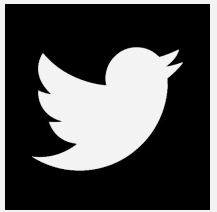
KAI HAYDEN

THE UNIVERSITY OF CHICAGO

Big Data Final Project

Covid-19 Twitter Analysis



2021

THE UNIVERSITY OF CHICAGO

Contents

Overview

Understanding the Task

01

THE UNIVERSITY OF CHICAGO

Executive Summary

- The most consistently retweeted top accounts are key opinion leaders and doctors, while influencer and political tweets get high "one off" retweet counts
- Twitterers that have high average retweets per post are more likely to be credible
- The majority of COVID-19 Tweets come from North America
- There are more tweets during the week than during the weekend, and more tweets during the afternoon than at morning or night
- Europe & Asia are most reactive to COVID-19 news while Oceania & Africa are the least reactive
- Only 2.5% of all tweets are unique, but about 76% of original tweets are unique

THE UNIVERSITY OF CHICAGO

Source Data

Volume

- Approximately 25,000,000 Tweets

Velocity

- Millisecond timestamps
- Data ranging from October 15th to December 12th

Variety

- JSON file with 37 fields, including strings, datetimes, numbers, booleans, and nested structs

THE UNIVERSITY OF CHICAGO

Methodology

Clean Data

- General preprocessing
- Datetime preprocessing
- Filling in missing values
- Geographical preprocessing

Top Twitterers

- Analyze tweet & retweet volume
- Analyze by twitterer category

Location

- Extract continents using keyword matching

Timeline

- Analyze time series data based on hours of the day & days of the week
- Plot the timeline of tweets by continent
- Calculate the coefficient of variance for each continent

Uniqueness

- Filter original and unique tweets
- Remove stopwords and links from tweets
- Vectorize tweets and apply MinHash LSH
- Use various jaccard thresholds to find unique tweets

Cleaning

Processing the dataset for analysis

02

THE UNIVERSITY OF CHICAGO

Cleaning - General

Initial Cleaning

1. Removed unusable and duplicate columns
2. Removed columns not needed for analysis
3. Remove unrelated tweets through keyword matching
4. Replace truncated tweet data with full tweet data

Created Columns

- **User and Tweet ID** for retweeted, quoted, and replied to tweets
- **Tweet** full text, hashtags, and location
- **User** ID, location, name, tweet count, description, and verification

Cleaning - Time

Converted Timestamp Data

- Original timestamp data given in milliseconds
- PySpark timestamp data type in seconds

Created Relevant Timestamp Columns

- Year
- Month
- Day of Month
- Hour
- Date (Y, M, D)
- Date Hour (Y, M, D, H)

THE UNIVERSITY OF CHICAGO

Cleaning - Retweets, Quotes, Replies

Retweets, Quotes, and Replies Are All 0

- Tweets are grabbed by the API at the time of posting
- Obviously, this data will all be 0 at the time of posting

Filling in the Missing Data

1. Count the number of tweets that retweet, reply, and quote the original tweet
2. Join the counts based on the status id of the original tweet
3. **Retweet**, **Quote**, and **Reply** counts are now populated

Cleaning - Geographical

Original Tweets with Location Data

- Before Extraction: 46,768
- After Extraction: 1,165,348 **(25x)**

Total Tweets with Location Data

- Before Extraction: 133,619
- After Extraction: 8,309,731 **(62x)**

Percentage of Tweets with Location Data

- Before Extraction: 0.56%
- After Extraction: 34.75% **(62x)**

User Location Extraction

User Location	Continent
Los Angeles, CA	NA
Ireland	EU
Sydney, Australia	OC
HKG, TPE, ULN, or 35,000 ft	AS
Chicago, IL	NA
kampala Uganda	AF
New Zealand	OC

Top Twitterers

The most prolific and influential Twitterers

03

Prolific Twitterers

Looking at Top 50 Twitterers by

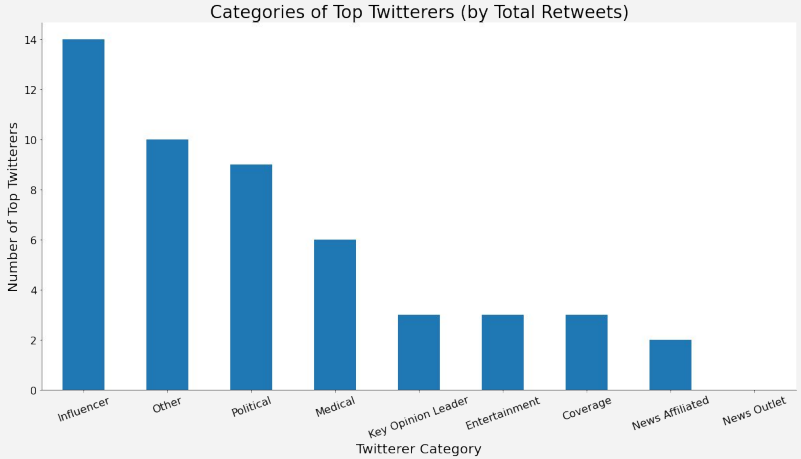
- Total Retweets
- Average Retweets

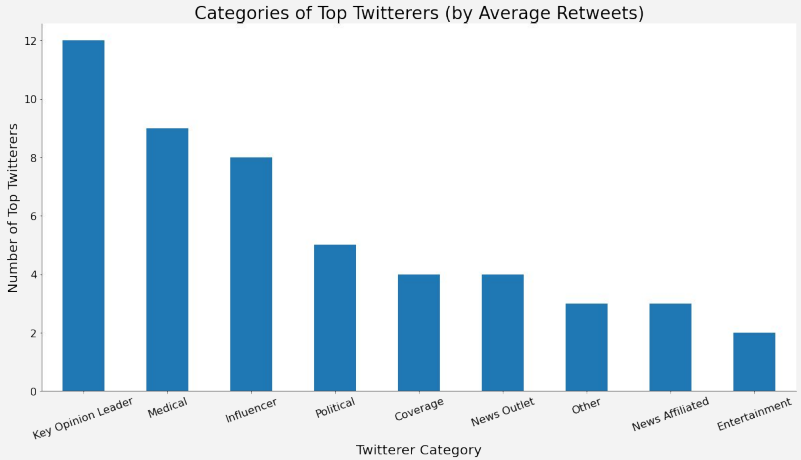
Viral Tweets

- **Influencer & Political** tweets get the highest retweet counts

Credible Sources

- The most consistently retweeted accounts are from **KOLs & Doctors**





THE UNIVERSITY OF CHICAGO

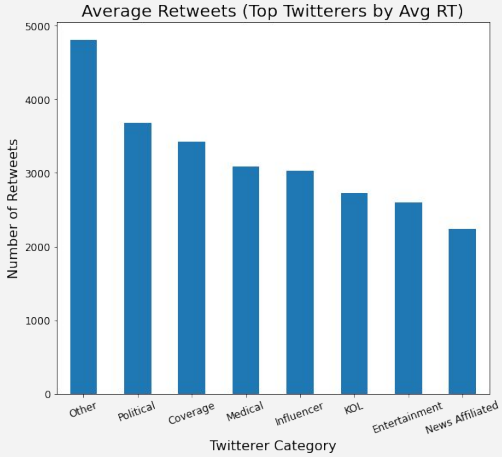
Prolific Twitterers

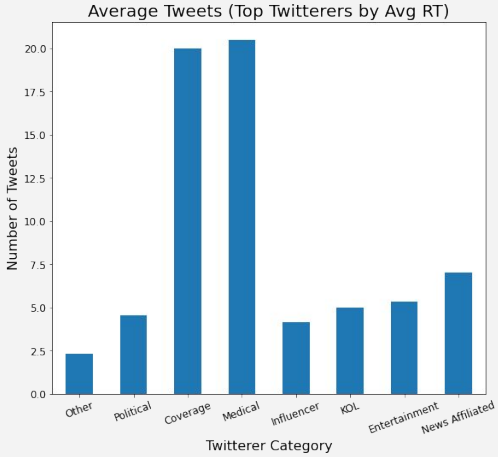
Top Twitterers (by Total RT)

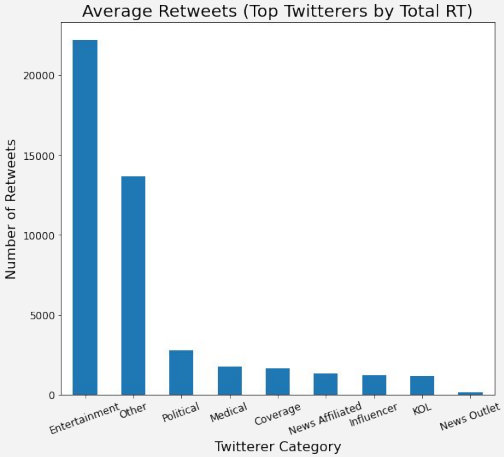
- **Independent coverage and medical** accounts maintain high average retweets while tweeting very frequently
- **Other, typically regular accounts** who have top tweets tend to get lucky with one viral tweet.

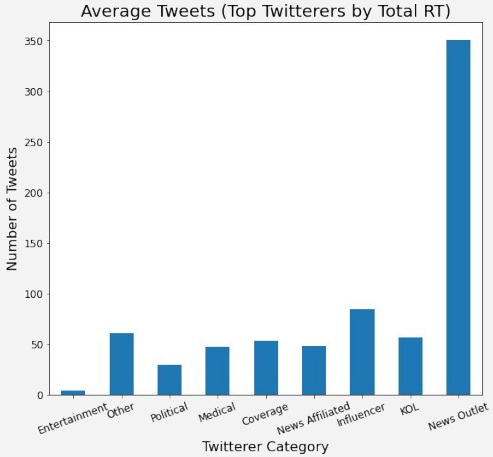
Top Twitterers (by Avg RT)

- **Entertainment** accounts get lots of retweets while having very few tweets.
- **News outlets** receive very few retweets per tweet, while tweeting the most frequently out of all categories









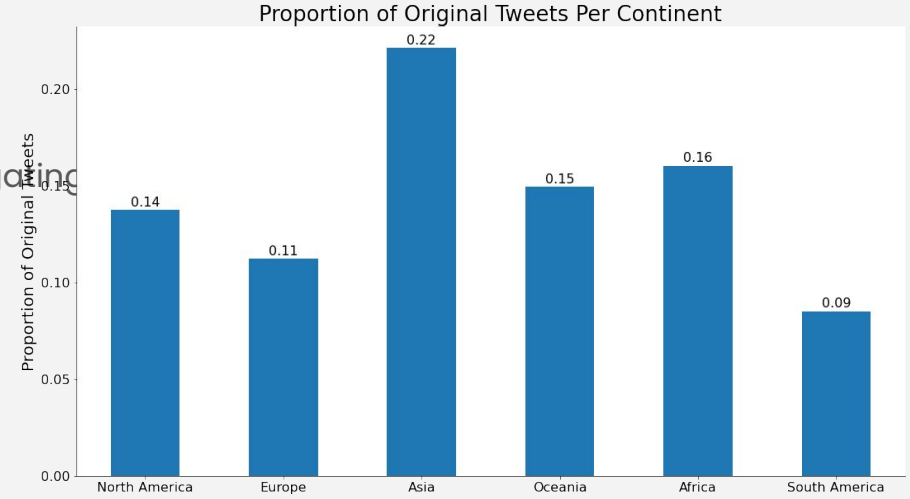
Location

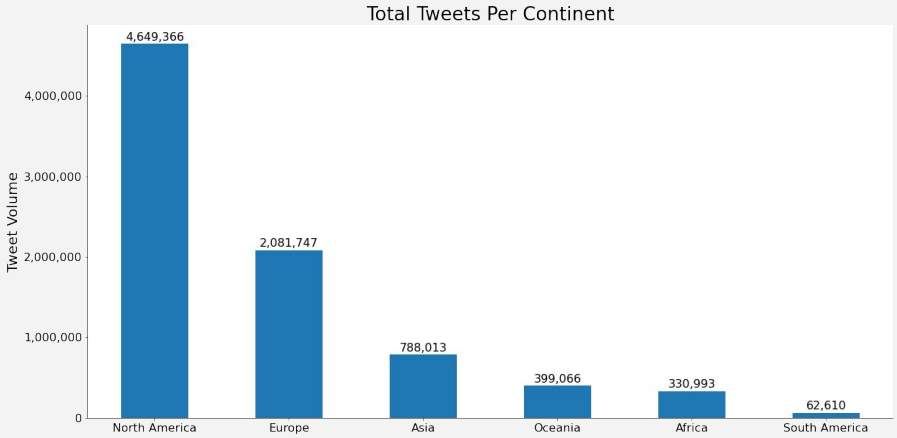
Where Twitterers are located

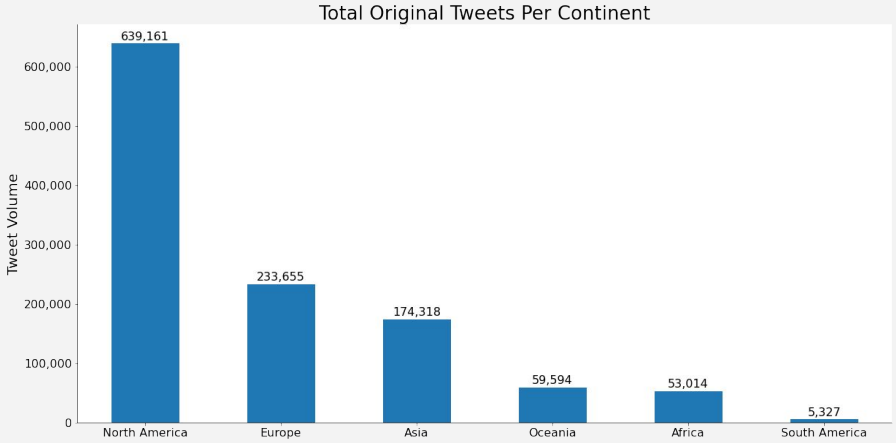
04

Where are Twitterers?

- Continents rank in the same order when aggregating total tweets and original tweets
- **North America** has more tweets than all other continents combined
- **Asia** has a large proportion of original tweets
- **South America** has a low amount of tweets, both total and original
- **Oceania & Africa** have similar tweet volume







Timeline

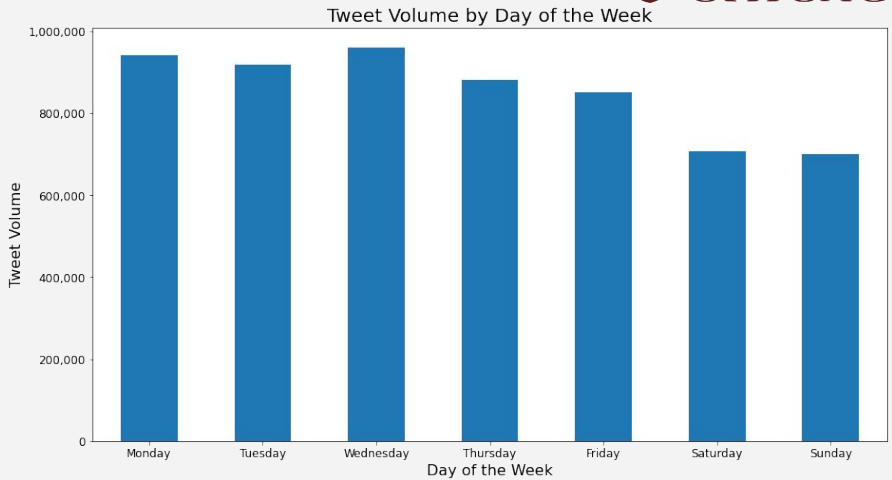
Analyzing Tweets over time

05

Timeline of Tweets

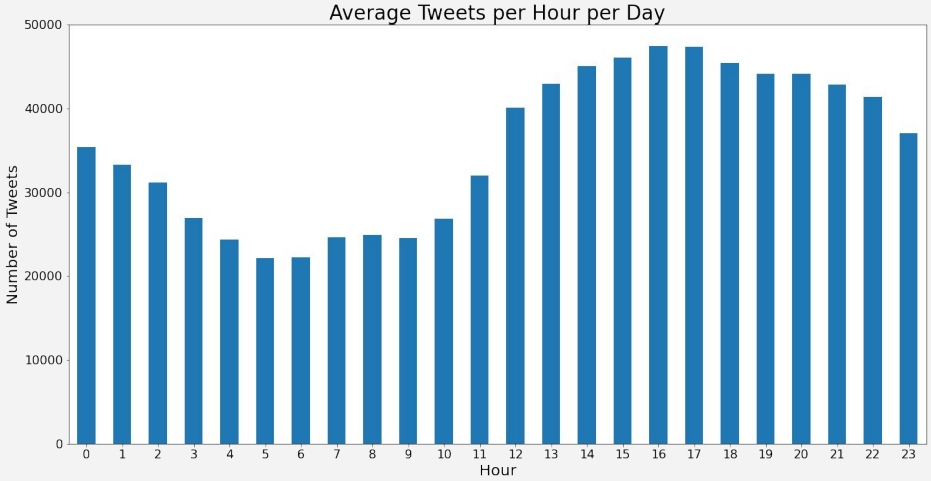
Tweet Volume per Day

- Tweet volume is significantly lower on weekends, roughly 24% less.
- Average **696,000 per day on weekends**
- Average **912,000 per day on weekdays**



Tweet Volume per Hour

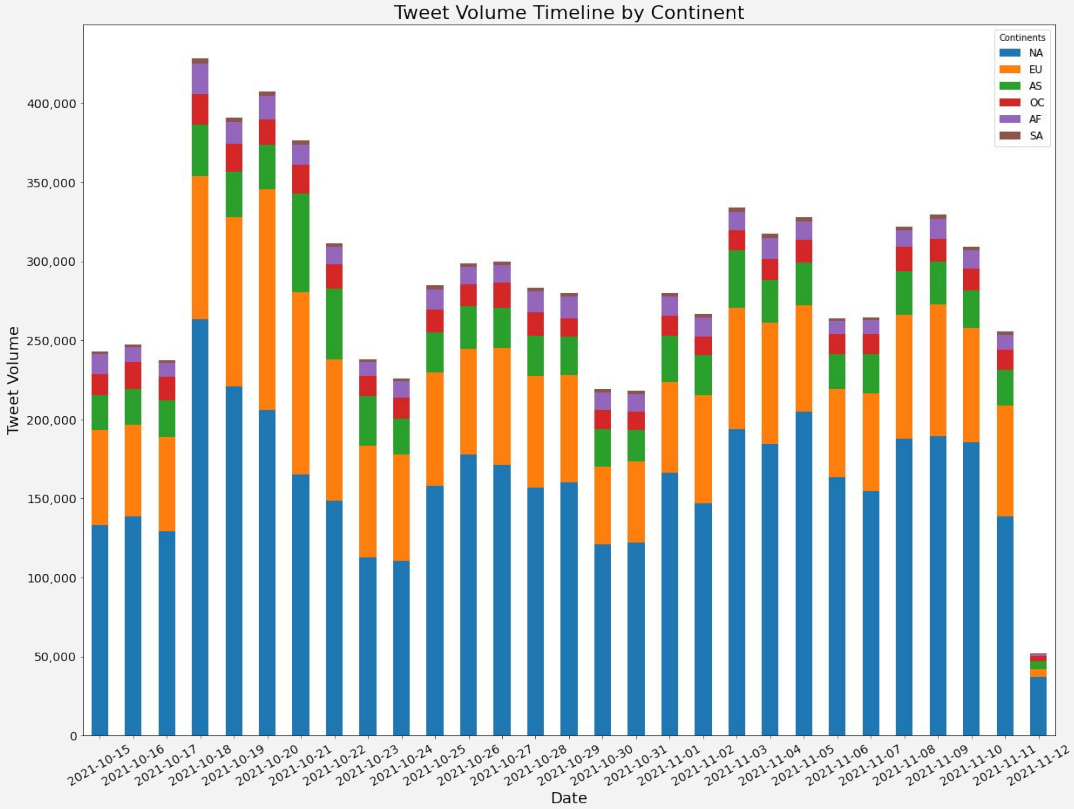
- Peaks during the afternoon around **4-5pm**, at roughly **47,000 tweets per hour**
- Troughs during the morning **5-6am**, at roughly **22,000 tweets per hour**
- Tweet volume throughout the day is a smooth fluctuation between the peak and trough



Pandemic Progression

- **Tweet volume in Europe & Asia** varies the most in response to news
- **Tweet volume in Oceania & Africa** varies the least in response to news

Continent	Coefficient of Variance
NA	0.261
EU	0.323
AS	0.340
OC	0.210
AF	0.215
SA	0.236



Uniqueness

Analyzing similarity of tweets

06

How Unique are Messages?

Finding Unique Tweets

- Retweets and quote tweets were not included
- Removed stopwords & links from 25,000 most recent original tweets.
- Vectorized the tweets and applied MinHash LSH
- Used various Jaccard thresholds to find unique tweets

Most Original Tweets are Unique

- Between 68-80% of original tweets are unique
- After checking samples of the result, Jaccard distance 0.4 seems to be the most appropriate
- **76%** of original tweets are **unique**
- **24%** of original tweets are **near duplicates**
- Assuming this holds for all original tweets, this implies roughly **641,000 unique tweets**, representing **2.5% of all tweets**.