Evaluating public options of COVID-19 vaccine using natural language processing

**Problem description**

2020 is an unprecedented time and the whole world were almost shut down during the pandamic. Recently, two types of vaccines have been approved by FDA and they are like the light in the dark. Hopefully, vaccine can help end the pandemic and we can go back to normal life in 2021. However, there are different options about vaccines. People may welcome, oppose, or be skeptical about Vaccinations. It is important to find what people think.

**Data sets**

My dataset is obtained from twitter API using stream filter method. The objective is to obtain the tweets related to covid vaccine. However, if I directly search ‘covid vaccine’, the obtained tweets may be misleading because covid related tweets can have different keywords other than ‘covid’. To tackle this issue, I will utilize two filters: the first filter is to filter vaccine related tweets and after that, utilize the second filter to filter the covid related tweets based on the tweets obtained from the first filter. The keywords utilized in the vaccine-filter is ‘vaccine’, ‘vaccinated’ and ‘vaccination’; the keywords of the covid-filter is 'covid' , 'covid-19', 'coronavirus', 'pandamic', 'covid19', 'social distance', 'cov2', 'quarantine', 'moderna', 'oxford', 'pfizer', 'reopen', 'social distance'. By filtering the tweets twice, I am able to obtain the datasets related to covid vaccine. The details about data import, data cleaning is described below.

1. Data import

As mentioned above, I utilized twitter API to obtain the tweets. After obtaining the API, I utilized the following function to obtain twitters.

tweets1 = tw.Cursor(api.search, q = SearchKeyWords, lang="en", since='2021-02-01', until ='2021-02-02').items(TotalSearchNum)

df1= pd.DataFrame([tweet.text for tweet in tweets1], columns = ['Tweets'])

This tweepy.Cursor function obtains the tweets that contain the key words in the “SearchKeyWords”. As mentioned above, as the first filter, the SearchKeyWords I utilized is ‘vaccine’, ‘vaccinated’ and ‘vaccination’. After filtering the tweets, the tweet text is saved into a datafame.

Api.search only allows to search for the tweets in the past week. In addition, I specify since and until date in the function which means I only search the tweets in a specific date. In order to obtain the tweets in the past weeks., I searched called tweepy.cursor function 7 times to obtain the tweets in the past weeks.

After obtained all the tweets, I combined all the tweets into a new dataframe called. The number of the total tweets obtained in this step is 2400.

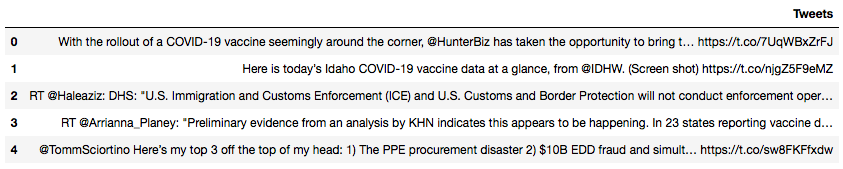


Figure 1: raw tweets

As shown in Figure 1, the raw tweets contain punctuation, web addresses, retweets, and etc.. These features need to be cleaned before data analysis.

1. Data cleaning

After importing the raw tweets, I utilized the following functions to clean the tweets by removing @, retweets, removing punctuation, changing to lower case, splitting sentences and joining words. Figure 3 shows the tweet text after cleaning.

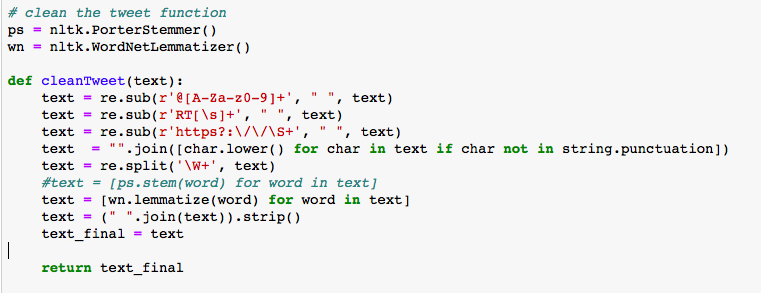


Figure 2: the tweet cleaning function

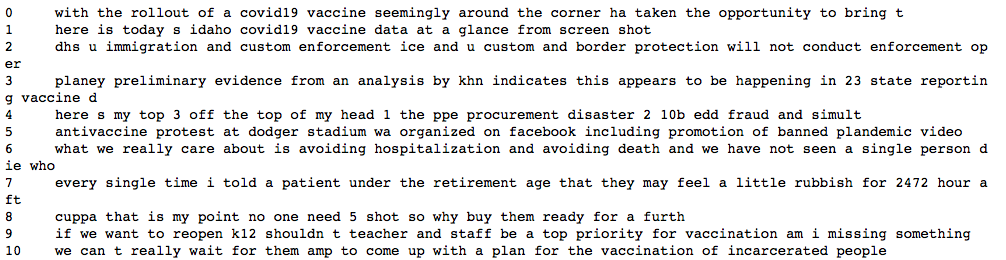


Figure 3: the tweet text after cleaning

The obtained tweets may contain duplicate tweets, and thus the duplicated tweets are removed after tweet cleaning. After removing the duplicated tweets, the remaining tweets is 1852. However, the tweets after all the previous data wrangling is the tweets only related to vaccine. Thus in the next step, I applied a second filter to filter the tweets related to covid. The following code shows the second filter that I utilized to obtain the tweets which are both related to covid and vaccine. The final number of tweets after all the data cleaning and filters is 548. Figure 5 shows the final tweets that will be utilized in the data analysis.

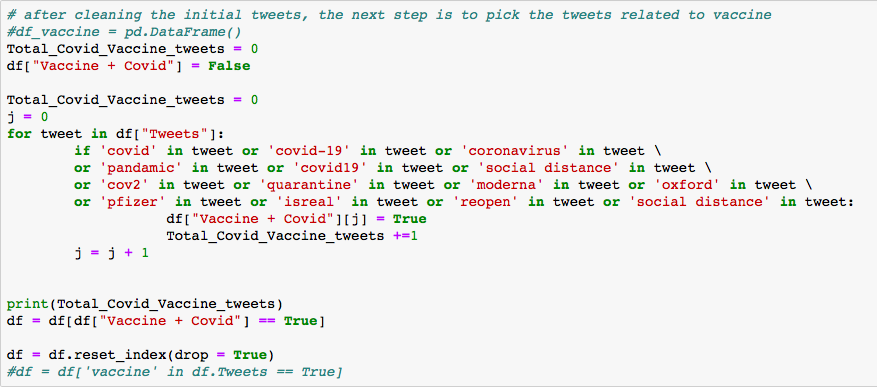


Figure 4: second filter to obtain covid related tweets

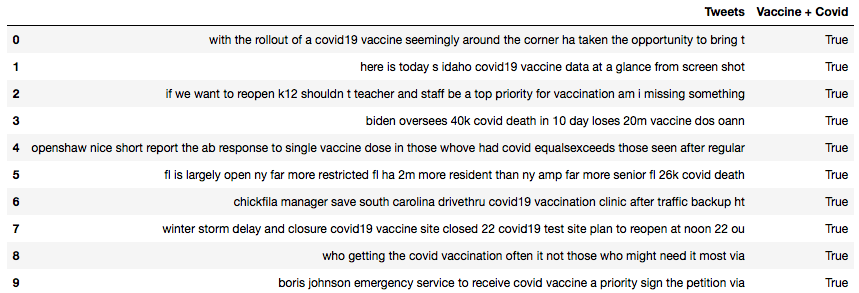


Figure 5: covid vaccine related tweets after cleaning and filters

**Data Analysis**

The topic in this project is related to twitter sediment. In the sediment analysis, it returns polarity and subjectivity. Subjectivity ranges from 0 to 1 which 0 is very objective and 1 is very subjective. Polarity ranges from -1 and 1 where -1 means negative statement and 1 means positive statement. People’s options about covid vaccine can be evaluated by evaluating the polarities. Figure 6 shows the code for twitter sediment analysis. Figure 7 shows the results of sediment analysis. The result shows that 44.34% is positive about the covid vaccine, while 37.23% is neutral about the covid vaccine and 18.43% is negative about the covid vaccine.

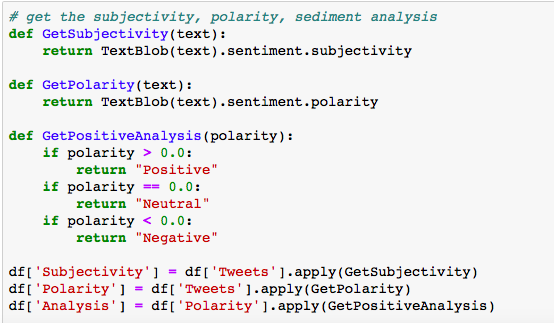


Figure 6: code for twitter sediment analysis

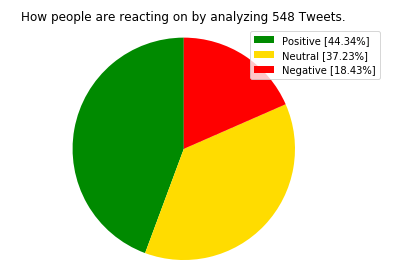


Figure 7: Pie chat of the sediment result

The subjectivity vs. polarity is also analyzed and the results are shown in Figure 8. The results shows that there is more people with objective options about covid vaccine than people with negative options. It also shows that if the tweet is more objective, the option is more neutral. However, if tweet is more subjective, there is more possibilities to see negative or positive options. In addition, the percentage of tweets whose options are strongly negative (<-0.7) or positive (>0.7) is relative low.

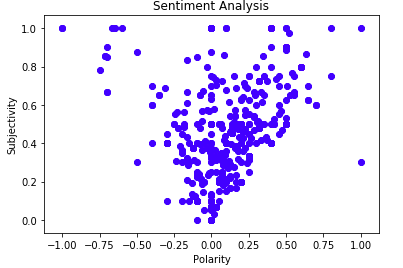


Figure 8: subjectivity vs. polarity

A word cloud chart is also generated to help visualize the text data and importance of each tag.

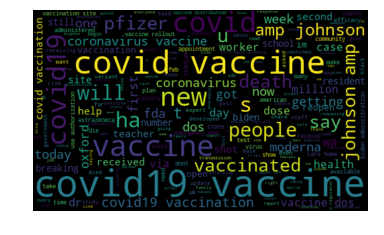


Figure 9: word cloud chart of covid vaccine related tweets

**Applications and clients**

1. It would help government/companies to decide how many vaccines are needed
2. According to WHO, Community immunity requires at least 80% of the population to be vaccinated. As a result, knowing percentage of people who welcome, oppose, be skeptical is important.
3. It would help the government to decide whether they need to put more effort to persuade people to get vaccination.

**Deliverables**

Code, a final report