Executive Summary

1. The analysis is performed on **15,018,765** tweets related to COVID-19 from **29-days** interval from 2021/10/15 to 2021/11/12.
2. Majority of tweets are from the United States and United Kingdom.
3. Higher tweets volume between 10/18/2021 and 10/21/2021 is likely due to government and health official announcements.

User groups and retweets:

- The project has categorized users into 5 groups: news organizations, government-affiliated accounts, health organizations, influencers and others (random users).
- Majority of the tweets (**95% of total tweets**) are from others (random users). Influencers have the second highest tweet volume, which accounts for **3%** of the total tweets.
- Only **5% of total tweets** can be considered as a credible source of information regarding COVID-19. Those tweets are from news organizations, government-affiliated accounts, health organizations and certain influencers. Those tweets are more likely to be the original tweets rather than retweets.
- **71.12%** of total tweets are retweets. News outlets have the least amount of retweets, while influencers have the most.

Near-duplicate and unique tweets:

- Majority of the near-Duplicated tweets are from random users, followed by influencers and news outlets.
- On average, **1 in 4** original tweets from a government affiliated account is a near-duplicate.
- **1 in 5** original tweets from a news outlet or a health organization is a near-duplicate.

Methodology and data source overview

Data Source 1

--Twitter data 10/15/2021 to 11/12/2021
--25M rows
--JSON format, stored in Google Cloud Storage

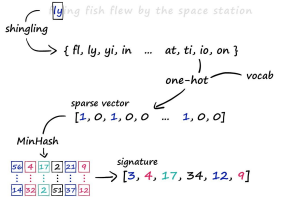
Data cleaning & filtering

--outcome: relevant rows with cleaned columns
--tool: Pyspark

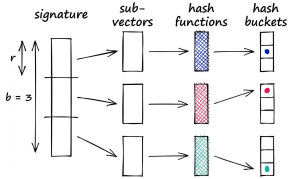
LSH for tweet similarity analysis

Texts / Document

Shingling
--Convert documents to sets
--A k-shingle (or k-gram) for a document is a sequence of k tokens that appears in the document



Minhash
--Encode sets using 0/1 vectors
--Find similar columns while hashing columns to small signatures
--Using Min-hashing as the hash function



Locality sensitive hashing

--Focus on pairs of signatures likely to be from similar documents

--Identify the ideal jaccard distance for near-duplicate identification

Candidate pairs

--Candidate pairs are those that hash to the same bucket

Data Source 2

--Daily COVID-19 confirmed cases worldwide
--CSV format, stored in John Hopkins github folder.

Visualization for location analysis:
--geopandas library for countries, plotly library for states.
--graph confirmed COVID-19 cases using JHU datasets.

Visualization for timeline analysis:
--pandas library
--graph tweet counts, grouped by dates and day of the week

Visualization for user analysis:
--pandas, matplotlib library
--graph original tweet counts, total tweet counts, duplicated and unique tweet counts, grouped by each user groups

**Source of images in the slide: https://www.pinecone.io/learn/locality-sensitive-hashing/

Data cleaning and filtering overview

Step 1. **Identify and select relevant columns**

Total number of columns in the data source: **37**

8 relevant columns were selected.

1. **'Created_at'**: tweet creation datetime.
2. **'Geo'**: location.
3. **'Id_str'**: tweet id.
4. **'Place'**: location.
5. **'Retweeted':** if the tweet has been retweeted.
6. **'Retweet_count'**
7. **'user'**
8. **'Retweeted_status' :** if the tweet is a retweet.
9. **'Text':** tweet body.

Step 2. **Removing duplicates and identify COVID related tweets** by filtering tweet that contains 'COVID' or 'coronavirus'.

Total number of tweets from data source: **25M** rows.

After removing the followings,

- Duplicates
- 'Id_str' or "Id_str" is null.
- 'Text' is null.

Tweets related to Covid-19 have **15M** rows. We also checked the number of nulls and unique values for non-nested column.

Step 3. **Flatten nested columns**

Those nested column types are struct. After flattening the following columns, they expanded into many columns.

1. 'user'
2. 'Retweeted_status'
3. 'Place'
4. 'Geo

Function to flatten columns:

```
flat_df = nested_df.select(
flat_cols +
[F.col(nc+'.'+c).alias(nc+'_'+c)
for nc in nested_cols
for c in
nested_df.select(nc+'.*').colu
mns])
```

Step 4. **Saving filtered record as parquet file** for faster access.

- **Df_5000**: 5000 rows of sample data to test run scripts.

-**Df_loc**: Geo, place, coordinates and id_str columns were flattened for location analysis.

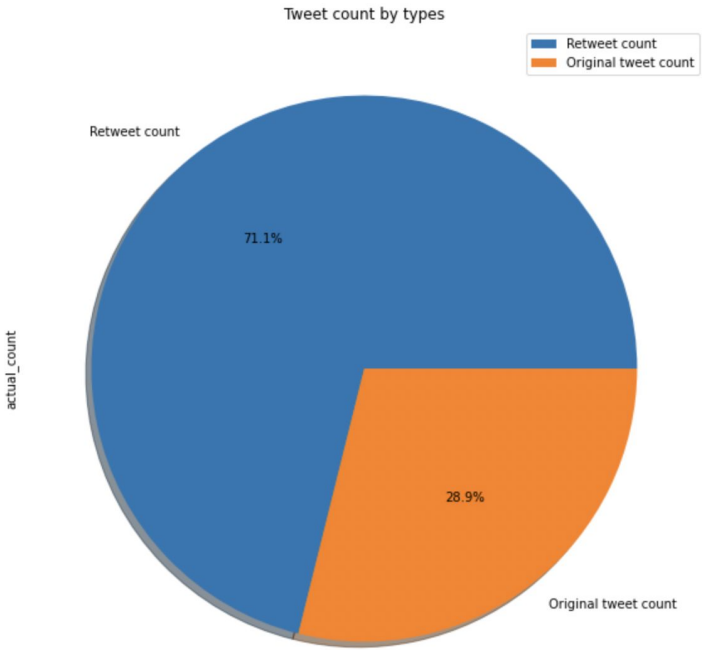
-**Df_time**: created_at and id_str columns were selected for timeline analysis.

-**Df_user_text**: id_str, text, retweeted, retweeted_count, retweeted_status and user columns were flattened for user and tweet uniqueness analysis.

EDA for retweets and user analysis

Step 1: EDA on retweets

- Number of total tweets = Number of retweets + Number of original tweets (including quote)
- **71.1%** of total tweets are **retweets**. Number of retweet is significantly higher than original tweets.
- We are curious to analyze if certain types of users retweets generate more original contents posts or vice versa.



Step 2: Use the 'Verified' status to identify certain user groups

According to Twitter, "The blue Verified badge on Twitter lets people know that an account of public interest is authentic." **The verified account represent or otherwise be associated with a prominently recognized individual or brand for the following categories.

- Government
- Companies, brands, and organizations
- News organizations and journalists
- Entertainment
- Sports and gaming
- Activists, organizers, and other influential individuals**

Verified status is a good indicator of an user representing government, news organizations and influential individuals.

** Source on verified user: https://help.twitter.com/en/managing-your-account/about-twitter-verified-accounts

Author Identification

News Media Outlets

-Users are verified.
-User descriptions contain ‘news’ or ‘channel’ or ‘magazine’.
-Each user tweeted **12 original tweets, the highest** across all groups.

Health Organizations

-Users are verified.
-User descriptions contain ‘hospital’ or ‘health’, or user URL contains ‘.org’.
-Each user tweeted 4 original tweets.

Government

-Users are verified.
-User descriptions contain ‘official’, or user URL contains ‘.gov’.
-Each user tweeted 5 original tweets.

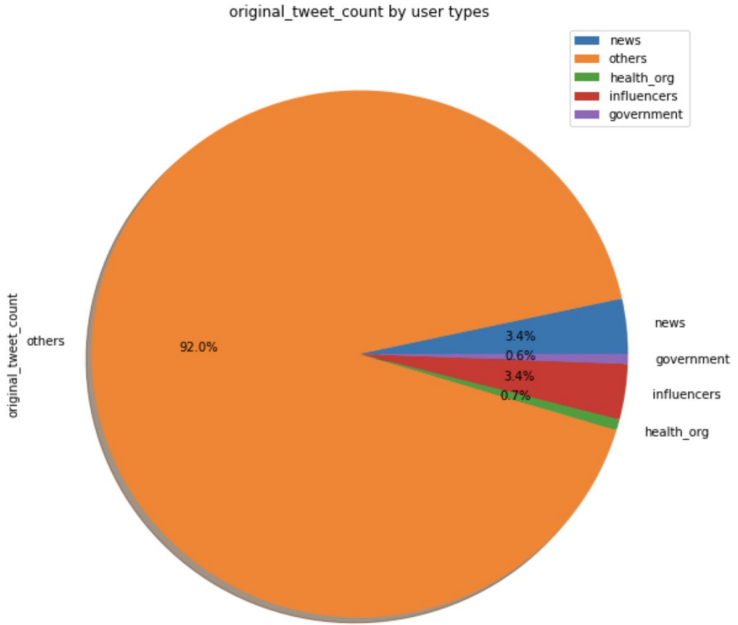
Influencers

- User has at least 1000 reach score**, and users have more than 10k tweets.
-Each user tweeted 4 original tweets, and **retweeted 9 tweets** on average.

Others

-All other users account for 95% of the total tweets, and 92% of the original tweets.
-Each user **tweeted 1 original tweet, but retweeted 3 tweets** on average.

type_user	total_tweet_count	original_tweet_count	retweet_count	number_of_user	original_tweet_per_user	retweet_per_user
news	179693	146213	33480	12288	11.898844	2.724609
others	14259820	3988948	10270872	3146835	1.267606	3.263874
health_org	50632	29218	21414	6659	4.387746	3.215798
influencers	488537	146998	341539	37216	3.949860	9.177209
government	40083	26251	13832	5532	4.745300	2.500362



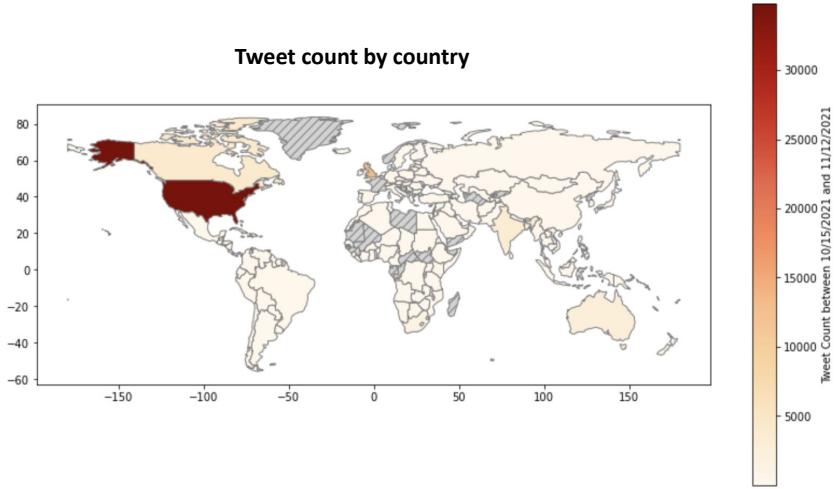
**Reach score = number of followers — number of friends

Location analysis

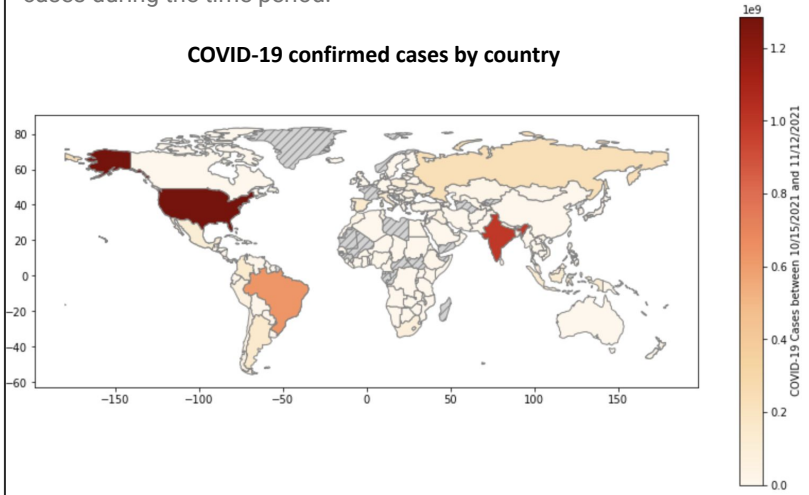
Overview:

Findings: There is no significant relation between the pandemic progression and location of the tweets. Majority of tweets are located in the US and UK, while COVID-19 hotspots are India, the US and Brazil.
Tools: geopandas library (Tweet and COVID-19 cases data is joined to geopandas country data using iso2 country code (e.g., US, CA, and FR)).

Tweet count by country:
Data population: Only 0.5% of all tweets are geotagged. Therefore, 69,591 tweets were used in this analysis. Grayed out countries do not have tweets.
Findings: US has the most tweets, followed by UK.



COVID-19 confirmed cases by country:
Data population: Daily COVID-19 confirmed cases from John Hopkins dataset from 10/15/2021 to 11/12/2021.
Findings: The US, India and Brazil have the highest confirmed COVID-19 cases during the time period.



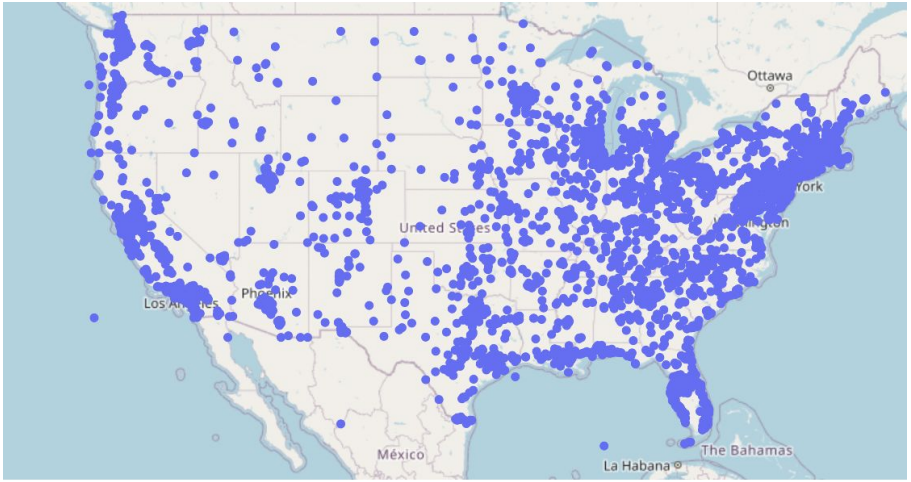
Location analysis - USA

Overview:

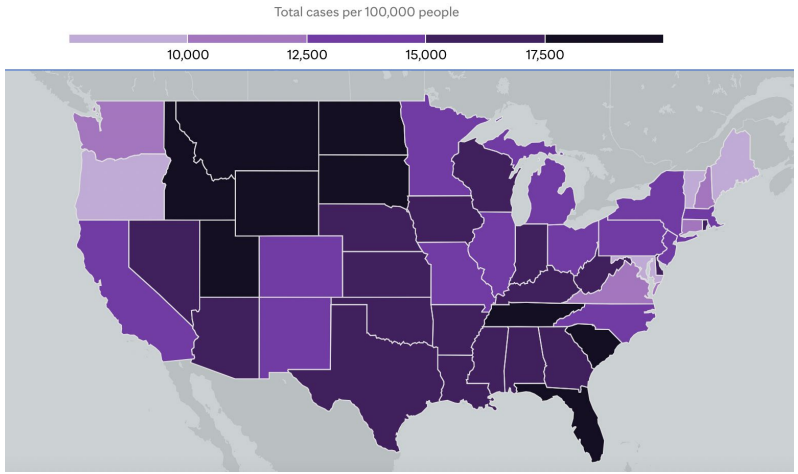
- Only 34819 tweets have geotags in the US.
- The tweet count are centralized around bigger cities, especially coastal cities and East North Central Region in the US. There are fewer tweets in the intermountain and west of the midwest region.
- Midwest region and Southeast have the most COVID-19 cases according to mayo clinic (https://www.mayoclinic.org/coronavirus-covid-19/map).
- The tweet volume does not reflect higher or lower COVID-19 cases geographically.

Tools: plotly.express library with mapbox=open-street-map

Tweet Count based on Latitude and Longitude in the US

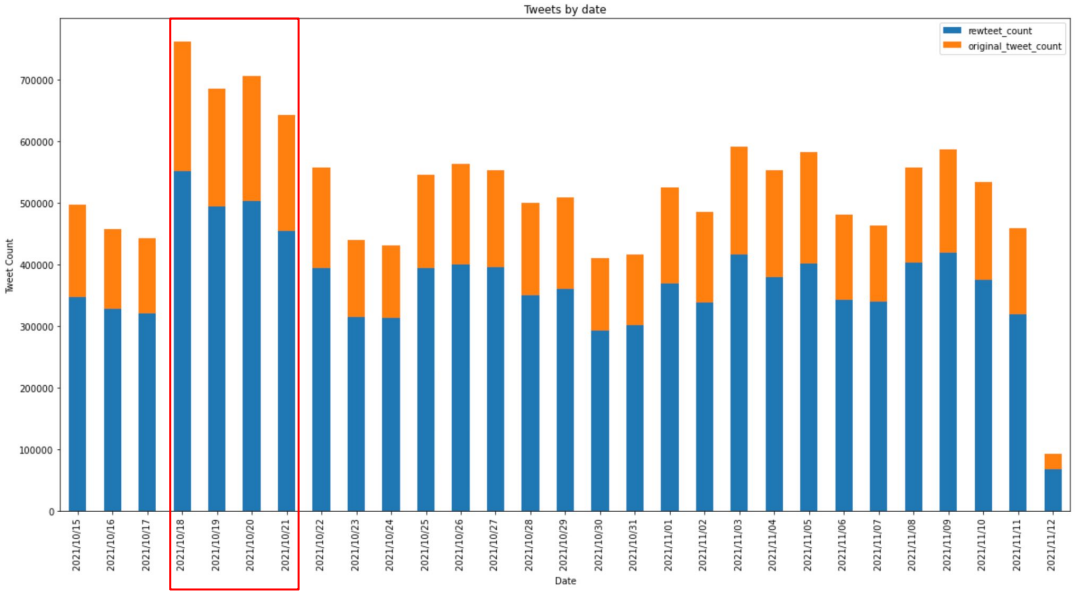


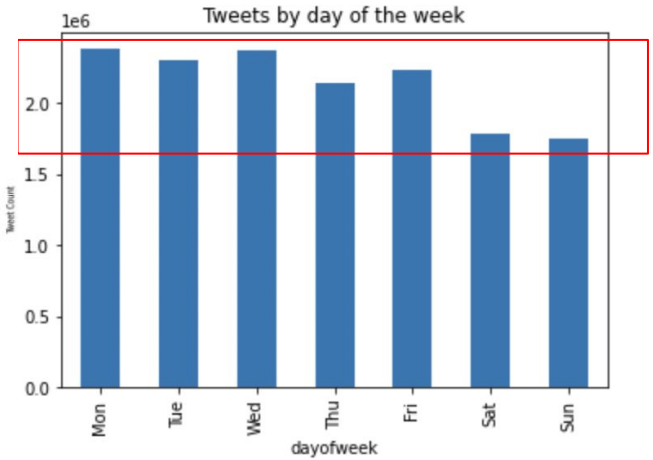
COVID-19 Cases from MayoClinic.org



Timeline analysis

- Tweets were collected from 29-days interval from 2021/10/15 to 2021/11/12.
- Data collection gaps:
  - Fewer tweets on weekends than on weekdays.
  - Missing data on 11/12/2021.
- Four-day period from 10/18 to 10/21/2021 have the most tweets related to COVID-19. The period has more retweets. Potential reason is that FDA approved 'mix-and-match' COVID boosters and White House released vaccination plan for children aged 5 to 11.





Message uniqueness analysis using MinHash and LSH

Step 1. Clean the data, remove stopwords and create index

- Change all strings to lowercase
- Remove stopwords
- Resplit documents and discard string length shorter than 8

Step 2. Fit countvectorizer to create word features

	list_of_words	id	features
0	[increased]	3	(0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 1.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...)

Step 3. Fit MinHash to create hash table

id	text	list_of_words	features	hashes
19	[influencers, 145...	[vaccinated, secr...	(58512,[0,14],[1....	[[4.7945584E7], [...

Step 4. Compare thresholds side-by-side and identify the Jaccard Distance threshold

- Compare Jaccard similarity = 0.3 and 0.5 to measure similarities.
- Jaccard similarity =0.3 is the most ideal. Found 776,952 near-duplicate tweets.

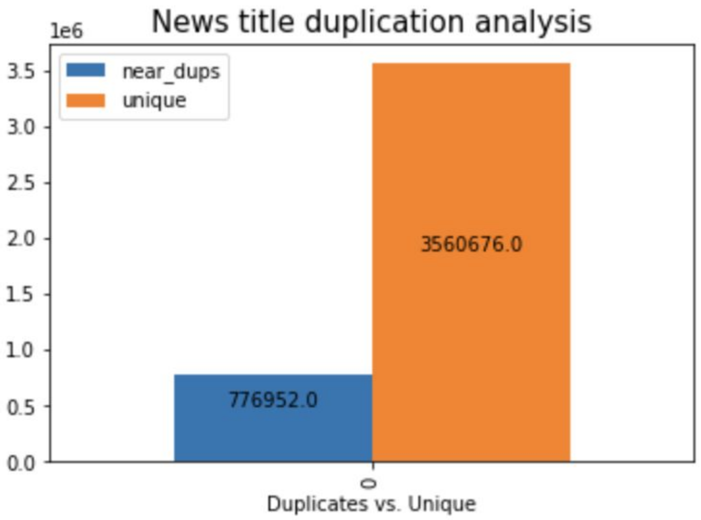
text_A	text_B	threshold_30	threshold_50	threshold_70
(others, 1452018996669157377, @ElijahSchaffer @McpeakRichie More Americans have died from Covid-19 so far in 2021 than in 2020, a sobering milest... https://t.co/7HmZwJwrkl)	(others, 1458991242277179396, @chauncy_bridges @gien_ed @GovTimWalz 1200 Americans died today of Covid-19.)	Non-Dup	Duplicate	Duplicate
(others, 1456699255968735236, @celliottability You should have expanded eligibility for a COVID-19 booster shot to Canadians at least 60+. The co... https://t.co/ll98O9RDOB)	(government, 1455898326524702724, Yesterday's update on the #COVID19 vaccine eligibility for 5-11 year-olds was monumental! We are working around the... https://t.co/bxNtlvfPUh)	Non-Dup	Non-Dup	Duplicate
(influencers, 1455646591919001601, #Pfizer's #COVID19 shot for children age 5 to 11 was backed unanimously by advisers to the #CDC in a vote that will... https://t.co/DBtaonmPVQ)	(others, 1451299871722516488, CDC panel unanimously endorses Moderna and J&J Covid boosters, sending to director for final approval https://t.co/iyJQr6igNj)	Duplicate	Duplicate	Duplicate

<u>Data population:</u>

Uniqueness analysis is performed on **4,337,628** original tweets. The population does not include retweets. The total number of tweets is the sum of original tweets and retweets.

<u>Results:</u>

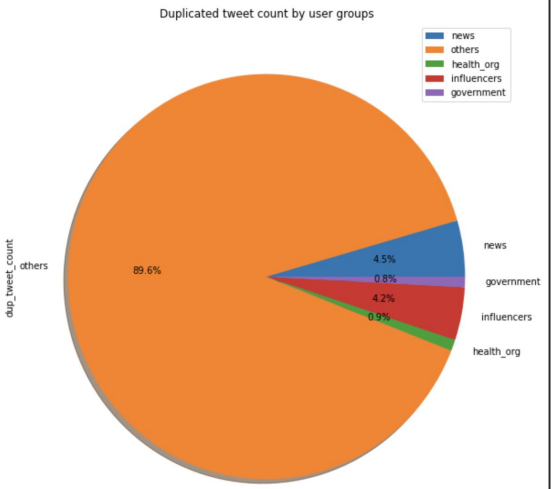
The analysis resulted in 3,560,676 unique tweets and 776,952 near duplicates. **21.8%** of all original tweets are near duplicates.



Message uniqueness analysis by user groups

Near-dup tweets Overview:

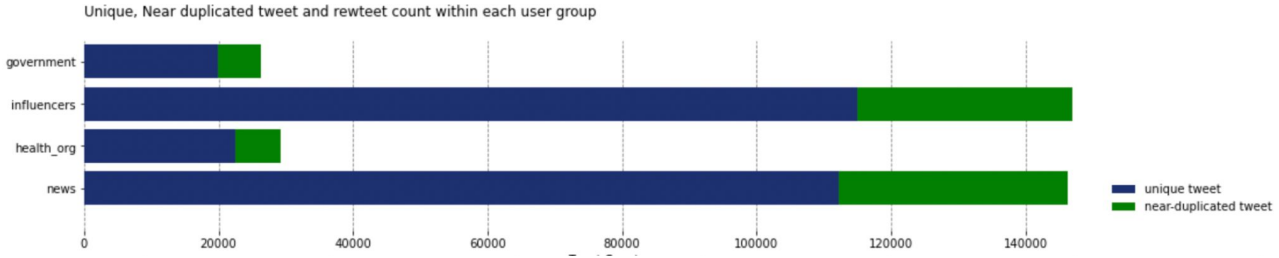
-**89.6%** of all near-duplicated tweets is from random users (labeled as 'others').
-**4.5%** of all near-duplicates is from new outlets, and **4.2%** from influencers.
-Government and health organizations accounted for the **least** amount of near duplicate tweets.



Near-dup/Uniqueness ratio within each user group:

- Near-dup ratio is calculated by (# of near-dup tweets)/(total number of original tweets). The unique tweet ratio is calculated by (# of unique tweets)/(total number of original tweets).
- Random users have the highest unique tweet ratio and lowest near-duplicate ratio. **17 in 100** tweets from a random user are near-duplicates.
- Government accounts have the highest near-dup ratios. On average, **1 in 4** original tweets from a government account is a near-duplicate.
- **1 in 5** original tweets from a news outlet or a health organization is a near-duplicate.

type_user	near_dup_ratio	unique_ratio
news	0.232332	0.767668
others	0.170550	0.829450
health_org	0.232494	0.767506
influencers	0.217915	0.782085
government	0.241629	0.758371



Conclusions and actionable recommendations

Conclusion:

1. Tweets from new organization, health organization and government accounts can be considered a credible source of real-time information regarding COVID-19.

2. Rather than the volume or location of tweets, tweets from certain user groups reflect higher COVID-19 risk.

3. Number of retweets and near-duplicate tweets reflect higher concerns over COVID-19.

Recommendation to improve the analysis:

1. Use the new government and state affiliated label in twitter to better identify user groups.
https://blog.twitter.com/en_us/topics/product/2020/new-labels-for-government-and-state-affiliated-media-accounts

2. Develop detailed metrics to profile influencer based on historical tweets. Potential metrics are popularity_score and reach_score.

3. Implement topic modeling in various languages to better filter COVID-related tweets. Topic modeling also allows us to identify the subtopics such as travel restrictions, vaccinations and market influence. By filtering on tweets that contain ‘COVID’, we only selected tweets from English speaking countries. Topic modeling in various languages will mitigate data imbalance.