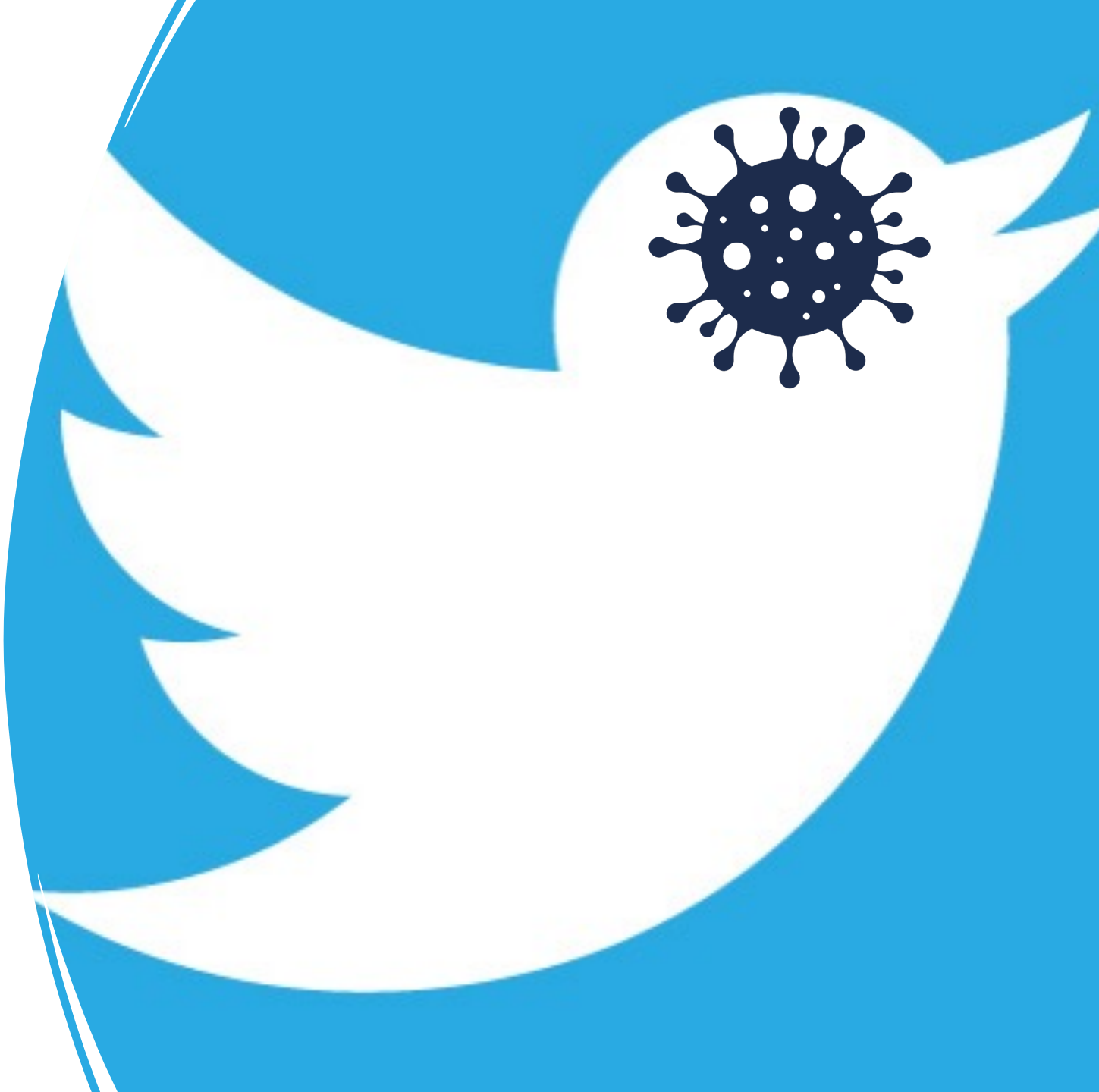
COVID-19 on Twitter

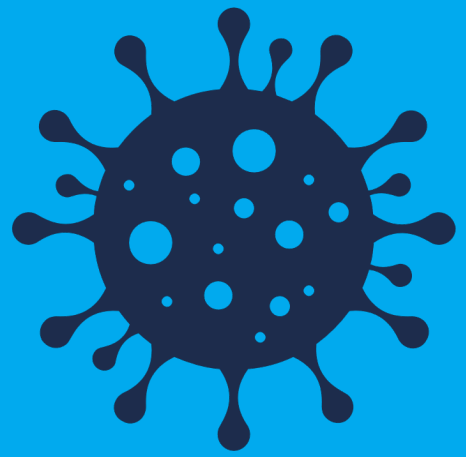
Nuka Gvilia

Big Data Platforms

Final Project



Executive Summary



Overarching Question: Can Twitter be considered a credible source of information, which reflects the risk of contracting the COVID-19 infection?

- Performing location, date, and similarity analyses led to the following conclusions:
  - Worldwide Tweet volume does not follow trends in daily Covid-19 cases
  - However, local Tweet volumes follow trends in daily Covid-19 cases in some countries
  - High Tweet volumes and retweet counts do not make for a credible source of Covid-related information, as most prolific users are non-verified accounts
  - Trends in Tweet volumes over time are the same for users across all organizations and categories
  - News and health organizations are the best source of original content, while most Tweets by non-verified accounts are simply retweets
  - News organization accounts have the highest percentage of duplicate Tweets, while influencers have the highest percentage of unique Tweets
  - The U.S. has the highest Tweet volumes, which coincides with its status as the top country with the most total Covid-19 cases*

* However, this conclusion should be treated with caution, as the U.S. has the most Twitter users in the world and high Tweet volumes could in reality be a result of that factor (source: statista.com)

Data Overview

- 16,847,876 total Tweets
  - 4,799,554 original Tweets
  - 12,048,322 retweets
- Tweet timeline: 10/15/2021 – 11/12/2021

- 3,372,761 total Twitterers
  - Verified: 58,081
  - Non-verified: 3,314,680

  Governmental: 6,934
  News: 25,025
  Healthcare: 4,697
  Influencer: 16,314
  Celebrity: 577
  Other: 3,319,676

Methodology

Filter COVID-19 Tweets

Only select the Tweets that contain COVID-19-related words

(e.g. ‘covid’, ‘quarantine’, ‘vaccine’, ‘isolation’, ‘mask’, ‘distancing’, etc.)

EDA

- **Explore the dataset to select relevant columns for analysis**
- **Handle rows with missing values**

Assign Organizations

Assign organizations and categories to Twitterers based on verification status, words used in their description and follower count

Organizations:
- Governmental
- News
- Healthcare

Categories:
- Celebrity
- Influencer
- Other

Determine Twitterer Influence

- **By Tweet volume**
- **By retweet count**
- **Assign influence score***

*See detailed explanation in the Appendix

Time and Location Analysis

- **Analyze Tweet volume patterns by date in the context of world Covid-19 cases**
- **Analyze Tweet volume patterns by country in the context of world Covid-19 cases**

Similarity Analysis

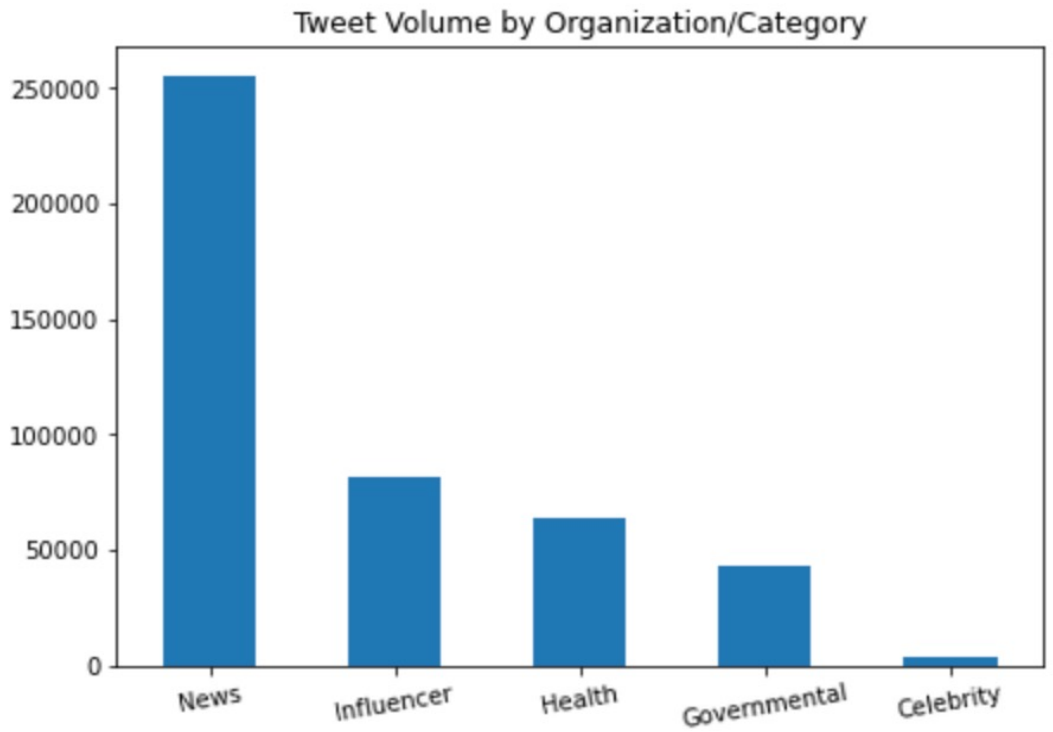
- **Analyze Tweets for proportions of unique vs. near duplicate for each category and organization**
- **Analyze Tweets for proportions of original Tweets vs. retweets for each organization and category**

Tweet Volume

- Top Twitterers with the highest Tweet volumes were non-verified users from the ‘Other’ category

Username	Organization	Verified	Volume
Nathan Joyner	Other	False	15349
iWeller.com	Other	False	4888
Galla Go	Other	False	4681
hiremaid.com.sg	Other	False	4336
Coronavirus Updates	Other	False	3684
Jeremy Hume	Other	False	3568
Paperbirds_Coronavirus	Other	False	2932
TittlePress	Other	False	2643
Andy Vermaut	Other	False	2599
(Isaiah)	Other	False	2355

- Out of the organizations/categories, news had the highest Tweet volume

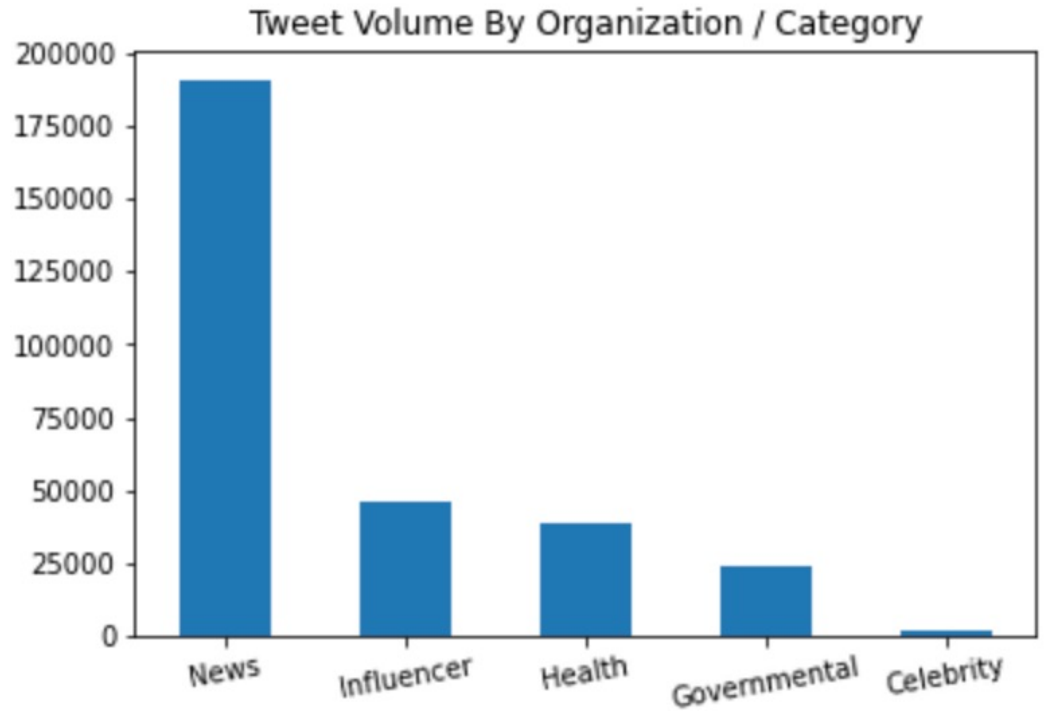


Retweet Count

- Top Twitterers with the highest retweet count were non-verified users from the ‘Other’ category

Username	Organization	Verified	Total Retweet Count
CovidOff Bot 🤖	Other	False	3187923.0
ContraTerrorismo 🇪🇸	Other	False	2208675.0
#StayHome	Other	False	1751666.0
...	Other	False	1034511.0
A Devoted Yogi	Other	False	981106.0
Lesley Gale	Other	False	905673.0
Doris Vazquez	Other	False	903358.0
Michael Hung Ming Lin, PhD	Other	False	881843.0
Barn owl enthusiast	Other	False	796891.0
Juan Roberto Barba Ribera	Other	False	772449.0

- Out of the organizations/categories, news had the highest retweet counts

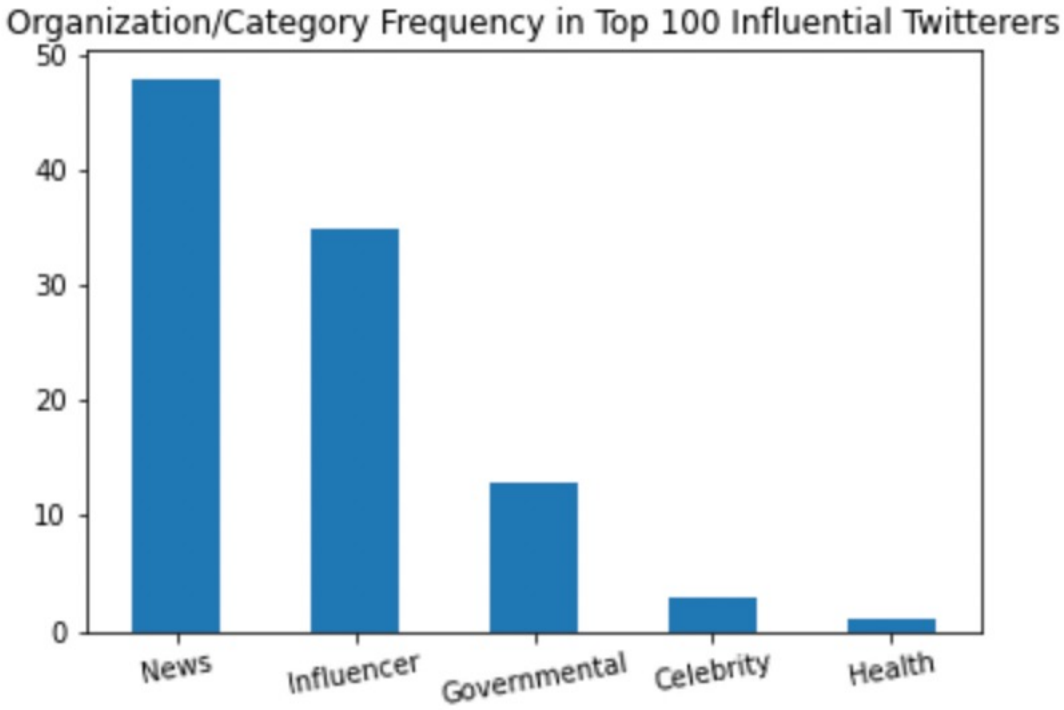


Most Influential Twitterers

- Top Twitterers with the highest influence score were non-verified users from the ‘Other’ category

Username	Organization	Verified
Drake stan account	Other	False
WambuiNjrg	Other	False
Saxon	Other	False
b	Other	False
huberb	Other	False
karmaisabitch	Other	False
BHR	Other	False
Jamet	Other	False
Georgia Zeagler	Other	False
Delphia Sammis	Other	False

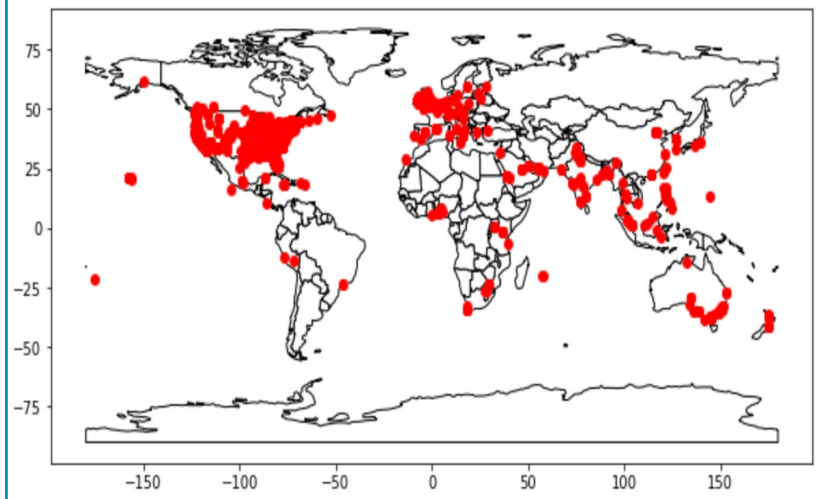
- Among organizations, news and influencer Twitterers appeared most frequently within the top 100 influential users



Location Analysis

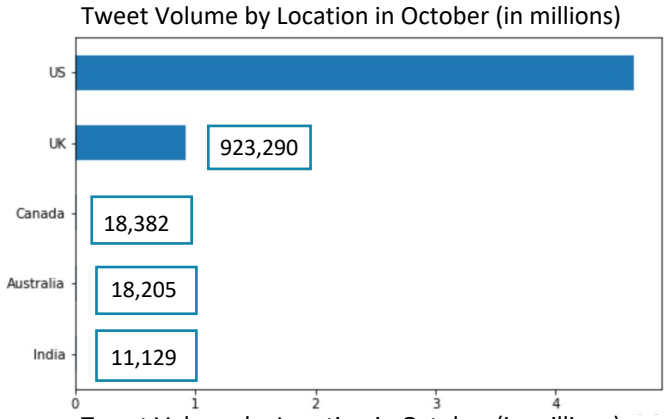
- Top Tweet locations were consistent with COVID-19 statistics*, as the US, UK and India dominate the world with the number of positive cases

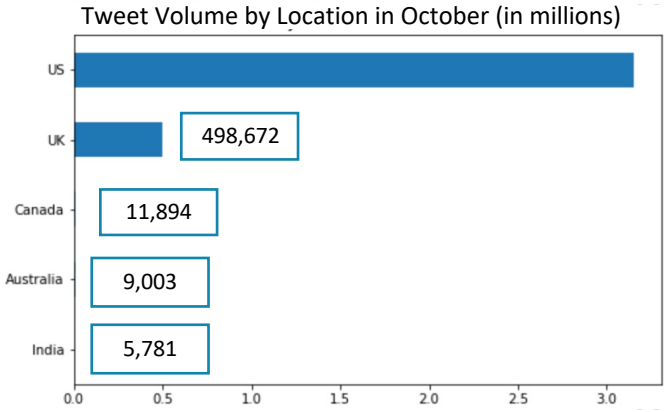
Twitterers on the World Map



The US	83.64%
The UK	15.25%
Canada	0.32%
Australia	0.29%
India	0.18%
Other	0.32%

* Based on WHO Coronavirus (COVID-19) Dashboard





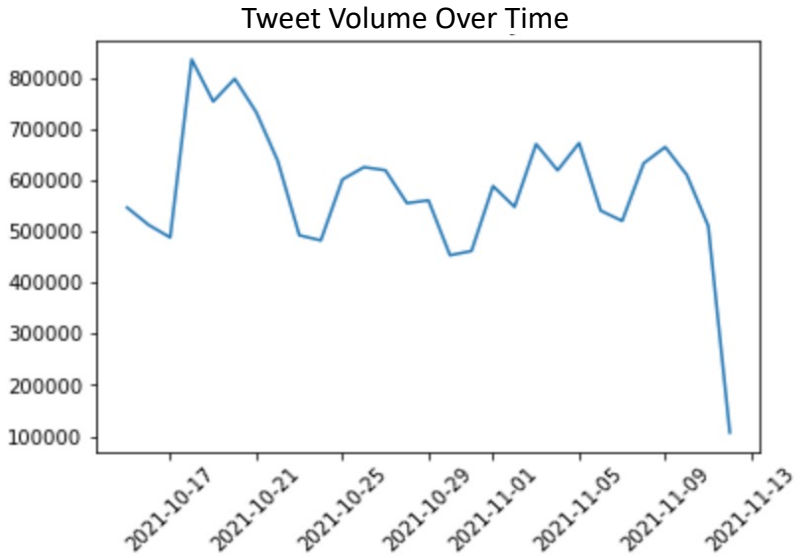
- Locations of top countries with Tweet volumes remained the same over the two months
- The drop in Tweet volume for UK, India, and Australia coincided with the drop in Covid cases in all three countries in the given time period*
- Total world Tweet volume was higher in October than in November, while daily Covid cases increased in November**

* Source, source2, source3

** However, this could be as a result of having data on five more days in October than in November

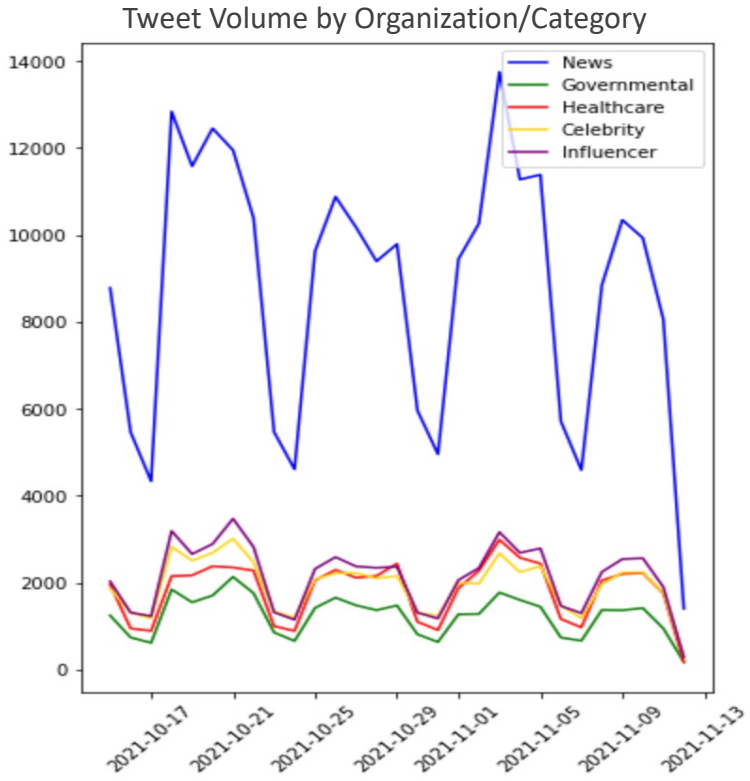
Time Analysis

- Total Tweet volume was the highest before October 21 and declined thereafter
  - This is inconsistent with the world COVID-19 statistics*: the average daily cases were on the decline before October 20th and started to pick up since



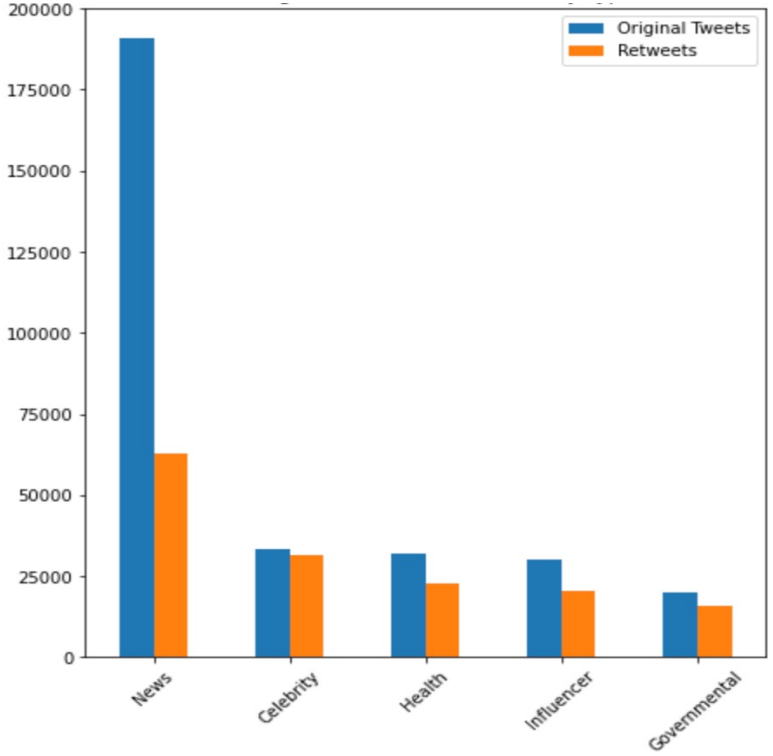
* Data from [Oxford Martin School reasearch]

- Tweet volumes across all organizations followed the same general pattern
- Weekdays had higher Tweet volumes than the weekends

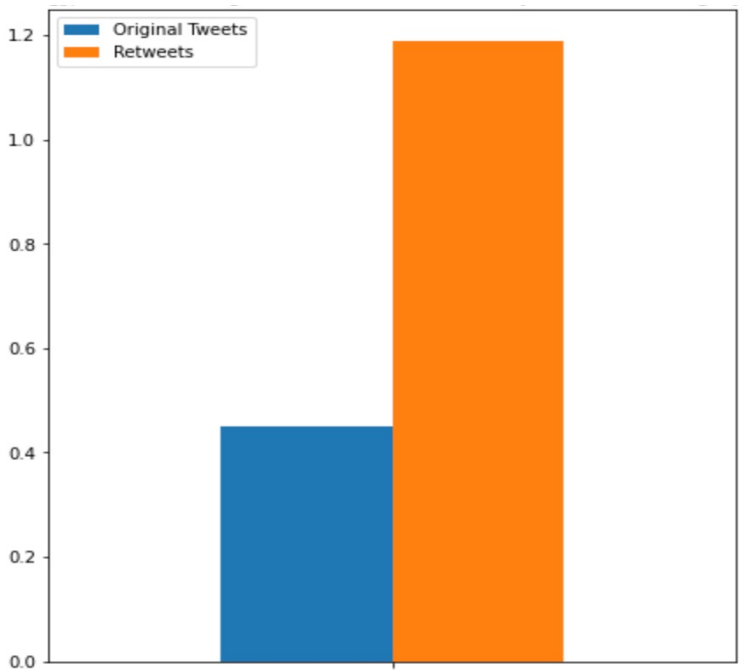


Original Content

Original and Retweeted Tweets by Organization/Category



Original and Retweeted Tweets for 'Other' Category (in MM)

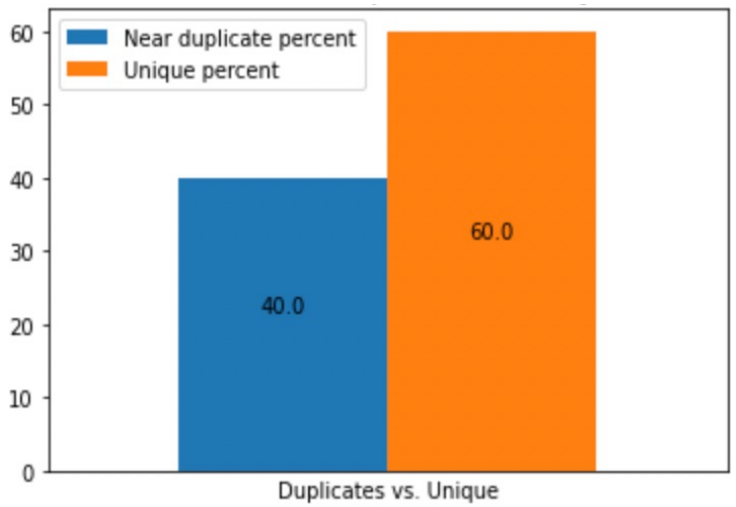


- News organizations had the highest original vs retweeted content ratio
- Out of organizations and categories, the celebrity category had the lowest original vs retweeted content ratio
- The majority Tweets by users in the 'Other' category were retweets rather than original content

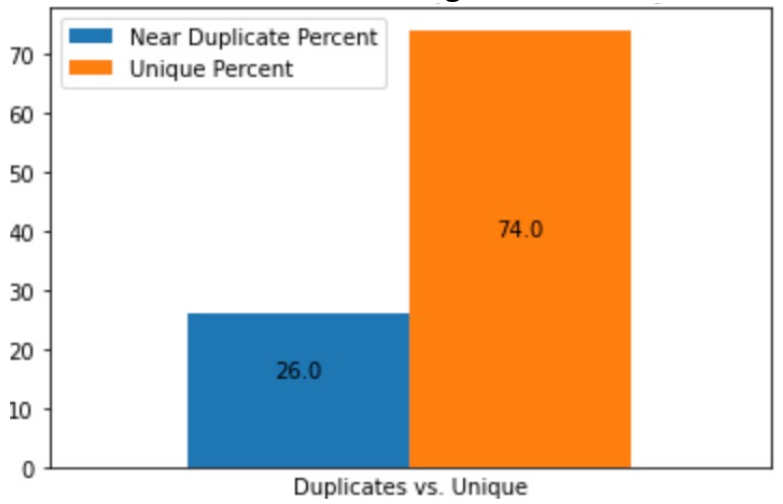
Tweet Similarity

- Celebrities and influencers have the highest percentage of non-duplicate unique tweets
- News have the lowest percentage of unique tweets among all organizations

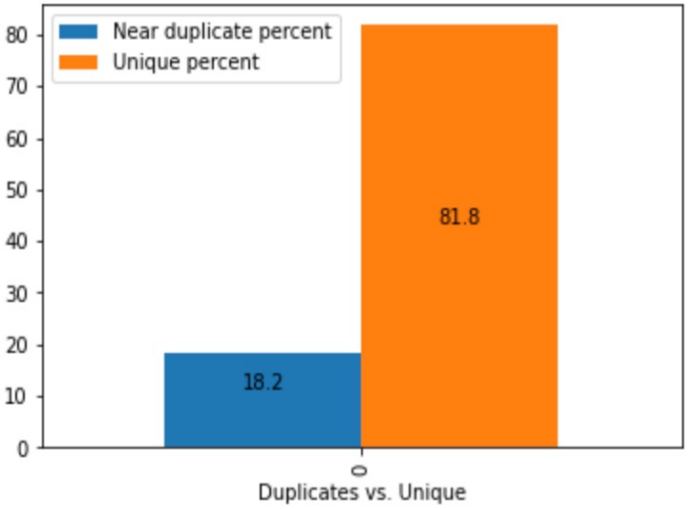
News Organizations



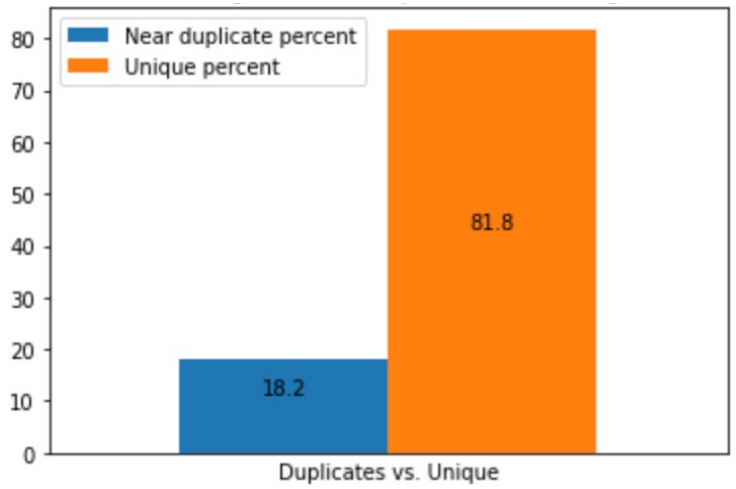
Governmental Organizations



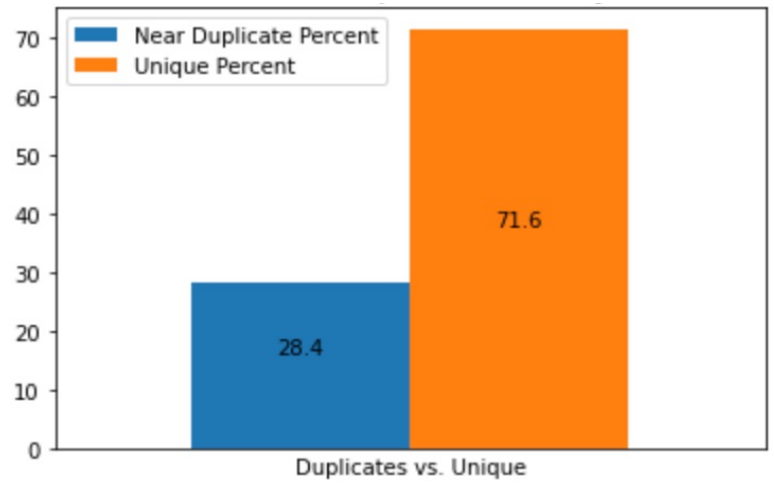
Influencer Category



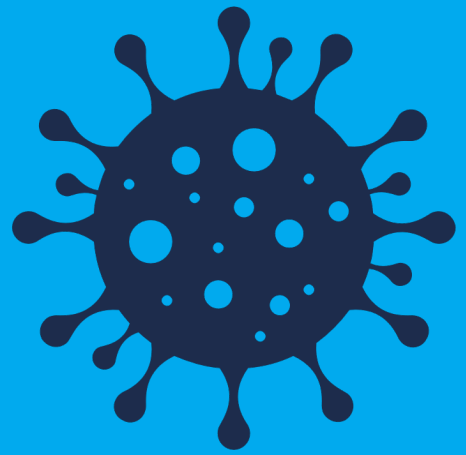
Celebrity Category



Healthcare Organizations



Final Insights and Recommendations



- When reading a Covid-19-related Tweet by a Twitterer with high Tweet or retweet volume, it is advised to check whether the Twitterer is verified. It is also advised to read their description to know whether they belong to a trustworthy organization. Otherwise, there is a risk of receiving false information from either an internet bot or an average person

- Total worldwide Tweet volumes might not be a good indicator of infection rates in the world, as high Tweet volumes do not coincide with high total infection rates. Low Covid-related Tweet volumes should not be taken as a sign of lower risks of contracting the virus. High worldwide Tweet volumes might not be directly related with high infection rates either and could be instead related to other events, such as new government actions

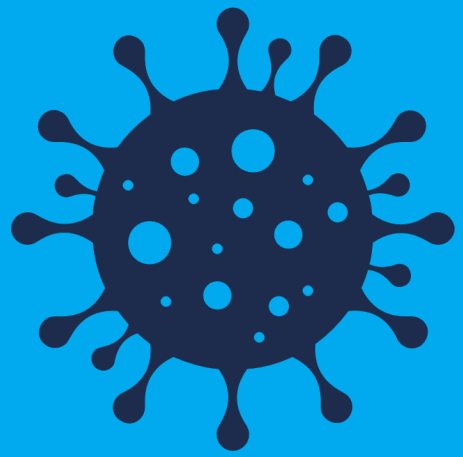
- Local analysis of Tweet volumes, rather than worldwide analysis, might produce more accurate insights into the progression of the pandemic, as Tweet volume trends for some countries seem to follow trends in infection rates. In addition, different countries have different numbers of Twitterers, and some countries will always dominate others with amount of Tweets

- News, healthcare and governmental organizations are the best source of original content, while non-verified users more often re-share information posted by others

- A considerable portion of news organization Tweets contain the same information, while more diverse information can be found from Tweets by governmental and healthcare organizations

* Statistics from *statista.com*

Appendix



1. Categories were defined as follows:
   - **Celebrity**: verified user who does not belong to either organization and has more than 30,000 followers*
   - **Influencer**: verified user who does not belong to either organization and has more than 10,000 followers*
   - **Other**: non-verified users

2. Influence score calculation:

$$\text{Influence Score} = \frac{3 * \textit{total retweets} + \textit{total likes}}{\textit{total tweets}}$$

   - Analysis showed Twitterers have, on average, three times as many likes as retweets. However, based on social media experience, higher retweet count shows more influence than high like count, thus, retweet count multiplied by a factor of 3
   - Original content was deemed more influential than retweeted content. Thus, if a Tweet was not a retweet, its influence score was multiplied by a factor of 3
   - An average of all Tweet influence scores was assigned as a final score to each Twitterer
   - Retweet count was calculated by adding the number of times a Tweet was retweeted and quoted

3. Tweets were deemed near-duplicate if they are 30% similar

*Based on article from AgilityPR