Name & Semester: Koushik Reddy Parukola (Spring 23)

Title: Topics Modeling and sentiment analysis on tweets related to COVID-19 Vaccine in India (Wordcount: 465)

Proposed questions:

1. What were the prominent topics on Twitter in India when the COVID-19 vaccine program initially began?

2. What impression did the COVID immunizations have on people in India at first (positive or negative)?

Background: On 16th of January, 2021, India started giving out COVID-19 shots. India had provided nearly 2.2 billion doses of all current vaccines licensed by the government as of 4 March 2023, including the first, second, and booster doses. The Press Information Bureau reports that two vaccines, Covishield (produced by Serum Institute of India Ltd.) and Covaxin (developed by Bharath Biotech Ltd.), have been used to start the COVID-19 vaccination program in India [1].

Data Collection: We use SNScrape to gather a minimum of 2000 tweets in total that are related to the subject we are discussing using hashtags like: #covaxine, #covisheild, #covidvaccineindia, etc. Since we are talking about tweets from the initial stage of vaccinating, we gather tweets from January 6, 2021 (the day the first vaccination was issued in India), until December 31, 2021. In order to look at how things are gradually evolving, we also retrieve data for each month separately. All our data sets will contain URL, date, rendered content, rely count, retweet count, like count, view count and user name. Finally, we examine the rendered content column for repeating tweets and those with only images and videos which will be removed.

Method: We use LDA (Latent Dirichlet Analysis) which is a topic modeling algorithm also an unsupervised algorithm to investigate some of the most important topics in monthly data sets and the one for whole time period. And we use VADER for getting the positive and negative polarities in the tweets for sentiment analysis. But before we feed the data to any of our models, we need to do some preprocessing. We will tokenize the text, change the capitalization to lower case, remove unnecessary information such as emoticons, URLs, and user names, etc. (this will only be done on the rendered content column in our data sets, which is the actual data for the tweet). We also calculate the coherence score to know the quality of the topics learned through the model along the accuracy score for the sentiment analysis to see the effectiveness of the analysis. The data we collected from twitter and the data we got from out LDA and VADER models are then used together to examine the topics that were discussed about the most and the feelings on vaccine effectiveness in people (positive or negative) over the course of the time period.

References

[1] Sanjeet Bagcchi. The world's largest COVID-19 vaccination campaign NewsDesk| Volume 21, Issue 3, p323, March 2021. DOI: https://doi.org/10.1016/S1473-3099(21)00081-5