IS Twitter a Credible Source of Educational Information?

Big Data Platforms

Final Project Education

Topic: Twitter Information Credibility Analysis

on Racial Injustice in Education

Presenter: Zhilin Yang

December 7, 2022

Executive Summary

1. There are a lot of non-verified twitters;
   Therefore, most of the twitters can not be recognized as a credible source of information

2. It is a wiser choice to use influence score instead of Retweeted/tweet counts when conducting the influential analysis.

3. The timeline analysis showed that the number of tweets is positively related with the trends or the level of discussion in education tweets. However, it is also very important to be careful when making such conclusion because the data results may cheat the data scientists sometimes

4. The similarity Analysis showed that most twitters are creating their original tweets.
   However, social media influences prefer just copy-pasting the text without generating their original tweets.

5. The location analysis showed that the United States has the greatest number of tweets related to racial injustice in education; There are indeed a lot of racial inequality issues in the United States.
   Therefore, the location analysis can be regarded as a credible source when conducting influential analysis

Methodology

Step One

Filter the Raw Data
By searching keywords

Step Two

Exploratory Data Analysis: Picking the features and processing the 'NA's in the data

Step Three

Setting the rules to categorize the Twitters' organization by keywords or number of followers

Step Four

Influential Analysis on Twitterers by different approaches, generated influence score to better evaluate influential level

Step Five

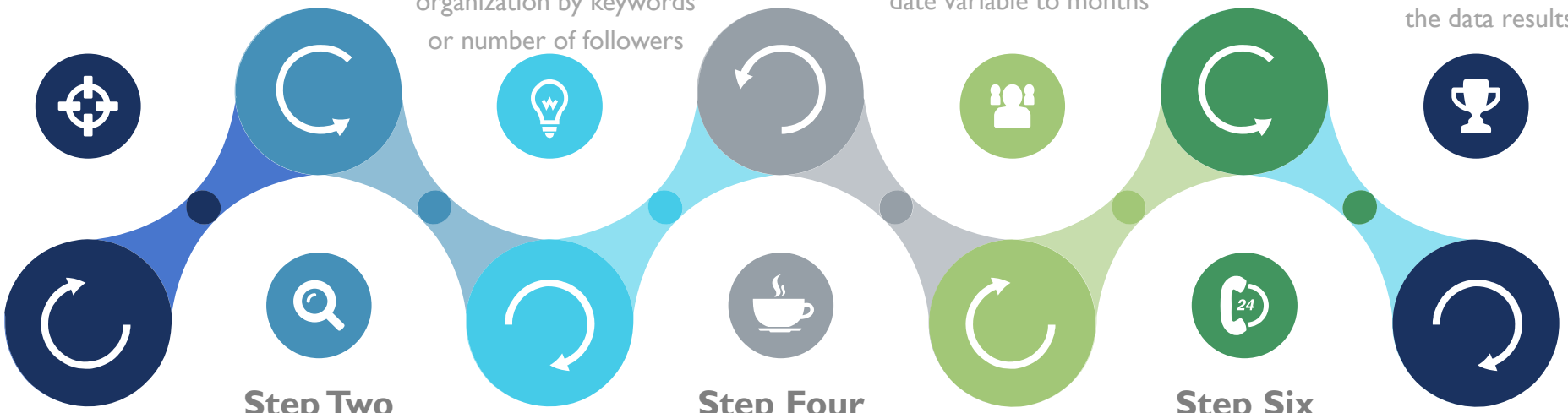
Visualization of Analysis Results, categorize the date variable to months

Step Six

Run the duplication analysis among original tweets to further investigate data

Step Seven

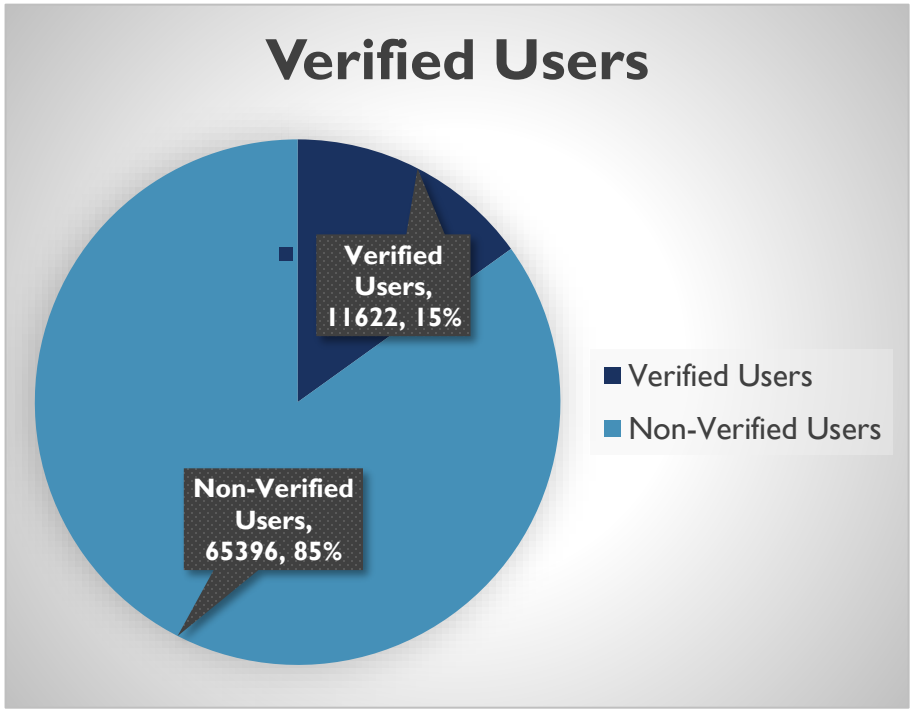
Cross Validate results; Research the trends and insights implied in the data results



Data Overview

- Filtered the data through searching for keywords
- Picked 19 variables and restored the intermediate results into the GCS bucket

Original Data	Filtered Data
99.99 Million Tweets	1.05 Million Tweets
April 05, 2022 to November 06, 2022	April 05, 2022 to November 06, 2022
Broad Education Tweets	Tweets Related to **Racial Inequality in Education**

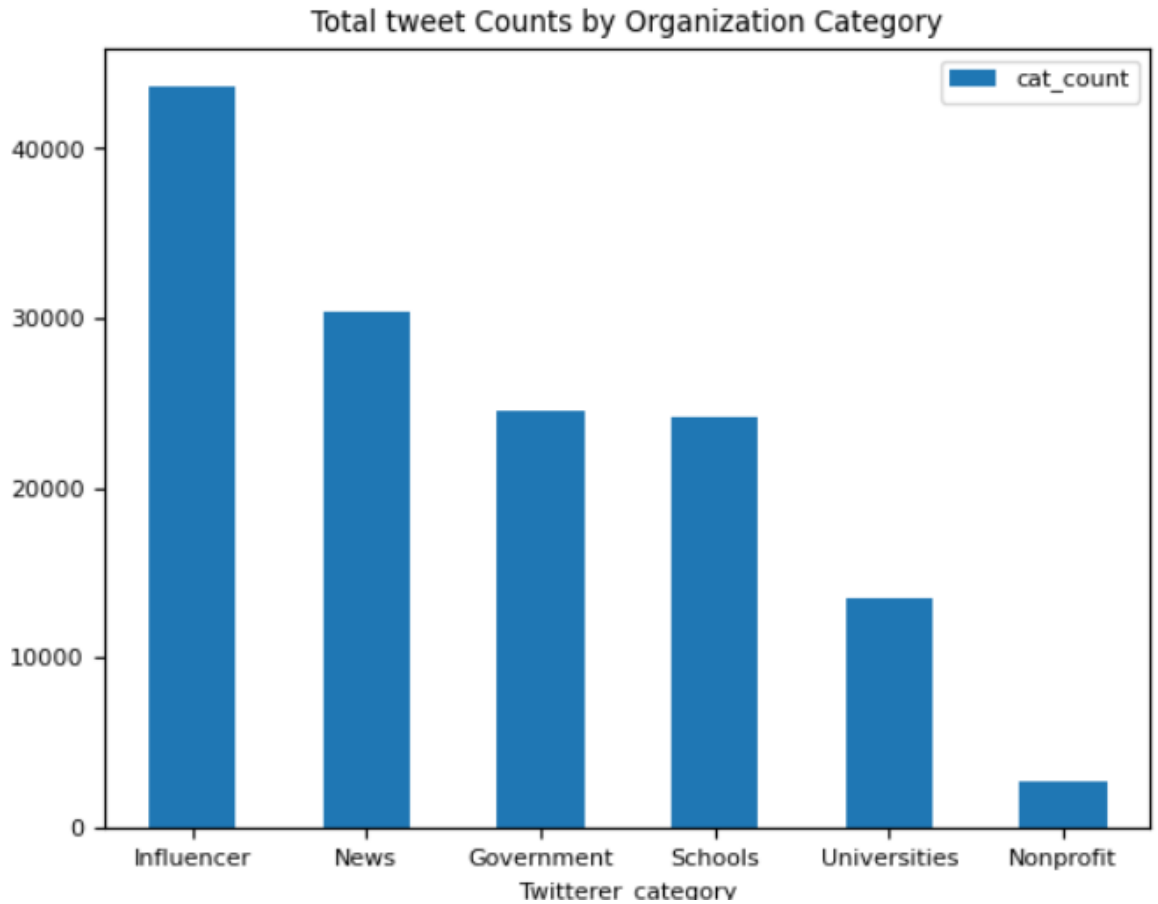



```
keyword = ['[Rr]acial','[Dd]iversity','[Rr]acial slur','[Cc]ollege racial discrimination',\
           '[Rr]acial preferences','[Aa]sian','[Aa]frican American','[Dd]iscrimination',\
          '[Ii]eauality','[Rr]acial [Dd]iscrimination','[Bb]lack [Ll]ives [Mm]atter', 'BLM','[Ee]thnic [Pp]reference(s)?',\
          '[Rr]ace','[Ee]quity','[Rr]acism','[Ww]hite [Pp]rivilege','[Ee]thnicity','[Bb]lackedu','[Ll]atinoedu','[Nn]ativeedu']
```

Influential Twitter Analysis – By original message volume

- The Social Media Influencer has the largest number of total tweets among organizations
- The Schools and Universities have a relatively small number of tweets:
  - The tweets closely-related to education are not coming from direct sources such as universities and schools

username	UserVolume
GraceSmartsBlog	1205
Grace Smart	931
Graceville WxSTEM	374
Abinaya Devi	303
Jordan Lindsey	282
Dr. Lena Gould, EdD, CRNA, FAANA, FAAN	264
Greg Grace	256
Mitch Zawaski	231
Jonathan Gelber, MD, MS	167
adam (READ PINNED) #buyingcontent	162



Influential Twitter Analysis – By message retweeted frequency

- Compared with ordering by total number of retweeted massages, ordering by the influence score will generate **results** with Higher Accuracy for the recognition of influential Twitterer
- Another interesting insight was found during EDA: The number of quotes should also be counted as a kind of retweets, because quote is in a 'retweets + comments' format but it still includes the content of retweet

username	influence_score	total_retweeted
Fansly.com/Littlesubgirl	1360.090952	195
Patreon.com/Littlesubgirl	2203.598327	195
James Finn	11492.967522	159
Alex 1984 vs 1776 us	2058.681947	108
Gulag Inmate 93271 us	354.922022	108
1984 vs 1776 (Alex) us	1485.242961	108
1984 vs 1776 us	253.048930	108
F* Your Short Memory 🏴 Anti-Cult of Absurdity	3419.837536	95
Babe Perales	7529.905456	86
Katarzyna Ski 1776	177.522990	81

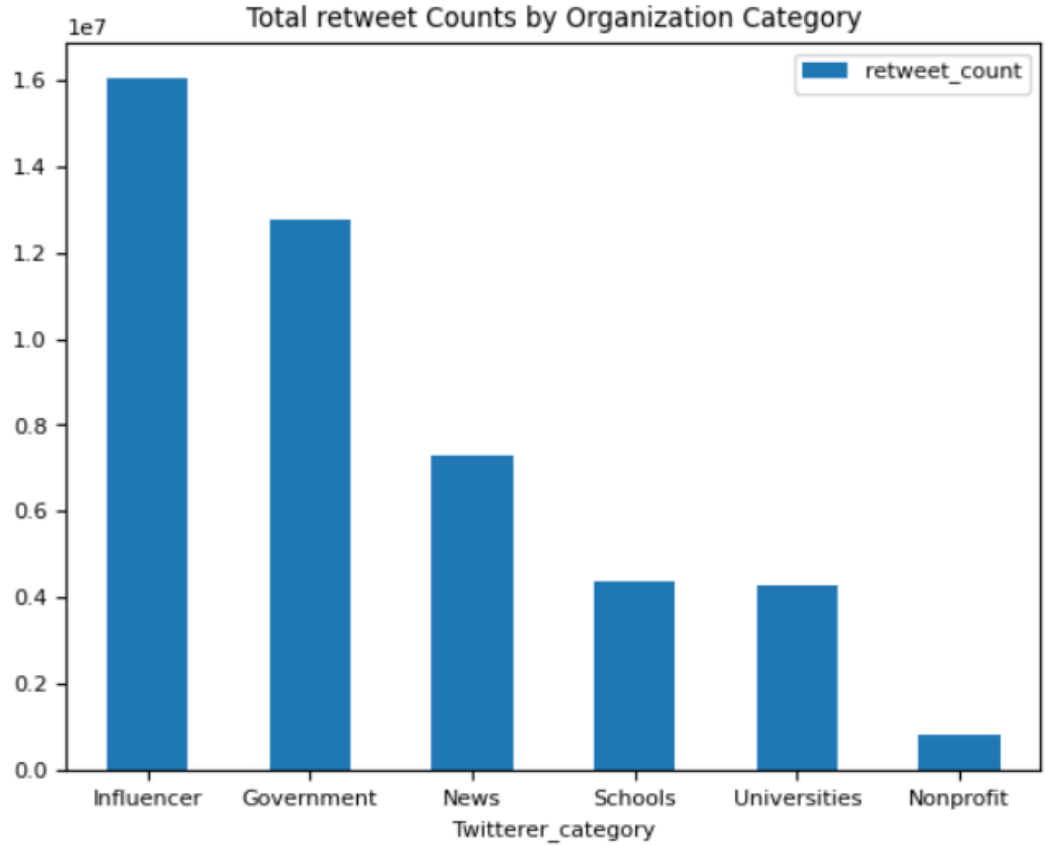
Order By total retweeted

username	influence_score	total_retweeted
James Finn	11492.967522	159
Babe Perales	7529.905456	86
Stigmabase | ORG	5167.500000	27
Marci Maier	4488.000000	66
UNDERCOVER MOTHER #ExposeNAIS	4032.242012	79
Carl Hilberg	3623.699063	29
F* Your Short Memory 🏴 Anti-Cult of Absurdity	3419.837536	95
Soph	3019.082171	45
Dan Kleinman of SafeLibraries® 🟧	2657.856797	74
James Cooper	2636.303662	30

Order By Influence Score

Influential Twitter Analysis – By message retweeted frequency

- In comparison with the histogram in the 5th slide (Total tweet count by organization category):
  - It is very interesting to notice that the ranking of news retweet count increased to the 3rd place;
  - while the ranking of school retweet count decreased to the 4th place In addition,
  - the ranking of all other categories remained the same; which might imply that the number of tweets and the number of retweets are positively related



Calculation Methodology of Influence Score

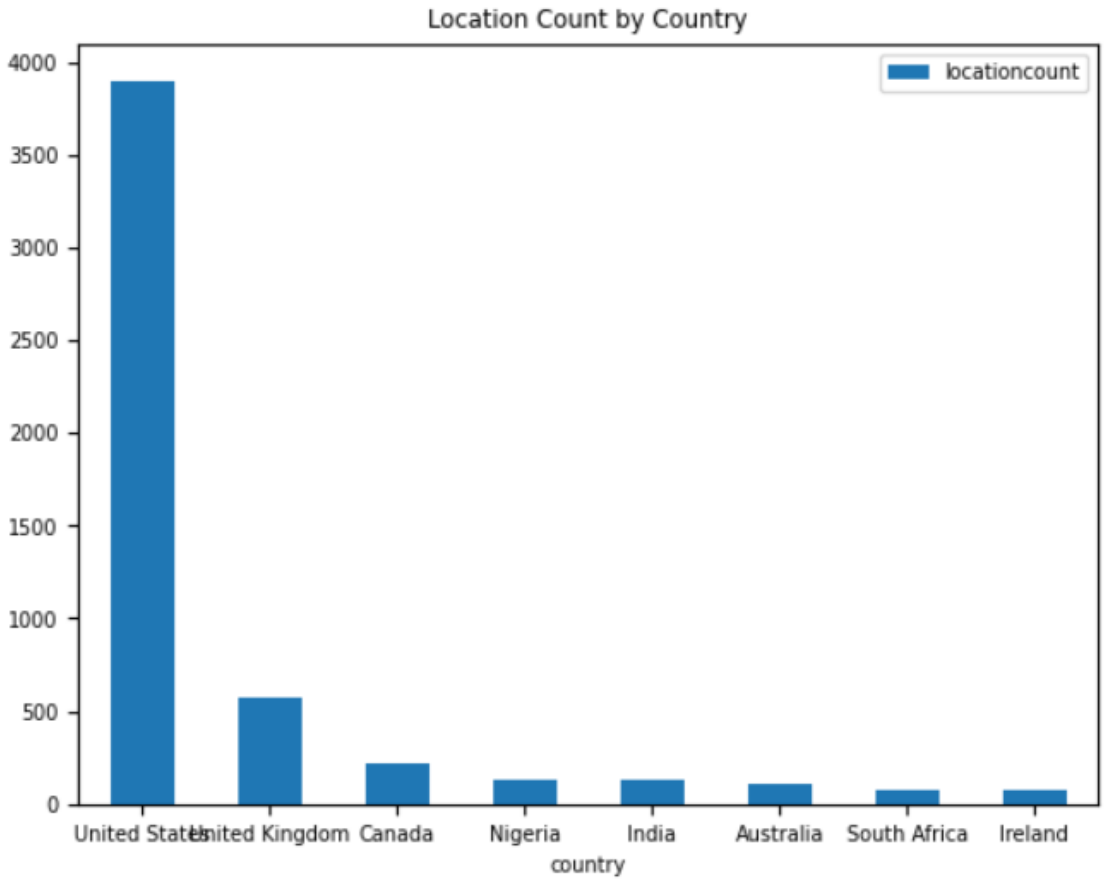
① $$Engagement\ Rate = \frac{(Quotes + Retweets)}{(Quotes + Retweets + Favourites)}$$

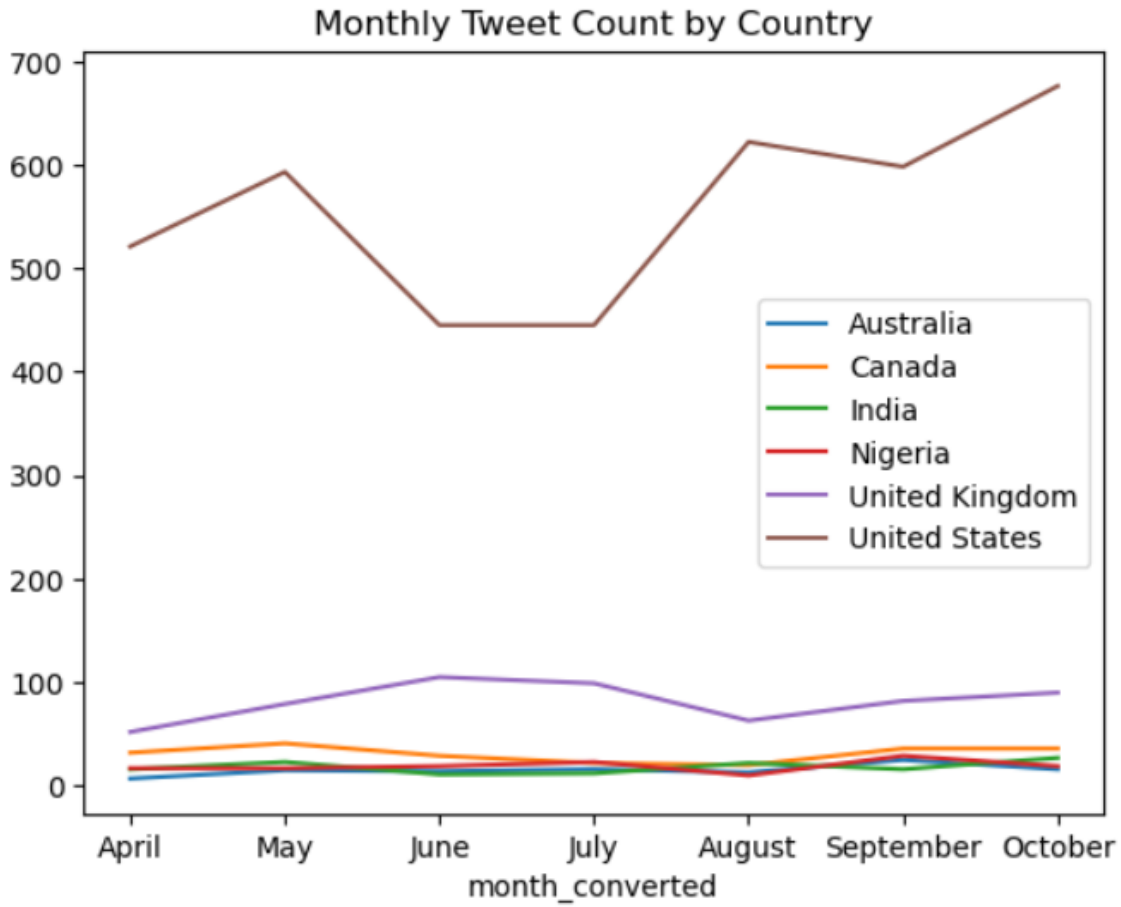
② $$Total\ Engagement = Engagement\ Rate \cdot (Retweet\ Count \cdot 2 + Original\ Count \cdot 1)$$

③ $$Influence\ Score = Sum\ of\ Total\ Engagement\ of\ each\ unqiue\ Twitterer$$

Location Analysis

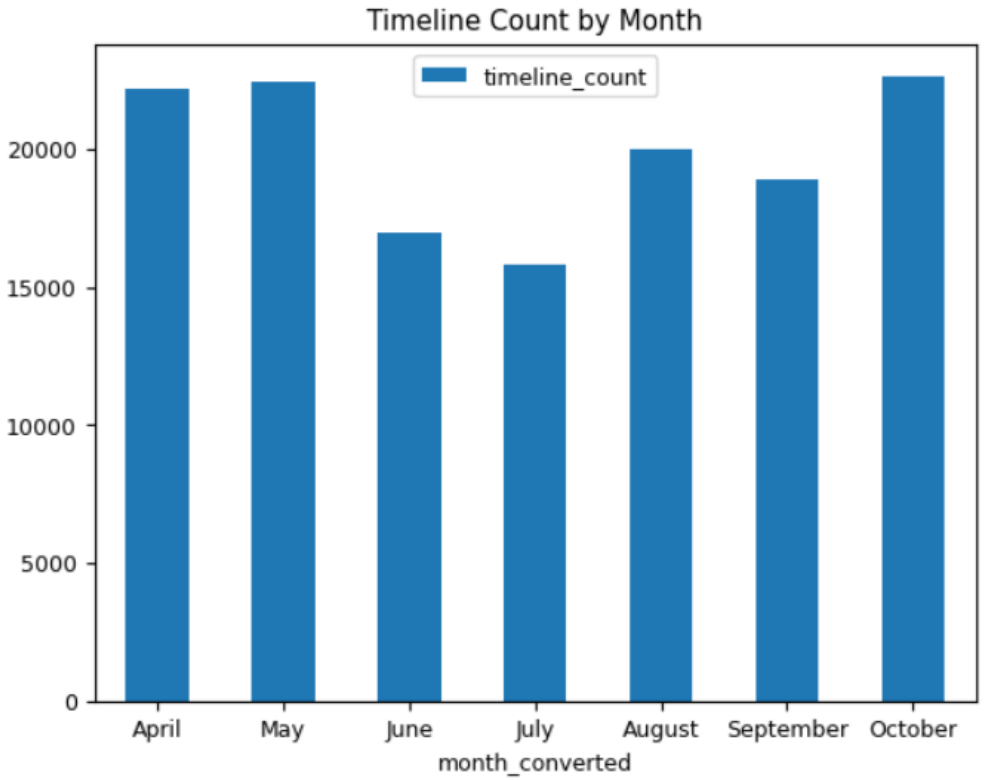
- The United States has the largest number of tweets related to racial injustice in education
- In May and August 2022, there are obvious boosts in the number of monthly tweet count in the United states
- The United states has the most diverse ethnicity among the top 8 countries, which matches the result that it has the largest racial injustice problem in education based on the number of related tweets published

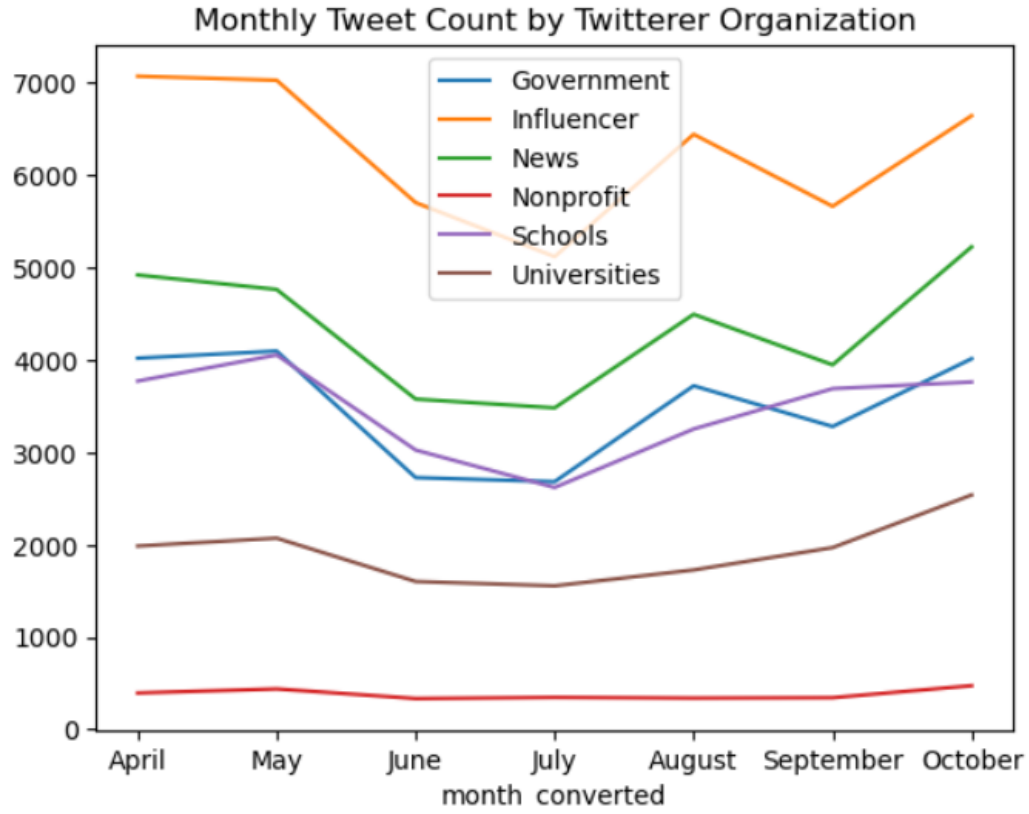




Timeline Analysis

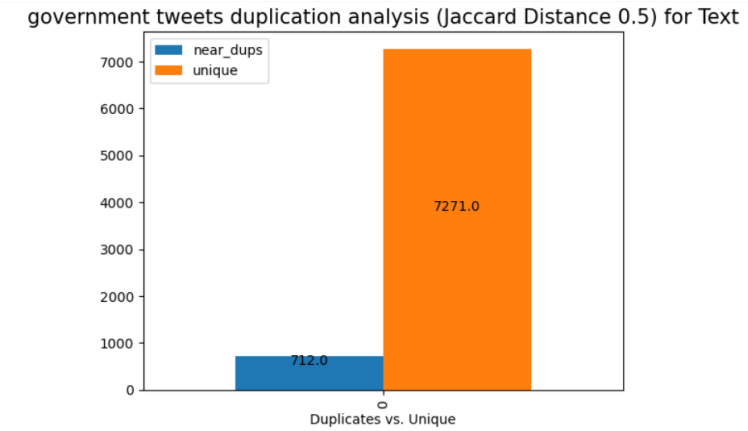
- May, August and October have the highest monthly tweet count in almost every organization.This result brought me to a further investigation
- I further investigated the hot topics in racial injustice in Education, I found that there exist a general upwards trends of discussion in May and August as well. These two months are the time that students will **leave or come back to school**. Here is one example of such topic happened in August 2022.
    - The United States Government signed the International Convention on the Elimination of All Forms of Racial Discrimination in 1966. In August 2022, the Committee on the Elimination of Racial Discrimination(CERD) will examine the reports by the United States on compliance with the Convention

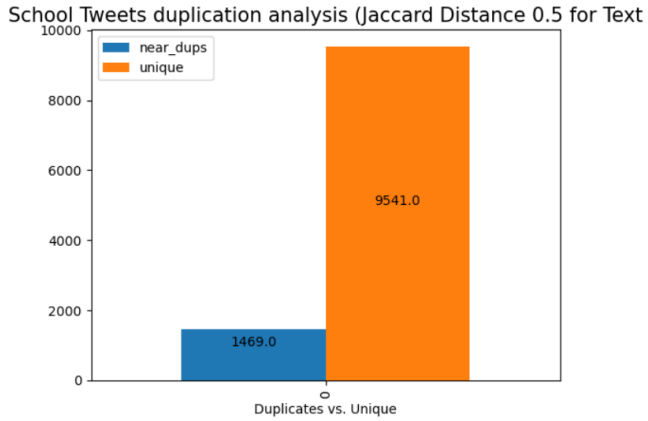


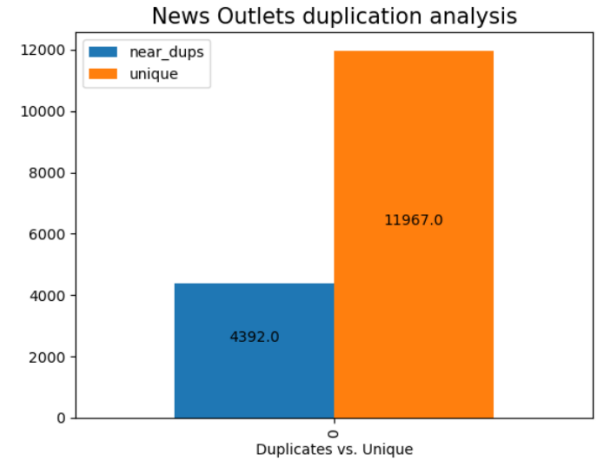


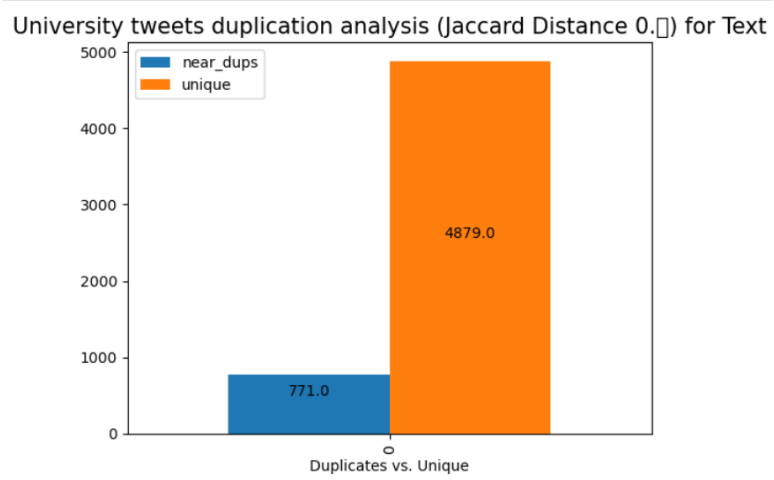
Message uniqueness analysis

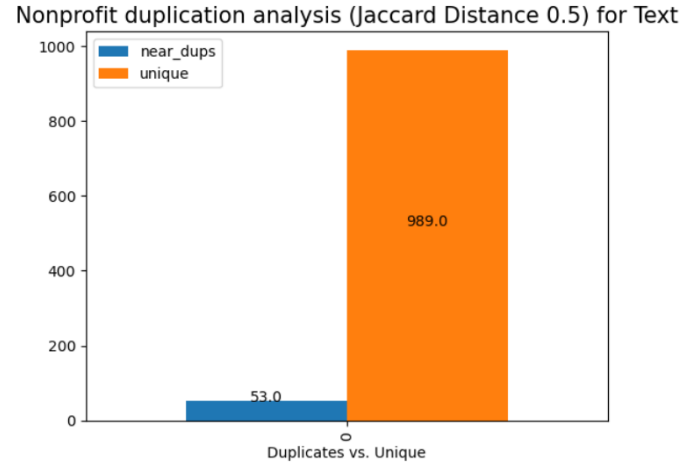
- All organizations expect social mediahave a much larger number of unique tweets than the number of duplicate tweets
- The duplication analysis is conducted by LSH method with a Jaccard Distance 0.5 ; the analysis is processed on all the original tweets
- Social Media has more duplicate tweets, which means that the social media twitters tend to just copy-paste the same text when they are publishing a tweet

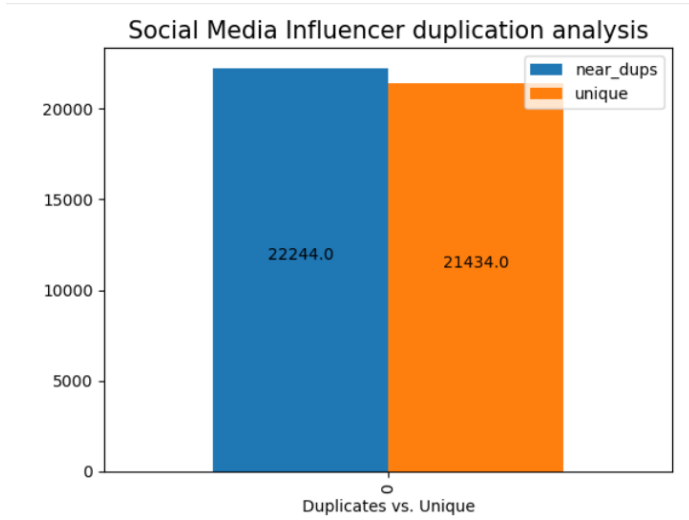












Conclusions and actionable recommendations

Conclusions:

1. Most of the twitters are non-verified twitters, so they can not be regarded as a credible source of information.
2. Social Media Influencers has the highest level of influence according to the tweet count and retweet count analysis histogram
3. It is a good choice to substitute the tweet/retweeted count with an influence score when conducting the influential analysis. Because it can generate more accurate results
4. Timeline analysis on the number of tweets by different organizations or countries is a good way to indicate the treads of new topics.
5. Most of social media influencers are sharing information based on others' thoughts instead of creating their own ideas. Therefore, it is not a good choice to trust social media influencers although they have a great level of influence

Recommendations:

1. Twitter can introduce a influence score when deciding whether to label an account as 'verified user' automatically.
2. Location and Timeline analysis can be included when determining the hottest trends or the most popular 'hashtags', which might generate more accurate recommendations for users
3. There should be an automated system or a badge which can help users clearly recognize the AI or bots accounts as they are normally not generating credible information
4. Twitter might need to generate a more developed hashtag system so that users can find their wanted information more quickly