Dissecting Twitter Data to Analyze Government & Public Attitude towards Covid and Vaccines

Shreyans Pathak
Graduate Student
spathak3@buffalo.edu

Alok Tripathy
Graduate Student
aloktrip@buffalo.edu

Harshad Arun Barapatre
Graduate Student
hbarapat@buffalo.edu

Taisia L Goncharouk
Undergraduate Student
taisiago@buffalo.edu

Department of Computer Science
University at Buffalo
Buffalo, NY 14260

Abstract

With the ongoing conversations around the Covid-19 pandemic on social media platforms like Twitter, it is essential to realize how politically significant individuals in a country affect the stance of the general population towards the virus and its vaccines. With opposing views of a population on the subjects of discussion, graphical representations around a topic of interest can help provide insights and capture the essence of these conversations.

1 Introduction

This report is part of the **Dissecting Twitter Data to Analyze Government & Public Attitude towards COVID and Vaccines** project being undertaken in the CSE 4/535 Information Retrieval course. Primary objective of this project is to apply IR concepts in detecting and analysing influence of twitter personalities on social sphere.

The project focuses on implementing a search engine with a web user interface on the tweets collected as part of Project 1 and provide various analysis on the basis of the context of the tweets for COVID and vaccine.

2 Project Requirements

The project has some basic requirements which are outlined below:

- Analyzing the attitude of the general population towards Covid vaccines.
- Analyzing the impact of Covid related political rhetoric on the common masses.
- Building a search engine.
- Developing a web user interface to present your content analysis.

3 Methodology

This section of the text describes the approach towards fulfilling the requirements of the project. Various options were considered and tried out of the whole corpus of tweets. The following sections provide details about each project requirement.

3.1 The Corpus of Tweets

The corpus used in the project contains about 180,000 tweets, including replies and retweets. It was built by collecting tweets using Twitter API v1.1 and v2. The tweets are based on keywords around COVID and its vaccines. Tweets from political influencers and medical agencies of the three countries - USA, India, and Mexico, were also collected. In total, around 20 POIs (Persons of Interest) were selected who were deemed to be politically influential in a country. The corpus dominantly has tweets in three languages - English, Hindi, and Spanish.The corpus was used to retrieve the results based on the queries.

3.2 Analysis of Attitude of General Population towards Covid vaccines

The attitude of any given tweet can be gauged by analysing the text and categorizing it into positive, negative, and neutral by using an algorithm. This process is usually known as sentiment analysis of text. Sentiment analysis of text helps us understand the underlying emotions, namely positive, negative, or neutral. The algorithm used for analysing the sentiment of the tweets in this project was VADER (Valence Aware Dictionary and Sentiment Reasoner.)

VADER is a rule-based analysis tool for sentiment analysis which is available in the *nltk* library. It detects the sentiments by analysing each token in the text which is done with the help of a lexicon sentiment dictionary. The words in the dictionary are usually labelled as positive or negative and the compound score of the text is calculated based on these scores of tokens. The algorithm gives about the positive and negative nature of the text and assigns a compound score between $[-1, +1]$ meaning that it does not just tell if the text is negative, positive, or neutral, but in fact, gives information about the severity of the nature.

The VADER algorithm requires input in the form of tokens of text, thus in this use case to analyze the sentiments of the tweets, the very first step was pre-processing the contents of the tweets. Since VADER performs best on English text, a tweet was first translated to English using the *deep_translator* package. The package contains various translation service functions which are based on translation services like Google Translator, My Memory Translator, Pons Translator, etc. Since Google's translation services are one of the very best in the industry, the GoogleTranslator functionality of the deep_translator package was used. Once the tweet is converted to English, the characters are converted to lowercase and all special characters from the tweet are removed. The tweet is then stripped off of any whitespace and then, it is tokenized. Next, stopwords from the tweet tokens are removed and finally, the tokens are stemmed using Porter Stemmer which is available through the nltk package.

The pre-processed tokens are sent to *polarity_scores()* in the Sentiment Intensity Analyser subpackage under VADER. The function returns the compound score for the tweet. These compound scores were then utilized to categorize the tweets into the following:

- *Positive:* sentiment compound score greater than or equal to 0.05
- *Negative:* sentiment compound score less than or equal to - 0.05
- *Neutral:* sentiment compound score between -0.05 and +0.05

The analysis and categorization of a tweet helps understand the stance of the tweet better. Given the limitations of the algorithm and the capability of deducing only so much from lexicon analysis, the categorization may certainly not always be accurate, but that cannot harm the general analysis of the tweets which gives a clearer picture of the overall rhetoric around a specific topic, which in this case, may be search queries.

3.3 Analysis of Impact of Political Rhetoric related to Covid

The politically important individuals of a country affect the general discussions about any given topic in that country. Similarly, the persons of interest taken into consideration for fetching the tweets can be generally seen to be responsible for the general population's perception around COVID and its vaccines. For example, if a POI tweets about the benefits of getting vaccinated, then the common masses have a positive perception about vaccinations. Or if a POI tweets about the pandemic being a hoax, the number of vaccinations can be seen to be going down while the number of active cases in the country increase. The two graphs do not necessarily have to be correlated, but it is a generally observed trend that more vaccinations lead to lesser active cases.

The observations stated above can be verified by plotting a graph of the number of active cases of coronavirus every day in a country and a graph of tweets by a POI on those days. The graph quickly shows that the two quantities appear to be somewhat correlated, meaning that whenever the POIs tweet about COVID and vaccines, the number of active cases go down and the number of vaccinations increase. Thus, from the observations, it is safe to say that the graphical representations provide a clearer picture of the impact of political rhetoric related to COVID and vaccines in a country.

3.4 Building a search engine

Apache Solr was used to index the data, and search queries are run on Solr which returns the results. In order to access solr and its functionalities using Python scripting language, the *pysolr* package was used. This package provides a comprehensive set of features to leverage the Solr indexing and querying capabilities. The retrieval system was built around the Apache Solr platform which is installed on an AWS EC2 instance for the data to be accessible from everywhere.

The retrieval system uses a unigram model with BM25 scoring in order to maximise the number of documents retrieved for a query. In general, the unigram model uses each token of the query in order to search and retrieve relevant documents. The documents are ranked on the basis of relevancy and the retrieval system returns documents on the basis of their scores. Documents with multiple tokens in the text are ranked higher and thus, appear on the top of the search. In the retrieval system, whenever the user inputs a query, it is converted to a unigram query before being sent to Solr to fetch results. This practice ensures that more query results can be retrieved without hurting the relevancy of documents retrieved as top results when the query is run as a whole on Solr. The results are relevant even for multiple keywords and phrases are entered as the query.

The retrieval system also incorporates the use of filters as a means to find specific tweets. The filters that are implemented are based on the following categories:

- POI
- Country
- Language

Each of these filters helps sort out the tweets for the selected option and displays only those tweets. For example, if POI “Joe Biden” is selected in the POI filter, all the tweets by Joe Biden will show up. Further, if Language filter is used to select “Hindi” on top of the previous filter, then the search may not provide any results because the POI may not have any tweets in the language.

3.5 Building a Web User Interface

The User Interface has been implemented using ReactJS, TypeScript, Material UI, CSS3. The whole implementation is split into multiple **components** which form the basic building block of React based applications. Various *npm* packages like *notistack*, *Recharts* etc. are used for showing the messages, rendering the graphs and many other. The user interface has been implemented in a way to give it professional and cleaner look and feel with intuitive components illustrated with data, legends etc. to make it easier for the user to comprehend the content rendered on the screen.

There are *4* main views in the application - home page, search results page, search result details page and insights. The user can enter the search query and the words being typed are also *autosuggested* to help user fill up the query efficiently. The backend servers the requests from the frontend using REST APIs which have been set up using *Flask* on the EC2 instance. Further details regarding the key features of the application are mentioned in the next few paragraphs.

Since the dataset we used is large, so it becomes imperative to retrieve the data from backend efficiently so that no slowness is caused when fetching data from queries which return results in thousands and more. To counter this, we have **paginated** the responses from the server and it is up to user to change to default number of rows to their liking. They can navigate forward or backward using the buttons given on the bottom right of the search result screen. This paginated approach makes the retrieval faster and the latency is reduced by a huge amount, which in turn increases the user experience overall.

Filters are also present in the application on the top right of the search result screen which facilitates in filtering the data retrieved. There are three main categories of the filters - POI based, language based and country based. The POI list is dynamically being generated using the API call that has been written on the backend for the same. User can choose from three different users from countries and three different records from language and filter out the results.

To generate a graph based on the tweets, the frontend sends a request to retrieve tweet data from the backend, which then receives the payload and returns the response using which we can plot a graph. A dedicated insights page was created to show, which gives visualisations for covid cases tracker, language wise and country wise distribution of tweets, covid status for each country and tweet counts of top POIs in our dataset. Another visualisation page is also created which is dedicated for providing insights for a specific tweets made by the POI. This visualisation page provides replies to the tweet and gives a visualisation for tweet counts of that specific POI according to dates.

An automated deployment pipeline has been set up using AWS CodePipeline, which is a tool for CI/CD that helps to automate release pipelines for fast and reliable application and infrastructure updates. The whole process is divided into 3 main steps - source, build and deploy. The source phase fetches the code from github repository, build phase runs the pre-specified *npm build* command to generate production version of the application and finally deploy stage hosts the application through S3 bucket.

4 Implementation

The homepage of the search application contains a search field where we can input query.

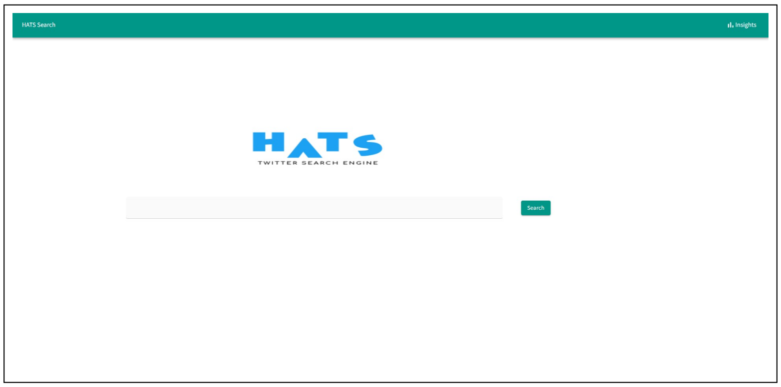


Figure 1: Homepage

At the top left is an Insights tab, once clicked, a series of graphs is loaded onto the page. The graphs are shown and described below.

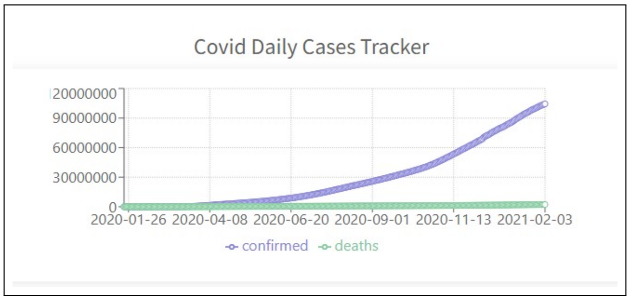


Figure 2: Covid Daily Cases Tracker

This first graph, Figure 2, displays the number of confirmed Covid cases against the number of Covid deaths over time.

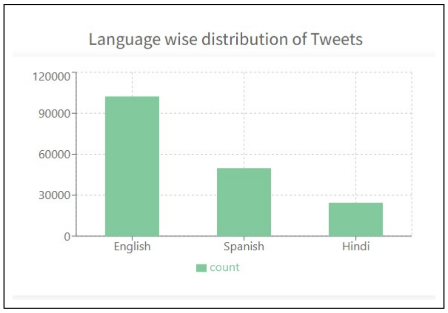


Figure 3: Language wise distribution of Tweets

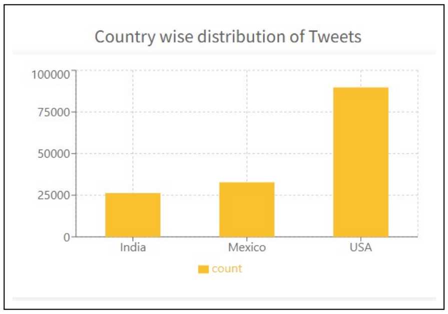


Figure 4: Country wise distribution of Tweets

Figure 3 and Figure 4 display the number of tweets per each language and per each country, respectively.

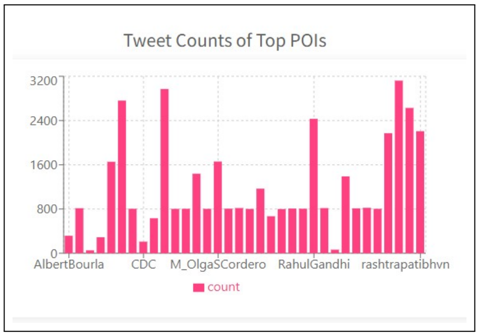


Figure 5: Tweet Counts of Top POIs

Figure 5 displays the total tweet counts for each POI. Hovering over a bar will display the POI name and exact count.

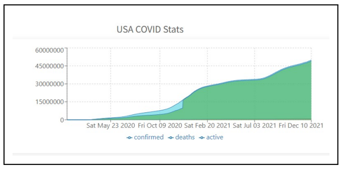


Figure 6: USA Covid Stats

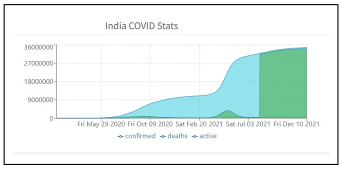


Figure 7: India Covid Stats

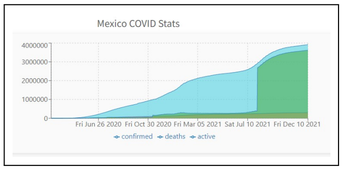


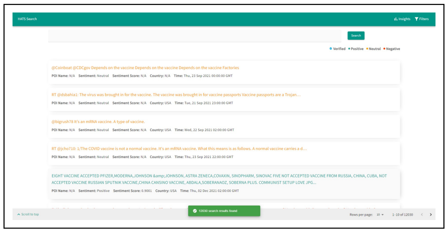
Figure 8: Mexico Covid Stats

Figure 6, Figure 7, and Figure 8 all display the Covid statistics for each country individually.

Now, back at the homepage, once we input a query and press 'search', the tweets for the specified query are retrieved. They are color coded based on sentiment, and have additional information attached, such as: poi name, sentiment, country, and time tweeted, as can be seen in Figure 9.

At the bottom, we are able to specify the number of tweets per page, and the page number. This allows us to view additional tweets. By increasing or decreasing the number of tweets per page, we can adjust how many tweets we view at a time. By specifying a page number, we can view less relevant tweets to the query.

As in Figure 10, once a query is entered, the filter option is displayed on the top right side of the application. When opened, the filter groups are seen as drop downs, with each option listed as check boxes. Checking 'Joe Biden' in the POI options, and 'USA' in the 'Country' option of these, the tweets are then updated.



@Coinboat @CDCgov Depends on the vaccine Depends on the vaccine Depends on the vaccine Factories
POI Name: N/A **Sentiment:** Neutral **Sentiment Score:** N/A **Country:** N/A **Time:** Thu, 23 Sep 2021 00:00:00 GMT

RT @dsbahia1: The virus was brought in for the vaccine. The vaccine was brought in for vaccine passports Vaccine passports are a Trojan…
POI Name: N/A **Sentiment:** Neutral **Sentiment Score:** N/A **Country:** USA **Time:** Tue, 21 Sep 2021 23:00:00 GMT

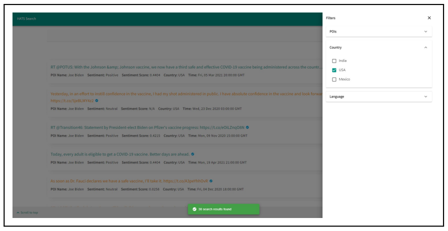
@bigrush78 It's an mRNA vaccine. A type of vaccine.
POI Name: N/A **Sentiment:** Neutral **Sentiment Score:** N/A **Country:** USA **Time:** Wed, 22 Sep 2021 02:00:00 GMT

RT @jcho710: 1/The COVID vaccine is not a normal vaccine. It's an mRNA vaccine. What this means is as follows. A normal vaccine carries a d…
POI Name: N/A **Sentiment:** Neutral **Sentiment Score:** N/A **Country:** USA **Time:** Thu, 23 Sep 2021 22:00:00 GMT

EIGHT VACCINE ACCEPTED PFIZER,MODERNA,JOHNSON &amp;JOHNSON, ASTRA ZENECA,COVAXIN, SINOPHARM, SINOVAC FIVE NOT ACCEPTED VACCINE FROM RUSSIA, CHINA, CUBA, NOT ACCEPTED VACCINE RUSSIAN SPUTNIK VACCINE,CHINA CANSINO VACCINE, ABDALA,SOBERANAOZ, SOBERNA PLUS. COMMUNIST SETUP LOVE JPG…
POI Name: N/A **Sentiment:** Positive **Sentiment Score:** 0.9001 **Country:** USA **Time:** Thu, 02 Dec 2021 02:00:00 GMT

12030 search results found

Figure 9: Searching for a query



□ India
☑ USA
□ Mexico

RT @POTUS: With the Johnson &amp; Johnson vaccine, we now have a third safe and effective COVID-19 vaccine being administered across the countr…
POI Name: Joe Biden **Sentiment:** Positive **Sentiment Score:** 0.4404 **Country:** USA **Time:** Fri, 05 Mar 2021 20:00:00 GMT

Yesterday, in an effort to instill confidence in the vaccine, I had my shot administered in public. I have absolute confidence in the vaccine and look forwa
https://t.co/5jeBLMY4z2
POI Name: Joe Biden **Sentiment:** Neutral **Sentiment Score:** N/A **Country:** USA **Time:** Wed, 23 Dec 2020 03:00:00 GMT

RT @Transition46: Statement by President-elect Biden on Pfizer's vaccine progress: https://t.co/eOiLZnqO8N
POI Name: Joe Biden **Sentiment:** Positive **Sentiment Score:** 0.4215 **Country:** USA **Time:** Mon, 09 Nov 2020 15:00:00 GMT

Today, every adult is eligible to get a COVID-19 vaccine. Better days are ahead.
POI Name: Joe Biden **Sentiment:** Positive **Sentiment Score:** 0.4404 **Country:** USA **Time:** Mon, 19 Apr 2021 21:00:00 GMT

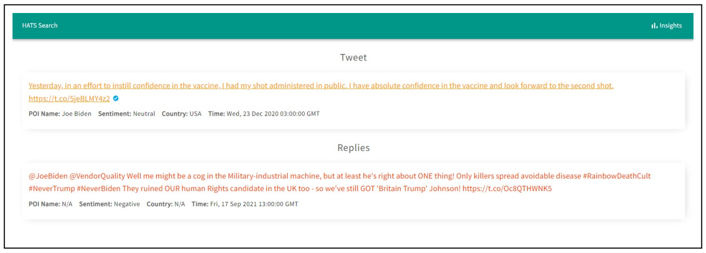
As soon as Dr. Fauci declares we have a safe vaccine, I'll take it. https://t.co/A3peYhhOvR
POI Name: Joe Biden **Sentiment:** Neutral **Sentiment Score:** 0.0258 **Country:** USA **Time:** Fri, 04 Dec 2020 18:00:00 GMT

38 search results found

Figure 10: Filtering

When a POI tweet from the results page is clicked, more information will be loaded about the tweet, if available. For example, in Figure 11, the POI tweet below displays the tweet text and the replies to that tweet if available. Following the replies are two graphs as shown in Figure 12. The first has the number of tweets over time by the POI. The second displays the rolling covid cases for the POI's country. Furthermore, clicking the tweet's text will load a twitter page of the original tweet.

Topic modelling was also implemented using LDA for better topic analysis based on the vaccine hesitancy but the module is currently not integrated to the web application due to inconsistencies but it can definitely be integrated with the application for better insights based on the topics of the tweets.



Tweet

Yesterday, in an effort to instill confidence in the vaccine, I had my shot administered in public. I have absolute confidence in the vaccine and look forward to the second shot.
https://t.co/5jeBLMY4z2
POI Name: Joe Biden **Sentiment:** Neutral **Country:** USA **Time:** Wed, 23 Dec 2020 03:00:00 GMT

Replies

@JoeBiden @VendorQuality Well me might be a cog in the Military-industrial machine, but at least he's right about ONE thing! Only killers spread avoidable disease #RainbowDeathCult #NeverTrump #NeverBiden They ruined OUR human Rights candidate in the UK too - so we've still GOT 'Britain Trump' Johnson! https://t.co/Oc8QTHWNK5
POI Name: N/A **Sentiment:** Negative **Country:** N/A **Time:** Fri, 17 Sep 2021 13:00:00 GMT

Figure 11: Joe Biden Tweet with Additional Info

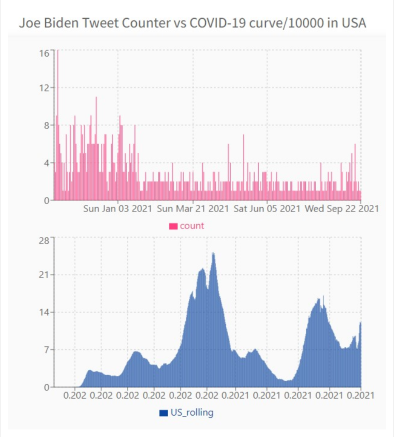


Figure 12: Joe Biden Tweet Graphs

5 Work Breakdown

- Frontend: Shreyans Pathak

- Backend: Taisia Goncharouk, Harshad Arun Barapatre
- Sentiment Analysis: Alok Tripathy
- Documentation: Harshad Arun Barapatre, Taisia Goncharouk, Alok Tripathy

6 CONCLUSION

In conclusion, information retrieval and visualization systems serve as means of understanding specific contexts from a large corpus of data. An end-to-end information retrieval system has several components related to indexing and querying to output relevant search results. These search results when used to draw graphical representations can help draw conclusions. Twitter is an information powerhouse in terms of the general population's interpretation and opinions about specific topics and definitive conclusions can be drawn out from these reactions. Furthermore, politically important personalities in various countries usually control the narrative around a topic. All in all, information retrieval systems cater to very specific information needs while graphical visualizations serve the purpose of fulfilling it.