Discovering Popular Covid-19 Myths Using Unsupervised Learning with Natural Language Processing

Aiman Chughtai

Abstract

The goal of this project was to use Natural Language Processing algorithms on Reddit comments and posts under the keyword search “Covid-19 Vaccine Myths”, in order to discover popular myths and conspiracies around the vaccine that may be circulating around reddit threads. Ultimately, my motivation for this project was to help NYC Health with their Covid-19 Ad Campaign which aims to inform New Yorkers about the effectiveness, safety and necessity of getting the vaccine. Therefore, this analysis not only provides valuable information for NYC Health to use for their next ad, but it might also help persuade those on the fence about the vaccine to finally get vaccinated. The dataset I used for this project was scraped using Reddits Praw API.

Data

The dataset contains 2834 rows of documents which are a combination of post titles, post bodies, and comments. I also scrapped features such as post upvote ratio, timestamp, comment upvotes and comment downvotes, but I didnt end up using that information for my final analysis.

Tools and Algorithms

- Python
- Pandas
- Numpy
- plotly
- NLTK
- Spacy
- TextBlob
- Matplotlib

Results/Design: **Modeling and evaluation**

For this project, I was interested in topic modelling in order to get a sense of what popular myths were circulating in reddit threads between the timespan of 2/2021 to 7/2021. I originally wanted to get a more broad timespan but because vaccines were primarily distributed around that same time, it appears that's when posts about the vaccine started popping up on reddit. I first began my analysis by doing some EDA in order to delve into what information post_score, post_timestamp, post_upvote ratio and comment ups/downs may give me. I found that Without knowing the specifics about the documents, none of these features were helpful in determining the overall sentiment around the vaccine. Next comes Topic modelling. Before topic modelling, I preprocessed my text by making everything lowercase, removing punctuation, removing emojis and profane language. Next I tokenized my corpus using WordTokenizer to break sentences up into words. After that I lemmatized my tokens using WordNetLemmatizer. Finally I added stop words which would be common in my corpus, such as vaccine, vaccination, corona, covid, as well as other stopwords I iteratively added after each run of vectorizing. Finally it was time for topic modelling. I started with NMF using CountVectorizer thinking it would give me a good baseline with decently picked topic words. I tuned parameters such as min_df and max_df, as well as n_grams. I then ran through the same steps on NMF using tfidf vectorizer and then LSA with both count vectorizer and tfidf vectorizer. I found that out of all of my iterations, NMF with CountVectorizor after having tuned min_df and max_df, left me with the most coherent set of topics that were both relevant to my use case but also showed a good amount of variability. Once I finished topic modeling and discovered that two popular topics for NYC Health to focus on was 1) skepticism about the government and their prioritization of economy over the health of their people and 2)covid vaccine adverse side effect, I moved on to sentiment analysis in order to see if the sentiments aligned with the topics. I found that in general, most redditors were at worst, curious about the vaccine and at best, encouraging others to get vaccinated. There were more positive and neutral sentiments around the vaccine than negative ones. For my recommendation to NYC Health, I suggest they focus on debunking the idea that the government doesn't prioritize its citizens' health and talk about covid vaccine side effects.

Communication

A presentation/ slides,