COVID-19 Vaccination Hesitancy Textual Analysis

Introduction

COVID-19, a type of coronavirus that provoked the global pandemic, has reshaped the entire world. Social distancing, travel restrictions, mask mandates, and remote working has become unprecedentedly common in our normal life. The negative impact of COVID is indeed severe: many individuals encounter both economical and mental issues. Since early 2020, everyone has had the same hope: end of pandemic. The good news is that vaccines were developed within several months, which can prevent infection and reduce spread of diseases. Thanks to cutting-edge pharmaceutical technology, we could now see a positive and promising turning point of a long-lasting health crisis.

Global pandemic, however, is not slowing down although various types of vaccines are deployed in the market. Recently, cases resurged everywhere in the US and other countries. Such increase of infection results from occurrence of two new variants: Delta, which went viral in the late summer, and Omicron, another subtype of COVID-19 that gave rise to new outbreaks. Another factor that pushed up daily COVID-19 is vaccination hesitancy. Although vaccines have been proven to be safe, a great percent of individuals are still reluctant to take their first shot.

Vaccination hesitancy could be harmful to people's health and extend the time of pandemic. To face the challenge of encouraging vaccination, we should discover its root cause, and one approach is utilizing vaccination-related tweets. Short texts and comments from twitter reflect personal beliefs. By deciphering perspectives from the public, government and specialists could further explain whether these thoughts are true. Also, they could set a more practical strategy to promote COVID vaccine and encourage vaccination.

Our research focuses on COVID vaccine tweets analytics, and aligns its results to COVID-19 time-series vaccination progress. We clean the dataset feature by feature to ensure high precision of our findings. For our analysis, we implement numerous sentiment models to decipher the feelings of the general public toward vaccines. Also, we extract the most common words among tweets to identify the most popular vaccine-related issues. Findings from both sentiment and most common-word analysis could help define the relationship between opinions about vaccines and action of vaccination, and provide constructive suggestions for vaccination encouragement.

Covid-19 Vaccine Tweets

We built our analytics based on the Covid-19 all vaccine tweets (Preda, "Covid-19 All Vaccine Tweets") dataset from Kaggle. The dataset encompasses 200,000 observations and columns **(Supplementary Table 1)** regarding vaccine tweets, user location, hashtags, etc. To break down the data and drive business insights in a more effective way, we utilize Natural Language processing. With this approach, we are able to quantify tens of thousands of textual data, which expedites the speed of the entire analytic pipeline. However, textual analytics could easily generate bias as these data are more unstructured compared to a quantitative one. Thus, a step-by-step exhaustive cleaning is required beforehand.

Data Preparation

We merely select data that are useful for our analysis. Among all extracted columns, 'id' identifies each user, 'date' represents the date the tweet is posted, 'user_verified' indicates whether the account is verified by Twitter, 'user_location' refers to the area of the user, and 'text' is, of course, the tweets. Within 'date' and 'id' variables, no major feature engineering is needed. Yet the 'user_verified' column requires more evaluation. In this variable, unverified users account for more than 90 percent of total users. Removing all unverified users could be unfeasible as less than 10 percent of data will be kept. The only way to maintain an ample number of data is to ignore this signal.

User Location

Values in 'user_location' are miscellaneous. Some rows appear to be cities, others are countries, and still others are random unrelated words. Location allows us to evaluate texts across different areas, and to compare data from one place to another. Yet messy locations bring about false classification, which ends up a biased analysis. To resolve this issue, we should unify all regions to the country level. We use the GeoPy module to obtain the exact location. We then search and output the country value in the exact location with the assistance of pycountry. Of course, data with opaque locations will be dropped. Once all locations are uniform, we could classify the data by them **(Fig. 1)**:

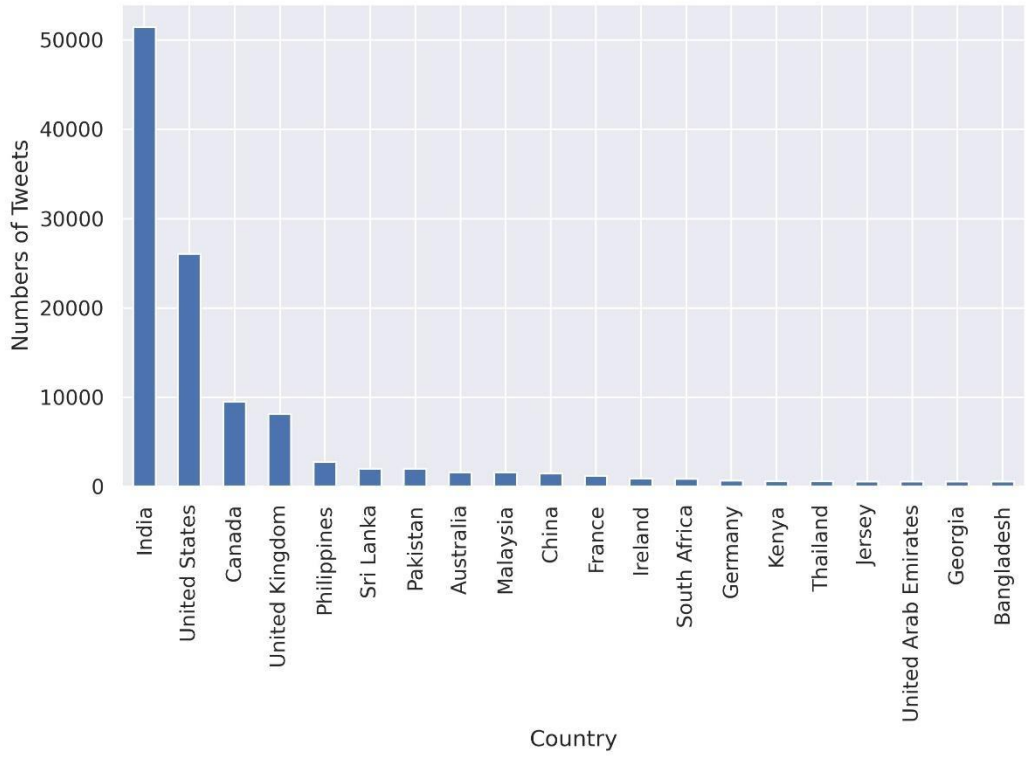

Fig. 1 Numbers of Tweets by Country

According to the figure above, data is extremely imbalanced by country. Tweets in India top 50,000, followed by the United States with more than 20,000, Canada and the United Kingdom with nearly 10,000 individually. Comparatively, many other countries possess less than 5,000 tweets, most of

them are even below 1000. In this case, a general analysis without considering country is not a suitable way as we are unable to balance data by resampling. Over-sampling is impractical as the numbers of rows across countries are too polarized. Under-sampling is unrealistic either since this approach will wipe out most rows. It is obvious that a more detailed and specific method is needed.

An alternative way is no doubt going over tweets country by country. Each country has its own culture, and thus, will probably shape its unique perspectives regarding vaccination. Country-level analysis could cover this characteristic, which could produce more customized results. In this project, we will select the country that includes ample numbers of tweets as only a large sample size could clearly represent the entire population. Four countries with the highest numbers of tweets, which are mentioned previously, will be selected for analysis. Other regions, which have insufficient data, will not be discussed at this time.

Text

Now, let's move on to the most complex feature—text. Textual cleaning is fairly hard to quantify given its unstructured traits. To prepare our text data, we commence our cleaning by lowercasing all words and removing noises (URL, hashtags, twitter id, line breaks, multiple spaces, and single character). Also, since few tweets are written in foreign languages, we should convert them to English. We detect non-English language with langdetect and translate them with Google Translator. All progress above makes the data well-prepared for a sentiment model, Flair. But for other sorts of analysis, such as TextBlob and Vader, as well as N-gram, which we will explain later, we should take further data preparation steps.

Data cleaning for other models relies on removal of punctuation and stopwords, tokenization, and lemmatization. Such models belong to a more-traditional and non-neural network analysis. They capture the patterns of text by counting numbers of words or evaluating polarity and subjectivity of each word. In this scenario, stopwords and various forms derived from one root word could interfere with the analysis, which makes further cleaning processes a must. In contrast, Flair, a pre-trained neural network model, covers the relationship between vocabularies, so we could keep the original sentence format.

Sentiment Analysis

Sentiment Analysis is considered one of the most popular ways to assess polarity of text. The output, so-called 'Sentiment score', could reflect whether a person is giving a positive or a negative tweet. Currently, there is a wide range of models to conduct such analysis and different models suit different types of textual data. Here, we demonstrate three sentiment models, Flair, TextBlob, and Vader, and evaluate the quality of their results by visualizing the distribution of scores and by, in an old-school way, reviewing some texts and scores manually.

Flair

Flair model, a pre-trained model that is built on top of PyTorch, captures forward and backward word patterns. It is developed in 2017 and has been regarded as 'state of the art' model (Saxena). Leveraging its large pre-trained library, we postulate that it could create a more precise sentiment score, and therefore, we feed our data to test the result. However, the model seems not to perform

as expected. The score is extremely polarized across tweets **(Fig.2)**. Divided results may be driven by one of two factors, either score is inaccurate or texts are extraordinarily disparate. We will review part of the data manually to confirm our assumption.

Diving deeper into sentiment scores by reviewing part of dataset manually, we found several mislabeled observations. Text *'singapore approves pfizer vaccine, becomes first south east asian country to approve the vaccine'* scores -0.6332, *'have released their interim guidance for use of covid vaccine'* scores -0.9644, and '*the -19 vaccine began leaving the company's factory on sunday*' scores -0.9996. Though these texts sound neutral, Flair model gives them extreme negative scores. A similar phenomenon appears in text labeled as positive. '*fact sheet for recipients and caregivers*' scores 0.9741 and '*pre screening form for vaccine'* scores 0.7736. It is clear that Flair model is over-sensitive.

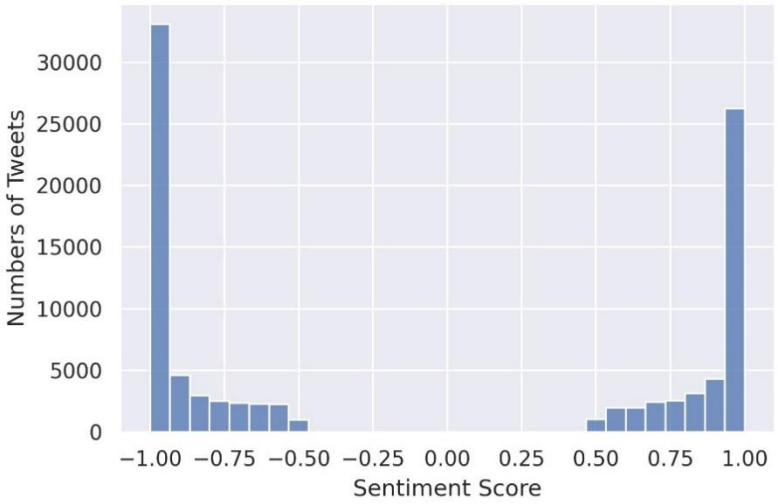

Fig. 2 Distribution of Sentiment Score from Flair Model

TextBlob

TextBlob calculates sentiment in a more traditional way. It gives the score for each single word and averages all scores together (Kuzminykh). In the opposite of Flair, TextBlob returns more neutral scores **(Fig. 3)** for our analysis. From the figure below, most data are grouped as 'completely neutral' with a score of 0. Only a few texts are considered positive and negative. Taking a more detailed look by manual review, some texts with slight positive view are regarded as neutral: '*privileged offered vaccine would encourage everyone take*' and '*grateful scientist basic scientist barney graham drew weissman*' score 0.

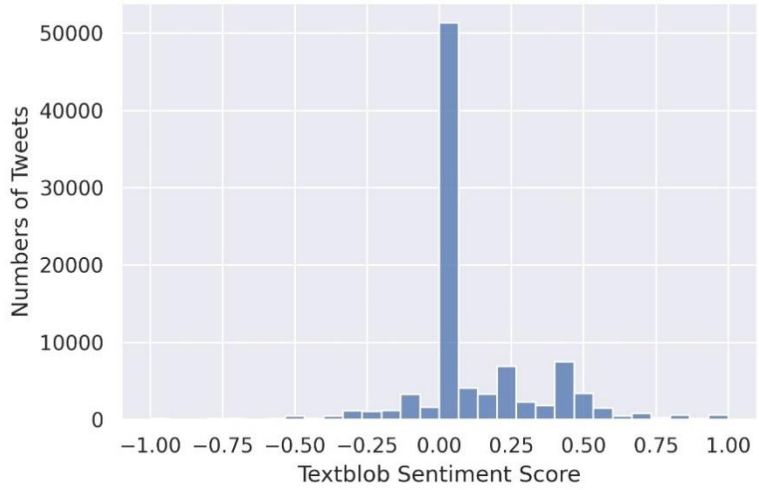

Fig. 3 Distribution of Sentiment Score from TextBlob

Vader

Vader, a sentiment library that works similarly to TextBlob, maps score to lexical features and average them together. Given that it is proven one of the best models for twitter analytics (Hutto and Gilbert), we applied this package for our analysis. Yet our result contradicts previous research. Compared to TextBlob, nearly all texts are classified as 'neutral' **(Fig. 4)**. For those with positive and negative scores, each tweet merely contains a word: 'scare' as -0.4939, 'shortage' as -0.25, 'glad' as 0.4588, and 'nice' as 0.6. We could no doubt see a poor performance of Vader model in this case.

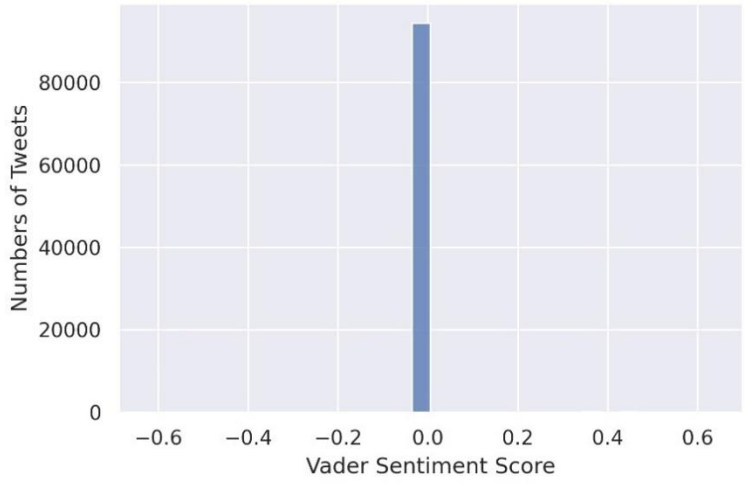

Fig. 4 Distribution of Sentiment Score from Vader

So, what caused the severe bias of sentiment analysis? Possible factors could be data in the pre-trained model or contents in the sentiment library. Within the first model we implemented, Flair, its database for pretrained model may include little to no text that is strongly related to COVID-19. Similarly, library in TextBlob and Vader may not contain strong positive or negative words related to pandemic. For example, 'allergic reaction' and 'side effects' should be marked as negative, yet these words could be undetectable by text models. As existing models could not adjust their assessment in COVID-19 related text, we should consider other approaches to break down text data.

N-gram and CountVectorizer

N-gram, the number of sequences in a text and sentence, is an alternative way to crack the textual data. It cuts the entire text into pieces with n-words: unigram for one word, bigram for two, trigram for three, and so on. Though it is hard to see its benefit when we introduce it individually, the power of n-gram becomes apparent once we combine it along with CountVectorizer, the module that counts the number of each different n-gram in a dataset. With the combination of two methods, we are able to rank each n-gram by how many times it appears, and therefore, find out the most common words or phrases in text.

We applied N-gram and CountVectorizer packages to datasets from 4 countries with the highest numbers of tweets: India, United States, Canada, and United Kingdom. We utilize unigram, bigram, trigram, and 4-gram to select words and phrases and count the numbers of grams that are existed in each dataset. Top 50 most common sequences in each gram will be selected for further discussions. We will review all top 50s manually to analyze whether it indicates any clues that trigger reluctance of vaccination.

Results of N-gram and CountVectorizer are more promising than sentiment analysis. It seems that this combination of techniques could detect key information from a large number of texts. Among all 4 countries, useful keywords and phrases are: *'free slot', 'paid slot'* in India, *'side effect', 'sore arm', '89-year-old man died', 'died day received vaccine', 'immune system make antibody', 'genetic code actually hacking'* in the United States, *'new case', 'open business', 'eating alone', 'side effect', 'stop politicizing'* in Canada, and *'side effect', 'blood clot', 'vaccine safe', '89-year-old man died', 'rare heart inflammation', 'died day receiving vaccine'* in the United Kingdom.

Taking a quick glance at all keywords, though we found that side effects are one of the major concerns, actual information indicated by those keywords still varies across borders. Instead of analyzing these texts on a general basis, we will discuss them based on characteristics of each country. We will connect keywords with vaccination progress, along with press releases, to check if there are any relationships among them. Only by integrating all reliable information together could we bring out a higher-quality analysis. But before moving onto these steps, we should prepare the vaccination progress dataset.

Vaccination Progress

The Covid-19 world vaccination progress datasets (Preda, "Covid-19 World Vaccination Progress"), which originates from Kaggle, covers the trend of vaccination progress from December 2020 to October 2021. Combination of 51308 rows and 15 columns **(Supplementary Table 1)** encompasses all vaccination data across different countries. As usual, we select the trend of 4 countries. For the column, we choose 'people fully vaccinated per hundred' as we believe percentage of fully vaccinated is the optimal way to demonstrate the progress of vaccination. We then visualize the trend and review its results **(Fig. 5)**.

Percentages of fully vaccinated differ across countries. In India, vaccination rate remained below 10 % until late-August and reached 20 % in mid-October. Numbers of fully vaccinated appears relatively more prospective in the USA. Vaccination rate jumped over 50% but slowed since October. Compared

to the United States, Canada lagged during the first few months yet later leaped ahead. Trends of vaccination rate in the United Kingdom shows some similarities with Canada. The most major difference is a less steep increase **(Fig. 5)**.

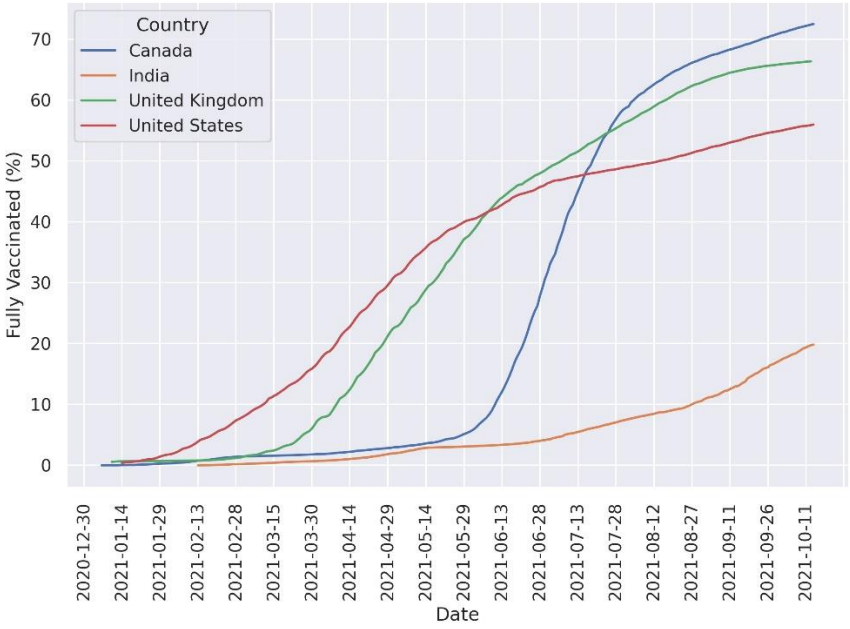

Fig. 5 Vaccination Progress across 4 Countries

Discussion

After gathering sufficient insights from two datasets, we can now put all pieces together. First, we will analyze the trend of vaccination rate for each country and deduce what stimulates changes in vaccination. Next, we will check if and how vaccination hesitancy appears in tweets as well as whether there are any connections between tweets and vaccination progress. Lastly, we will provide suggestions driven by information from dataset and a few of those from press releases. Customized analysis will be given for each country.

India

The vaccination rate in India is surprisingly low. Fully vaccination rate failed to reach 20 % until early October. According to latest press release, it is reported that “70 % of India’s population have received at least a single dose of vaccine, but just 30 % are fully vaccinated with both doses” (Mukherji). Such a low full vaccination rate may be driven by several reasons. One factor is that government asked citizens to schedule a second dose through app while low penetration prevents accessibility to the next shot. However, door-to-door vaccine service could resolve this issue (Mukherji).

Another factor that prevents the increase of vaccination is the availability of free slots. Breaking down the tweets data, we spot that *‘free slot’* and *‘paid slot’* are two of the most popular phrases. This crucial message indicates a lack of free vaccines, which reflects the truth of Indian nation. “State governments in the country have had to ration Covid-19 vaccines doses and shut down free vaccination centres, while private centres, where Indians can pay to get the vaccine, have plenty of

shots available" (Kapur). Concerns about shortage of free vaccines could be clearly found in both textual data and press releases. Therefore, government should increase the availability of free shots.

In regard to vaccination hesitancy, there is no hint in tweets data. Yet media reported a great percentage of general public that are hesitant to take a jab. Among all unvaccinated populations, 23% are worried about side effects and 16 % are awaiting more data to prove enough protection ("LocalCricles"). However, given that there is no further evidence of reluctance to take vaccines from our dataset, we could not prove that side effects and concerns about protection bring about a stagnant vaccination rate.

United States

Fully vaccination rate in the United States falls between 50 to 60 percent in October. That is, a great number of Americans have not yet received their first or second dose and vaccination hesitancy, which has been frequently mentioned by press releases, encumbers the increase rate. If we take a look at the tweet data in the US, worries of potential harm is everywhere. Twitter users mention *'side effect', 'sore arm', '89-year-old man died', 'died day received vaccine', 'immune system make antibody', 'genetic code actually hacking'*. Fortunately, CDC is addressing the hesitancy in the correct way—giving more detailed explanations regarding mechanisms and side effects of vaccines.

Canada

In Canada, full vaccination rate stays below 10 % until June, followed up by a steep rise for two months. Yet it later decelerated. Slowing speed of vaccination triggered by higher vaccination hesitancy in the south, according to experts (Aziz). Our tweets data also shows that *'side effect'* appears the most common word. This could be a possible element that leads to a slowing vaccination rate. Also, the phrase 'stop politicizing' may indicate that making vaccination political could hinder the process of injecting all vaccines into every single person.

Apart from detecting vaccination hesitancy issue, we could also drive specific solutions from the most common words. *'eating alone'*, *'new cases*', and *'major business'* are very popular topics on twitter. Promoting vaccination based on these points could probably attract more populations to schedule their doses. To be more specific, we could emphasize the fact that: "Without vaccination, cases will surge, which will affect major businesses. We even have to eat alone more often to prevent the spread of diseases." We could then mention that increase in vaccination is the best way to eliminate pandemic. This could help encourage vaccination.

United Kingdom

Vaccination rate in the United Kingdom jumped over 60 % between mid and late August, showing a promising progress. Yet it does not mean the problem of vaccination hesitancy is completely ameliorated. All useful most common tweets are all-around common awareness regarding COViD-19 vaccines: *'side effect', 'blood clot', 'vaccine safe', '89-year-old man died', 'rare heart inflammation', 'died day received vaccine'*. From these sequences, we could see some similarities with tweets from the US. Therefore, we could recommend a similar strategy—asking more health experts and government to explain mechanisms and side effects of vaccines.

Challenges

Though we eventually brought out some results and business suggestions, challenges occurred when moving on the analytics flow. We have collected both tweets and vaccinations progress datasets and interpreted them individually, yet it is hard to build a strong link between the two. Changes of increases in vaccination rate depend on several factors: supply chain, regulations, etc. Given that vaccination hesitancy is not the only component that drive the trend, we could not confirm a causal impact of personal thoughts on real action. Also, when it comes to twitter analytics, tweets may not represent the entire public voice. Such issues may negatively affect our accuracy of analysis.

Conclusion

Throughout our analysis, we have cleaned the tweet datasets and selected useful features. We have transformed all location variables to countries, classified tweets with this variable, and filtered out countries with lower than 5000 tweets. We then measured the polarity of tweets with different sentiment models and found none of them are suitable for COVID-related texts. Alternatively, we leveraged a combination of n-gram and bag-of-words to deal with our textual data and brought out some useful findings. In the last part, we processed the vaccination progress dataset and strove to align the trend to results derived from all tweets although the relationship is weak. All steps are necessary for providing practical suggestions to fight vaccination hesitancy.

With the help of our analytics pipeline, we are able to give customized recommendations to encourage vaccination. As we found that twitter users in the United States and the United Kingdom both focus more on detailed information of side effects, we recommend government and health experts to provide further interpretation of safety and mechanism of vaccinees. In India, as general public is concerned about lack of free slots, we suggest government to make more free vaccines accessible to every single nation. In Canada, as citizens are strongly aware of how life is affected by pandemic, we propose government highlight how vaccination could bring everything back to normal.

Of course, there are biases in our analysis. We are incapable of matching vaccination progress with common tweets. We are unable to discover factors that stimulates each change of vaccination rate. Within our textual data, we are unsure if tweet information could precisely reflect an overall opinion in public. Yet despite these biases, we still brought insights that will probably be beneficial for government to encourage vaccination, and thus, help pushes up vaccination rate. Such actions aim for a goal—the goal that all of us shares—end the pandemic as soon as possible.

References

"LocalCircles estimates 11.5 crore Indian adults are currently hesitant to take the COVID vaccine." LocalCircles. https://www.localcircles.com/a/press/page/localcircles-vaccine-hesitancy-survey. Press Release.

Aziz, Saba "Canada's 2nd dose vaccinations surpass U.S. as Americans grapple with COVID-19 surge." Global News, 17 July 2021, https://globalnews.ca/news/8036558/covid-vaccine-second-dose-canada-passed-us/. Press Release.

Hutto, C., & Gilbert, E. "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text." *Proceedings of the International AAAI Conference on Web and Social Media*, 8, 216-

225. 2014.
Kapur, Manavi “India’s vaccine supply is a curious mix of abundance and shortage.” Quartz India, 2 Aug. 2021, https://qz.com/india/2041072/indias-cowin-has-many-paid-vaccination-slots-despite-shortage/. Press release.
Kuzminykh, Natalia “Sentiment Analysis in Python With TextBlob” Stack Abuse. https://stackabuse.com/sentiment-analysis-in-python-with-textblob/
Mukherji, Biman “100 million Indians have skipped their second vaccine dose, leaving the country vulnerable to a third COVID wave.” Fortune, 2 Nov. 2021, https://fortune.com/2021/11/02/india-covid-vaccine-covishield-skip-second-dose-third-wave-cases/. Press release.
Preda, Gabriel “COVID-19 All Vaccines Tweets.” Kaggle. https://www.kaggle.com/gpreda/all-covid19-vaccines-tweets
Preda, Gabriel “COVID-19 World Vaccination Progress.” Kaggle. https://www.kaggle.com/gpreda/covid-world-vaccination-progress
Saxena, Sharoon “Introduction to Flair for NLP: A Simple yet Powerful State-of-the-Art NLP Library.” Analytics Vidhya. 11 Feb. 2019. https://www.analyticsvidhya.com/blog/2019/02/flair-nlp-library-python/