MS Project: COVID-19 Vaccination Social Media Trends

Kadie Clancy Advisor: Dr. Song Jae Hwang

October 15, 2021

1 Motivation

At the inception of this project in June 2021, the United States was experiencing an odd phenomenon; although the race to find a vaccine to address the deadly COVID-19 pandemic had been rapid and successful despite all odds, a significant percentage of the vaccination-eligible population was (and still is) hesitant to receive it [Bru21]. As noted by many experts, high vaccination adoption rates are critical for herd immunity and a return to some semblance of pre-pandemic life [Org, Bru21].

The COVID-19 pandemic is an unprecedented and ever-changing situation [ajm]. Public opinion, news events, and vaccine developments were rapidly evolving from the start of the pandemic. Factors such as these were planting seeds of doubt in the COVID-19 vaccine before clinical trials were even underway. For this reason, we look to social media to gauge real-time public perception and investigate trends related to COVID-19 vaccine adoption. We propose to use social media posts concerning the COVID-19 vaccine to identify common factors or trends relating to hesitancy or refusal of vaccine adoption. We will hereby refer to these factors as social trends. The identification of trends from a social media source in real-time would allow public officials to address issues related to not only initial COVID-19 vaccination adoption, but also similar situations that will continue to arise as we battle this pandemic (i.e. booster shot distribution).

2 Methods

2.1 Webscraper

Top Tweets containing the word "vaccine" were scraped at daily intervals for the full period of time from the official talk of vaccine (May 21, 2020) to the project start date (June 9, 2021) using the Scweet Github repository [Jed21]. On May 21, 2020, the United States and AstraZeneca formed a vaccine deal [BFH20, ajm]. Top Tweets are the most relevant Tweets for a given search, and according to Twitter "relevance" is determined by a number of popularity factors. Additional parameters used to scrape Top Tweets are as follows: location = anywhere, and language = English. Due to the nature of the COVID-19 pandemic, we can be sure that a vast majority of Top Tweets containing the word "vaccine" are referring to the COVID-19 vaccine. In total, the Twitter Dataset consists of

18,799 Tweets. For each tweet, the following information was collected: UserScreenName, UserName, Timestamp, Text, Embedded text, Emojis, Comments, Likes, Retweets, Image link, and Tweet URL.

2.2 Data Cleaning

In order to use the Tweet text from the set of scraped Tweets to investigate social trends, the data was first appropriately preprocessed. As many Tweets in the Twitter Dataset are in reference to the Embedded Text (or RT), both Tweet Text and Embedded Text were utilized. Using standard Natural Language Processing (NLP) techniques, the text and embedded text from each tweet was cleaned, stemmed, and tokenized for later use.

2.3 Topic Formulation

We used topic modeling on our corpus of Tweets to identify the underlying topics present in the Twitter Dataset. Specifically, we employed Latent Dirichlet Allocation (LDA), an unsupervised approach to extract topics that best describe a set of documents [BNJ03]. LDA is a generative probabilistic model for text corpora where the latent layers represent the topics and topics are composed of word tokens. In the context of this project, we considered each Tweet a document and used LDA to uncover the underlying topics present. As these topics come from social media posts, we hypothesized that the topics formed could concern social trends allowing us to make inferences about relating to vaccine hesitancy.

The LDA topic model was trained to uncover 60 topics. The number of topics was decided upon empirically, as the typical metrics of model goodness for LDA like topic coherence and Jaccard similarity and are not appropriate to determine optimal topic number in this context. The LDA topic model was trained on a chunk size of 350 Tweets for 50 passes.

PyLDAVis was used for visualization of Topics [Mab18]. Topics were plotted in a shared space via metric multidimensional scaling. This form of dimesnionality reduction preserves relationships so we expect closer topics to be more similar and topics that are farther away from one another to be less similar. We specifically noted the 5 closest neighbor topics for each topic.

For each tweet in the Tweet Dataset, the distribution of underlying topics present in the tweet is generated using the LDA topic model. The most prevalent topic in the distribution will be referred to as the Dominant Topic of the tweet. For each of the 60 topics, the subset of Tweets from the Tweet Dataset with Dominant Topic X form a topic's Tweet subset.

2.4 Twitter Dataset

Additional NLP algorithms were used to provide more insight about the set of Topics generated by the LDA model. Using the TweetEval standard, pretrained models were used to generate additional information about each Tweet in the Twitter Dataset [BCCNEA20]. Specifically, each Tweet has sentiment, emotion, irony and offensiveness prediction information included. The pretrained models

for each task can be found here [NLP20]. The sentiment of each Tweet is classified as positive, negative, or neutral [RFN17]. The emotion label for each Tweet can be anger, joy, optimism, or sadness [MBMSK18]. Each Tweet can be labeled as either ironic or not ironic [VHLH18]. Each Tweet can be labeled offensive or not-offensive [ZMN+19].

2.5 Topic Dataset

As each Tweet has a corresponding Dominant Topic, we can explore each topic as a subset of Tweets. From this corpus of Tweets representing each topic, a WordCloud and EmojiCloud were created to aid in further visualization and analysis of each topic. A WordCloud is a visualization of the most frequent words in the topic's Tweet subset where more frequent terms are displayed in a larger font. Using only the emojis from the topic's Tweet subset, an EmojiCloud was similarly generated. This Github repository was used for both visualizations: [Mue20]. To compactly summarize each topic, two topic summaries were generated. The first summary generated was an extractive summary, which used a PageRank-type algorithm [XG04] to find the most representative Tweet from a topic's Tweet subset. An abstractive summary was also generated using BART model [LLG+19], which is trained on the CNN/Daily Mail News Dataset. The BART model uses a standard seq2seq/machine translation architecture with a bidirectional encoder and a left-to-right decoder to produce a summary of text. The summarization of long corpus of text (i.e. long subset of Tweets) is still an active research question, so a simple workaround was employed.

As we generated additional features for each Tweet in the Twitter Dataset (sentiment, emotion, irony, and offensiveness), we can aggregate the results to explore these per topic as well. Additionally, the Twitter Dataset contains the timestamp of each Tweet so we can also view the prevalence of the topic over time. The Topic Dataset consists of the following features: Topic Word Tokens, WordCloud, EmojiCloud, Extractive Summary, Abstractive Summary, Closest Topics, Set of Topic Tweets, Overall Sentiment, Overall Emotion, Overall Irony, and Overall Offensiveness.

2.6 CDC Datasets

To show the social trends in the broader scope of the pandemic, two publicly available CDC datasets were employed. The first follows COVID-19 vaccination rates in the United States [fDCP21a]. The second tracks total cases and deaths from COVID-19 in the United States [fDCP21b].

2.7 Tableau Visualizations

Topics and corresponding Tweet information mentioned above is visualized as Tableau Dashboards for analysis.

3 Discussion

The exploration of the Twitter Dataset by means of latent topics has led to several interesting inferences related to social trends and the Covid-19 vaccine. The first example of a vaccination-related social trend identified by the LDA topic model is the clear emergence of "don't worry about whats in the vaccine if ..." meme as a topic. Topic 30 contains many tweets that pertain to this specific social trend. The main idea behind this trend is that people have taken bigger risks with their bodies and/or health than receive a safe and effective vaccine. It should be noted that this social trend is pro-vaccine.

Another interesting inference deals with political events that were coinciding with the pandemic timeline. Topic 38 suggests lack of trust in (then president) Donald Trump's handling of vaccine development and overall Covid-19 response, most notably an idea that the vaccine was "rushed." Looking at the prevalence timeline for Topic 38, we see that the peak prevalence hovers around September 2020 and proceeds to fade as Trump leaves office.

According to [Fra21], leading causes of vaccine hesitancy/refusal stem from two main incorrect theories, both of which are present in the exploration of topics. The first is the incorrect link between autism and vaccinations in general. This is evident in Topic 27 where tweets containing discussions of autism, vaccinations, and mercury are present. The second is the COVID vaccine-specific source of misinformation that COVID-19 vaccines contain a microchip created by Bill Gates intended to track the American public. Bill Gates and the concept of "microchip" or "chip" is very present throughout the Twitter Dataset and in the exploration of topics. Topic 7 seems to capture this microchip theory and Topic 40 seems to be more focused on Bill Gates, specifically.

Finally, the exploration of topics also uncovered an interesting pattern related to vaccine passports. Topic 16 captures the idea and certain negative feelings towards the idea of a "vaccine passport," or the concept of requiring proof of vaccination for travel or admittance into events. The emotion predictions for Topic 16 are overwhelmingly angry and the sentiment is majority negative. A further inspection of the topic's Tweet subset shows individuals who do not agree with this approach. It is interesting that this trend was prevalent before the vaccine was even released, supporting the idea that the stance of hesitancy or refusal of the vaccine may not have to do with the actual vaccine at all. Rather for some, it seems the choice of receiving a Covid-19 vaccination comes down more to a point of personal freedoms rather than a mistrust in the vaccine itself. While this fact is unfortunate, the emergence of this pattern may better inform public policy to address individuals such as the authors of these Tweets and support measures like vaccination mandates.

In conclusion, we scraped a large dataset of top social media posts from the web to investigate social trends related to Covid-19 vaccine hesitancy/refusal. In the future, this work could be improved by the addition of location information as public opinion of the vaccine has been shown to be related to geographical region [KRRM21]. The Twitter Dataset could also be updated to include real-time Top Tweets containing the word "vaccine" so that social trends may be actively monitored in to address the fluid nature of the Covid-19 pandemic.

References

[ajm] A timeline of covid-19 developments in 2020.

[BCCNEA20] Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. *arXiv preprint arXiv:2010.12421*, 2020.

[BFH20] Aakash B, Guy Faulconbridge, and Kate Holton. U.s. secures 300 million doses of potential astrazeneca covid-19 vaccine, May 2020.

[BNJ03] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[Bru21] Geoff Brumfiel. Vaccine refusal may put herd immunity at risk, researchers warn, Apr 2021.

[fDCP21a] Centers for Disease Control and Prevention. Covid-19 vaccinations in the united states,jurisdiction. `https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-Jurisdi/unsk-b7fc`, 2021.

[fDCP21b] Centers for Disease Control and Prevention. United states covid-19 cases and deaths by state over time. `https://data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-Deaths-by-State-o/9mfq-cb36`, 2021.

[Fra21] Kathy Frankovic. Why won't americans get vaccinated?, Jul 2021.

[Jed21] Yassine Ait Jeddi. Scweet. `https://github.com/Altimis/Scweet`, 2021.

[KRRM21] Wendy C King, Max Rubinstein, Alex Reinhart, and Robin J Mejia. Time trends and factors related to covid-19 vaccine hesitancy from january-may 2021 among us adults: Findings from a large-scale national survey. *medRxiv*, 2021.

[LLG+19] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

[Mab18] Ben Mabey. pyldavis. `https://github.com/bmabey/pyLDAvis`, 2018.

[MBMSK18] Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 1–17, 2018.

[Mue20] Andreas Mueller. wordcloud. `https://github.com/amueller/word_cloud`, 2020.

[NLP20] Cardiff NLP. tweeteval. `https://github.com/cardiffnlp/tweeteval`, 2020.

[Org] World Health Organization. Coronavirus disease (covid-19): Herd immunity, lockdowns and covid-19.

[RFN17] Sara Rosenthal, Noura Farra, and Preslav Nakov. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, pages 502–518, 2017.

[VHLH18] Cynthia Van Hee, Els Lefever, and Véronique Hoste. Semeval-2018 task 3: Irony detection in english tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 39–50, 2018.

[XG04] Wenpu Xing and Ali Ghorbani. Weighted pagerank algorithm. In *Proceedings. Second Annual Conference on Communication Networks and Services Research, 2004.*, pages 305–314. IEEE, 2004.

[ZMN+19] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, 2019.