QSS 20 Final Project Milestone 1

Group: Seba Maccias; Emma Nguyen

Last updated: May 17, 2024

1 Project Option

- COVID-19 Twitter: 0.1% sample

2 Your project's main questions

We aim to understand the correlation between the timing of Twitter account creation and the level of engagement regarding COVID-19 in the platform. This could help us determine if certain world events, such as the COVID-19 pandemic, influenced the creation of new accounts or if accounts created during specific periods are more active in discussing COVID-19.

3 Your project's data sources/relevant fields

We would be using the *acc_age* field, which represents the age of the accounts in days and the verified status to measure credibility which potentially affects the account's engagement and reach. Additionally, by using the accounts' creation date, we can categorize the accounts by year and find in which year/period were created the highest number of accounts. This could help us determine if accounts created during the pandemic were more engaged in COVID-19 related tweets than older accounts or vice versa. Other relevant fields include *date*, which indicates when each tweet was posted, allowing us to temporally analyze tweet volume over time. Finally, we would have engagement metrics like *rt_rt_count* (retweet count), *rt_reply_count* (reply count), and *rt_fav_count* (like count), which will help us measure the influence and reach of the tweets from these accounts. By looking at these data fields, we can understand how Twitter accounts created at different times contribute to the discourse on COVID-19, exploring whether significant world events like the pandemic have increased account creations or if these newly created accounts engage differently with the topic.

4 Learn from an earlier project on your research topic

Measuring Improvement in Medical Students' Understanding of Intellectual and Developmental Disabilities by Gitlin et al assesses educational outcomes with the application of

natural language processing (NLP) techniques to medical student training about intellectual and developmental disabilities (IDD). This approach quantifies the effectiveness of training modules, and pinpoints areas for potential enhancement; the study found significant improvements in the students' understanding of how to care for patients with IDD. The use of NLP allows for analysis of changes in knowledge and understanding before and after the training, providing a reliable method to evaluate and improve educational content based on quantitative data. However, a significant weakness in the study's approach is the small sample size in the post-training assessment (49 observations pre-training to 16 post-training), which may not provide a comprehensive or trustworthy view of the training's effectiveness. This limitation could lead to biased results that do not accurately reflect the experiences of a larger group of medical students. Future research should aim to increase participation rates throughout the training process, perhaps by integrating incentives or simplifying the commitment required to complete post-training evaluations. Additionally, ensuring a larger sample size could help reduce the influence of outliers and provide more reliable data.

5 Your project's anticipated challenges

Our primary concern is accurately measuring the influence of account creation times on COVID-19 discourse patterns, especially in determining causality between the timing of account creations and their engagement with COVID-19 on Twitter. We could encounter issues with missing data, particularly in user metadata such as account creation dates or the deletion of accounts, which can skew analysis of long-term trends. Another challenge is the potential confounding factors that could affect the relationship between account age and engagement, such as algorithm changes, varying user engagement levels over time, or external events influencing user behavior.