

Is Twitter a Credible Source for **COVID-19** Information?

Anisha BharathSingh
MSCA 31013 4
Big Data Platforms
Final Project



Executive Summary

Is Twitter a credible source for information on COVID's spread and risks?

The Facts:

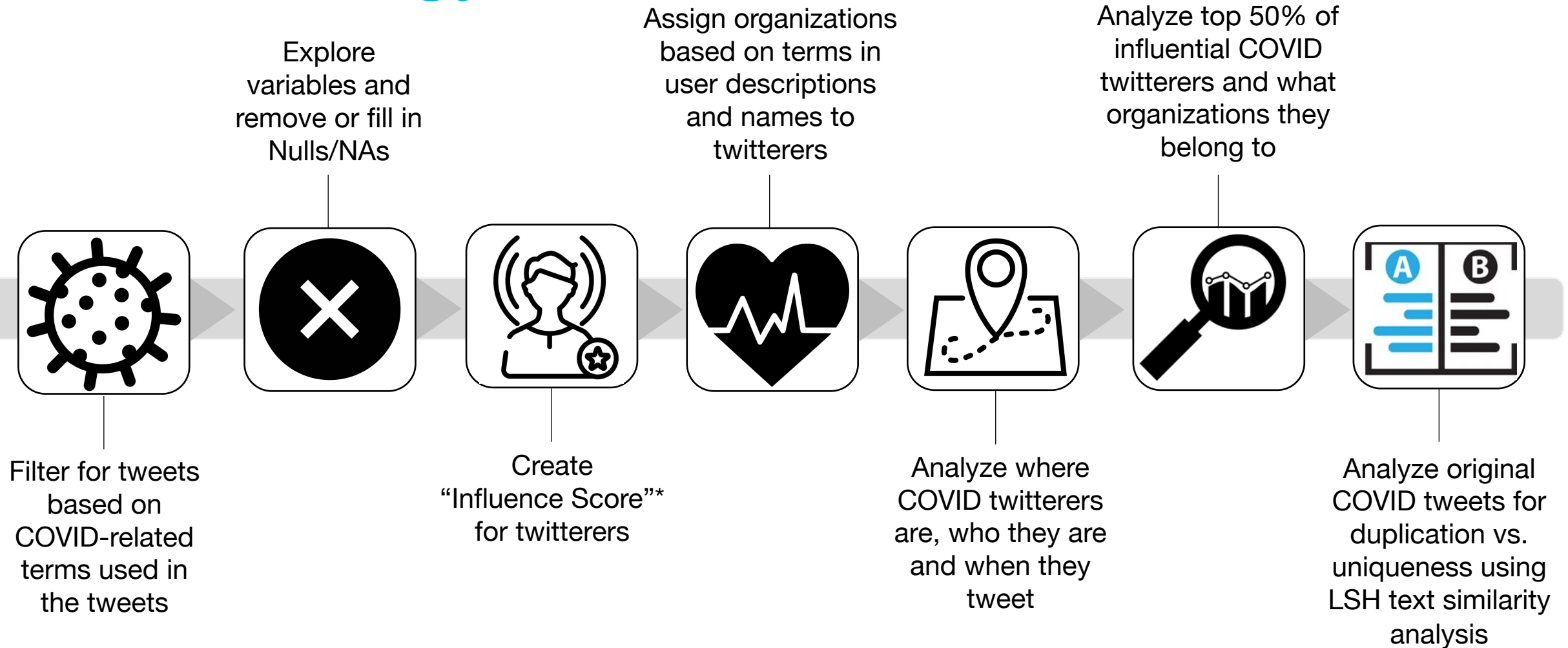
- The majority of COVID twitterers and Top COVID twitterers are **non-verified** and **do not belong to a credible organization**, so all COVID tweets should not be considered a reliable source of COVID information.
- **Influence Score is better** for determining influential and credible twitterers rather than original tweet count or retweet count.
- Time series analysis of COVID tweets over time show that **tweets do not align with worldwide COVID spread or risks**.
- Similarity analysis of original content showed that many credible organizations shared original content that are similar to other tweets, so **similar tweets should not be a sign of an unreliable source**.

Key Initiatives:

- **De-prioritize tweets from non-verified twitterers** that don't belong to an established organization in the Twitter algorithm, while **prioritizing those of influential and credible twitterers** determined by their influence score and background based on their profile information.
- **Add badges** to tweets by credible COVID twitterers so users know what information to trust.
- **Analyze Trending COVID tweets by country** or location to better understand if there is a correlation between COVID tweets and COVID spread or risks.



Methodology



*Influence Score = ((sum of all retweets/total engagements) * total tweet count) *3 for original tweets only

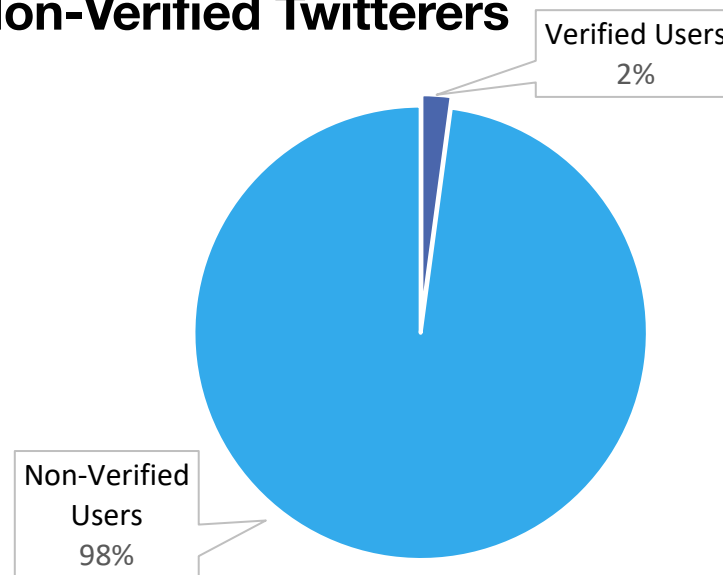
See Appendix for further explanation

Some Background

- **13.6M+ COVID Tweets** extracted from the Twitter API from October 15, 2021 - November 12, 2021.
- **2.7M+ Twitterers** talking about COVID-19 on Twitter.

Organization/ Twitter Type	Number of Twitterers
Other	2,684,856
Social Influencers	16,841
News	16,128
Government	6,081
Health	3,862
Celebrity	1,583

**Percent of Verified vs.
Non-Verified Twitterers**

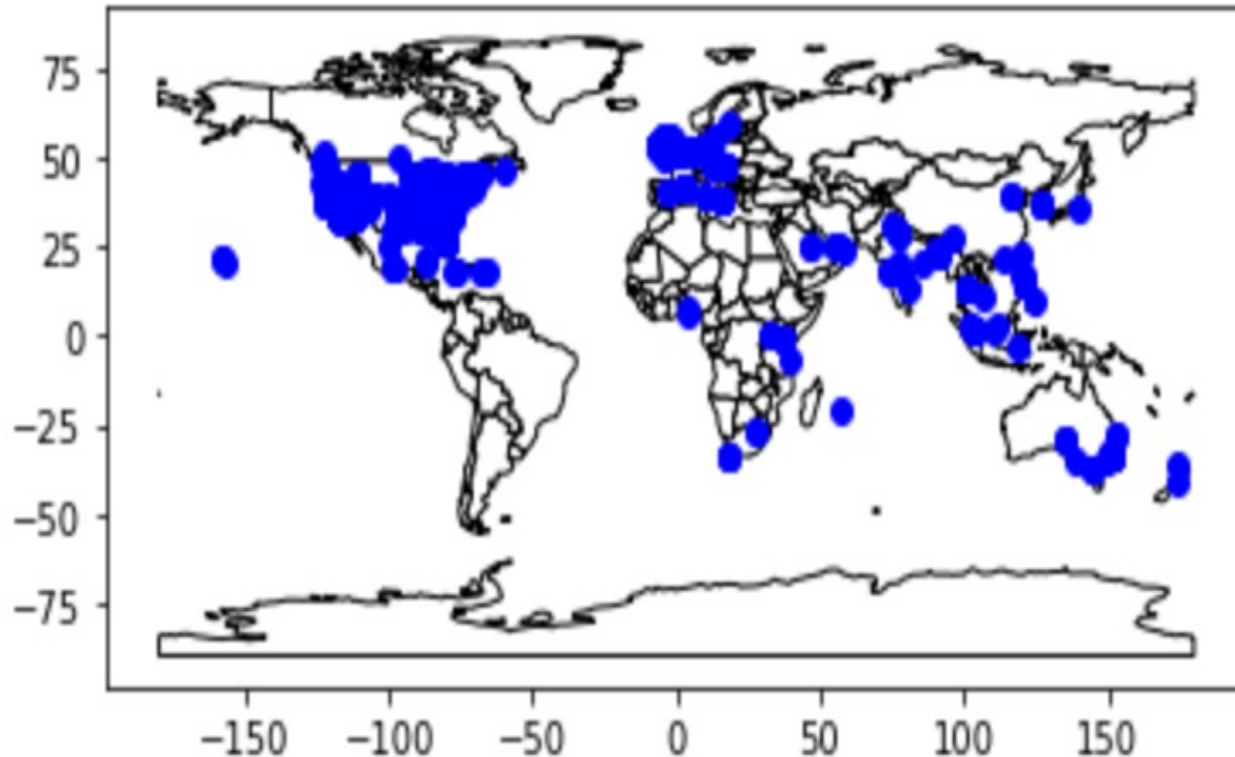


Top 6 Countries by % of Twitterers	% of Twitterers
United States	14.7%
United Kingdom	5.1%
Canada	1.5%
India	1.3%
Australia	0.9%
South Africa	0.5%

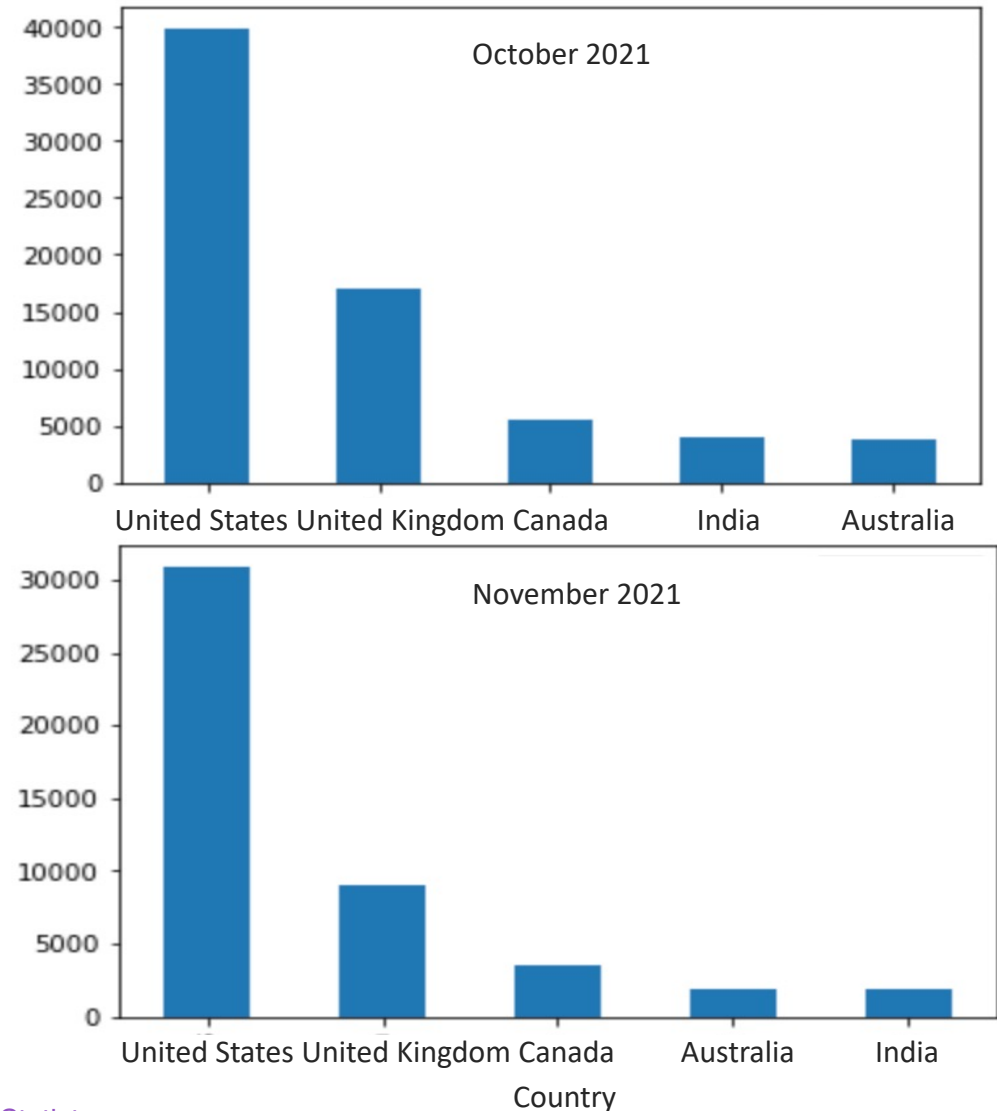
Where are People Tweeting From?

- The majority of COVID tweets were from the U.S., UK and Canada which align with the top countries with the most Twitter users.*
- Comparing October to November, counts were slightly lower** in November and Australia took over India with more tweets.
 - Australia had a spike in mid-October which could explain the increase in tweets, while India cases have been on a downward trend.***

COVID Tweets by Location



Tweet Counts by Top Countries in October vs. November 2021 (See Appendix)



*[Statista](#)

**There were fewer days of data for November which could explain the lower count

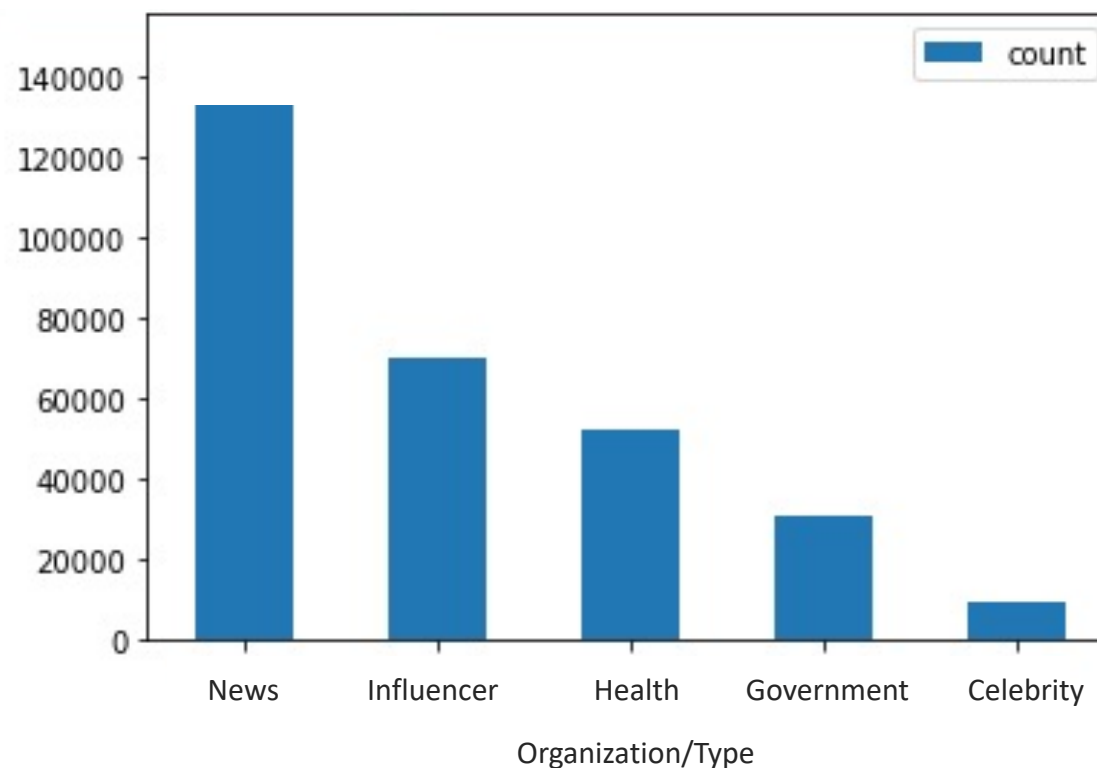
***Country COVID cases information taken from [Our World In Data](#)

Top Influential Twitterers* By Original Tweet Count

- The top influential twitterers by tweet count are all non-verified and belong to the “Other” category.
- The “Other” category has the most tweets, but by organization/type -> “News” has the most COVID tweets.

Twitterers	Tweet Count	Avg. Influence Score
Hiremaid.com.sg	6,739	2391.5
Workout Solutions	3,268	311.7
Vonvictor Valentino Rosenchild	3,107	994.6
Hubbucket	2,530	841.8
Tomthunkit	1,942	5.3
Jeff Lee	1,723	159.5
Progressive Truth Seekers	1,684	11.8
Evelyn Corr Spring	1,609	598.1
Peter K Burian	1,489	8.9
Jaguarbeesoz	1,482	444.3

Tweet Counts by Organization/Type

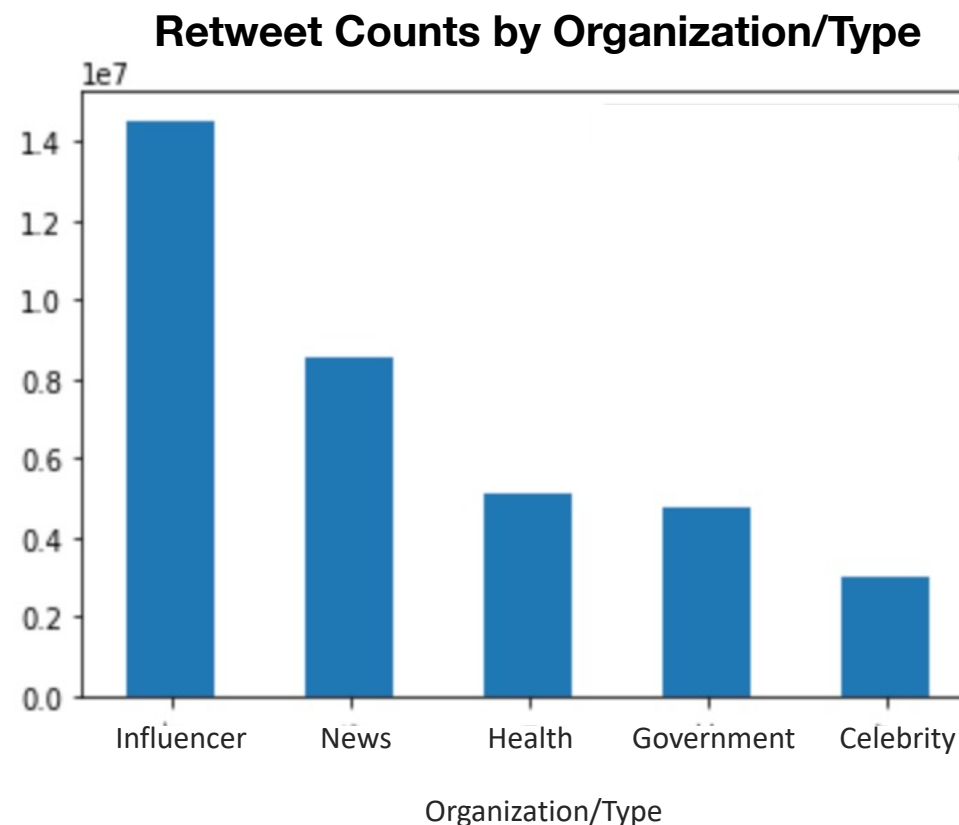


Note: dropped the “Other” Category from chart for visibility of the other organizations
*Analysis done on Top 50% of twitterers based on average influence score

Top Influential Twitterers* By Retweet Count


- The top influential twitterers by retweet count are all non-verified and belong to the “Other” category.
- The “Other” category has the most retweets, but by organization/type -> “Influencers” have the most COVID retweets.

Twitterers	Retweet Count	Avg. Influence Score
Covidoff Bot	3,112,127	2934.2
Contraterrorismo	2,238,941	84.2
A Devoted Yogi	942,016	178.5
Michael Hung Ming Lin, PHD	907,016	337.1
Juan Roberto Barba Ribera	820,966	59.6
Barn Owl Enthusiast	796,891	1.5
Doris Vasquez	719,178	30.2
Seeking Truth & Justice	650,576	101.6
Barb Tomko	630,696	77.6
Sally Fereday Cfirm	624,864	763.5



Note: dropped the “Other” Category from chart for visibility of the other organizations
*Analysis done on Top 50% of twitterers based on average influence score

Top Twitterers by Influence Score For Each Organization

- Top twitterers for each organization are relevant to their category and COVID.
- All top twitterers by influence score for each organization are Verified. 



News Twitterers	Avg. Influence Score
The Straits Times	366.5
Hindustan Times	137.5
News18	137.1



Health Twitterers	Avg. Influence Score
Liz Szabo	166.3
Lucy Johnston	139.7
Cleavon MD	130.5



Government Twitterers	Avg. Influence Score
Heerak Christian Kim for US Congress	133.6
John Bel Edwards	74.3
Martin Cowen	64.9



Celebrity Twitterers	Avg. Influence Score
Morgan Fairchild	97.8
Daniel Horowitz	56.0
Montgomery Granger	52.3



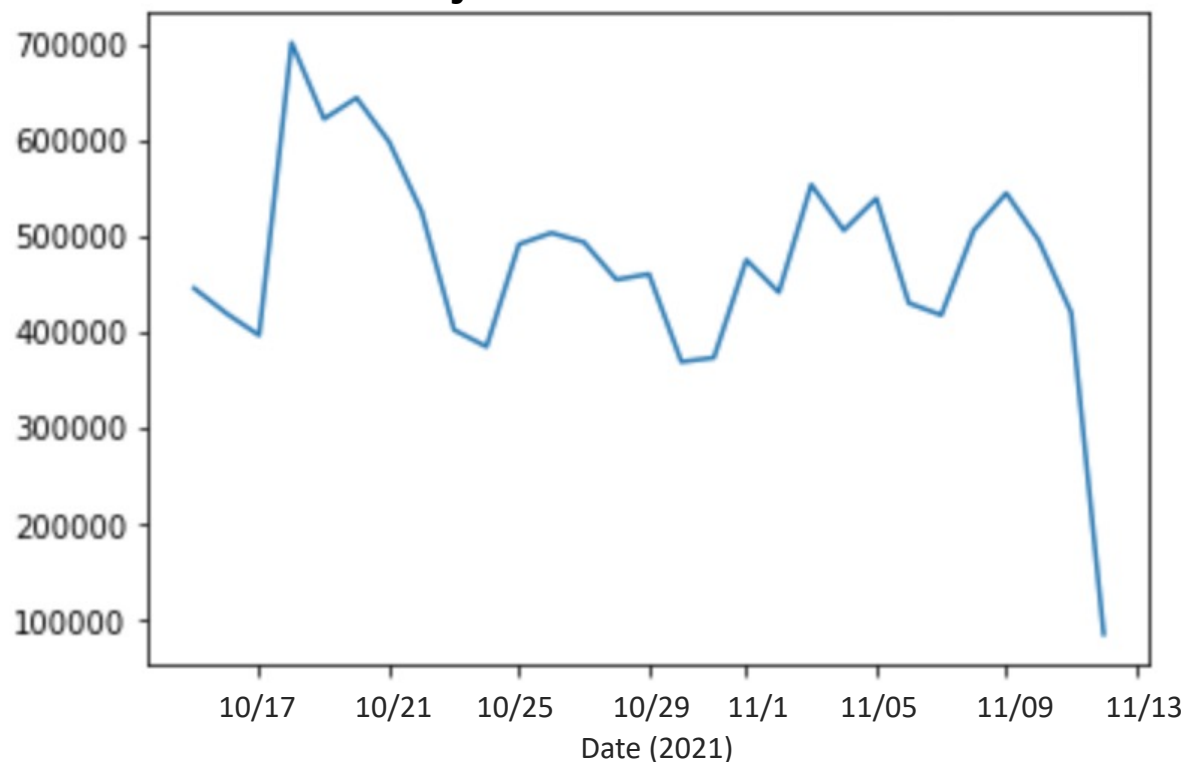
Social Media Influencer Twitterers	Avg. Influence Score
Dorit Reiss	167.3
Alan W. Silberberg	157.9
BFM89.9 Radio	93.9

- Top 2 “News” twitterers aligned with the UK and India being top countries with COVID tweets
- Liz Szabo is a Senior correspondent at Kaiser Health News
- Dorit Reiss is a professor of public health law

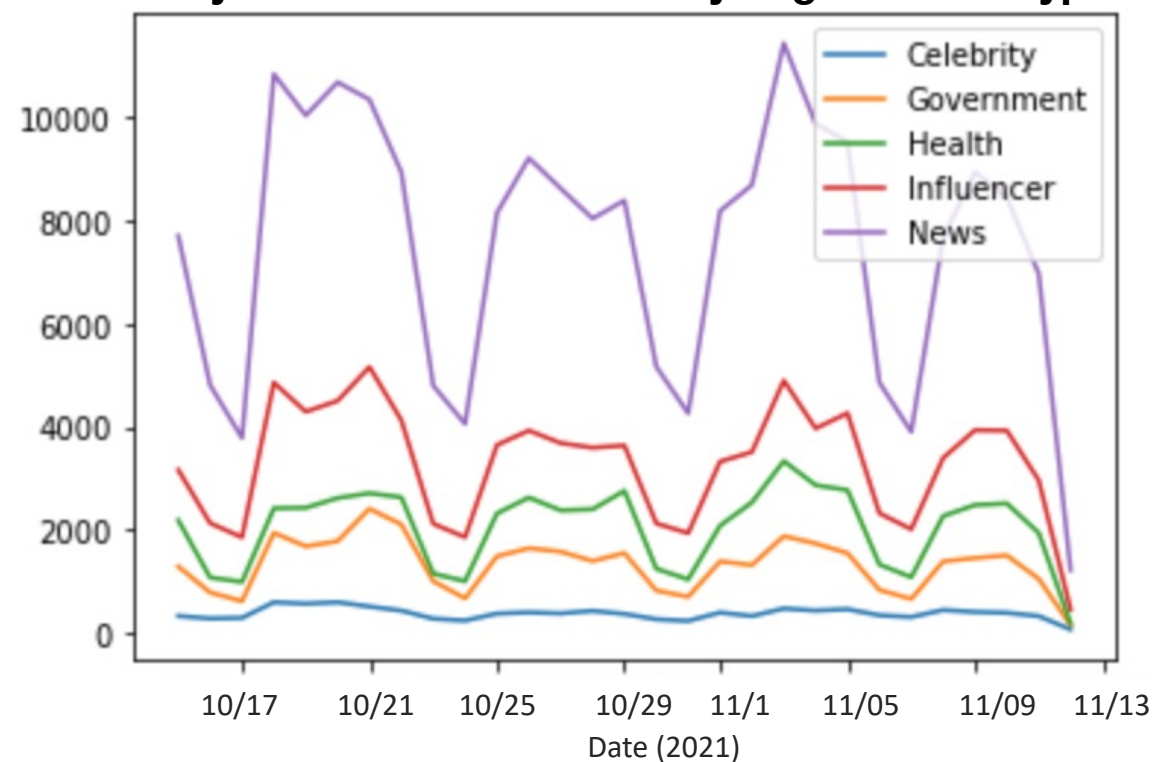
A Look at COVID Tweets Over Time

- Daily COVID tweet counts had a large spike in mid-October and were consistently lower after that.
 - This is not aligned with world COVID cases* which started increasing in mid-October.
- Otherwise, spikes are higher on the weekdays meaning that people are tweeting less about COVID on weekends.
- Organization/type had a similar weekday trend with “News” having the highest counts/spikes aligned with the fact that they report on everything COVID-related.

Daily COVID Tweet Counts



Daily COVID Tweet Counts by Organization/Type

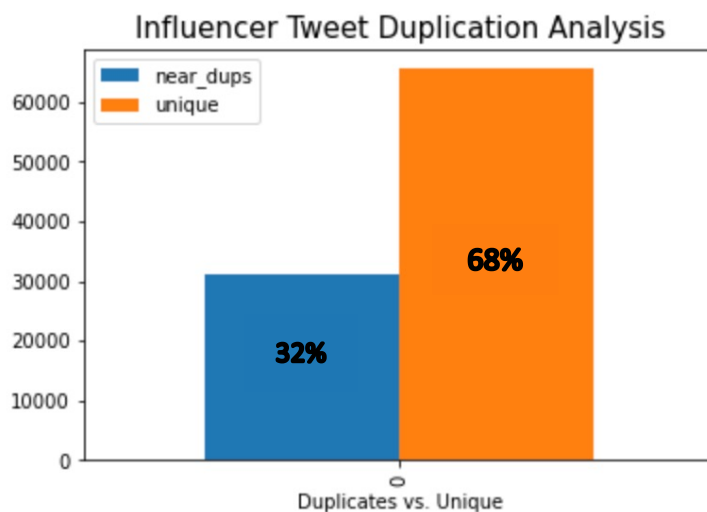
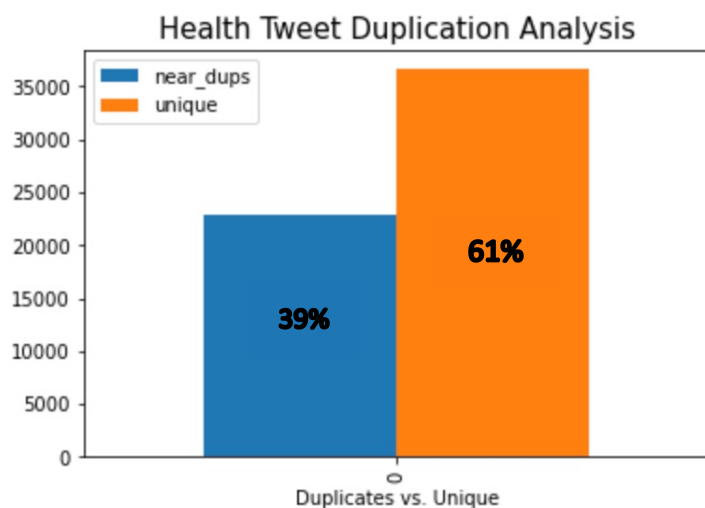
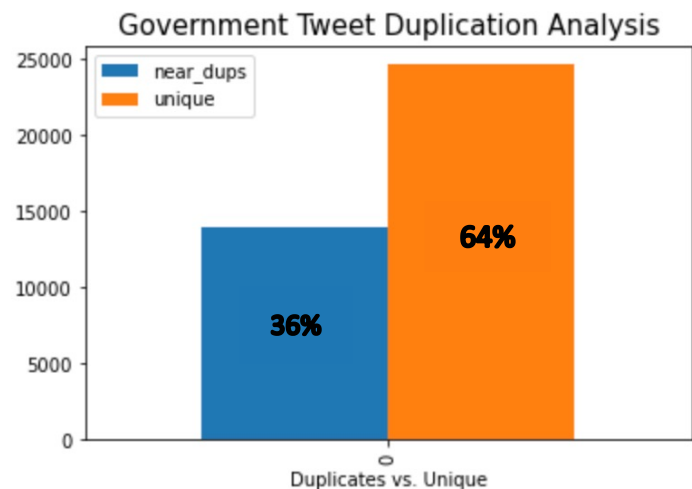
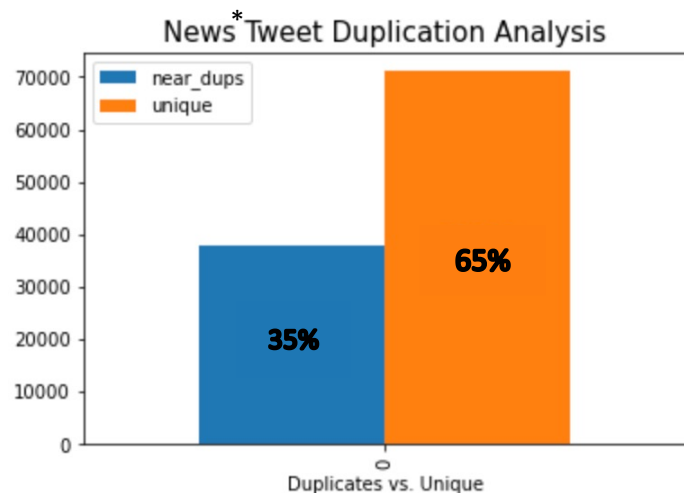


Note: dropped the “Other” Category from chart for visibility of the other organizations

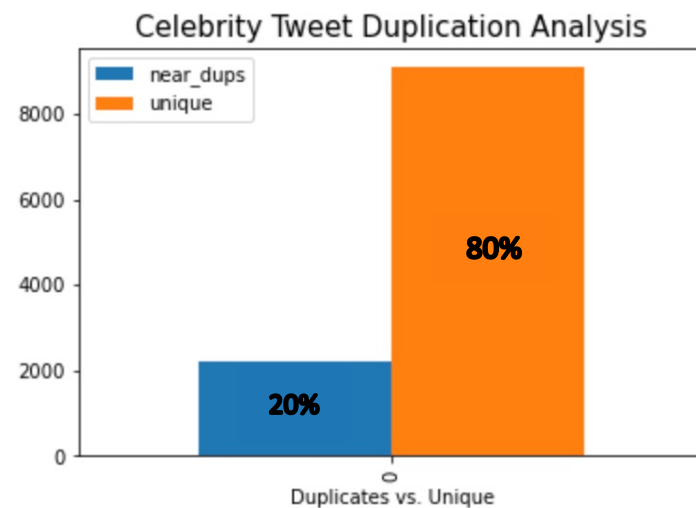
*Country COVID cases information taken from [Our World in Data](https://ourworldindata.org/)

Original COVID Content Similarity Analysis

All organizations/types had a higher percentage of unique tweets which verified that this is original content.



- Health, Government and News had higher percentages of duplicate content which could be a sign that they are sharing similar COVID content.
- Influencers and Celebrities have more original content as they don't have the similar expectation of credible organizations to share all important COVID information.



*Similarity analysis was done on a random sample of 50% of News tweets due to the high data volume

Conclusions and Recommendations

Conclusions:

- The majority of COVID twitterers are non-verified and don't belong to a credible organization. Even top twitterers by original tweet count and retweet count are made up of this group of users.
- By organization, "News" and "Influencers" have a big impact as their content is retweeted the most.
- Using Average Influence Score within each organization is a reliable way to find the most influential and credible twitterers regarding COVID.
- Analyzing the counts of tweets over time does not correlate with world COVID cases, but rather comparing at a more local or country-level might have some indication of spread.
- Similarity analysis shows duplicate content is not an indicator of unreliable COVID information, but rather could be a sign that credible twitterers are sharing similar COVID content.

Recommendations:

- Twitter should de-prioritize non-verified users that don't belong to a credible organization within the Twitter algorithm.
- Twitter should use an influence score combined with whether the user is a credible COVID source based on their profile information and prioritize their COVID tweets within the algorithm.
- Twitter could add badges to COVID tweets from authors identified as credible COVID Twitterers.
- While COVID tweets are not a good indicator of cases worldwide, having COVID trending tweets by country could be something to analyze and a potential indicator at a more local-level.
- Twitter should not use similarity analysis as a means of determining if a COVID tweet is credible since many credible organizations could share similar COVID information.

Appendix

Methodology Details:

- Filtered for COVID tweets based on words like covid, coronavirus, pandemic, booster, pfizer and more.
- Influence Score:
$$\left(\frac{\text{Total Retweets (Retweets+Quotes)}}{\text{Total Engagement (Total Retweets+Favorites)}} \right) * \text{Total Tweet Volume}$$
 - Where tweets were original (retweeted_status = Null), influence score was multiplied by a factor of 3.
 - Influential twitterers (Top 50%) were anyone with an average influence score above the median average score.
- Organizations
 - Looked for users based on names and descriptions with terms related to health, government and news.
 - Used terms related to sports, acting, music and where follower count was above 30,000 for celebrity twitterers.
 - For general social media influencers, used twitterers where follower count was above 10,000.
 - Everyone else and non-verified twitterers went into the “Other” category.
- Location Analysis
 - For all COVID tweets, visualized location of tweets by their coordinates.
 - For COVID tweet counts by top countries, the country variable was used. Every tweet does not have an assigned country, but the top counts are representative of the the top countries with the most twitterers ([Statista](#)) so the trend makes sense, but the counts are lower.
- Analyzed the Top 50% of influential twitterers by original tweet count, retweet count, organization and influence score.
- Time Series Analysis was conducted on all COVID tweets by total count and by organization counts.
- Similarity Analysis was conducted on original COVID content by organization.
 - For ‘News’, the dataset volume was too large to process, so a random 50% sample was taken for analysis.