Research Proposal

Jakob Schlierf: 3147567

May 4, 2022

Title of research: (work in progress):
Comparative Investigation into the Extent, Formulation, and Arguments of confidence in- and against - COVID-19 Vaccines in the US on Reddit

Research Question (work in progress)

1. What are the main topics of opponents & proponents of COVID-19 vaccines on Reddit?
2. Does the behavior of these groups indicate the existence of echo chambers? How big are these groups & echo chambers in the US on Reddit?
3. How did this structure and the topics evolve in concurrence with the development & distribution of the COVID-19 vaccines?

Objectives

- Identify the main topics of opponents & proponents of vaccinations against COVID-19
- Compare the vocabulary and propensity to hesitancy to the vaccines for each group
- Quantify the prevalence of opposing viewpoints for each group and conversely whether these groups exist in echo chambers as well as determining the size of the relevant groups and echo chambers
- Evaluate temporal changes of the group sizes, topics, and vocabulary in relation to important events (ex. Vaccine trial failures, start of vaccinations, waves of Covid-19 variants)

Methods
To achieve these objectives, the following methods will be utilized (methods referring to the objectives in the same order of appearance):

- Natural Language Processing Techniques, collectively known as Topic Modeling (most prominently LDA (Blei, Ng, & Jordan, 2003)), Contextual Topic Modeling (Bianchi, Terragni, Hovy, Nozza, & Fersini, 2021) or LSI. Lastly, utilizing Pathfinder Networks (PFNETs) (Quirin, Cordon, Guerrero-Bote, Vargas-Quesada, & Moya-Anegon, 2008) if possible due to volume, could provide a unique opportunity to identify the differences in topics associated with each group (Amith, 2020)
- For the vocabulary analysis, compare the heuristics of keywords, potentially using other NLP techniques as possible
- Several methods to identify and delineate echo chambers exist. Options include weighting individual opinions then evaluating the overall distribution of leanings (Cinelli, De Francisci Morales, Galeazzi, Quattrociocchi, & Starnini, 2021) or identifying a starting seed (post, user, or subreddit) and its close relationships (Conover, 2011)
- Utilize temporal subsets of collected data with the aforementioned methodologies to observes changes, especially around important events, using sliding windows

Justification for Research Objectives

Along with tools such as mask wearing, social distancing, or increased hygiene sanitization, vaccinations seem to be among the most prolific and arguably sensible measures to combat the spread of - and death from – SARS-Cov-2-induced COVID-19. Vaccination rates in most countries, however, remain lower than expected by experts to eradicate or severely lower the spread of COVID-19 (Pandey, Sah, Moghadas, & Mandal, 2021). The spread of misinformation, primarily through social media sites like Reddit, presents a key challenge to achieving this level of vaccination (Melton, Olusunya, Ammar, & Shaban-Nejad, 2021). The uninterrupted flow of information due to the structure of sites like Reddit is influenced by algorithms, sorting certain users into distinct pathways. In extreme cases, this directing creates echo chambers, only showing users content they agree with, supports their present opinions, or furthers their path into more extreme views (Cinelli, De Francisci Morales, Galeazzi, Quattrociocchi, & Starnini, 2021). Understanding the arguments which justify hesitancy towards vaccines, or which urge vaccination efforts could provide vital information to encourage more people to get vaccinated, which provides an opportunity to save lives and prevent further spread or variants of Covid-19.

Data Acquisition

This project will utilize data from Reddit, focusing on users in the US, in American English. This data procurement will target the approximate two-year period between the beginning of the Covid-19 pandemic in March 2020 to the height of the Omicron variant wave in the US around March of 2022. Data will be acquired from the pushshift API and from Reddit's API through the following three steps:

1. Acquisition of available via the pushshift API through August 2021
2. Identification of the relevant subreddits for both groups (pro- & anti-vaccine)
3. Utilizing the Reddit API to retrieve the remaining data for the period