CSC110 Project Proposal: A Study of Online Conspiracy theories regarding COVID-19

Aaron Ma, Benjamin Liu, Vishnu Nittoor

Friday, November 5, 2021

Problem Description and Research Question

Research Question What is the nature of the circulation and development of conspiracy theories surrounding COVID-19?

Problem Description Since the advent of the COVID-19, the world has witnessed the circulation and development of conspiracies surrounding the origin of the virus, the efficacy and nature of the new vaccines, and the political motivations behind COVID-related policy decisions. Although these conspiracy theories may be seemingly innocuous, they originate significant public unrest surrounding the vaccine, masking policies, the origins of COVID, and more.

We observe that the anti-mask movement, the anti-vaccine movement, and nationalistic narratives that suggest that the virus was engineered are key products of such conspiracy theories. Examples of certain ideas originated by these conspiracy theories involve the perception of COVID-19 as an attempt at biological warfare, that the vaccine causes and helps proliferate diseases such as autism, and that the pandemic was part of an elaborate business strategy for pharmaceutical firms to generate abnormal amounts of revenue.

Several accounts such as **PeopleWh91:online** report that conspiracy theories severely impact the well-being of those that believe and spread them. Further, they pose emotional and physical risks for others around them - **10.2307/27040260** suggests that ideologically diverse extremist groups who are especially reactive to these conspiracy theories, pose the threat of violent terrorism in the public domain. On a global scale, these theories have escalated political tensions, promoted anti-masking and anti-vaccination, and have acted against the dissemination of reliable information about the pandemic and the preventative measures that are relevant, according to **COVID19c99:online**.

We are motivated by these problems to investigate the bidirectional relationship between COVID-19 case data and conspiracies surrounding the pandemic. Specifically, we would like to examine when these conspiracy theories gained prominence relative to the spread of the virus, how they are related in content to the pandemic, and how emotionally charged they are, and the popularity of certain topics over time.

Dataset Description

We are going to be investigating conspiracy theories surrounding COVID-19 using post data from the subreddit **2conspir6:online**, a page on the popular message board/forum website Reddit. Being an incredibly active page with 1.6 million members, each post contains a title, a score representing net "upvotes" versus "downvotes", comments, and a text body. We will only consider text-based posts for this project. We will be extracting data from the subreddit using a Python script seen on the website OSRSBox **OSRSBoxB42:online**.

The data output of the script is a JSON file containing data that corresponds to a variety of attributes. The API used by the script is PushShift Reddit (API **pushshif39:online**), which provides a large array of attributes for each post. Out of the several, we are only considering each post's title, number of comments, score, post body, and the created time in seconds since the Unix epoch (detailed in **Epoch0:online**).

An example (filtered) entry in the dataset produced by this script is as follows with the below attributes:

- `title (str): "Covid-19 Conspiracy, found this and I would like to know what people think about it."`
- `score (int): 1`

- `num_comments (int): 0`
- `body (str): Am I the only one wondering if the Coronavirus in China is biological launch from the US due to the trade war?`
- `created_utc (str): 1583906293`

Computational Plan

Our computational plan consists of the following this list of steps.

- **Data Collection, Aggregation, and Filtering**

  First, we will run the aforementioned script to scrape subreddit data. We will then filter this data using the JSON library in Python (**jsonJSO77:online**) to only contain the attributes we are going to use (mentioned above). We will save this data in a CSV file for later use.

  For our analysis to be well-informed, we are also going to be working with worldwide case data for the Coronavirus pandemic in terms of infections. We plan to get the latter from the Our World in Data COVID-19 dataset as shown in **covid19d6:online** in a CSV format.

  We will filter the CSV file on the COVID-19 data repository to only load rows that that correspond to worldwide case data in terms of infections, process the datetime string to Unix epoch time, and use 7-day smoothed values for the new confirmed cases.

- **Computation of various metrics and variables**

  1. Relation of the post content to COVID-19 and topic categorization

     To determine whether a post is related to COVID-19 or not, we will analyse the post title and post body for mentions of keywords relating to the pandemic, a few of which include `COVID-19`, `mask`, `vaccine`, `China`, `Wuhan`, `biological warfare`, etc. Posts are simply categorized into topics by the class of keywords that they contain.
  2. Sentiment analysis

     We will use the VADER sentiment analysis model to characterize the emotional polarity of each post. Using the `vaderSentiment` Python library (**cjhuttov0:online**) , we will use the `polarity_scores` method of the `SentimentIntensityAnalyzer` class to obtain a compound polarity value for each concatenated post title and post body.
  3. Popularity of each post

     We will simply add the score of each post to the number of comments it has to obtain a numerical value representing its ‘popularity’.

- **Reporting Results**

  We will report results as follows, aiming to answer each of the below questions with a relevant figure that contains information on the variables in question.

  1. **Does the popularity of conspiracy theories depend on how negatively charged they are?**

     We will plot post popularity against post intensity and use a polynomial regression model from the `numpy` library (using the function `numpy.polyfit` to calculate coefficients of the model and `numpy.polyval` to sample them for the figure).
  2. **To what extent are COVID-19 conspiracy theories dependent on the proliferation of the virus?**

     We will plot both post frequency and number of confirmed cases against time on parallel vertical axes. We will compute the Pearson correlation coefficient using `numpy.corrcoef`.
  3. **What recurring topics of discussion develop over time amongst the conspiracy theories?**

     We will plot multiple curves, each representing a topic class, against time. This, with a graph in parallel representing COVID-19 cases, would help determine the temporal relationship between key events in the pandemic and concurrent topics of discussion.