TWITTER ANALYSIS

Harsha Muthukuru

Prepared for Big Data Class- 9 March 2023

Agenda

1 Executive Summary

2 Methodology and source data overview

3 Tweet cleaning and filtering

4 EDA

5 Author identification

6 Location analysis

7 Timeline analysis

8 Message uniqueness analysis

9 Conclusion

IS TWITTER A CREDIBLE SOURCE?

- Twitter is a social media platform that allows users to share short messages called “tweets” with over 300 million active users and is available in more than 40 languages.
- To what extent can Twitter data be considered a reliable and safe source for educational purposes, given the diverse user base, the absence of credibility and limitations, and the possibility of noise resulting from fake accounts and bots?
- The project concentrates on whether to identify twitter can be considered as a credible source of information, which reflects the emergence of important trends or topics in education and profiling twitterers.
- The objectives of this project include answering the following research questions:
  - Who are the Twitter users relevant to a particular topic of interest?
  - What are the most prolific and influential Twitter users within this community?
  - Where are these Twitter users located geographically?
  - What is the timeliness of tweets, including significant peaks and valleys in activity? and
  - How unique are the tweets in terms of their content and perspective?

METHODOLOGY

- Started with finding the required variables and filtered the data based on the keywords generated by ChatGPT
- Based on the keyword filtering of user_name, user_description and Followers count segregated them to organization named NEWS/ Universities/ Schools/ Influencers/ Government Entities/ NGO's
- To find the location I had to group by user_location which resulted in half a million locations which is quite huge for the API that I am used Hence, took the top 300 locations where 50% of the users are located and found the country of that location

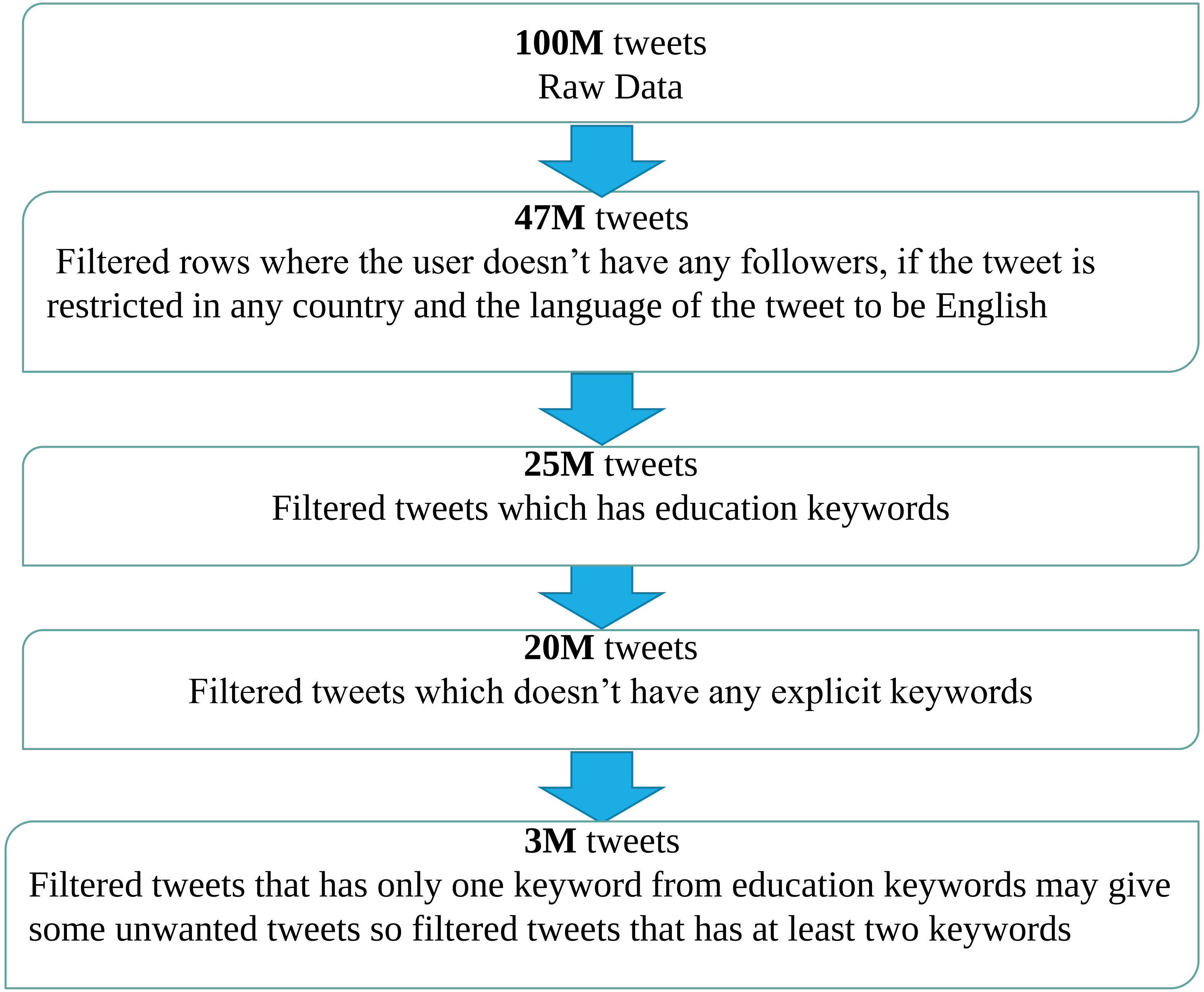
Technologies used:

- The project utilizes a range of technologies
  - Google cloud
  - PySpark for coding
  - Pandas for small data analysis plotting
  - LSH for text analysis
  - GeoCoder for country filtering, matplotlib, nltk, geopandas etc for the analysis and plotting

Source Data Overview:

- The source data for the project is collected periodically from the Twitter API and spans from Feb 2022 to Feb 2023, comprising over 100 million tweets.
- The data is stored in Google Cloud and contains 41 columns, with most columns consisting of null values that do not provide any useful information.

FOCUSING ON EDUCATION AND ENGLISH TWEETS



100M tweets
Raw Data

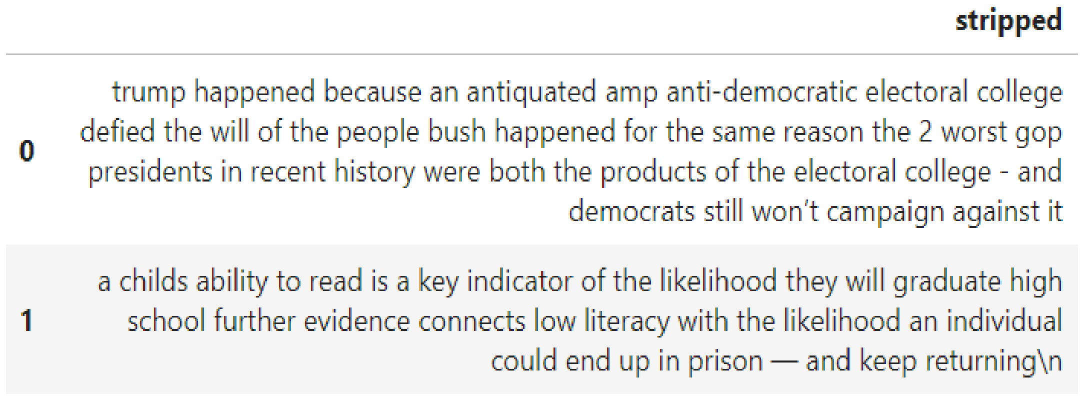
47M tweets
Filtered rows where the user doesn’t have any followers, if the tweet is restricted in any country and the language of the tweet to be English

25M tweets
Filtered tweets which has education keywords

20M tweets
Filtered tweets which doesn’t have any explicit keywords

3M tweets
Filtered tweets that has only one keyword from education keywords may give some unwanted tweets so filtered tweets that has at least two keywords

Sample Tweets after filtering

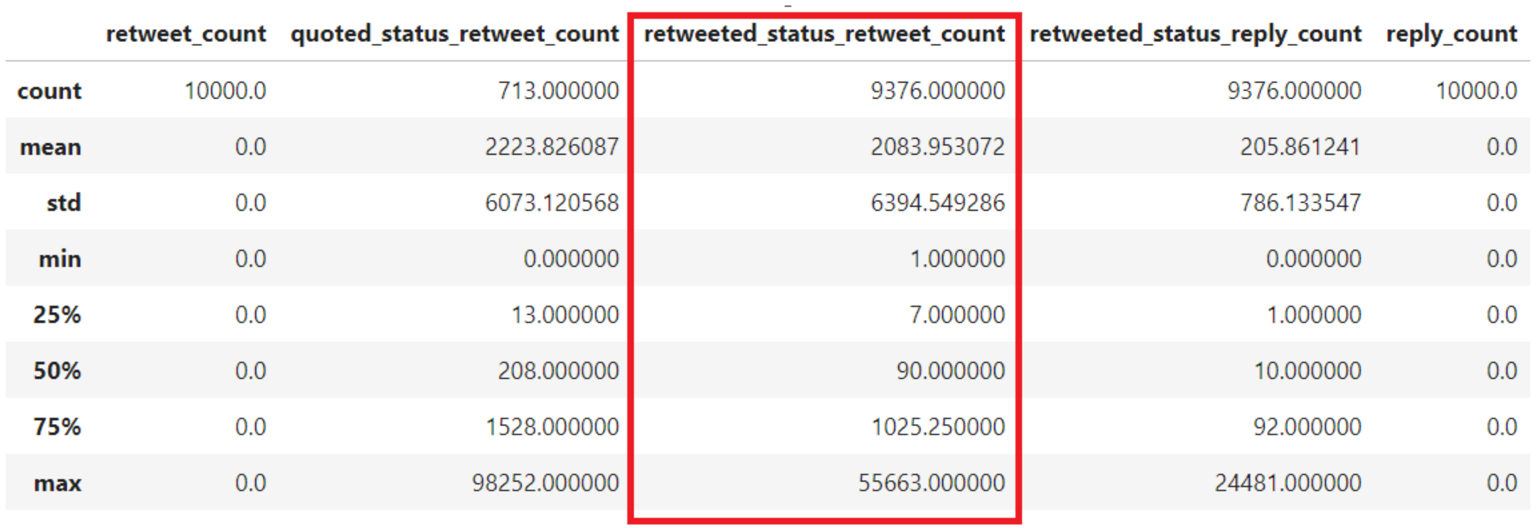


	stripped
0	trump happened because an antiquated amp anti-democratic electoral college defied the will of the people bush happened for the same reason the 2 worst gop presidents in recent history were both the products of the electoral college - and democrats still won’t campaign against it
1	a childs ability to read is a key indicator of the likelihood they will graduate high school further evidence connects low literacy with the likelihood an individual could end up in prison — and keep returning\n

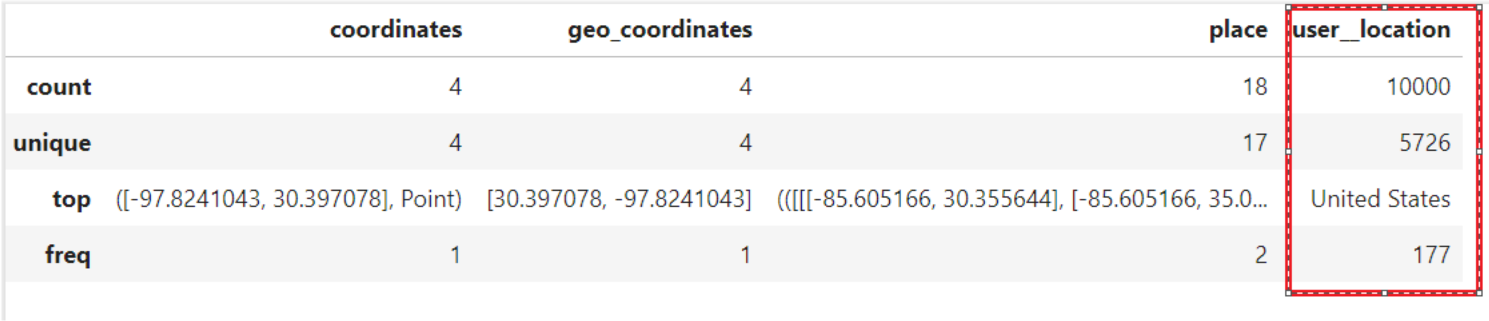
Note: Keywords used are generated by ChaptGPT

EDA: VARIABLE SELECTION

- The analysis is focused on timeline, geographical and profiling twitterers we must find columns which provide meaningful information. As there are columns with no useful information I had eliminated without any checks
- To locate Twitter users, various fields were considered, and preliminary checks were conducted to identify the number of null values in each field, providing an initial indication of their reliability in providing accurate location data.



	retweet_count	quoted_status_retweet_count	retweeted_status_retweet_count	retweeted_status_reply_count	reply_count
count	10000.0	713.000000	9376.000000	9376.000000	10000.0
mean	0.0	2223.826087	2083.953072	205.861241	0.0
std	0.0	6073.120568	6394.549286	786.133547	0.0
min	0.0	0.000000	1.000000	0.000000	0.0
25%	0.0	13.000000	7.000000	1.000000	0.0
50%	0.0	208.000000	90.000000	10.000000	0.0
75%	0.0	1528.000000	1025.250000	92.000000	0.0
max	0.0	98252.000000	55663.000000	24481.000000	0.0

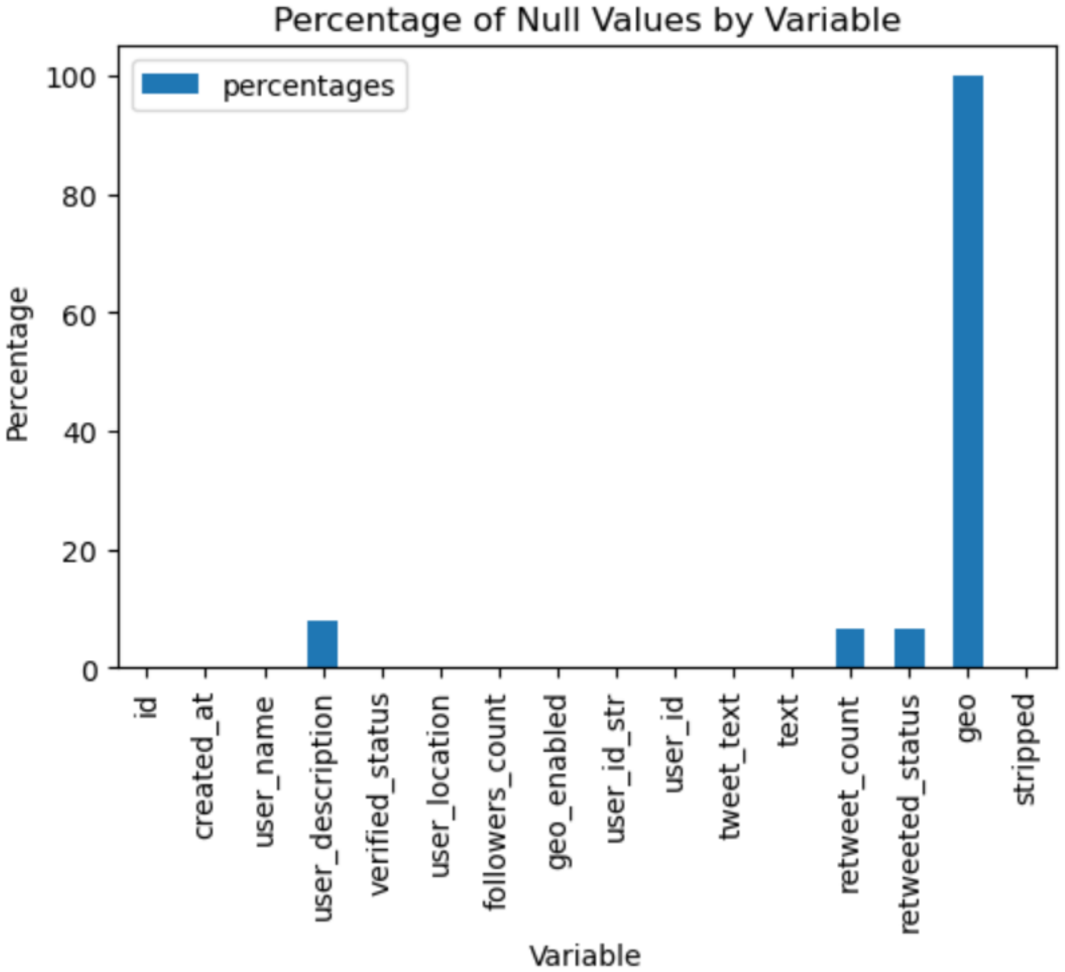


	coordinates	geo_coordinates	place	user_location
count	4	4	18	10000
unique	4	4	17	5726
top	([-97.8241043, 30.397078], Point)	[30.397078, -97.8241043]	(([[[-85.605166, 30.355644], [-85.605166, 35.0...	United States
freq	1	1	2	177

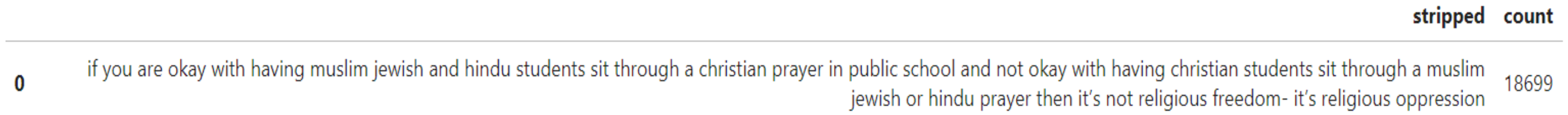
- Based on the lowest count of null values, "user_location" and "retweeted_status_retweet_count" were chosen as the most suitable fields for analyzing geographical location and retweet activity, respectively
- The "tweet_text" field was selected for message analysis over the "text" field as it provides a cleaner and more focused view of the text content by excluding embedded media or links

EDA: MOST TWEETED TWEET AND NULL VALUES ANALYSIS

- A null count check was conducted on selected fields to profile Twitter users. The resulting table indicates the completeness of the data for each variable, helping to identify potential gaps in the data for further analysis.

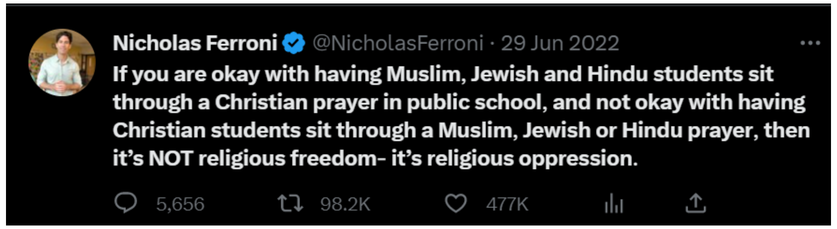


- The null values in retweeted_status represent that they are original tweets
- The "geo" field has the highest null count, with 99.96% of data missing, which is reasonable as many Twitter users choose not to share their location information
- user_description can be null as not all the users are interested to write about description



	stripped	count
0	if you are okay with having muslim jewish and hindu students sit through a christian prayer in public school and not okay with having christian students sit through a muslim jewish or hindu prayer then it's not religious freedom- it's religious oppression	18699

- The above tweet was the most tweeted tweet in the data and was tweeted by ‘Nicholas Ferroni’ on Jun 29 2022.
- This was tweeted after the “Harvard can’t keep Muslim and Hindu students in the Basement” – thecrimson



Nicholas Ferroni @NicholasFerroni · 29 Jun 2022

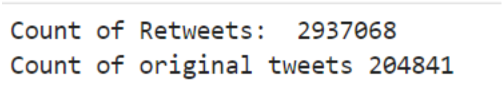
If you are okay with having Muslim, Jewish and Hindu students sit through a Christian prayer in public school, and not okay with having Christian students sit through a Muslim, Jewish or Hindu prayer, then it's NOT religious freedom- it's religious oppression.

5,656 98.2K 477K

Source: Harvard Can’t Keep Muslim and Hindu Students in the Basement | Opinion | The Harvard Crimson (thecrimson.com)

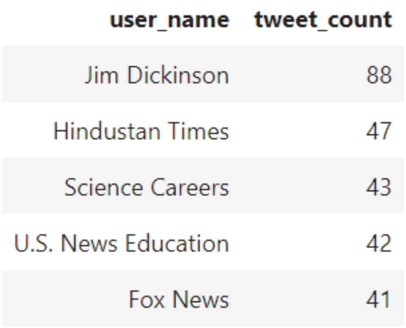
VERIFIED USERS ARE NOT GETTING MOST NUMBER OF RETWEETS

- The data shows that a significant proportion of Twitter users prefer to retweet rather than create original content, which may impact data analysis results. Thus, it is crucial to consider this trend to prevent biased or inaccurate interpretations.



```
Count of Retweets:  2937068
Count of original tweets 204841
```

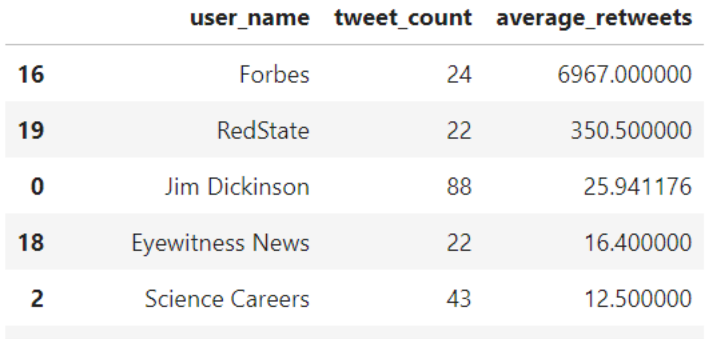
- Original tweets were identified by analyzing tweets with a null "retweeted_status" field and verified_status as true, indicating that they were not retweets. These original tweets were then grouped by "id" to identify the tweets posted by verified profiles, producing the results provided.



user_name	tweet_count
Jim Dickinson	88
Hindustan Times	47
Science Careers	43
U.S. News Education	42
Fox News	41

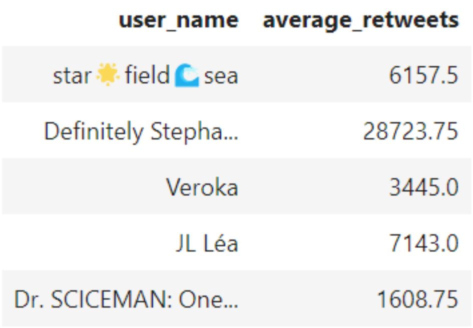
Jim Dickinson is an Associate editor at WONKHE who takes particular interest in student experience, university governance, .. Which makes sense on why he topped in the prolific twitterers. Source link

- Lets see on average how well these tweets are being retweeted for the prolific twitterer tweets. Looks like only the tweets from NEWS organizations are being retweeted more as it is a credible source and it makes sense



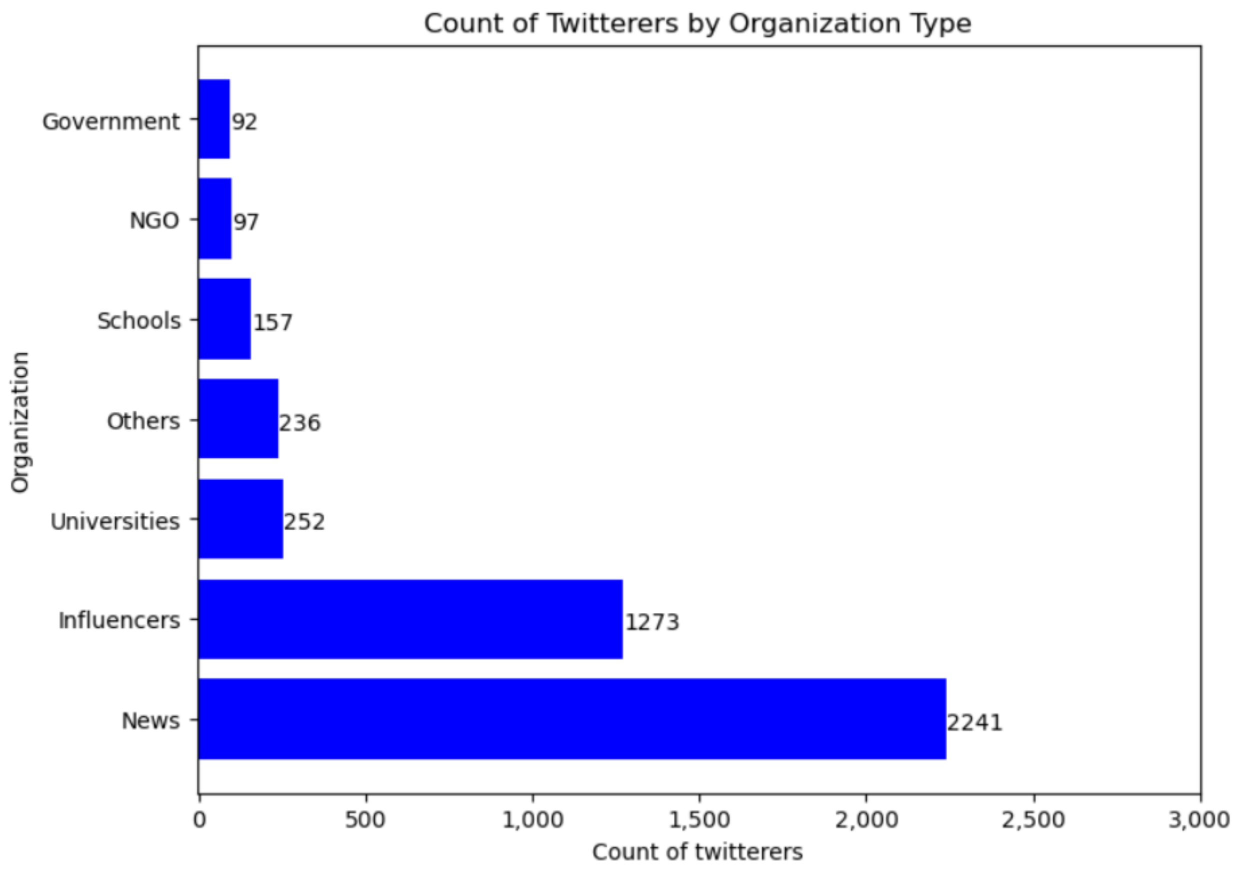
	user_name	tweet_count	average_retweets
16	Forbes	24	6967.000000
19	RedState	22	350.500000
0	Jim Dickinson	88	25.941176
18	Eyewitness News	22	16.400000
2	Science Careers	43	12.500000

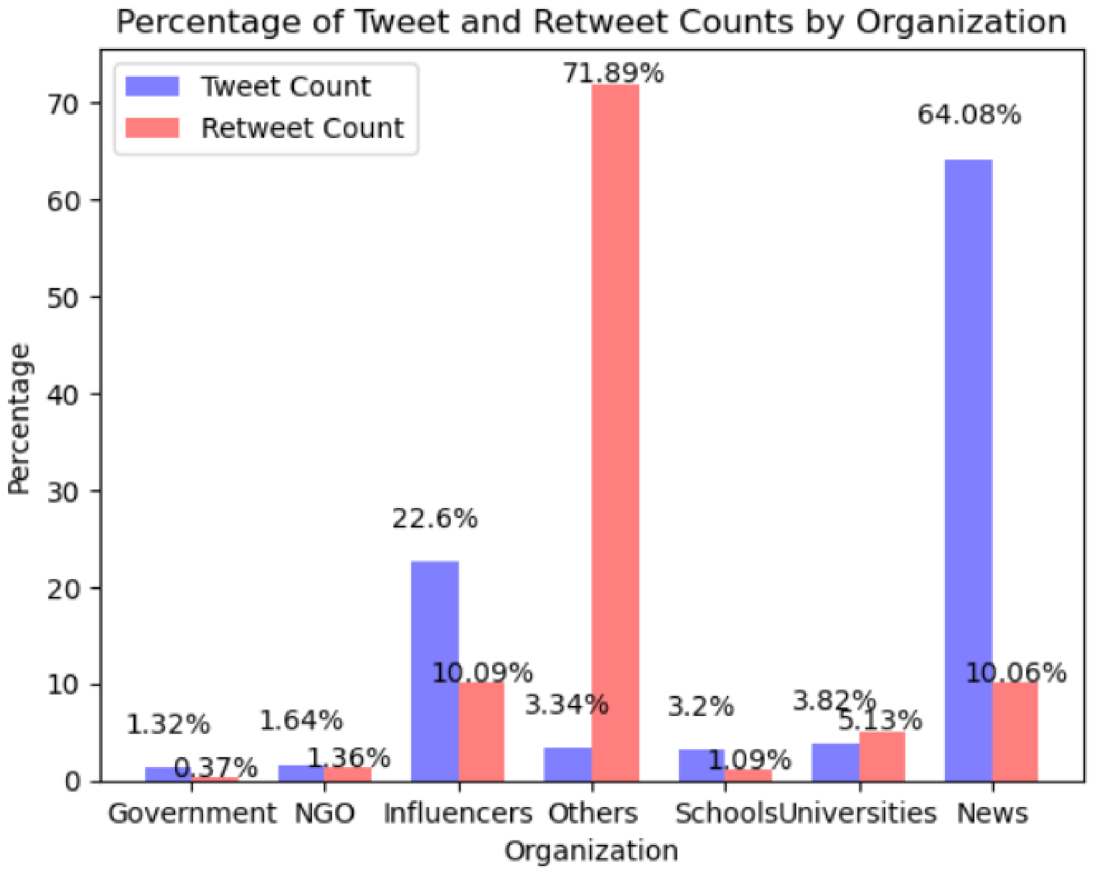
- However if we check the average retweets for all the users they are high when compared to the verified twitterer tweets



user_name	average_retweets
star☀️field🌊sea	6157.5
Definitely Stepha...	28723.75
Veroka	3445.0
JL Léa	7143.0
Dr. SCICEMAN: One...	1608.75

MAJORITY OF THE TWITTERERS BELONG TO NEWS CATEGORY



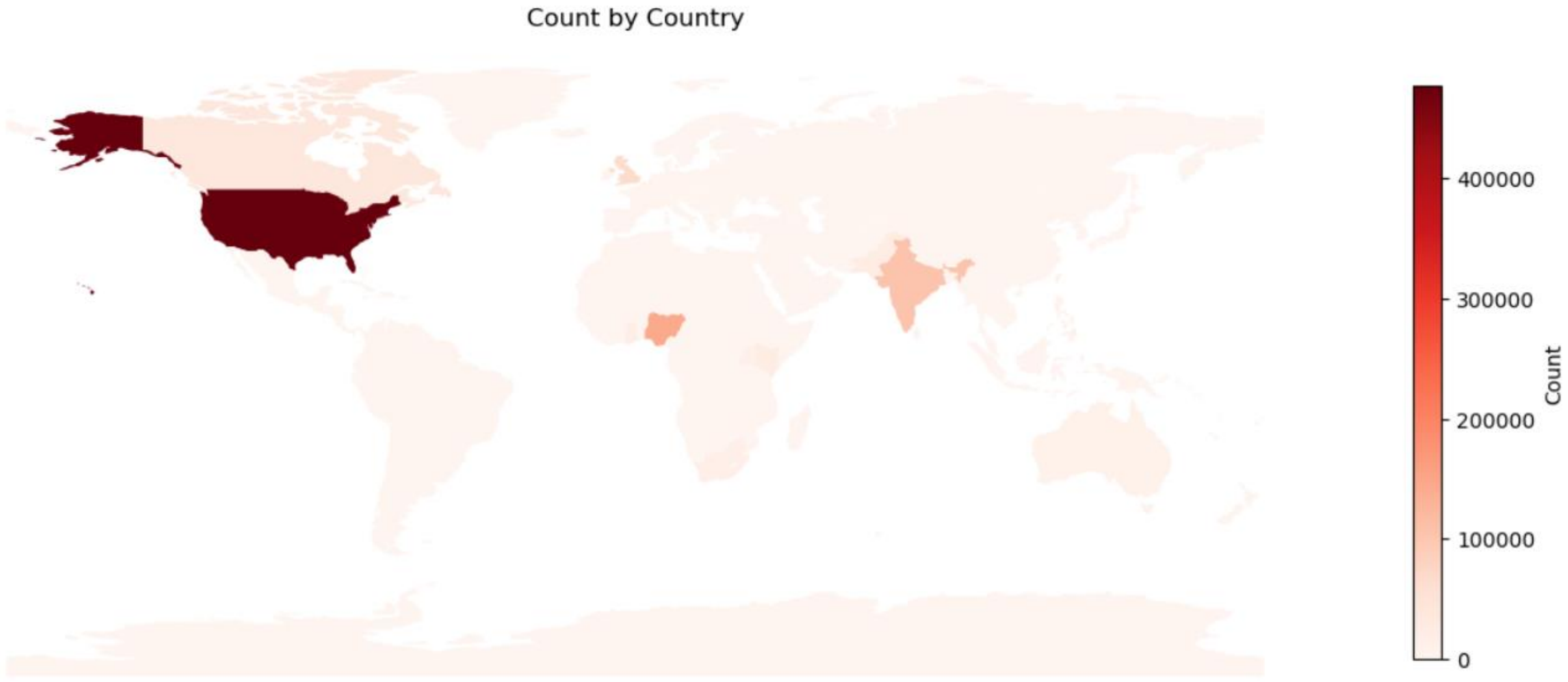


- Of the prolific twitterers (Verified users) I tried to check to which organization the twitterers belong to
- It is not surprising to see NEWS on the top as 4 out of top 5 twitterers are NEWS organization handlers
- Government twitterers are the least with NGO's being next

- From the above graph we can say that individuals would be retweeting more as the than posting original content related to education
- One conclusion from this is the people in organizations would concentrate more on original content which makes individuals engage

MAJORITY OF THE TWITTERERS ARE IN US

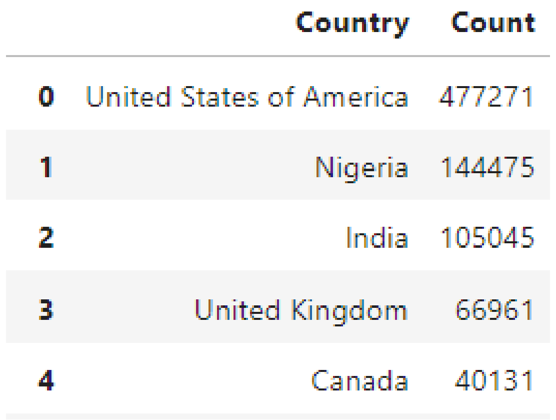
- Due to the limitations with API (GeoCoder) in cleaning the locations I had only took the countries where 50% of the users are located
- Below is the count of users in top 5 countries and the geo graphical distribution of the users country wise
- We can say that people in US are more engaging in twitter when compared to other countries
- The reason on why the Nigerian twitters were engaging more is due to twitter Ban. User's used VPN to bypass the Nigerian network
  Source link



Country wise twitterers distribution

Nigeria's Twitter ban: The people risking arrest to tweet

8 June 2021

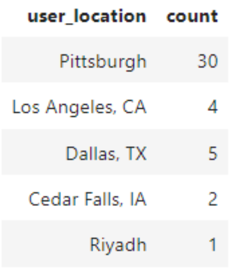


	Country	Count
0	United States of America	477271
1	Nigeria	144475
2	India	105045
3	United Kingdom	66961
4	Canada	40131

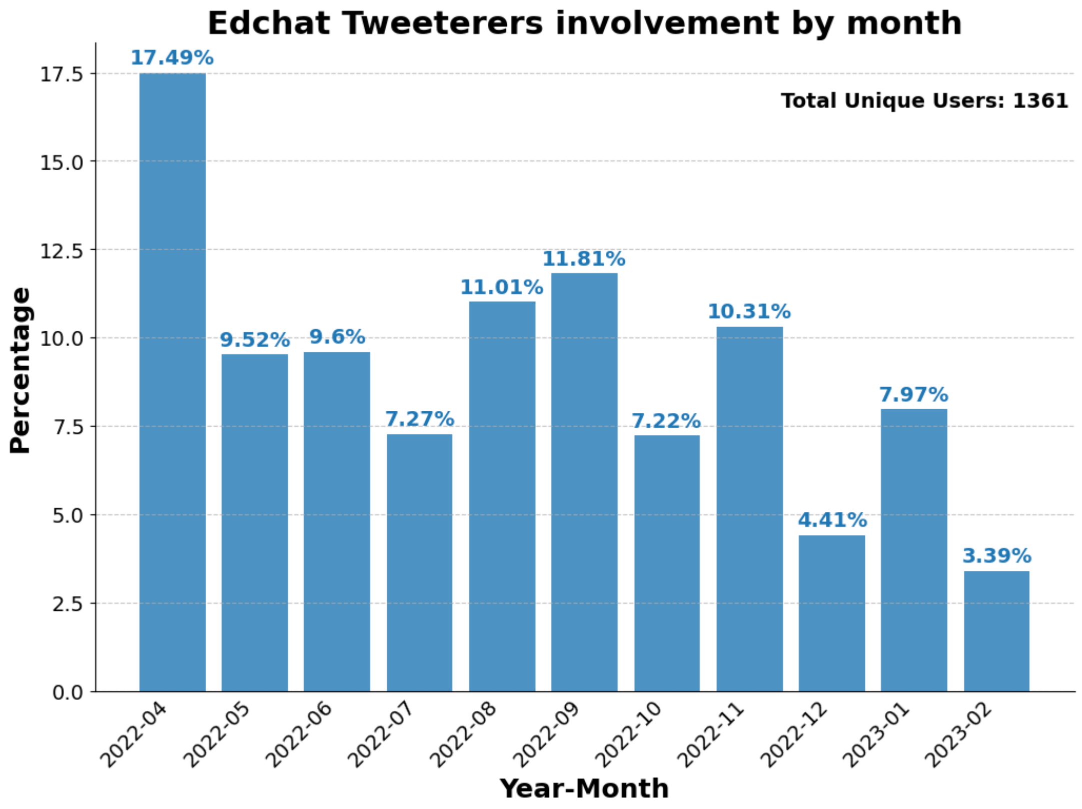
Country wise Count of twitterers

TRENDS IN EMERGENCE OF NEW TWITTERERS

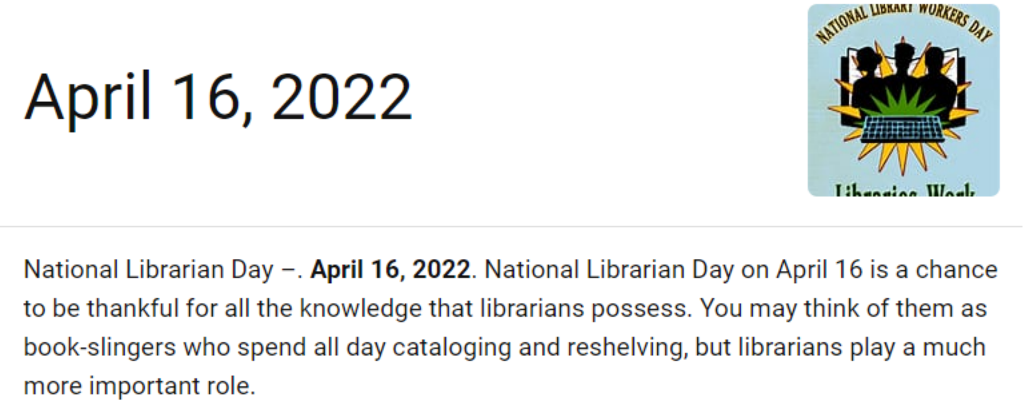
- There are around 2k tweets regarding edchat. Due to the limitation of API where it can take handle 400 requests I was unable to narrow down the location of these twitterers. The table give info about the top 5 locations of twitterers for edchat and we can say that US topped in the twitterers.
- When I did the timeline analysis for the edhcat I found interesting results



user_location	count
Pittsburgh	30
Los Angeles, CA	4
Dallas, TX	5
Cedar Falls, IA	2
Riyadh	1



- There were a total of 1361 unique users that tweeted about edchat
- We see that many of the twitterer's engagement is more in April month
- Possible reasons: April is National School Library Month in the United States, which may lead to an increase in discussions about education and teaching and other can be Spring Break! (Which I am looking forward to ☺) giving more time to participate in discussions and chats



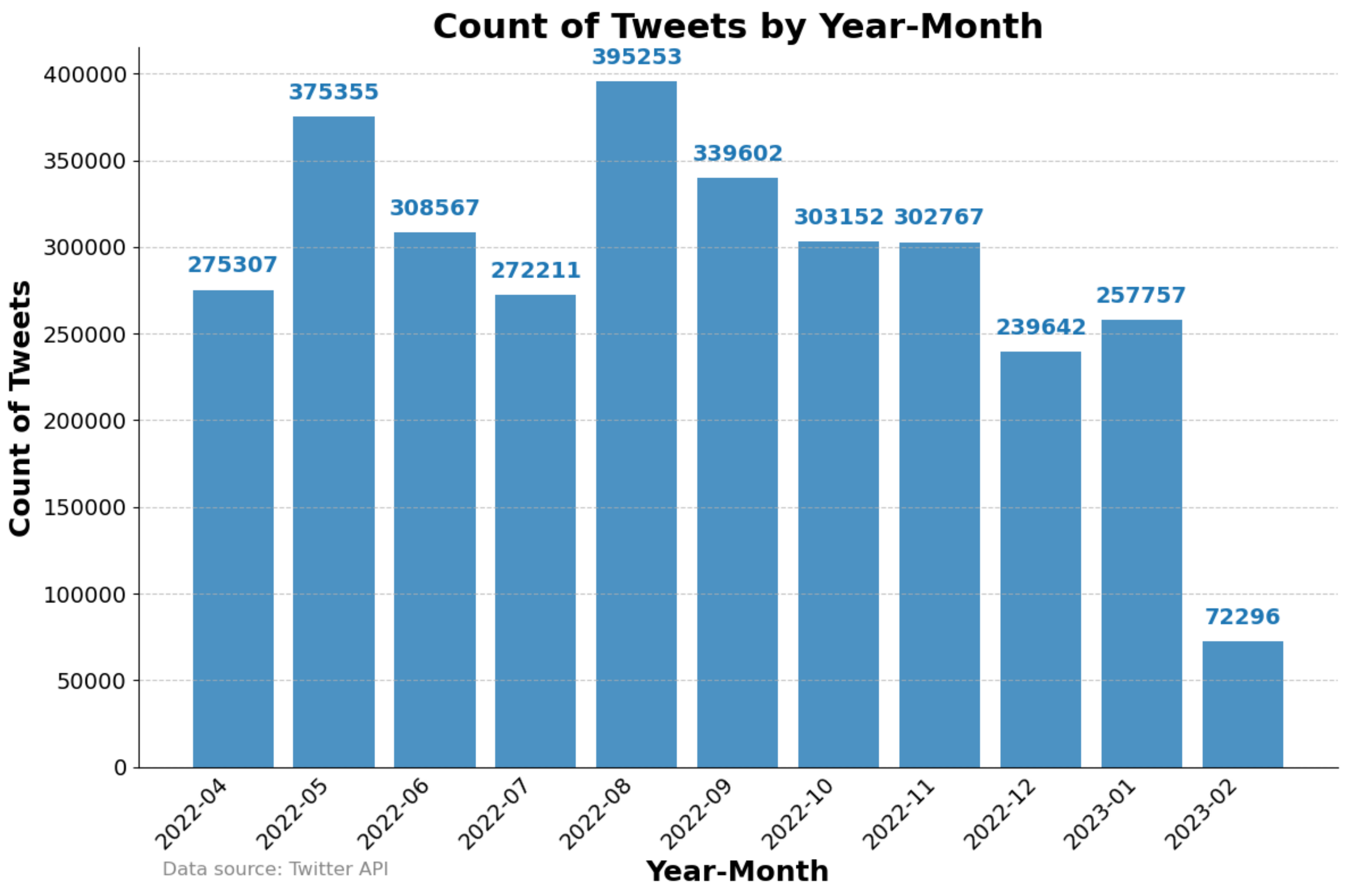
April 16, 2022

National Librarian Day –. **April 16, 2022**. National Librarian Day on April 16 is a chance to be thankful for all the knowledge that librarians possess. You may think of them as book-slingers who spend all day cataloging and reshelving, but librarians play a much more important role.

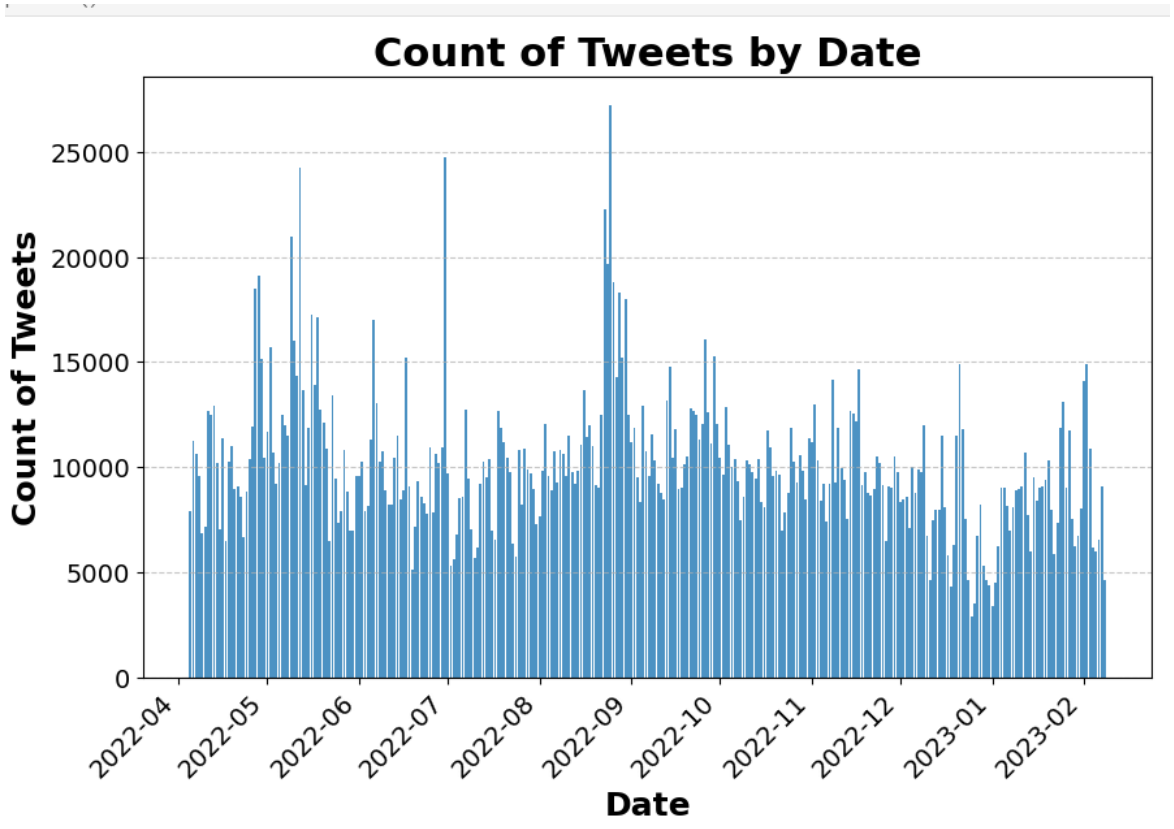
Source: NATIONAL LIBRARIAN DAY - April 16, 2023 - National Today
national librarian day 2022 - Search (bing.com)

MOST NUMBER OF TWEETS ARE ON AUGUST 25 2022

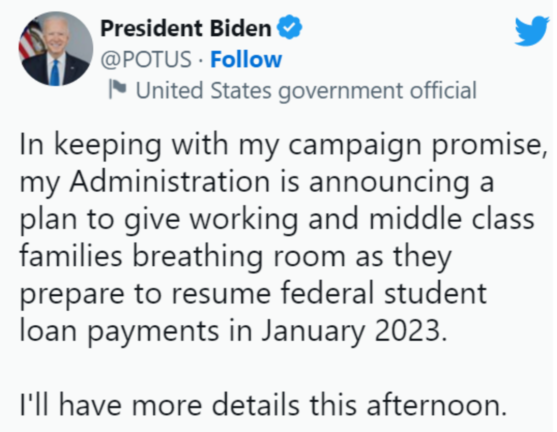
- The data from first three months of 2022 was missing which skewed some of the results



Data source: Twitter API



- From the monthly trend of tweets we can see that Aug got most number of tweets. Primary reason is because of Biden announcing that he kept his campaign promise i.e., to cancel some student debts
- It was announced on Aug 24 2022 in the Morning without any concrete details
- When we check the tweets on Aug 24 2022 people were engaging and few details were leaked due to which the most number of tweets were on Aug 25



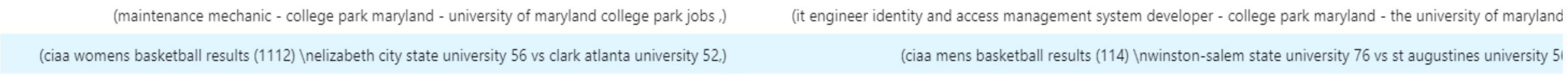
President Biden
@POTUS · Follow
⚑ United States government official

In keeping with my campaign promise, my Administration is announcing a plan to give working and middle class families breathing room as they prepare to resume federal student loan payments in January 2023.

I'll have more details this afternoon.

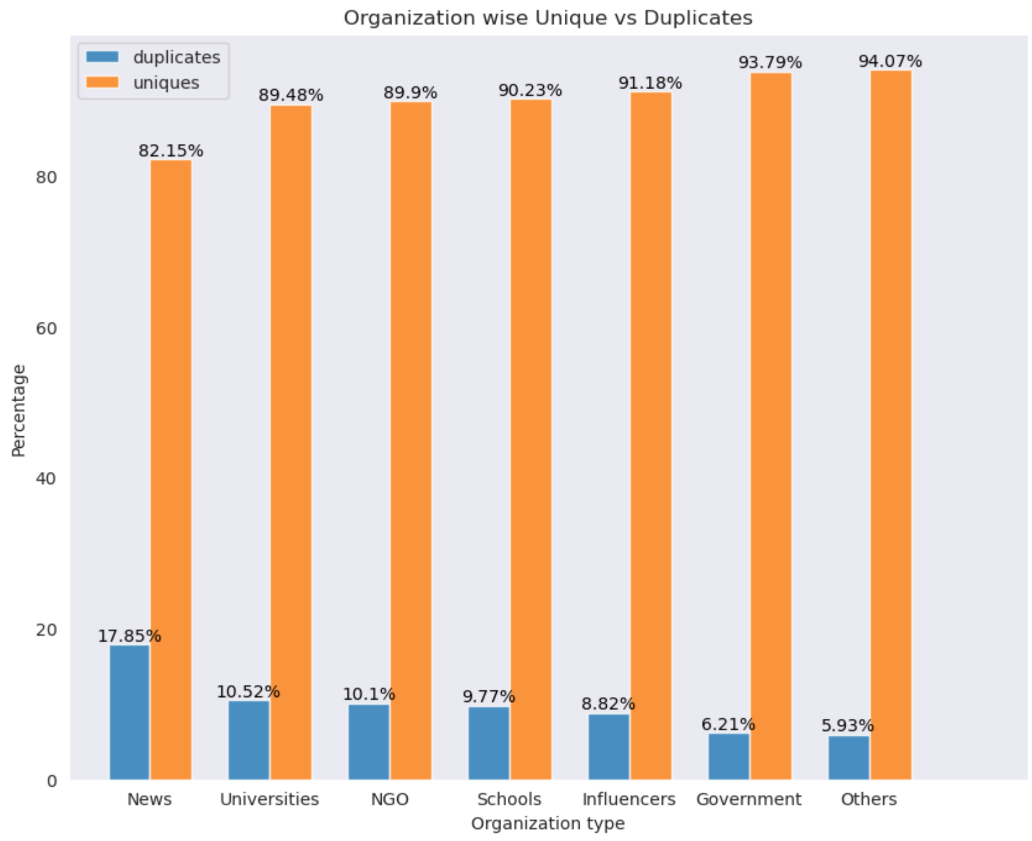
MAJORITY OF THE MESSAGES ARE UNIQUE

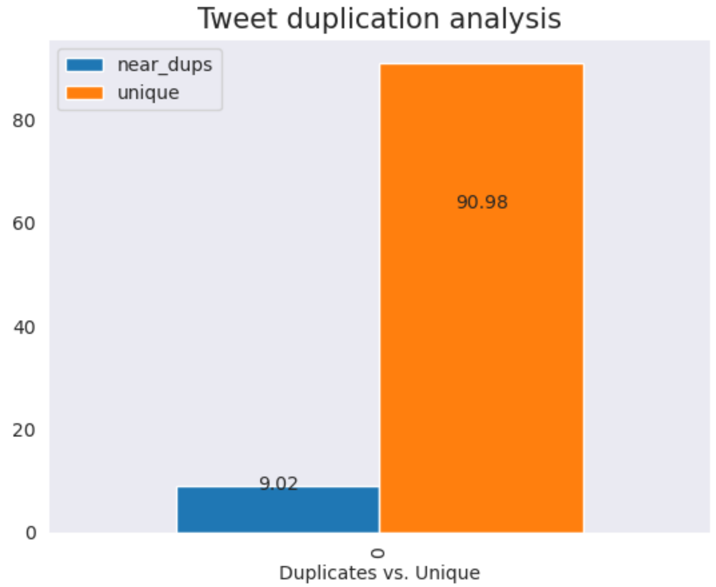
- To analyse the message uniqueness I had filtered the retweets and analysed on original tweets as to check whether the users are actually writing content or copy pasting other tweets
- From plot we can say that most of the messages are unique and not near duplicates
- When we see organization wise NEWS has most duplicates and it realistic as they tweet the same news more than once by adding more details to it
- Below are some sample tweets where they are conveying the same message but added few more details
- The first tweet talks about jobs in University of Maryland and the duplicate is about a particular job in University of Maryland



(maintenance mechanic - college park maryland - university of maryland college park jobs ,)	(it engineer identity and access management system developer - college park maryland - the university of maryland
(ciaa womens basketball results (1112) \nelizabeth city state university 56 vs clark atlanta university 52,)	(ciaa mens basketball results (114) \nwinston-salem state university 76 vs st augustines university 5

- Others have most unique messages with 94.07%





TWITTER CAN BE CONSIDERED AS A CREDIBLE SOURCE

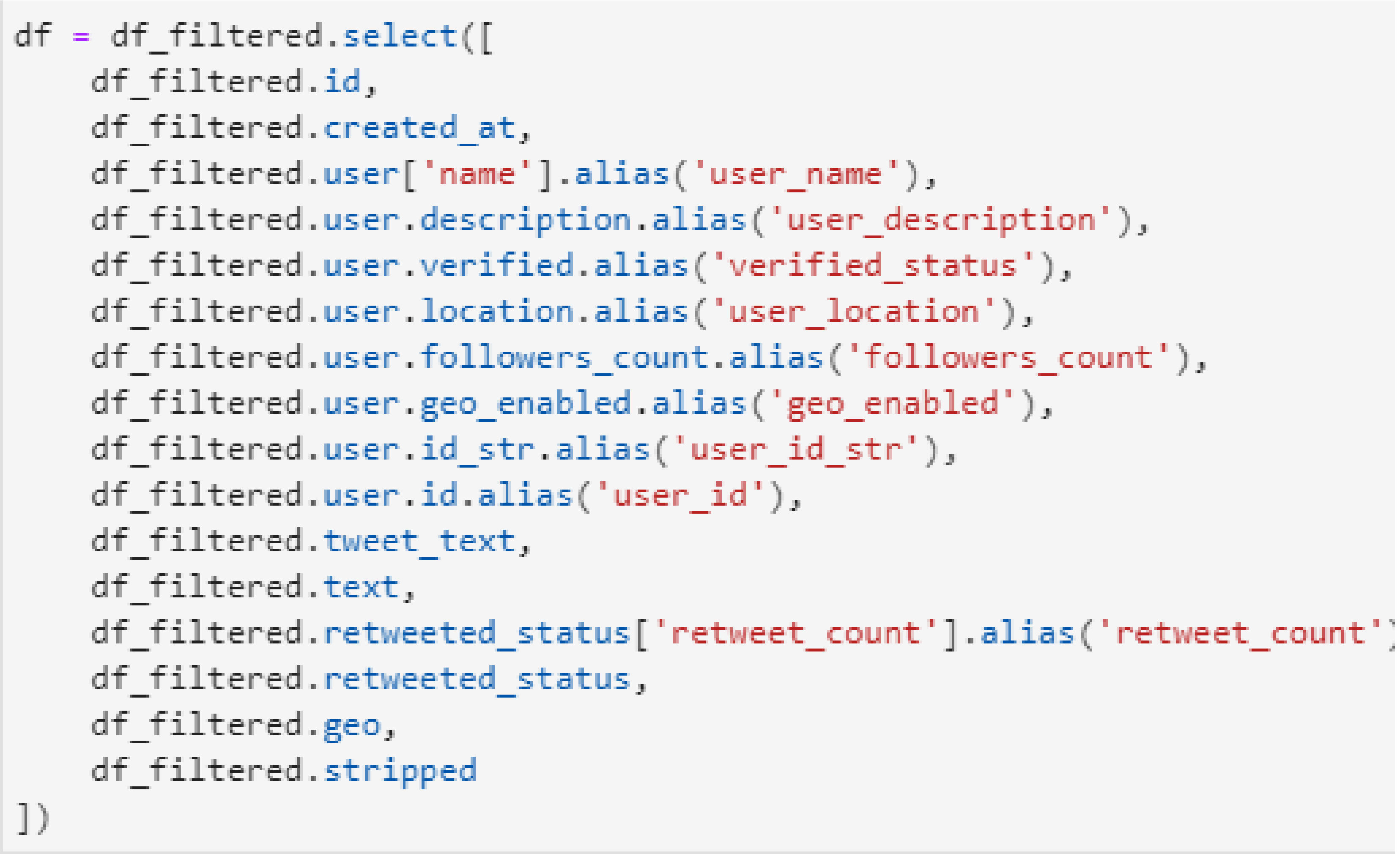
- Data in twitter is quite noisy and can be very tricky to figure out required columns, as there are more than two columns that gives us the right information
- We see that the verified users tweet majority of the original content but the retweet count of their tweets is quite questionable.
- Most retweets are for the individuals whose profile is not verified and has no information that the information that post is valid or whether the account is an individual or bot
- When we try to analyse based on a trend we see that people engage on twitter but the verified profile twitterers are not quite engaging as individuals whose account is not verified which comes to question the trend
- When we see the timeline analysis of the trends when a major news (Biden in our case) was announced the twitterers were quite engaging
- Content posted in the twitter is quite unique which means that the users post original content or either engage to a topic that was posted
- From this we can say that we can not entirely rely on twitter as a credible source for the topics that are going around. Unless the post was from verified profiles we cannot say that the information was credible

RECOMMENDATIONS

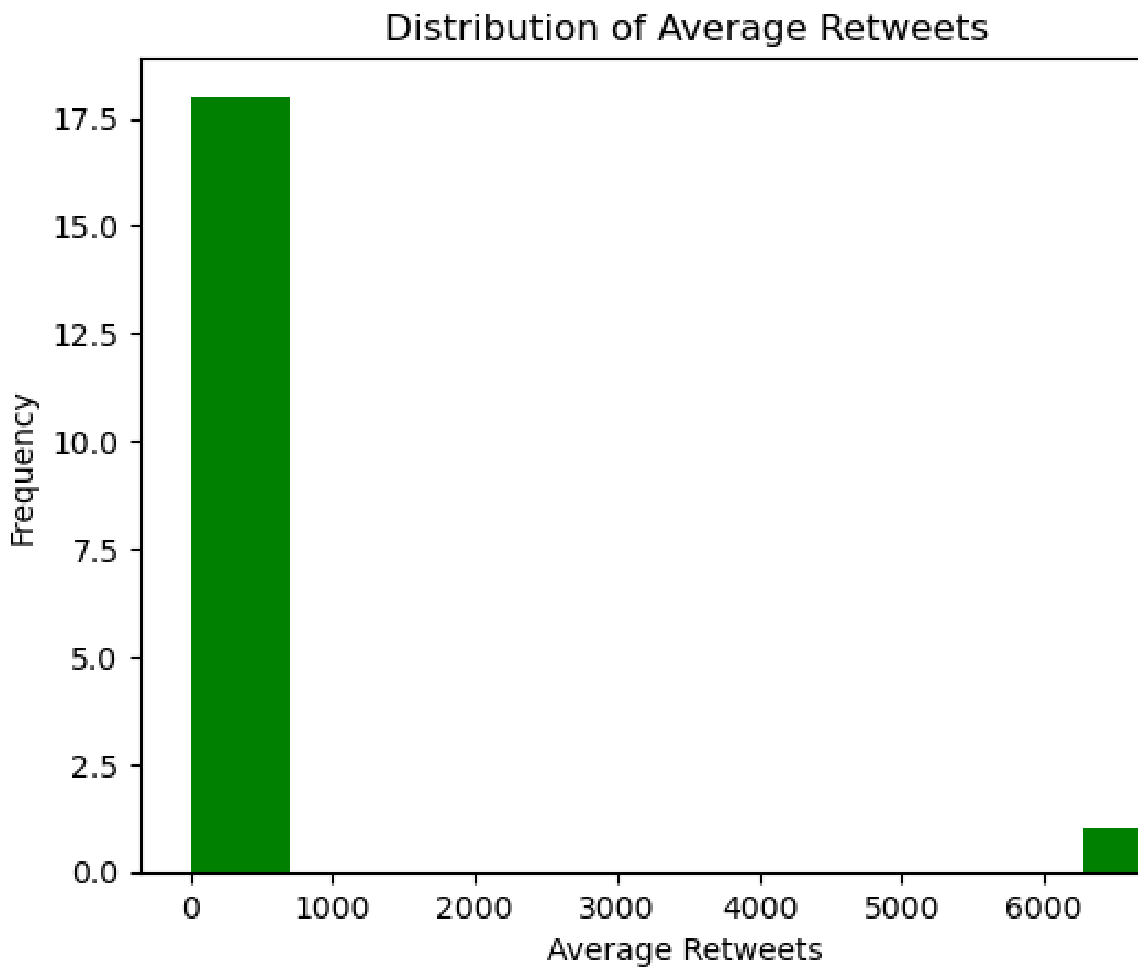
- Doing the sentimental analysis on the tweet can give us a broader picture on the tweet and better segregate user
- Geographical limitation is one thing to consider since the focus was on English tweets which narrowed down to handful of countries and eliminated majority of the users

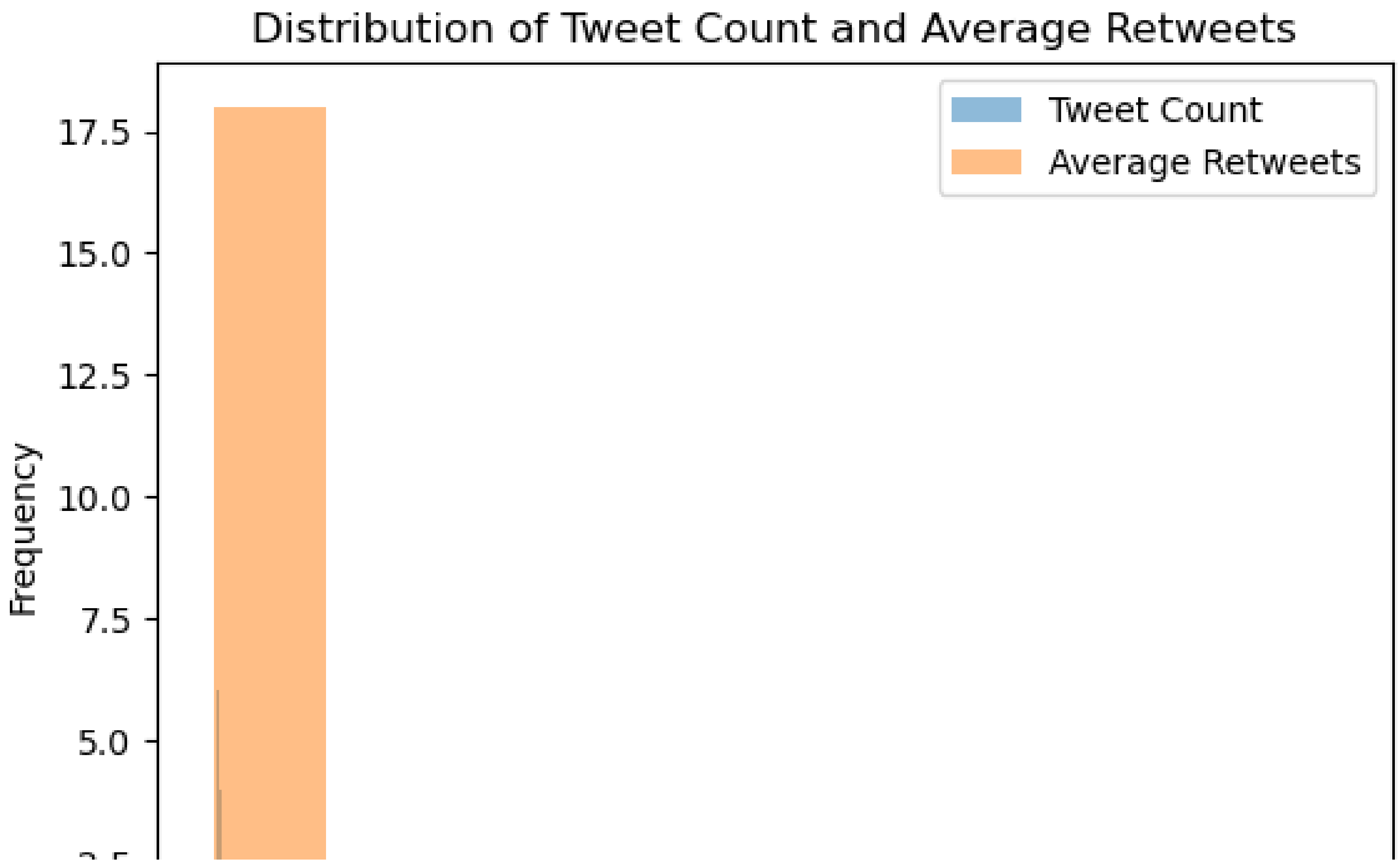
APPENDIX

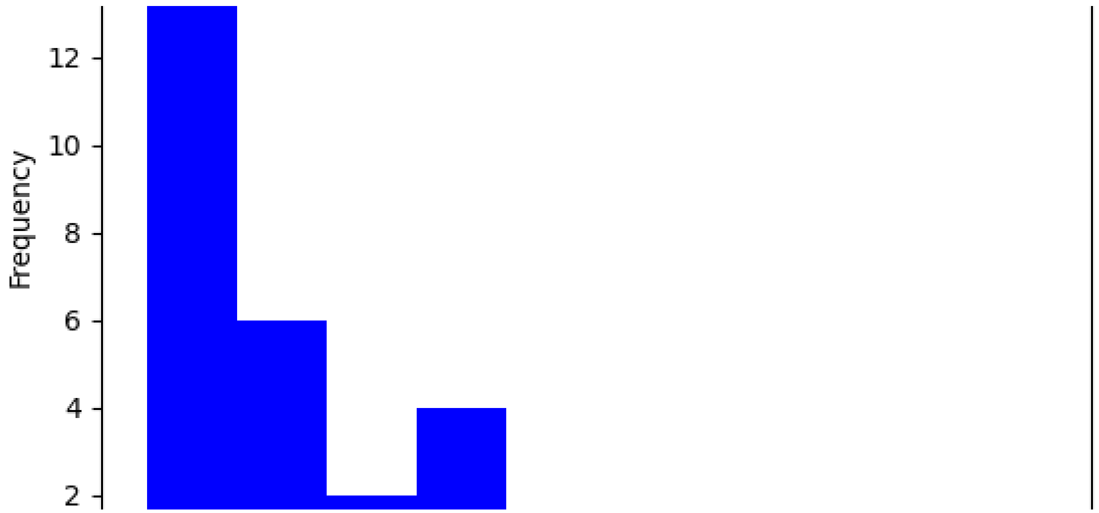
SELECTED VARIABLES



```
df = df_filtered.select([
    df_filtered.id,
    df_filtered.created_at,
    df_filtered.user['name'].alias('user_name'),
    df_filtered.user.description.alias('user_description'),
    df_filtered.user.verified.alias('verified_status'),
    df_filtered.user.location.alias('user_location'),
    df_filtered.user.followers_count.alias('followers_count'),
    df_filtered.user.geo_enabled.alias('geo_enabled'),
    df_filtered.user.id_str.alias('user_id_str'),
    df_filtered.user.id.alias('user_id'),
    df_filtered.tweet_text,
    df_filtered.text,
    df_filtered.retweeted_status['retweet_count'].alias('retweet_count'
    df_filtered.retweeted_status,
    df_filtered.geo,
    df_filtered.stripped
])
```



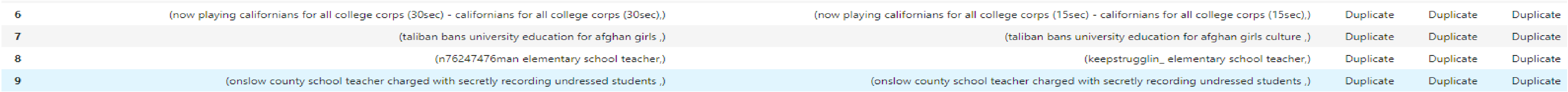




TWEET VS RETWEET DISTRIBUTION

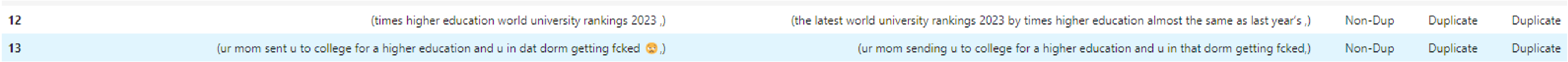
FINDING CORRECT THRESHOLD FOR JACCARD SIMILARITY

- Comparing the thresholds 0.3, 0.4 and 0.5 to see for all tweets



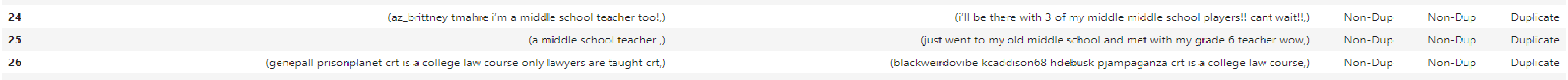
6	(now playing californians for all college corps (30sec) - californians for all college corps (30sec),)	(now playing californians for all college corps (15sec) - californians for all college corps (15sec),)	Duplicate	Duplicate	Duplicate
7	(taliban bans university education for afghan girls ,)	(taliban bans university education for afghan girls culture ,)	Duplicate	Duplicate	Duplicate
8	(n76247476man elementary school teacher,)	(keepstrugglin_ elementary school teacher,)	Duplicate	Duplicate	Duplicate
9	(onslow county school teacher charged with secretly recording undressed students ,)	(onslow county school teacher charged with secretly recording undressed students ,)	Duplicate	Duplicate	Duplicate

Sample tweets for the threshold 0.3 which is good



12	(times higher education world university rankings 2023 ,)	(the latest world university rankings 2023 by times higher education almost the same as last year's ,)	Non-Dup	Duplicate	Duplicate
13	(ur mom sent u to college for a higher education and u in dat dorm getting fcked 😩,)	(ur mom sending u to college for a higher education and u in that dorm getting fcked,)	Non-Dup	Duplicate	Duplicate

Sample tweets for the threshold 0.4 which is good

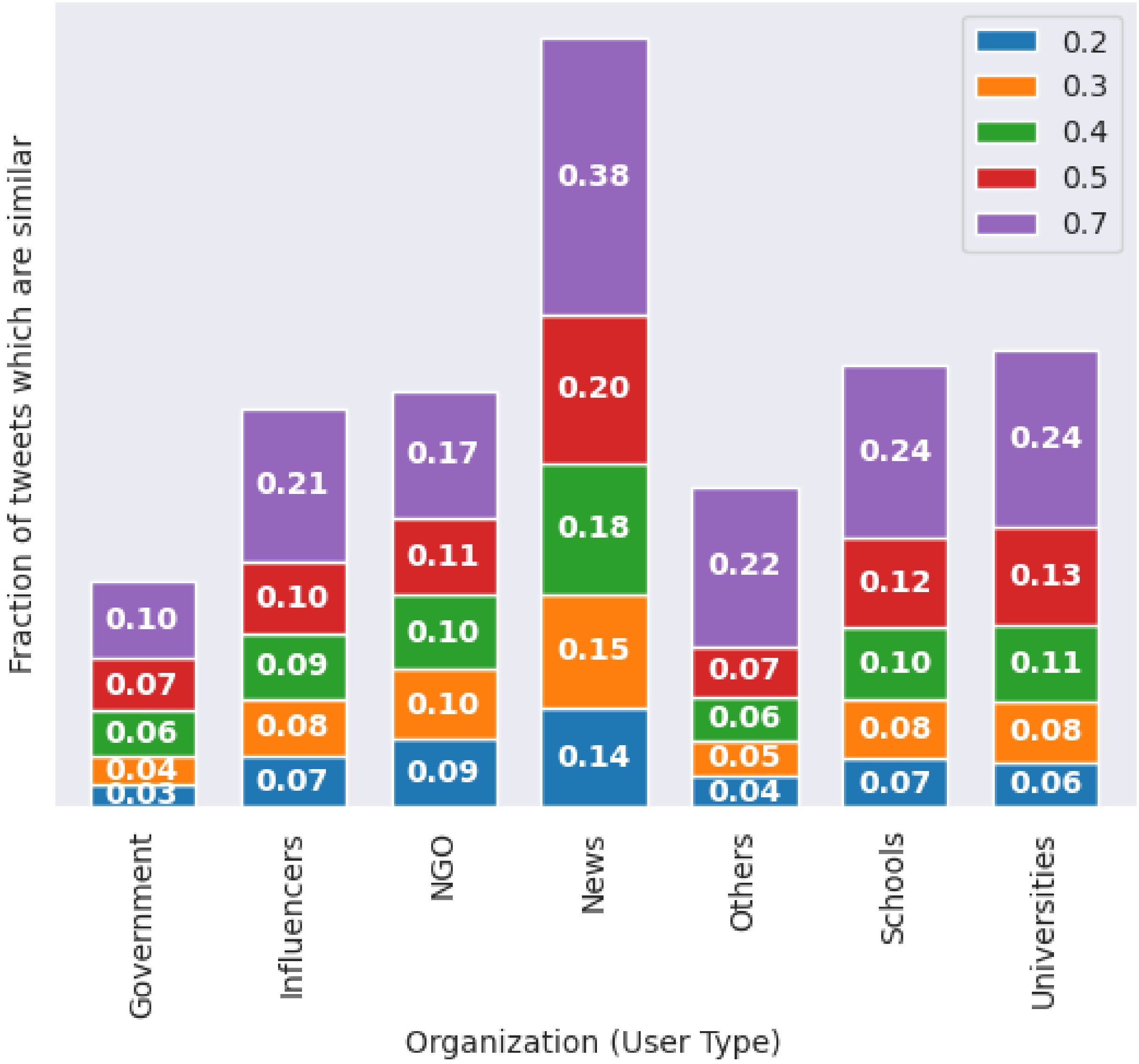


24	(az_brittney tmahre i'm a middle school teacher too!,)	(i'll be there with 3 of my middle middle school players!! cant wait!!,)	Non-Dup	Non-Dup	Duplicate
25	(a middle school teacher ,)	(just went to my old middle school and met with my grade 6 teacher wow,)	Non-Dup	Non-Dup	Duplicate
26	(genepall prisonplanet crt is a college law course only lawyers are taught crt,)	(blackweirdovibe kcaddison68 hdebusk pjampaganza crt is a college law course,)	Non-Dup	Non-Dup	Duplicate

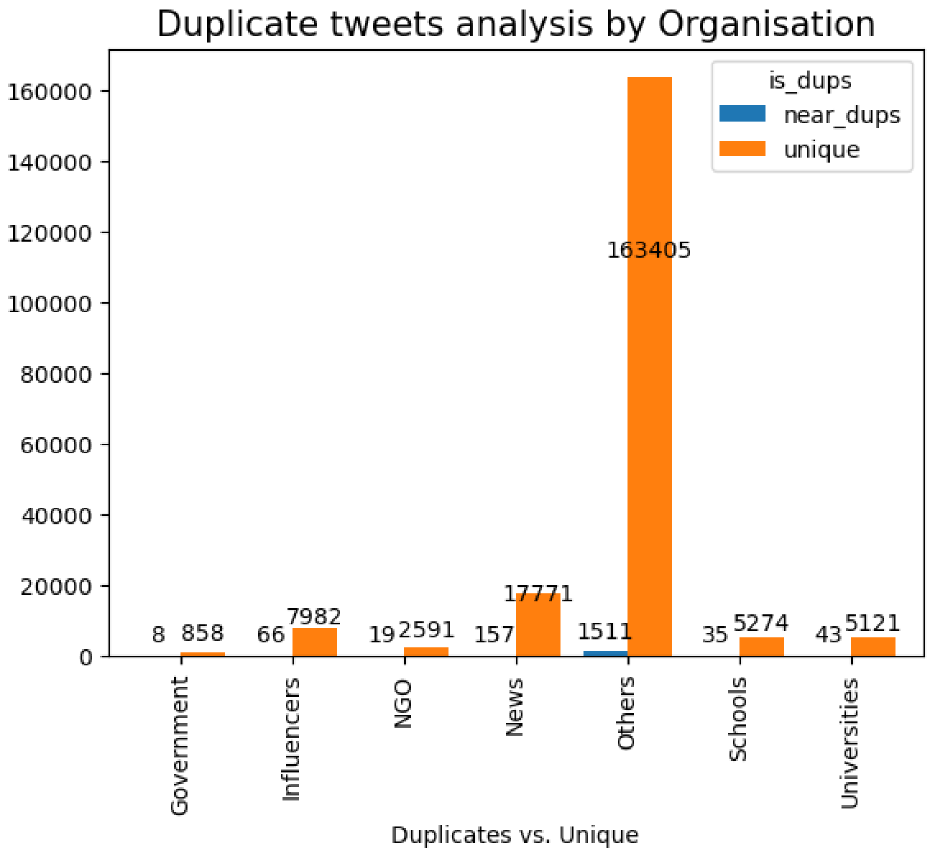
Sample tweets for the threshold 0.7 which is good

Based of the sample tweets we see that Jaccard similarity of 0.4 gives better near duplicates

ORGANIZATION WISE JACCARD SIMILARITY



- The graph says the percentage of the duplicates for each Jaccard similarity



- The above graph is the Duplicate Vs Unyque's for Jaccard Similarity of 0.4