Sentiment Analysis on Vaccination Tweets

Anand Vadavelli
San Diego State University
San Diego, USA
RED ID 823389639
avadavelli2394@sdsu.edu

Mayur Bagwe
San Diego State University
San Diego, USA
RED ID 825903748
mbagwe4673@sdsu.edu

***Abstract*—The objective of this project is to perform sentimental analysis on COVID-19 vaccination using Twitter as a data source and tweets from the users as a dataset. A worldwide vaccination campaign is underway to achieve herd immunity and bring an end to the SARS-CoV-2 pandemic; however, its success relies heavily on the actual willingness of individuals to get vaccinated. Social media platforms such as Twitter may prove to be a valuable source of information on the attitudes and sentiment towards SARS-CoV-2 vaccination that can be tracked almost instantaneously. In the project we performed sentiment analysis on vaccination tweets data from twitter.**

I. Introduction

The practice of vaccination dates back hundreds of years. For example, the first smallpox vaccine was developed in 1798 and over the 18th and 19th centuries, systematic implementation of mass smallpox immunization culminated in its global eradication in 1979. Any vaccination development would take years to be administered to people. However, in the case of covid-19, for the first time, some of the vaccinations were developed using mRNA techniques. Although scientifically there were no negatives with this approach, some people are hesitant towards covid-19 vaccination. In other scenarios, some people are against injecting a dead virus with the regular type of vaccination.

A certain section of people was against the vaccination because covid infections were also happening among the vaccinated people. Although most people are pro for vaccination some were against forced vaccination. When the AstraZeneca vaccine was administered in the U.K. there were cases with people having side effects. This created panic among some people. In India, when vaccination was first offered, people were against it and were under the impression that their natural immunity would handle covid. This led to the second wave in India where cases reached the peak and overburdened health care systems. Even my family members were affected by this. Some people wanted to take vaccination only after proper data on the efficacy has been published.

However, such efficacy data was not fully published which resulted in lower vaccination rates in some countries. In U.K. covishield vaccination administered in India was not recognized for a long time. This made people think if taking vaccines is right for them. Although these are some of the scenarios, where people were against the vaccination, it has brought things back to normal. Without vaccination, things wouldn't have been back to normal. This motivated me to sentiment analysis on the vaccination. In our sentiment analysis, we found that hashtags like stoptheshot, pandemicofvaccinated have been some of the frequently used negative hashtags. Some of the hashtags that had positive responses included GetVaccinated, CancelCovid, and VaccinePassport.

- Why it's worth researching?
  The discussion of vaccination progress, accessibility, efficacy, degree of trust in vaccines and side effects is ongoing. Perception on vaccination and mRNA techniques has been both positive and negative based on people's religious believes, geographical regions, lack of trust in government and philosophical beliefs, and it is permeating through news stories on Twitter each day. However, as online users, our visibility is limited to our own echo chambers. Thus, the motivation for this project is to widen my perspective on the state of the global pandemic by harnessing the power of Twitter data

II. Approach

We have used Tweepy API library to receive real time data about vaccination tweets from twitter. We performed data cleaning and pre-processing using natural language processing techniques such as lemmatization, stemming and removing stop words. We used pandas data frames and PySpark data frames along with PySpark SQL. We have also used libraries like Textblob, genism, vaderSentiment, nltk, spacy for analyzing the data. Furthermore for visualizations we have used word cloud, seaborn and matplotlib.

In our research project we choose vaccination topic and performed sentimental analysis on this topic. Twitter provides a tweepy library in order to get started with tweepy, we install the python package with the command

"pip install tweepy"

In order to use tweepy library, you will need to have a twitter developer account. Request a twitter developer account from https://developer.twitter.com/ . Once you have a twitter developer account, create a new App and provide a unique project name to get your keys tokens than can be used along tweepy. Copy and save the API Key, API Key Secret and Bearer Token that will be provided with the new App.

Authenticate with the tweepy API by providing the **consumerKey, consumerSecret, accessToken and accessTokenSecret**

- authenticate = tweepy.OAuthHandler(consumerKey, consumerSecret)
- authenticate.set_access_token(accessToken, accessTokenSecret)
- api = tweepy.API(authenticate, wait_on_rate_limit = True)

Note: We will be using wait_on_rate_limit=True in order to handle twitter limit exception and wait dynamically.

Tweepy provides a search_tweets method that allows to search for the tweets. A lot of meta data is returned along with the tweet. But for our sentimental analysis we are only interested in tweet, date and hastags. One request to search_tweets method only returns a maximum of 100 tweets, so we need to call this method multiple times in order to extract large amount of data. For this project we extracted 30,000 tweets. The timestamp of the tweet is returned in GMT and we are only interested in date so we converted the timestamp to yyyy-mm-dd

We used the below hashtags to query the data from twitter.

hastags = ['vaccination', 'vaccine', 'mrna', 'boostershot', 'covid19', 'Antibody', 'Allergy', 'Antigens', 'Antiviral', 'pandemic', 'efficacy', 'immunity', 'infectious', 'outbreak', 'placebo', 'quarantine', 'side effect', 'strain', 'pfizer', 'moderna', 'jj', 'johnsonjohnson', 'covaccine', 'covishield', 'biontech', 'janssen', 'zycov-d', 'astrazeneca', 'sputnik', 'convidecia', 'sinopharm', 'coronavac', 'novavax', 'epivaccorona', 'CancelCovid', 'HealingStartsHere', 'CovidVaccine', 'ImVaccinated', 'IGotVaccinated', 'VaccinesSaveLives', 'VaccinesWork', 'TheTruthAboutCovid', 'CovidTruths', 'CovidMyths', 'LifeInAPandemic', 'VaccinesSaveLives', 'trustscience', 'vaccinated', 'frontlineworkers', 'cdc', 'hospital', 'doctor', 'HealthForAll', 'Antivax', 'longtermcare', 'nurse', 'vax', 'deadvirus', 'virus', 'Coronavirus', 'mandatevaccine', 'omicron', 'deltavariant', 'spread', 'walgreens', 'cvs', 'pharmacy', 'Antivaxxer', 'forcedvaccination', 'adultvaccination', 'kidsvaccination', 'vaccinationrate', 'mutation', 'maskmandate', 'n95', 'flu', 'delta', 'ppe', 'ventilator', 'incubation', 'communityspread', 'Asymptomatic', 'Presymptomatic', 'NovelStrain', 'Pathogen', 'Socialdistancing', 'SelfIsolation', 'SelfQuarantine', 'SuperSpreader', 'Epidemic', 'SurgicalMask', 'HerdImmunity', 'SARS', 'SARS-CoV-2', 'covid-19', 'Cluster', 'Transmission', 'IncubationPeriod', 'DropletTransmission', 'FlattenTheCurve', 'DriveThruTesting', 'RTPCR', 'rapidtest', 'faceshield', 'covidtesting', 'lockdown', 'reopen', 'deltaplus']

III. Evaluation

Data Processing

Any data from social media platforms like twitter has to properly cleaned to analyze the data We will do the following on this data

1) Removing the ulrs from the tweets
2) Removing twitter handles
3) Removing twitter reserve words
4) Removing punctuations
5) Removing single letter words
6) Removing blank spaces
7) Removing stop words
8) Removing numbers
9) Tokenization
10) Stemming

We evaluated our analysis by calculating polarity scores for the tweets and by comparing different algorithm techniques such as topic model analysis.

A. Polarity/VADER sentiment:

With this project we wanted to analyze people's opinion towards vaccination across the globe. Classifying the sentiments either as positive or negative is called polarity. The sentiment is expressed in 3 ways. positive, negative and neutral based on polarity scores. Supervised and unsupervised techniques such as knowledge bases, ontologies, databases, and lexicons can be used for predicting the sentiment. In this project we used lexicons. In simple terms lexicon is a dictionary, vocabulary or a book of words. A lexicon has a list of positive and negative polar words with a score associated with it. The score is based on the position of the words, surrounding words, phrases, parts of the speech, context and so on. With the aggregation of these scores, we get a final sentiment.

We found the VADER lexicon is more efficient for this project. VADER is a lexicon and a sentiment analysis tool that is specifically used to find out sentiments expressed in social media. Vader can handle emojis and slangs. It also considers capitalization and punctuations when giving the scores. For example, Vader understands that HAPPY!! is more positive than 'happy'. Vader gives 4 types of values, positive, negative, neutral and compound. In our code we considered the compound score which is a combination of other three scores.

B. Topic model analysis:

Topic model analysis is an unsupervised statistical model that can be used to find out topics that occur in a document based on text mining. It identifies the topics by detecting patterns, distance between words and recurring words. It groups similar words to infer topics with unstructured data. There are two main topic modeling methods. Latent Semantic Analysis and Latent Dirichlet Allocation. In this project we used LDA. Although we did not use LSA in our project, lets understand LSA. LSA computes how frequently words occur in a document and the whole corpus using term frequency-inverse document frequency(tf-idf). This model further creates document-term matrix which shows tf-idf value of each word in the document.

The purpose of LDA is mapping each tweet in our corpus to a set of tweets which covers a good deal of the tweets in the document. LDA assigns topics to arrangement of words and treats as bag of words. It assigns a probability of that these words belong to a particular topic. LDA assigns two hyper parameters alpha and beta for topic similarity. A higher value of alpha will assign more topics to a tweet. A low value of beta will use fewer words to model a topic whereas a high value will use more words, thus making topics more similar between them. A third parameter will be set when implementing LDA for the number of topics. In our project we use pyLDAvis. It is a package used for interactive chart and is used visualizing the topics-keywords. Each bubble that is present in the graph represents a topic. The larger bubble represents that it is more prevalent topic. From our graph we observed the model is having big overlapping bubbles scattered. This suggested is a good topic model. A model with many topics is having many overlaps and all the small bubbles are grouped into a cluster. When you move the cursor over the bubbles then you can see the bars on the right side gets updated accordingly with the important keywords in that selected topic.

Related Work:

A supervised machine learning model would have given us a better sentimental analysis on our dataset. For this project we used a reference from the kaggle project.

Conclusions

In our sentiment analysis, we found that hashtags like stoptheshot, pandemicofvaccinated have been some of the frequently used negative hashtags. Some of the hashtags that had positive responses included GetVaccinated, CancelCovid, and VaccinePassport. Overall we noticed that many tweets had neutral score and slightly more positive responses.

References

[1] T. Vijay, A. Chawla, B. Dhanka and P. Karmakar, "Sentiment Analysis on COVID-19 Twitter Data," 2020 5th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE), 2020, pp. 1-7, doi: 10.1109/ICRAIE51050.2020.9358301.

[2] N. S. Sattar and S. Arifuzzaman, "COVID-19 Vaccination Awareness and Aftermath: Public Sentiment Analysis on Twitter Data and Vaccinated Population Prediction in the USA," Applied Sciences, vol. 11, no. 13, p. 6128, Jun. 2021.

[3] https://www.kdnuggets.com/2018/08/emotion-sentiment-analysis-practitioners-guide-nlp-5.html

[4] https://monkeylearn.com/blog/introduction-to-topic-modeling/

[5] https://www.kaggle.com/keplaxo/introductory-eda