

**DEEPAK SAI NURGOOR** [A20300845]

**NITIKA DWIVEDI** [A20294429]

**SRINIVAS KOTAMRAJU** [A20295847]

**SRUJANA TALLA** [A20301022]

**DECEMBER 2, 2021**

**TEAM**

**COVID-19 VACCINE SENTIMENT ANALYSIS**

Table of Contents

[**EXECUTIVE SUMMARY** 1](#_Toc89378820)

[**STATEMENT OF SCOPE** 2](#_Toc89378821)

[**Project Objectives** 2](#_Toc89378822)

[**Unit of Analysis** 2](#_Toc89378823)

[**Variables** 2](#_Toc89378824)

[**DATA PREPARATION** 3](#_Toc89378825)

[**Data Collection** 4](#_Toc89378826)

[**Data Reduction** 5](#_Toc89378827)

[**Data Consolidation** 6](#_Toc89378828)

[**Data Dictionary** 6](#_Toc89378829)

[**DATA PRE-PROCESSING** 7](#_Toc89378830)

[**Tokenization** 7](#_Toc89378831)

[**Stop Words Removal** 7](#_Toc89378832)

[**Stemming** 8](#_Toc89378833)

[**Lemmatization** 8](#_Toc89378834)

[**TF-IDF (Term Frequency and Inverse Document Frequency)** 8](#_Toc89378835)

[**SENTIMENT ANALYSIS** 9](#_Toc89378836)

[**Stanza** 9](#_Toc89378837)

[**Topic Modelling Using LDA** 17](#_Toc89378838)

[**Topic Modelling for Positive Tweets** 18](#_Toc89378839)

[**Topic Modelling for Neutral Tweets** 18](#_Toc89378840)

￼

[**Topic Modelling for Negative Tweets** 19](#_Toc89378842)

[**Named-Entity Recognition Analysis** 20](#_Toc89378843)

[**Emotion Detection Using Text2emotion Library** 20](#_Toc89378844)

[**Text2emotion** 20](#_Toc89378845)

[**Classification Model** 22](#_Toc89378846)

[**CONCLUSION** 25](#_Toc89378847)

[**APPENDIX** 26](#_Toc89378848)

[**Codes** 26](#_Toc89378849)

# **EXECUTIVE SUMMARY**

On 31 December 2019, several pneumonia cases were reported in Wuhan, China. The cause “**novel coronavirus**” (**2019-nCoV, renamed as SARS-CoV-2**) was declared as a pandemic by the WHO in March 2020 when more than 118,000 cases were registered in 114 countries. As on November 30th, 2021, **263 million Covid-19 cases** with more than **5 million deaths** have been registered across the globe.

For nations to come out of the pandemic and make their way back to “pre-Covid normal”, it was imperative for them to **administer vaccines** to their citizens and to do it as fast as possible. On December 11, 2020, FDA issued the first **Emergency Use Authorization for the use of Pfizer – BioNTech Vaccine** in the people aged 16 years and older. Since then, several other vaccines like **Moderna, Sputnik V, Oxford-AstraZeneca, Covaxin**, etc. have been administered across the globe. President Biden’s administration laid out an ambitious plan to get at least 70% US adult population vaccinated by July 4, 2021. However, as on October 22nd, 2021, **70.2 % of adults** are partially vaccinated and **59.4% of adults** are fully vaccinated. Not just the United States, but other countries like India struggled to ramp up their vaccination rates. Despite the availability of the vaccines, a slow vaccination rate led to a pronounced question: **“Are people hesitant to take the Covid-19 vaccine?”** This question was the trigger for our team as we identified a need of analyze the public’s sentiments towards Covid-19 vaccines.

Our team used Twitter to scrape tweets based on different keywords related to vaccines. Analyzing the Twitter content will **aid health experts and policymakers to learn more about public’s reaction to vaccines and hence devise efficient vaccination strategies and policies**. In addition to saving so many lives, such strategies will also help nations reopen their economies and boundaries.

# **STATEMENT OF SCOPE**

## **Project Objectives**

1. **Analysing Sentiments** of publictowards Covid-19 vaccination in the **United States** over the latest period
2. Generate Actionable Insights by narrowing down on the **list of key factors** driving the vaccine hesitancy.
3. Develop a meaningful report summarizing our Insights from the data collected and recommend meaningful next steps to our stakeholders (Health Experts and Policy Makers)

## **Unit of Analysis**

* **Twitter Data:** Tweets using certain keywords will be used/scraped to perform sentiment analysis.

## **Variables**

All the data collected by scraping tweets can be processed to create a dataset. The team would be working with the following columns – **User Id, Tweet Date, Tweet Content, Hashtags etc.**

**PROJECT SCHEDULE**

Graphical user interface, table

Description automatically generated

**Team’s Gantt Chart Snippet**

Please click on the below link to access the complete Gantt chart of the project.

<https://docs.google.com/spreadsheets/d/1XeAWEwE2rEYLL4C8RsrrBs2ooXCOcbeB-U02BTHyAUc/edit?usp=sharing>

Also, the team will upload the spreadsheet as a submission.

# **DATA PREPARATION**

**Collect Data from Twitter**

**Perform Data Pre-processing**

**Generating Bag of Words**

**Sentiment Analysis**

**Draw Actionable Insights**

**Process Flow Diagram**

**Tokenization**

**Data Cleaning**

**Stop Words Removal**

**URLs**

**Data Cleaning**

**Punctuation Marks**

**Data Cleaning**

**Normalization**

**Special Characters/Digits**

**Stemming**

**Lemmatization**

**White Spaces**

**Data Pre-Processing**

## **Data Collection**

Twitter’s API does provide access to the public tweets by random sampling in near real time. Although there might be concerns around how biased or imbalanced data from the sample tweets would be, **the data extracted from the samples of tweets obtained via API and the full tweet dataset reflect the same sentiment percentage with very little deviation.** The team set up multiple Twitter Developer Accounts, which enabled us to request access tokens and keys.

We used **tweepy** library in python to access the twitter APIs. Using the access keys and tokens received, we initialize the API and utilize its functionalities to extract the tweets over a period based on the related keywords.

Few keywords used to extract relevant tweets for the analysis are: **Covid, Covid-19, Coronavirus, Vaccine, Vacc, vaccination, vaccinated, Moderna, Pfizer, Johnson and Johnson, vaccine – hesitancy-willingness, variants, etc**.

From using the above API we have collected about 40,000 tweets.

The tweets obtained through the iterative search process are collected along with the meta data such as author, location, tweet date and hash tags, and are organized into a data frame to carry out the pre-processing steps.

**Data Cleaning**

In the data cleaning we will focus on two aspects, which are **extracting the hashtags from the tweet text** and **cleaning the text from the elements** which doesn’t contain any meaningful information related to the context.

To perform sentiment analysis, only text is required. Hashtags are extracted out of the tweet text and are analyzed separately; only relevant hashtags are retained in the tweet text.

In text cleaning, we remove the following elements:

* Punctuation marks like: ['%','/',':','**\\**','&amp;','&',';']
* Special characters and digits
* URLs
* Unnecessary white spaces

We remove the redundant tweets and remove the rows, which contains empty cleaned-text cells before carrying out the further analysis.

## **Data Reduction**

Initial data collected, based on the keywords related to covid19 and its vaccine, has the data from around the globe. This initial pull generated massive dataset, which upon further examination was narrowed down to the data limited to United States for the below reasons.

We aim to provide health care experts the analysis on the public’s reaction to vaccination to make proper decisions regarding vaccination policies. We wanted to restrict our analysis just for the United States as it would be tailored specifically to one country and its policies. Considering multiple countries and vaccines available across the globe will result in vague conclusions.

*Why United States?*

The United States was one of the worst affected countries in terms of the daily reported cases. Thus, the need for amplifying the vaccination rate is high. The United States is currently the home to three most effective vaccines and stands as one of the countries that have the highest vaccination rates. Since the vaccination rate is higher and is also available for population of various age groups, there is much scope for discussions about these vaccines and their effectiveness on twitter. This reduction in data has improved the data quality.

For reducing the data based on the geographical location we have used **Geopy** a python client on the obtained tweets from around the world. Geopy can make API requests to Nominatim (a geocoding software used for OpenStreetMap), Google Maps and many others. Geopy does not provide this information, but it simply helps the user to connect to the services. We Checked the country based on the location column previously available in the data frame. If the country is United States then the particular row is added to a new data frame and a new column Loc\_Country is created specifying the country as United States.

## **Data Consolidation**

Twitter allows to collect the data only up to week in the past, so we have collected the data over several weeks and with the search queries “Covid, Covid-19, Vaccination”. The different csv files are preprocessed and filtered with the location and then consolidated into one single csv file before carrying out the sentiment analysis. The final csv file is uploaded in GitHub and the code for the same can be seen in Appendix.

The final data on which we have performed the sentiment analysis was about 8100 tweets which was obtained after the data cleaning and location reduction.

## **Data Dictionary**

Since we are dealing with the tweets i.e., text, which is unstructured data we do not have the variables that directly explains the data or the model. However, we have organized the preliminary data into a data-frame for all the pre-processing purposes.

INPUT DATA

Table 1

|  |  |  |  |
| --- | --- | --- | --- |
| **Column Name** | **Description** | **Data type** | **Source** |
| **Tweet\_id** | Unique\_id to identify the tweet | Int | Internal assignment |
| **Username** | Author of the tweet | Str | [www.twitter.com](http://www.twitter.com) |
| **Location** | Location from which the tweet was generated (City, Country) | Str | [www.twitter.com](http://www.twitter.com) |
| **Text** | Content/text within the tweet. | Str | [www.twitter.com](http://www.twitter.com) |
| **Tweet\_date** | Date and time when the tweet were generated | Str/Date-time | [www.twitter.com](http://www.twitter.com) |
| **Hashtags** | The hashtags associated with the tweet | Str | www.twitter.com |

OUTPUT DATA

|  |  |  |  |
| --- | --- | --- | --- |
| **hashtags\_extracted** | The cleaned hashtags extracted from the original hashtag column | Str | Output data |
| **loc\_country** | United States – Data reduced based on the location | Str | Output data |
| **text\_cleaned** | Preprocessed tweets removing stopwords , punctuation etc | Str | Output data |
| **stanza\_score** | The sentiment score given to each tweet based on the sentiment of the tweet | Int | Output data |
| **stanza\_sentiment** | Classified as Positive, Negative and Neutral based on stanza\_score | Str | Output data |
| **emotion** | A dictionary with key as emotions and the score of each emotion in the tweet as key | Dictonary | Output data |
| **Happy** | Score of the emotion happy in the tweet | Float | Output data |
| **Angry** | Score of the angry happy in the tweet | Float | Output data |
| **Surprise** | Score of the surprise happy in the tweet | Float | Output data |
| **Sad** | Score of the sad happy in the tweet | Float | Output data |
| **Fear** | Score of the fear happy in the tweet | Float | Output data |

# **DATA PRE-PROCESSING**

Text Mining, also known as Text Data Mining, corresponds to the process of transforming unstructured texts into a structured format to identify meaningful Patterns and new Insights. However, a very important step before we can start deriving Insights is pre-processing our corpus. Output of this step is generally used for Topic Modelling, Sentiment Analysis, Sentiment Classification etc.

Following are some steps that the team carried out as a part of the data pre-processing step:

**Tokenization:** Tokenization is a process of converting meaningful and sensitive data element into random string of words called “tokens”, which has no meaningful or exploitable value. It is a unique identifier that retains all the suitable information relevant to the raw data without compromising data’s security. This step is vital for developing model for NLP. Python offers various effective library for tokenization. The most popular and widely used library is NLTK, which has been used in our project as well.

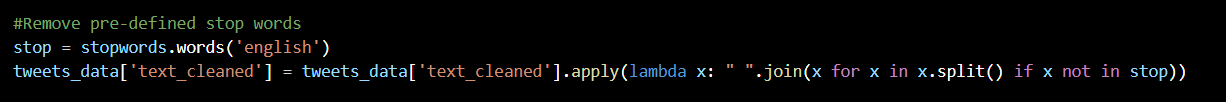
**Stop Words Removal:** Tokenization process used to remove non-relevant words such as if, but, we, she, he, etc. Removal of these words does not change the semantics of text but contributes to improving the performance of sentiment analysis model by removing low-level information from the data. NLTK is an effective library, which is used for stop words removal in our project. In addition, we also removed words like “Covid”, “Vaccine”, “get” etc., which are repeating across all the documents in the corpus

Fig 1: Code for Stop Word Removal

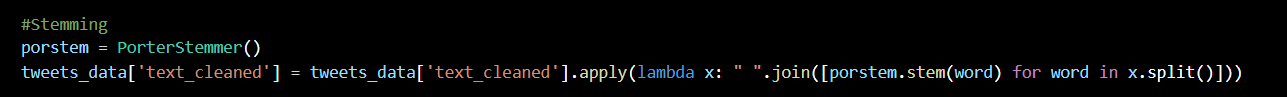
**Stemming:** Stemming is a rule-based approach of reducing inflection from words by removing suffix or prefix from the same, which in turn reduces the dimensionality of the data. The output depicted as a group of relatable words.

Fig 2: Code for Stemming

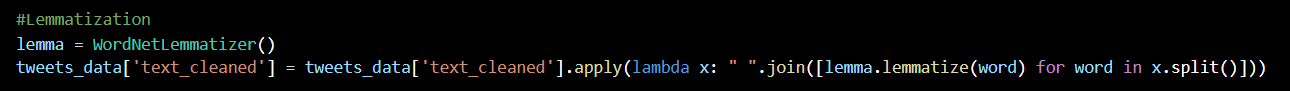
**Lemmatization:** Stemming and Lemmatization go hand in hand. In many cases, reducing inflection from words and grouping them under similar stems can diminish the actual meaning of the word. Each inflected word has a root-base word form known as lemma. Lemmatization reduces the words to its lemma using vocabulary and morphological analysis of words. We have used NLTK for Stemming and Lemmatization to achieve the desired results.

Fig 3: Code for Lemmatization

**TF-IDF (Term Frequency and Inverse Document Frequency):** Term Frequency accounts for the count of the frequency of the term in the document (Document in our project refers to a tweet). Words that occur frequently are assigned frequency count value. Higher is the frequency of words, the higher will their value weightage. However, words like “the”, “is” etc. can appear multiple times in a tweet or text data, therefore, can possess higher weightage. Inverse Document frequency evaluated the frequency of words across all the documents. Thus, even if word has a high-frequency count in one document, it is still assigned with a low F value. On the other hand, words with high frequency across all the documents are assigned with higher IDF value. Assigning these values helps in ranking each tweet and assessing the importance of a particular word. While bag of words creates a set of vectors with the count of frequently appearing words, TF-IDF assigns weightage to words and provides information on words with high and low importance value. Thus, our team preferred TF-IDF over the bag of words.

A screenshot of a computer

Description automatically generated with medium confidence

Fig 4: Code for TF-IDF

# **SENTIMENT ANALYSIS**

There are techniques under **Natural Language Processing** which enables us to overcome the challenges of understanding the meaning of a text. One such technique is the **Sentiment Analysis**, which is contextual mining of the text to systematically identify, extract and quantify subjective text-based data on emotional (Positive, Negative or Neutral) or other categories with valences (strong, weak, powerful, powerless). This technique helps us get an understanding of the attitude, opinion, or views people have towards a product, service, organization, brand, politician, and anything else that is written about (Covid Vaccines in our case). We have used the **Stanza library** in Python for Sentiment Analysis.

## **Stanza**

Python provides many libraries for text and sentiment analysis. One of such effectual libraries is Stanza, consisting of a compilation of effective tools for text and sentiment analysis.

Built on highly accurate neural network components, Stanza provisions efficient data training and evaluation of approximately 66 human languages, using **Universal Dependencies formalism**. Universal Dependencies (UD) is a framework for consistent annotation of grammar (parts of speech, morphological features, and syntactic dependencies) across different human languages. Furthermore, Stanza maintains stable Python interface to CoreNLP Java Package, inheriting constituency parsing, coreference resolution, and linguistic pattern matching and other such functionalities.

Following is an example how Stanza can be used

(Refer [***https://stanfordnlp.github.io/stanza/sentiment.html***](https://stanfordnlp.github.io/stanza/sentiment.html) )

Graphical user interface, application, Teams

Description automatically generated

Fig 5: Stanza Code Example

Text

Description automatically generated

Fig 6: Code for Stanza Sentiment Analysis

Following is the distribution of Positive, Negative and Neutral Sentiments across the different tweets recorded in our dataset:

Chart, bar chart

Description automatically generated

Neutral sentiment can be dominantly seen across the corpus of documents (61 %) as compared to the Positive (14.2 %) and the Negative sentiments (24.8 %).

Fig 7: Distribution of Positive, Neutral & Negative Tweets

Fig 8

Fig 9

Fig 10

We have generated word clouds for the respective sentiments using the WordCloud2 Package available in R Programming Language.

***Following is the Word Cloud generated for Positive Sentiments across the tweets:***

A picture containing shape

Description automatically generated

Fig 11

‘**Protect**’, ‘**Safe**’, ‘**good**’, ‘**best**’ ,’**help**’ are some words which are reflective of people’s Positive Sentiments towards vaccines. They find the vaccine safe and believe that it is the best option they have in addition to masks to protect themselves against Covid.

Following is the Word Cloud generated for Neutral Sentiments across the tweets :

Text

Description automatically generated

Fig 12

A Neutral Sentiment generally refers to the announcements or discussions around the vaccines and their availability. ‘**Child’** , ’**dose**’ , ’**age**’ ,’**avail**’ are some key words which we can see in the above word cloud. These tweets are probably around different vaccines now being available for children, hence increase the scope of vaccinations around the world.

Following is the Word Cloud generated for Negative Sentiments across the tweets:

Text, shape

Description automatically generated

Fig 13

‘**death**’, ‘**effect**’, ‘**mandat**’, ‘**dont**’ ,’**hesit**’ are some words which are reflective of people’s Negative Sentiments towards vaccines. They don’t find the vaccine safe primarily due to the aftereffects of the vaccine. Another major discussion happening these days is around the Vaccine Mandate, which are being introduced across different places to increase the rate of vaccinations. This plan has faced a lot of criticism from different sections of the population and hence can be seen here clearly in the word cloud.

## **Topic Modelling Using LDA**

Topic Modelling is a method for unsupervised classification of documents, similar to clustering on a numeric data which finds some natural group of items even when we are not sure of what we are looking for.

Text

Description automatically generatedWe used LDA to perform Topic Modelling on our corpus of data. Latent Dirichlet Allocation (LDA) is commonly used to discover a user-specified number of topics shared by documents within a text corpus. Each observation (tweet in our case) is a document, the features are the presence (or occurrence count) of each word, and the categories are the topics. The topics are learned as a probability distribution over the words that occur in each document. Each document, in turn, is described as a mixture of topics.

Fig 14

***Following is the result of Topic Modelling:***

### **Topic Modelling for Positive Tweets**

* **Topic -1**

['commun', 'mask', 'best', 'amp', 'health', 'good', 'test', 'provid', 'proof', 'requir']

* **Topic -2**

['thank', 'peopl', 'today', 'dose', 'recommend', 'clinic', 'covid', 'receiv', 'shot', 'booster']

* **Topic -3**

['hesit', 'peopl', 'covid', 'immun', 'amp', 'good', 'best', 'child', 'rate', 'protect']

**Interpretation –** Topic 1 in our view talks about 2 things. One, vaccines along with the masks being good for the health of the community. Topic 2 talks about people being thankful for receiving their booster shots which has very recently commenced in the United States, Topic 3 talks about how vaccines are best for children and their protection. This makes sense, primarily because vaccines like Pfizer etc. are now being administered to children as well.

### **Topic Modelling for Neutral Tweets**

## 

* **Topic -1**

['shot', 'health', 'visit', 'test', 'amp', 'dose', 'booster', 'clinic', 'age', 'child']

* **Topic -2**

['feder', 'peopl', 'employe', 'requir', 'hospit', 'hesit', 'state', 'rate', 'covid', 'mandat']

* **Topic -3**

['pfizer', 'coronaviru', 'immun', 'appoint', 'ny', 'avail', 'new', 'variant', 'read', 'clinic']

**Interpretation –** As stated earlier, neutral tweets are mostly around the discussions which happen around vaccines and different vaccine policies. One the same lines, Topic 1 talks about a general advice to people around visiting hospitals to get the booster shot and get even the children vaccinated. Topic 2 is a discussion around the vaccine mandate where federal employees are required to get vaccinated before resuming offices. Topic 3 is a discussion around the new variant and around Pfizer talking about their vaccine against the new variant.

### **Topic Modelling for Negative Tweets**

* **Topic -1**

['polici', 'report', 'covid', 'new', 'proof', 'neg', 'mandat', 'requir', 'test', 'dose']

* **Topic -2**

['risk', 'covid', 'immun', 'low', 'case', 'infect', 'flu', 'death', 'rate', 'effect']

* **Topic -3**

['want', 'know', 'like', 'die', 'amp', 'dont', 'hesit', 'peopl', 'covid', 'health']

**Interpretation –** Topic 1 very clearly talks about people’s negative sentiments about the new vaccine mandate. The mandate would require people to show the proof of vaccination or get vaccination before they could resume offices. Topic 2 is more reflective of fear people have about the side effects of the vaccine. There are issues here and there about possible side effect of every vaccine out there. Topic 3 on some similar lines highlight the vaccine hesitancy among the anti-vaxers.

## **Named-Entity Recognition Analysis**

## **Emotion Detection Using Text2emotion Library**

We used the Text2emotion Library in Python to be able to relevant emotion from out corpus of data.

### **Text2emotion**

Emotions are neurophysiological reactions associated with thoughts, feelings, behavioral responses causing an affective state of consciousness causing humans to experience feelings such as joy, sorrow, fear, hate, love, etc.

Humans are intellectual being with highly effective neural network that enables them to easily identify as well as experience these emotions physically and in textual form as well.

Unlike humans, machines do not have this capability to comprehend emotion right out of text. To achieve the same, python provisions text2emotion package, which is capable to extract relevant emotion from the textual data.

Chart, bar chart

Description automatically generated

Fig 15

From the visualization, we can see that there is more fear among people regarding vaccinations as compared to other feeling, i.e., joy / anger / sad etc. The fear might be stemming out of the possible side effects of the vaccine doses. The fear might also stem out of the possible ineffectiveness of the vaccines against a stronger and a more transmissible variant like a ‘Delta’ variant. There might also be fear among the population for the young children who are the most vulnerable given the limited number of vaccines approved for them. Some tweets in the corpus can also be attributed to general attitude/sentiment of the population towards this pandemic situation rather than just vaccines.

## **Classification Model**

We built a classification model to predict the sentiment of the public towards the vaccination for Covid-19 by giving the text from the tweets as the input. As mentioned earlier, we have defined the sentiment as ‘Positive’, ‘Negative’ or ‘Neutral’ based on a Normalized Stanza score. A score of 1 is for Positive, 0 for Neutral and -1 for Negative.

The text from these tweets is converted into numerical data based on Term Frequency-Inverse Document Frequency values (TF-IDF Matrix ) with 2500 maximum features, 7 absolute counts as minimum document frequency and 0.8 proportion of the document as maximum document frequency.

The original data was split into 80% for training and 20% Validation and the Random Forest Classifier model was used for classification. A Random Forest is a meta estimator that fits several decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is controlled with the max\_samples parameter if bootstrap=True (default), otherwise the whole dataset is used to build each tree.

The Overall accuracy of the model was **79%** on the validation data. The Classification report is given below. The f1-score for the Negative Sentiment is 0.66, for the Neutral Sentiment is 0.86 and for Positive Sentiment is 0.63.

Table

Description automatically generated

Chart, treemap chart

Description automatically generated

Most of the tweets in the dataset is Neutral and the classifier model is doing a good job in predicting the Neutral Tweets with high accuracy. The Number of True Neutral Tweets which our model predicted as Neutral is 956. Similarly, the number of the True Positive tweets predicted as Positive, and the number of True Negative tweets predicted as Negative is on the higher side in terms of accuracy.

We see this as an important step in the analysis. For federal governments, policy makers to take any decision on the vaccine policy, it is important for them to understand the true sentiment among most of the population. This will help in addressing the right issues at the right time.

# **CONCLUSION**

# This has been an interesting analysis for the team. Having been following the policies around Covid-19 very closely, this analysis has helped the team to understand the gradual shift of sentiments among the population towards vaccinations. There was a lot of hesitancy around vaccines when it was first rolled out in December 2020 with FDA approving the vaccines by Pfizer and Moderna. The news of vaccines having unusual side-effects only fueled the hesitancy. The

# **APPENDIX**

## **Codes**