Twitter's Credibility in Topics of Education 🎓

Han Jiang

Final Project Presentation

MSCA 31013 Big Data Platform

December, 2022

Executive Summary

Key findings:

- **Verification Issue**: Less than 2% of twitter users are verified users and don't belong to a credible organization, including top twitter accounts by retweet
- **Influence Measurement**: The influence scoring method performs better than just using retweet count to find most influential twitterers
- **Content Uniqueness**: Text similarity test shows that many organizations are sharing unique content, except News & Media are sharing more similar content which is comprehensible due to the nature of their role
- **Reflection of trending topics**: breaking news get spread more easily, but less hot topics can be flooded under vast irrelevant information. A narrow scope like state-wise location analysis might help to reduce noise

Recommendations:

- **Improving Verification and Fact Checks**: before streaming popular tweets, Twitter needs to validate popular twitters' identity and credibility. For largely spreading tweets, more fact checks are needed
- **Influence Scoring and Measurement**: when deciding what is to be recommended and prioritized for streaming, Twitter needs to improve its algorithm by incorporating more credible users/organizations
- **Disinformation List**: Users with multiple recent records of spreading misinformation or rumours will be tagged with disinformation warning

Methodology*

01 | **Filtering**: use keywords to filter for topic-related terms in the tweets (i.e., HomeSchooling)

02 | **Data Processing**: convert data types like time and numbers; tag twitterers with corresponding organizations

03 | **Exploratory Data Analysis***: explore variables and remove or fill the Null values

04 | **Influence Analysis**: for each user, calculate the number of original tweets, retweets, & times of being retweeted

Influence Score: for each user id, calculate its influence per tweet and give a score for overall influence

05 | **Location & Time Analysis**: analyze the geological distribution of twitterers, who they are and when they tweet

06 | **Most Influential**: Who is the most influential group of twitterers, and which organization do they belong to

07 | **Tweet Duplication**: Analyze original tweets about education for duplication vs. uniqueness using LSH text similarity

* More details about the methodology is available in the Appendix

Background & Data Source Overview

Background: the objective is to identify whether Twitter can be considered a credible source of information, which reflects the emergence of important trends or topics in education.

Data Source: Approximately **100 Million Tweets** about education are extracted from Twitter API from April 2022 to November 2022.

Filtering & Cleanup: We intentionally control the number of filtered tweets (about **7.4 million**) by filtering topics such as ***racial equality, literacy, tech & digital, special need, school curriculum, higher education*** to see if the tweet reflect the trending real-world online discussion*. Also, **only tweets in English** (lang='en') are considered

Total Twitter Usesr

3.1M

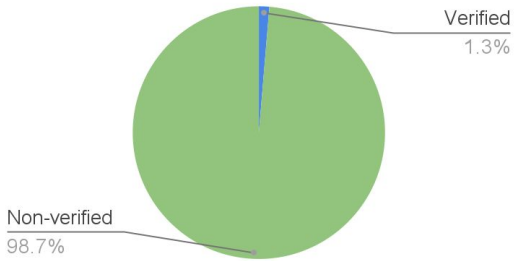
<2% of them are verified user

Total Tweets

7.4M

(Original + Retweet)

Verified vs Non-Verified Users by Percentage



Country**	Number of Twitter Users
United States	34226
United Kingdom	4038
India	2299
Canada	1364

Organization Type	Number of Twitter Users
Government Entities	7918
NGOs	2339
Schools	1206
Universities	8598
Celebrity & Influencer	11095
News and Media	53359
Other***	61570765

* More details are in appendix

** This shows the filtered twitter account associated with our education topic. We expect higher numbers if using a more loose filtering standard

*** Other indicates users who are not verified users or belong to other organizations than what is listed above

EDA and Selection of Variables

We selected variables that are: (1) related to retweets/favorite/quote, (2) the same variables (retweet, quote count, source_user_id) under retweet_status, and (3) basic information: userid, user name, location, follower counts etc. The findings are:

- Basic Information like user name, id, and verification don't have null values
- Retweet, quote, and favorite count are all not nulls, but they are **all zeros** (which means missing data)
- Retweet, quote, and favorite count under retweet_status has some null values (which is normal because some tweets are original and don't have these data)
- Quoted_status has a similar problem (over 95% null) and its information is removed due to limited slide space
- 99% of geological information is lost, but we still have ~ 0.7 million to use

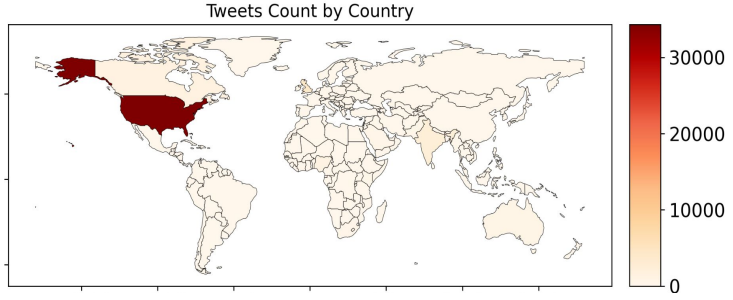
Variables Basic info	craeted_at	id	Retweet count	Favorite count	Quote count	retweeted	verified
Null (%)	99	0	0		0	0	0

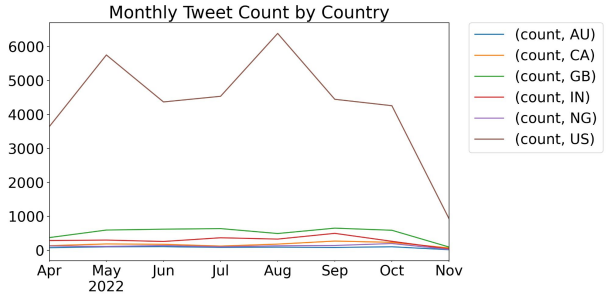
Variables Basic info	country	user_name	followers_count	User description
Null (%)	99	0	0	0

Variables Retweet_status	rt_retweet_ count	rt_quote_count	rt_favorite_count	source_id
Null (%)	27	27	27	27

Overall, I chose the basic information (id, created_at, verified, country, etc.) and those under retwet_status (retweet, quote, favorite, source user id)

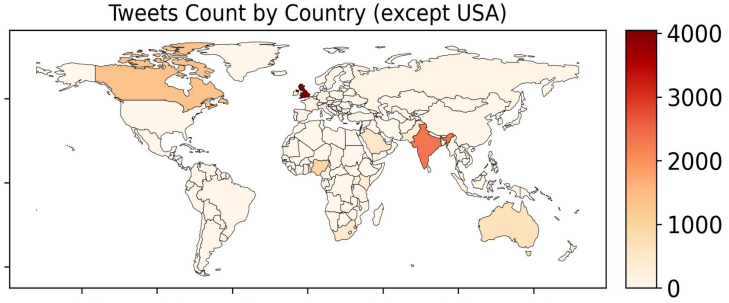
Location Analysis

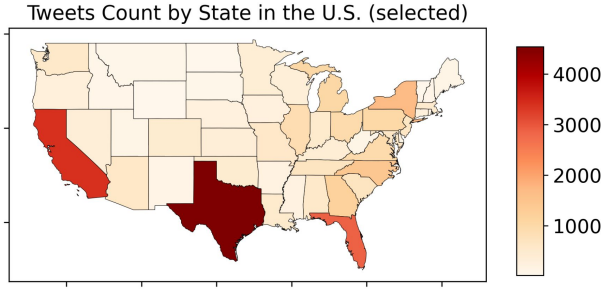




Although the filtering session has limited the tweets to the language of English and topics related to education, we still find:

- The majority of tweets are from the U.S., the U.K, and India, which is in accordance with real-world data of the countries with most twitter users[1]
- Twitter users are mostly from the U.S.
- Texas, California, and Floria are the states with most tweets[2] in the U.S.



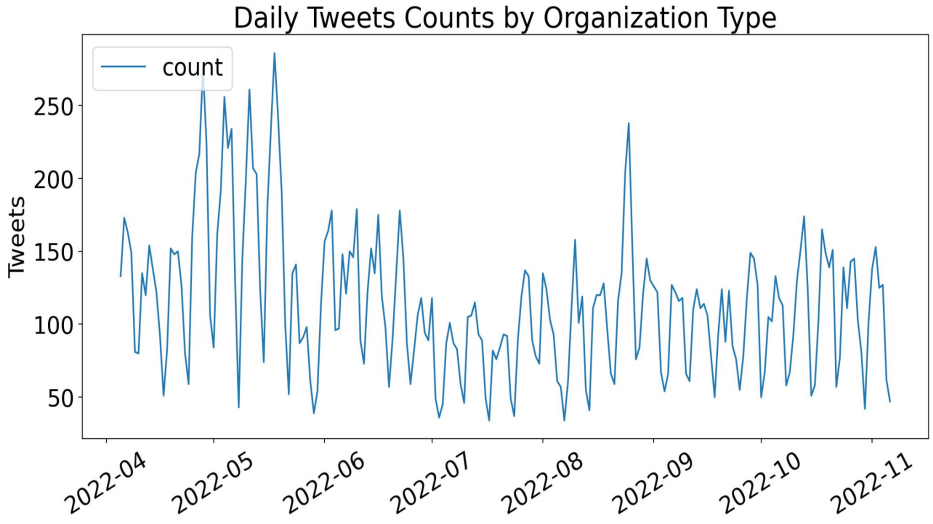


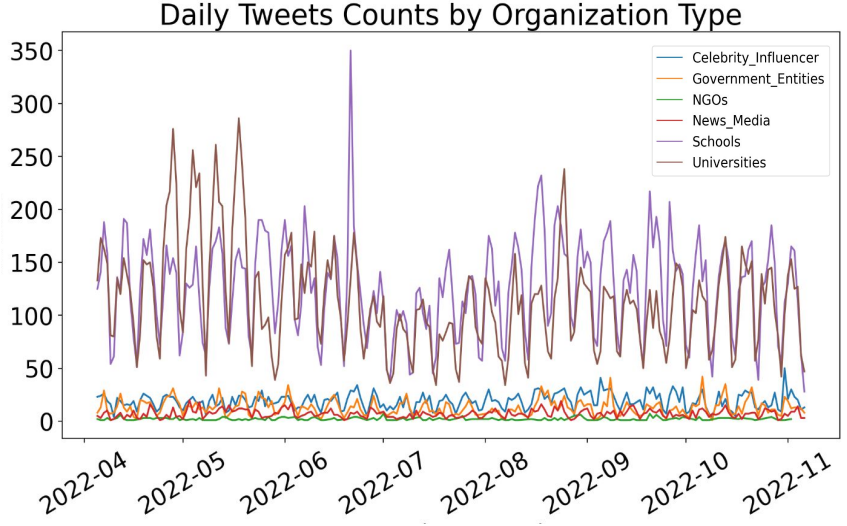
- Recall that we included certain heated topics while filtering the data and EDA. There is a obvious spike in the U.S.'s tweets in August, 2022. The corresponding news is that the Biden Administration announced a $10 Billion student debt relief[3] This was a series of news announcements in August, which could explain part of the spike in tweets.

[1] Sources: Oberlo, note that our results only consider English speaking countries.
[2] Alaska and Hawaii are removed for ease of plotting. Their records are low.
[3] U.S. Department of Education. Aug, 2022. The Administration Announces Public Service Loan Forgiveness Surpasses $10 Billion in Debt Relief.

Trending Tweets Over Time

- The daily education tweet counts had large variations. There are two spikes: one in May and another in mid-August. Potential rationales are:
  - The school shooting incident in 05-24-2022 at Robb Elementary School in Texas had risen a huge online disgruntled in school gun control
  - The $10 billion Student Debt Relief announced by the U.S. administration had also caused a rise in online discussion about student loan
  - **Both spikes cross-validate the real-world news**, as well as previou slide. For example, *Texas* is where the shooting incident happened, and it has the highest tweets count in our tweet count map by each state in the U.S.
- The tweets also has **a pattern of a weekly cycle** (tweets on weekday is more than tweets on weekends). Considering the News & Media tweets and retweet the most, it aligns with the fact that the News agencies are taking days off during weekends*



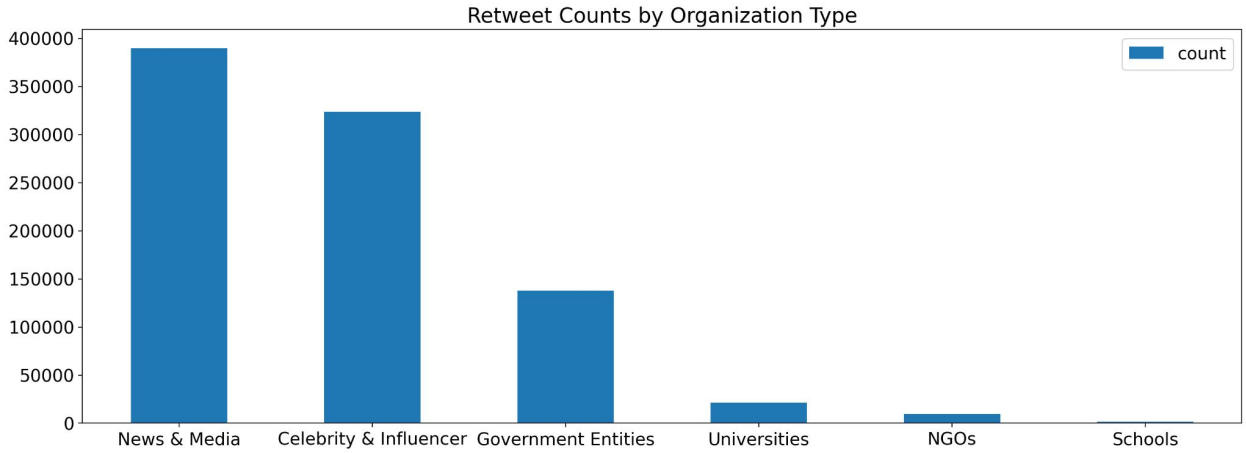


* The Other category is dropped for visibility of the rest of the organizations

Most Influential Twitter By Retweet Count*

- The most influential twitterers by retweet count are non-verified users under the organization category of “Other”
- Beside the “Other” category that has the most original tweets, **News & Media** has the most retweet count about education, and **Celebrity & Influencer** have the second most number of times being retweeted

Top Twitterers	Count	Influence Score
Libs of TikTok	82541	160714.1
John-Carlos Estrada	21806	5160.7
Jo	21043	61092.1
Christian Christensen	19672	7083.7
Michael Warburton 🇺🇦	18129	3059.2



Note: [1] the “Other” category is dropped for better visibility

[2] Definition of “News & Media” consider (1) news and the press, (2) account of or related to social media platforms with 100k+ followers (like YouTube and Tiktok)

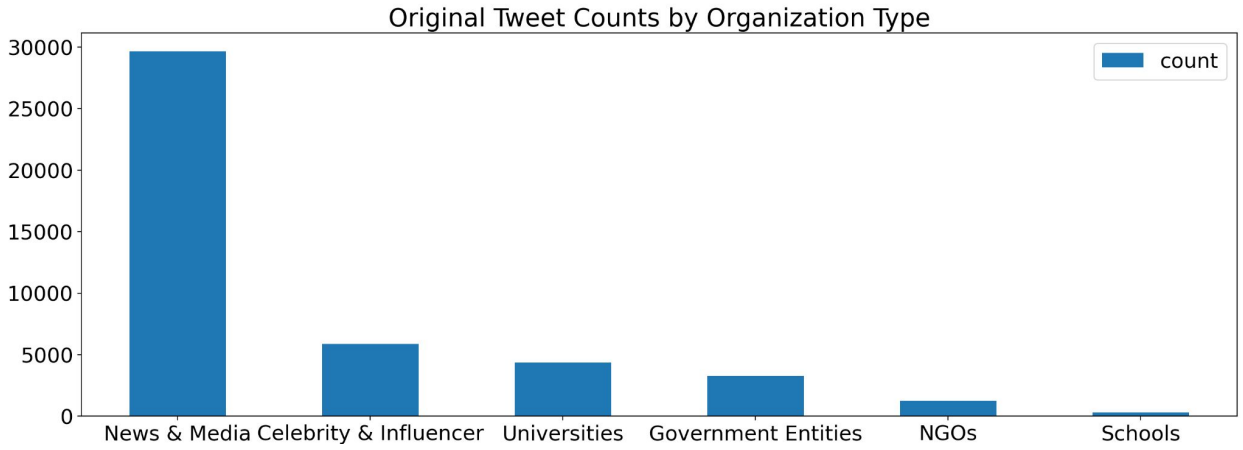
[3] Definition of “Celebrity & Influencer” consider (1) celebrities (i.e., actors, athletes), (2) social media influencers with 100k+ followers)

* Retweet Count means the number of times being retweeted. I didn’t use count of retweet and quote due to concerns of duplicated counts. See more details in appendix

Most Influential Twitter By Original Tweet Count

- The most influential twitterers by original tweet count are non-verified users under the organization tag of “Other”
- Beside the “Other” category that has the most original tweets, **News & Media** has the most original tweets about education
- News & Media has significantly larger amount of original tweets than others , which is different from the previous slide about retweets. One possible reason is that news agencies & media are constantly covering education topics, while education is less popular among other groups.

Top Twitterers	Count	Influence Score*
NJSchoolJobs.com	5400	5399.7
Kevin Edwards	3118	3422.8
Larry L. Robinson - Free - Education - University	2362	1426.3
MrJBTM	2045	1135.1
LocalHeadlinesNow	1882	1135.1



Note: [1] the “Other” category is dropped for better visibility

[2] Definition of “News & Media” consider (1) news and the press, (2) account of or related to social media platforms with 100k+ followers (like YouTube and Tiktok)

[3] Definition of “Celebrity & Influencer” consider (1) celebrities (i.e., actors, athletes), (2) social media influencers with 100k+ followers)

* Influence Score: weighted score from total tweets and retweets times the average tweet engagement influence of one twitter user. See more discussion and definitions in the appendix

Most Influential Twitter (Top 3) by Influence Score and Organization

- Previous slides implied that using only retweet/tweet count is not a reasonable method to find influential accounts. Hence, a redesigned Influence Score standard was used to identify influential accounts with a weighted score by engagement and number of followers
- **Most results are checked by real-world facts**. For example, Obama, PM Modi, and Sen Sanders are the most influential accounts in the Government Entities group. Another good example of identification is the users in NGOs–this group has less data and thus has less noise
- However, some categories have less accuracy rate. Schools are highly mixed with high school sports, plus numbers of school accounts are low
- All users with organization are verified users

Government	Influence Score
Barack Obama	13198605.7
Narendra Modi	8220988.4
Bernie Sanders	1555190.0

News & Media	Influence Score
YouTube	7593304.2
CNN	5984844.7
New York Times	5317510.3

Universities*	Influence Score
Jill Biden	413454.7
Harvard University	148964.1
Stanford University	92496.1

* Some results under University doesn't fit the fact. Two cells are removed and the reasons are in the appendix

NGOs	Influence Score
Bill Gates	6059230.9
Kiran Bedi	1225880.2
ACLU	200484.0

Schools	Influence Score
Josh Shapiro	19988.4
Texas HS Football	15301.6
SportsDayHS	13645.5

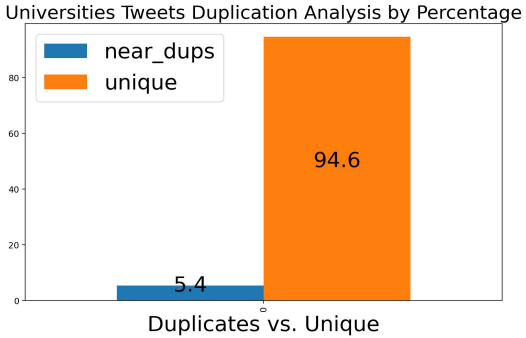
Celebrity & Influencer**	Influence Score
Selena Gomez	6589736.3
NASA	6487937.6
ESPN	6487937.6

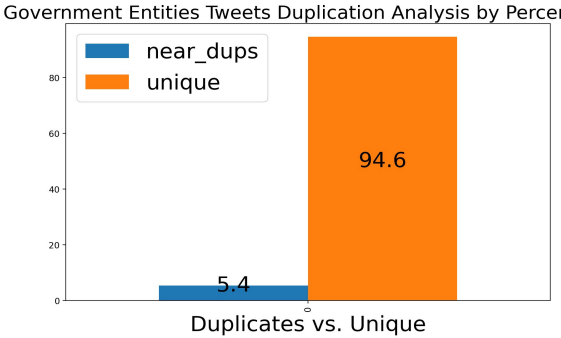
** This category also had minor identification error since we include celebrities and influencer accounts with 100K+ followers. Languages and emojis in user descriptions also disturb identification.

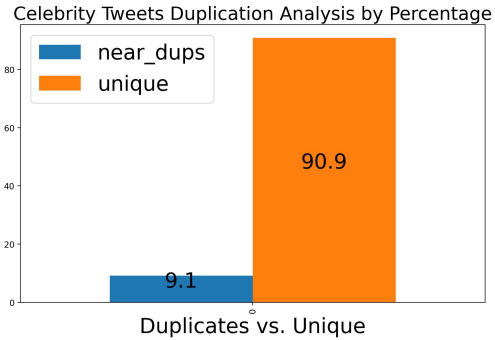
Original Content Similarity Analysis

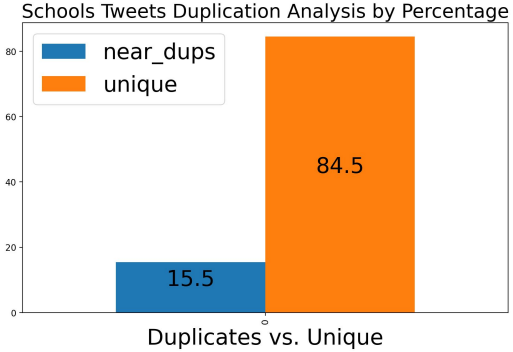
All organizations had a high percentage of unique tweets, which indicates that they are posting original content.

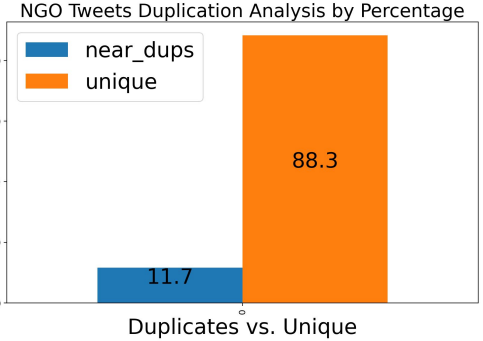
- **News & Media** had the least percentage of duplicate content, which indicates that they might be sharing similar education information. **Schools** and **NGOs** also have lower results of unique content
- **Universities**, **Government**, **Celebrities** (celebrities & social influencers) had higher percentages of original content, which is more likely to tweet about their own content

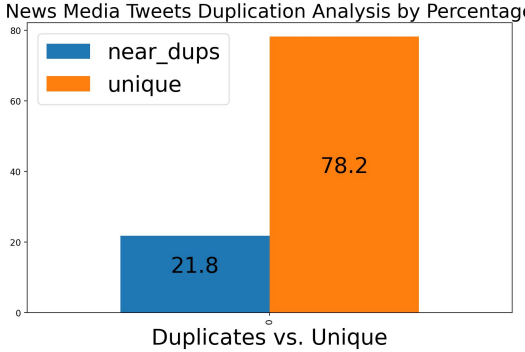












* Since we intentionally controlled the volume of filtered data, we are able to run a full LSH similarity test for most groups (groups with over 30k data will randomly take 50% of its data)
** Jaccard Similarity Distance = 0.5

Conclusions

- **Verification**: Less than 2% of twitter users are verified users and they don't belong to a credible organization. Most content from raw dataset is irrelevant
    - The top twitter accounts by retweet/original tweet are mostly not verified. **News & Media** and **Celebrity & Influencers** are the groups that get most retweets
- **Influence Measurement**: The adjusted method of measuring influence has better performance in identifying most influential twitter accounts. It also cross-validates that twitter accounts with less retweet amount still have huge influence
    - For example, @BillGates has lower retweets but much higher influence scores than a TikTok bot. However, improvements in classification of user's organization are needed for groups like highschool and influencer
- **Content Uniqueness**: Text similarity test shows that all organizations are creating unique content, except News & Media are sharing more similar content which is comprehensible due to the nature of their role
- **Reflection of trending topics**: The daily tweet count and geological distribution of twitter users in the U.S. check that extreme tweet volume correspond to new hot topics (i.e., student debt and school shooting). However, for new topics with less societal influence like #SpecialNees and #StudyfromHome, they are likely to be flooded by other hot topics like sports and a vast volume of irrelevant data. Overall, it is easy to find breaking news on twitter, but more difficult to find less heated topics

Recommendations

- **Reinforcing more resources on verification and fact checks**: topics under the trending and recommendation should have more validation on twitterers' identity and credibility. The news that get widely spread needs in-time fact checks, especially from unverified users
- **Influence Scoring and Steaming**: our preliminary influence scoring method indicates that tweet/retweet counts cannot reflect true influence and a better algorithm is needed to prioritize valuable information to obtain social attention. Verified and official accounts (i.e., WHO, CDC) deserves priority when Twitter streams news and topics
- **Disinformation List**: it is difficult to accredit a twitter account as "true-teller" given the vast user population, however, Twitter could tag twitter accounts that have records of spreading misinformation or hated speech.
    - For example, Twitter could add the tag under account description: "Please be aware that this user has recently posted more than 5 times of misinformation or hated speech"

Appendix

Details of methodology:

1. **Terms of subject.** This is an individual assignment but the writing sometimes use the subject term “we”.
2. **EDA**. Results of EDA includes counts of total tweet and twitter users (by country and location). Exploration of usable variables are in the Jupyter Notebook called “final_2 EDA”.
3. **Filtering Tweet**. We used keywords for hot topics in 2022 that covers *racial equality, literacy, tech & digital, special need, school curriculum, higher education*. For example, keywords for special needs are *'ece', 'specialneeds','dyslexia','tck'*
4. **Tagging Organizations**. We use user name and user descriptions to identify the organization.
    a. For example, terms for government entities include variations of words like *state, senate, ministry, and department*
    b. Celebrity and Social Influencers. We limit to users with over 100K followers and filter terms like player, musician, athlete. However, this condition also collect users that are not people, such as tutoring organizations or companies
    c. Non-verified user are tagged as “Other”
5. **Location Analysis**. A huge amount of tweets have null values in their geological information, thereby we dropped them.
    a. Country. Names of countries might be other than English, iso_2 code from twitter API is used for geopandas and plotting
    b. States. Name of states in Twitter API is not consistent like (Texas, USA) or (Huston, Texas). Names are unified and tweets with unidentified locations are dropped
6. **Similarity Test**. Original text content was used. Due to strict filtering, the dataset for each group is small at an average of 15K level. In case of huge dataset, a random sample of 50% will be extracted for LSH Similarity test.

Appendix: Most Influential Account and the Influence Score

With the help of our TA sessions, it is suggested that we can have our own standard of evaluating influence rather than just number of retweets. Some notes are made here to explain how do I addressed this question.

1. Most Influential Account by Influencer Score. I removed two accounts to avoid keyword filtering errors:
   - Recording Academy is arguably to be an university. It's famous for the Grammy Award
   - Rauf Klasra is a journalist and a columnist who writes about language. He is barely a staff at university
   - I kept Jill Biden because she is an educator who have been worked for universities in decards

Even we removed the two accounts, we can see the influence score standard is able to identify Harvard, Stanford, and Oxford as the most influential accounts in the organization type of university

Universities(Top 6)	Influence Score
Jill Biden	7593304.2
Recording Academy	6589736.3
Rauf Klasra	6487937.6
Harvard University	138561.9
Stanford University	92496.1
University of Oxford	88169.2

2. How do we calculate the influence score?

I used the Engagement Method to calculate influence.

Firstly, we want to know the influence of each tweet. It is called *the Influence Coefficient (alpha)*, which is the level of engagement. Retweet and Quote are considered as actions of spreading information, thereby they are seen as results of being influenced. All retweets and quotes are recorded in the source tweet, which is why we are using source user id and these variables.

$$\alpha = \frac{Retweet + Quote}{Favorites + Retweet + Quote}$$

$$Influence\ Score = \bar{\alpha} \cdot (1 \cdot Total\ Retweet + 2 \cdot Total\ Original) \cdot 99\% + Followers \cdot 1\%$$

As we are able to calculate how much engagement one tweet of each user can lead to, we will also calculate the average influence coefficient of all of their tweets. Then we calculate the total Influence Score for a user by using the number of original tweet and retweet (whether retweeted is RT).

Note that original tweet is rewarded with a weight of 2 because it takes more effort to do so–otherwise, a tweet bot will be more influential than content creator. The number of followers are also used as a weighted factor of the influence score.

* The PowerPoint file does not support Markdown format. A detailed discussion on methods of calculating influence score can be found in the ipynb file called '*final_3b Influence use original_id*'