were additionally identified. The category, subcategory, and term text for each of these terms were stored within an individual MySQL table, search_keyphrases for purposes of identifying how represented they are within collected tweets.

Geographic Data: To highlight geographic differences in both quantity of tweets and sentiment we extracted twitter user locations. The two letter country codes, latitude, longitude, and country names for 244 countries and territories found within a public dataset were added to an additional MySQL table [5]. This table was created to help generate a visualization within the final dashboard showing the country locations of collected tweets on a world map. In order to plot geographic data outside of the U.S. in a AWS map visualization, geographic coordinates are required. As the country location of tweets is determined by the processing of user-entered location descriptions and is not expected to contain coordinates, this table was created with the intention of being used with the country name data extracted from the user-provided text in plotting points on a world map.

C. Data Ingestion

To collect real time tweets a Java application is used to interface with the Twitter Streaming API. We use Twitter4j to interface with Twitter and JDBC to interface with our MySQL

Fig. 2. Entity Relationship Diagram

relational database. To perform continuous tweet ingestion our application is run on an AWS Linux EC2 instance.

Keyword Filtering: To determine the set of keywords to filter on we initially query the search_keyphrases table in our database. Then, using the keyword "vaccine" and all keywords that do not contain it we construct our filtering and initiate our stream.

Feature Extraction: Besides tweet_text the specific features selected for each tweet include status_id, user_name, user_id, published_date which serve as unique identifiers, retweet_count and favorite_count which serve as measures of tweet popularity and location which is a user defined string denoting the user location at account creation.

Handling Retweets: Many tweets are in fact retweets of popular tweets. To avoid overloading our analysis system we handled these retweets by updating the retweet-count and favorite_count in our database. A retweet was only inserted into the database if it was not found. By tracing back the old retweet we were able to decrease the load on our system as well as getting a larger time horizon of tweets.

D. Analysis

An AWS CloudWatch Events rule, which allows for the scheduling of automated actions, was created. This rule is configured such that at regular intervals of 15 minutes, it triggers the initiation of tweet analysis. Specifically, it triggers an AWS Lambda function that performs sentiment analysis, keyword extraction, and location extraction. AWS Lambda is a service that allows for the running of code without any management of servers [11]. Within the created Lambda function triggered to run every 15 minutes, a collection of unprocessed tweets are first read from the MySQL table to which the Twitter4j application writes collected raw tweet data. After the Lambda function has performed sentiment analysis, keyword extraction, and location extraction for a specific tweet, it writes the three individual sets of results to the sentiment-scores, tweet-keywords, and tweet-locations tables, described previously.

Sentiment and Keyword Detection: Within the Lambda function, the AWS Comprehend API is used in performing analysis of individual tweets. AWS Comprehend is a natural language processing service that is useful in extracting insights from unstructured text data, in this case the content of tweets [11]. The AWS Comprehend API is first used to detect the dominant language of an individual tweet's content, and if the dominant language of the tweet is found to be English, the AWS Comprehend API is then used to detect sentiment of the tweet's content and to detect keyphrases within the tweet's content. In detecting sentiment, the Comprehend API returns scores for Mixed, Positive, Neutral, and Negative sentiment along with a label determined by which of the one of four scores is the highest. In detecting keyphrases, the Comprehend API returns information including the text for each identified keyphrase and a corresponding score for each keyphrase, indicating the Comprehend service's confidence in the accuracy of its detection.

Location Extraction: Also within the Lambda function, the public user-provided location text associated with each individual tweet is processed using the geotext python library [18]. Geotext is specifically used to identify city names and country names indicated by each user-provided location. However, due

to the goal of creating a U.S. map visualization displaying the locations of vaccine-related tweets and the fact that geotext does not detect U.S. state names and state name abbreviations, an additional step of location text processing was added. The steps for extracting the location from user-provided location text are as follows:

1) The city, state, and country for the tweet are intialized to "None". This is considered the unset state.
2) Geotext is used to identify the names of the countries referred to in the location text. If exactly one country is identified by geotext, the country for the tweet is set.
3) Geotext is used to identify the names of the cities referred to in the location text. The identified results are looped through, and for each the following conditions are checked:
    a) If the city for the tweet is currently unset, it is set to the current geotext-identified result.
    b) Otherwise, if the state is currently unset and the current geotext-identified result is in the set of predefined U.S. state names, the state for the tweet is set to the geotext-identified city. (This condition covers cases such as "Buffalo, New York", where geotext identifies both "Buffalo" and "New York" as cities.)
    c) Otherwise, city and state are both set back to "None", as the location text contains more than one city, and all subsequent steps are skipped.
4) If the state is unset and either the country is unset or the country has been set to "US", the location text is split at every ',' character to generate several smaller tokens. The generated tokens are looped through, and for each the following conditions are checked:
    a) If the current token is in the set of U.S. state names and state is no longer unset, a previous token within step 4 has also been identified as a state. Therefore, the city and state are set back to "None" and all subsequent steps are skipped.
    b) Otherwise, if the token is not equal to the currently-set city and the token is in the valid set of U.S. state names, the country is set to 'US' and the state is set equal to the token.
5) If the state is unset, the city has been set, and the token that the city is currently set to is in the set of U.S. state names, the state is set equal to that token and the city is set back to "None". (This condition covers cases such as "Michigan, USA" or "New York, USA" where geotext will identify "Michigan" or "New York" as the city and the state will remain unset until step 5 is reached.)

III. Visualizations

Performing analytics and visualizations of our results is critical in providing valuable insights to a wide range of healthcare professionals. Using the AWS QuickSight platform we constructed visualizations of keyword mentions, sentiment, and locations.

A. Categories

To support monitoring of the mentions of keywords in our curated list, visualizations that refresh on a daily basis and represent the results of searching for each of the keywords within tweets from the past seven days were created. A tweet found to contain a specific keyword was designated as belonging to the subcategory (e.g. companies, key names, key people) associated with that keyword. The chart in Figure 3 was then created, demonstrating the number of tweets found to belong to each of the subcategories. Using the overall sentiment labels for tweets within each subcategory, a visualization demonstrating the percentage of tweets within each subcategory with each sentiment label was created (Figure 3). Additional visualizations were created to monitor mentions of individual keywords within the subcategories containing the greatest amounts of tweets. For example, a visualization for the keywords contained within the "Companies" subcategory is shown in Figure 5.The number of tweets containing each of these keywords is demonstrated within this visualization, as well as the breakdown of sentiment labels for tweets containing each keyword.

B. Time Series

One of the features of vaccine sentiment that we are interested in is its dynamic nature. By plotting the sentiment proportions for each daily sample of tweets we can see its fluctuation in time (Figure 4). To give researchers and healthcare professionals insight into the temporal fluctuations we have decided to plot the sentiment proportions over the last 3 weeks. Interestingly no large changes in sentiment proportions were found over this span currently, additionally when including time periods with more daily tweets we found no differences in proportions. Despite these findings we think that over a longer period of time there will be deviations from the current pattern. As new research indicates high efficacy and a strong safety profile of initial COVID-19 vaccines we expect there to be a positive surge in overall sentiment. However, as the world comes closer to initiating a mass vaccination program we forsee a variety of changes in sentiment. If everyone is able to receive the vaccine in a timely manner and there are limited side effects then we expect the daily sample of tweets to reflect a strong positive overall sentiment as individuals' lives go back to normal. However, difficulties with distribution, political concerns and health effects are some factors that could cause a negative trend in overall sentiment during 2021. To better understand the temporal changes in sentiment we propose a time series plot that includes a normalized topic score for a variety of different categories.

C. Geographic Distribution

United States and world maps were created to support study study of the geographic distribution of collected tweets (Figure 6). Both of these maps refresh on a daily basis and represents extracted locations for tweets from the past 7 days. For the United States map, tweets are grouped together by their determined state, while the number of tweets associated

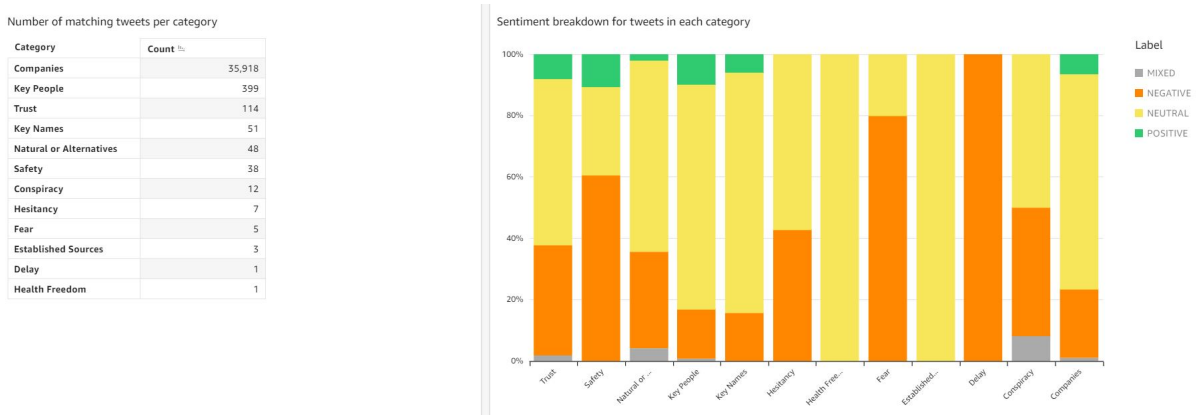

Fig. 3. Category Breakdown

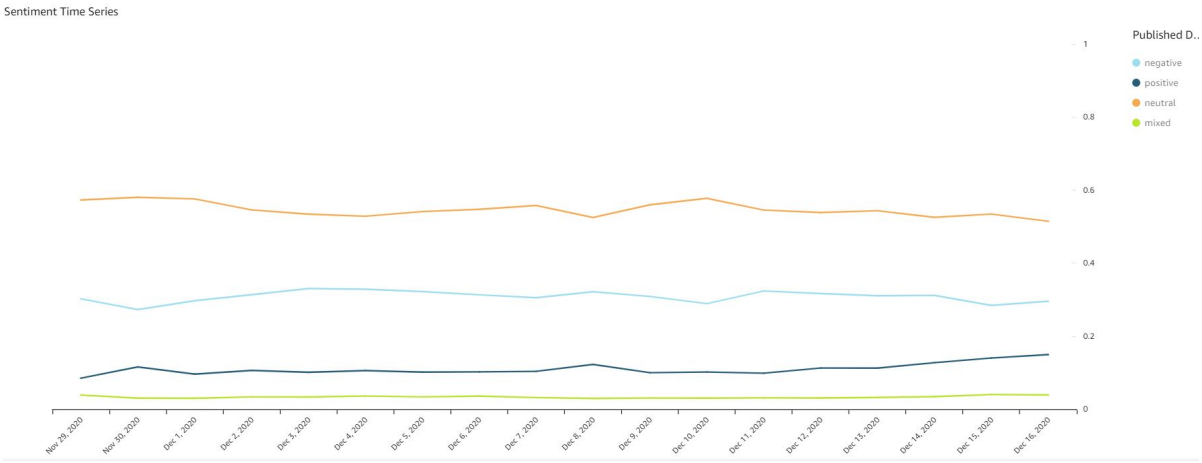

Fig. 4. Sentiment Time Series

with individual cities within a state can be viewed by hovering over a state's marker within the QuickSight dashboard. For the world map, it is important to acknowledge that the tweets being collected, and therefore represented in the visualization, are those containing a set of predetermined English terms. Additionally, as the location data is extracted from optional user-specified textual information, not all collected tweets are expected to have a determined location after analysis. Therefore, to provide users with additional insight into the relative amount of collected tweets reflected in these maps, a visualization showing the percentage of tweets from the past 7 days for which location data was successfully extracted was created.

IV. Conclusion

Our Sentiment Analysis system was built to aid healthcare professionals and researchers in the goal of overcoming vaccine hesitancy and responding to problems arising from public opinion. By dividing our system design into ingestion, analysis and visualization we have provided a flexible and robust modular framework which can be adjusted to gain valuable insights regarding overall public sentiment, as well as potential causal factors. To follow this work we would like to focus on the spatiotemporal differences and the causal factors behind sentiment breakdowns. By getting access to more specific location data we would be able to note the differences between regions and potentially how certain narratives spread. In addition to sentiment scores we believe that applying topic modelling methods like Latent Dirichlet Allocation to data over a long training period would allow fine tuned category breakdown which could then be used to inform causal factors for temporal fluctuations.

Acknowledgment

We would like to acknowledge all members of the Predictive Analytics & AI Lab of NYU Courant for working on this project. Special thanks to Matthias Heymann for always getting us to ask the right questions and helping through every step of the process.

References

[1] A. Syal, "Wearing a mask has become politicized. Science says it shouldn't be," *NBC News*, Jul-2020. [Online]. Available: https://www.nbcnews.com/health/health-news/wearing-mask-has-become-politicized-science-says-it-shouldn-t-n1232604. [Accessed: 28-Nov-2020].

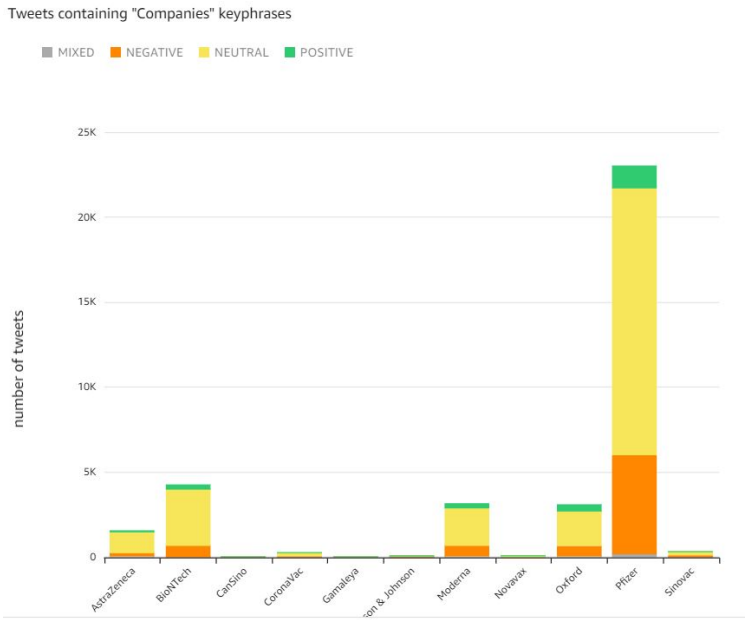

Fig. 5. Subcategory Breakdown: Companies

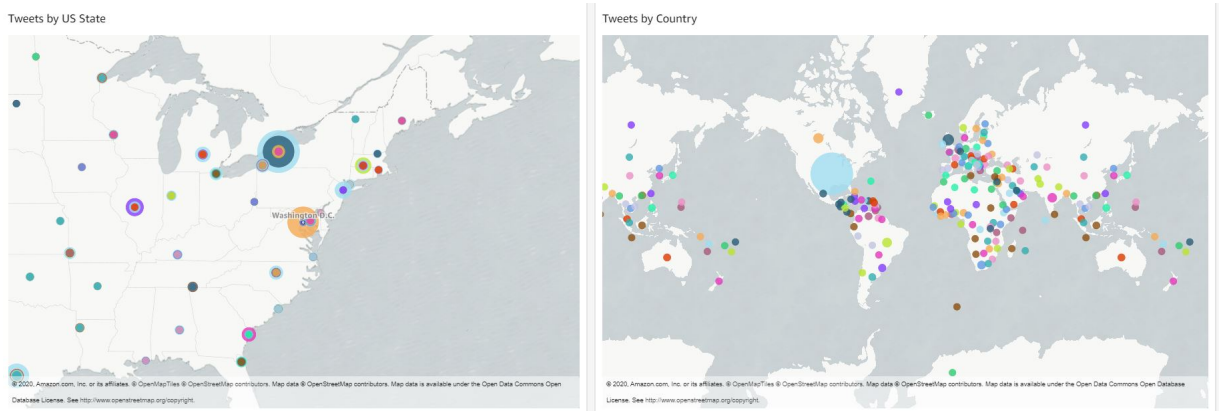

Fig. 6. Tweet Locations

[2] A. Tyson, C. Johnson, and C. Funk, "U.S. Public Now Divided Over Whether To Get COVID-19 Vaccine," *Pew Research Center*, Sep-2020. [Online]. Available: https://www.pewresearch.org/science/2020/09/17/u-s-public-now-divided-over-whether-to-get-covid-19-vaccine/. [Accessed: 28-Nov-2020].

[3] C. Aschwanden, "The false promise of herd immunity for COVID-19," *Nature*, vol. 587, no. 7832, pp. 26–28, 2020.

[4] C. Kent, "How to combat vaccine hesitancy in the age of Covid-19," *Pharmaceutical Technology*, Nov-2020. [Online]. Available: https://www.pharmaceutical-technology.com/features/how-to-combat-vaccine-hesitancy-in-the-age-of-covid-19/. [Accessed: 28-Nov-2020].

[5] *countries.csv*, Google Developers, 2012 [Dataset]. Available: https://developers.google.com/public-data/docs/canonical/countries_csv [Accessed: 20-Nov-2020].

[6] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *J. Comput. Sci.*, vol. 2, no. 1, pp. 1–8, 2011.

[7] J. Cheng, "The Psychology and Political Orientation of Social Distancing Compliance and Attitude toward Mask-Wearing during the COVID-19 Outbreak in the U.S.", 14-Aug-2020. [Online]. Available: psyarxiv.com/5k4ve.

[8] J. Darling, K. Thomas, A. Kapteyn, and F. Perez-Arce, "Vast Majority of Americans Support Wearing Masks, But a Deeper Look at Behavior Reveals Troubling Lack of Adherence," *Evidence Base - USC Schaeffer*, Aug-2020. [Online]. Available: https://healthpolicy.usc.edu/evidence-base/vast-majority-of-americans-support-wearing-masks-but-a-deeper-look-at-mask-wearing-behavior-reveals-troubling-lack-of-adherence-to-social-distancing-recommendations/. [Accessed: 28-Nov-2020].

[9] J. Du, J. Xu, H.-Y. Song, and C. Tao, "Leveraging machine learning-based approaches to assess human papillomavirus vaccination sentiment trends with Twitter data," *BMC Med. Inform. Decis. Mak.*, vol. 17, no. S2, 2017.

[10] J. Ulloa, "There's a big obstacle looming for coronavirus vaccines — a strong antivaccine movement," *The Boston Globe*, Nov-2020. [Online]. Available: https://www.bostonglobe.com/2020/11/28/nation/theres-big-obstacle-looming-coronavirus-vaccines-stronger-antivaccine-movement/. [Accessed: 29-Nov-2020].

[11] "Overview of Amazon Web Services," AWS, Aug-2020. [Online]. Available: https://docs.aws.amazon.com/whitepapers/latest/aws-overview/introduction.html. [Accessed: 30-Nov-2020].

[12] S. Sanche, Y. T. Lin, C. Xu, E. Romero-Severson, N. Hengartner, and R. Ke, "High contagiousness and rapid spread of severe acute respiratory syndrome Coronavirus 2," *Emerg. Infect. Dis.*, vol. 26, no. 7, pp. 1470–1477, 2020.

[13] V. Raghupathi, J. Ren, and W. Raghupathi, "Studying public perception about vaccination: A sentiment analysis of tweets," *Int. J. Environ. Res. Public Health*, vol. 17, no. 10, p. 3464, 2020.