Twitter Credibility Study

Zeyu Chen

Executive Summary

Twitter has emerged as a popular platform for educators to share and exchange information, ideas, and resources. However, the credibility of information shared on Twitter has been a topic of debate.

Key Points:

1. The majority of prolific twitterers in education sector are sports related and unverified by Twitter with relative low follower base. So original posts are not credible source of information.
2. With number of retweets, about half of the most retweeted accounts are verified social media influencers that are unrelated to education topics, which is still not credible enough for education related information.
3. Based on the influence scoring system, most of the top influential twitterers are verified with a considerable proportion of political entities, so the most influential twitterers can be considered credible.
4. Most of the twitterers and tweets come from the US, for regional education related events, Twitter provides a timely public opinion platform for relevant users.
5. Verified accounts of education entities share similarities in their tweets, while most of the tweets from verified government accounts are unique. Uniqueness should not be considered a metric for credibility, higher uniqueness meant to indicate credible source of information was being shared.

In conclusion, **Twitter is a not a credible enough source of information** in education due to the large amount of unrelated original posts and retweets that dilute the spread of credible information from government officials and news outlet, even though Twitter is a timely and engaging platform for education topics. But **verified accounts in this sector can be considered credible**.

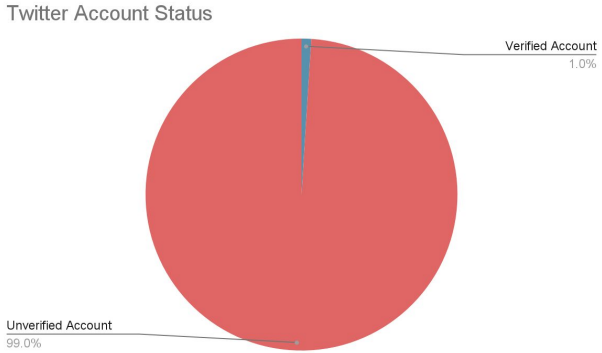
Methodology

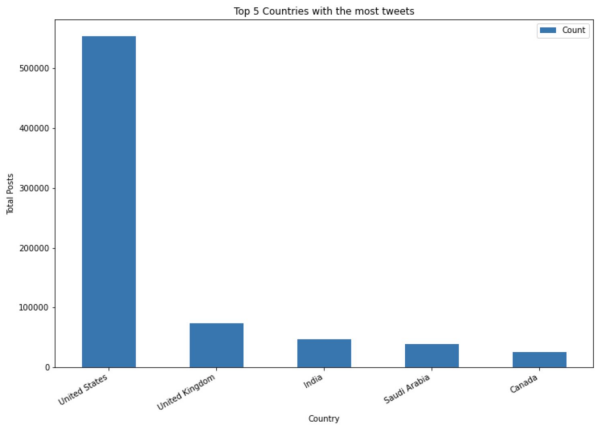
1. Author Identification and Influence Scoring
    a. Understand the most prolific twitter users with the most original posts about education topics.
    b. To discover the most retweeted authors with the highest amount of retweeted posts.
    c. Design Influence scoring systems and identify the most influential twitter users of this subject.
2. Geographical & Timeline Analysis
    a. Identify the locations where most tweets were coming from.
    b. Identify the trend of daily/monthly number of tweets and investigate the potential cause of the fluctuations.
    c. Combine geographical with time series output to explore user engagement and tweets timeliness.
3. Uniqueness of Tweets Analysis
    a. Group tweets data based on user names and descriptions into different sectors to understand the information dissemination characteristics.
    b. Use Jaccard similarity of ½ as the threshold to measure similarity.

Source Data Overview

The data source contain roughly 500 Gb, with 100 million tweets objects in nested JSON format.

- Data Structure
  - Tweets, user info, text, extended entities that store media data, geographical data of the location of tweets and status of retweeted or quote tweets compose the first layer of the data structure.
  - Retweets/Quote tweets Info have similar structure as the Tweets Info, with original tweets' retweet status being null, nested under the first layer.
- 1% of the twitter users of education data are verified account, which is considerably higher than the whole twitter user dataset, which has 0.2% verified account, which may suggests that Twitter can be more credible in education related topics.
- The majority of tweets do not have locations, the top countries with the most tweets came out are English speaking countries with a surprising Saudi Arabia.





Reference: https://www.statista.com/chart/28633/verified-users-on-twitter/

Variable Usage

- text, created_at, id: To keep track of the content of tweets for uniqueness, timeline analysis
- place: keep track of the location and country of tweets
- user
    - user_id: unique identifier for search
    - user_name: use ‘screen_name’ under user for identifying
    - verified: imply the status of the user as a credibility indicator
    - followers_count: used as credibility indicator
    - user_location: used for geographical analysis
    - user_description: used as source of segmentation for geographical engagement study
- retweeted_from: identify the original poster of the tweets
- retweet_status: null being original posts
    - retweet_count: to keep track of the number of retweets when retweeted by the user

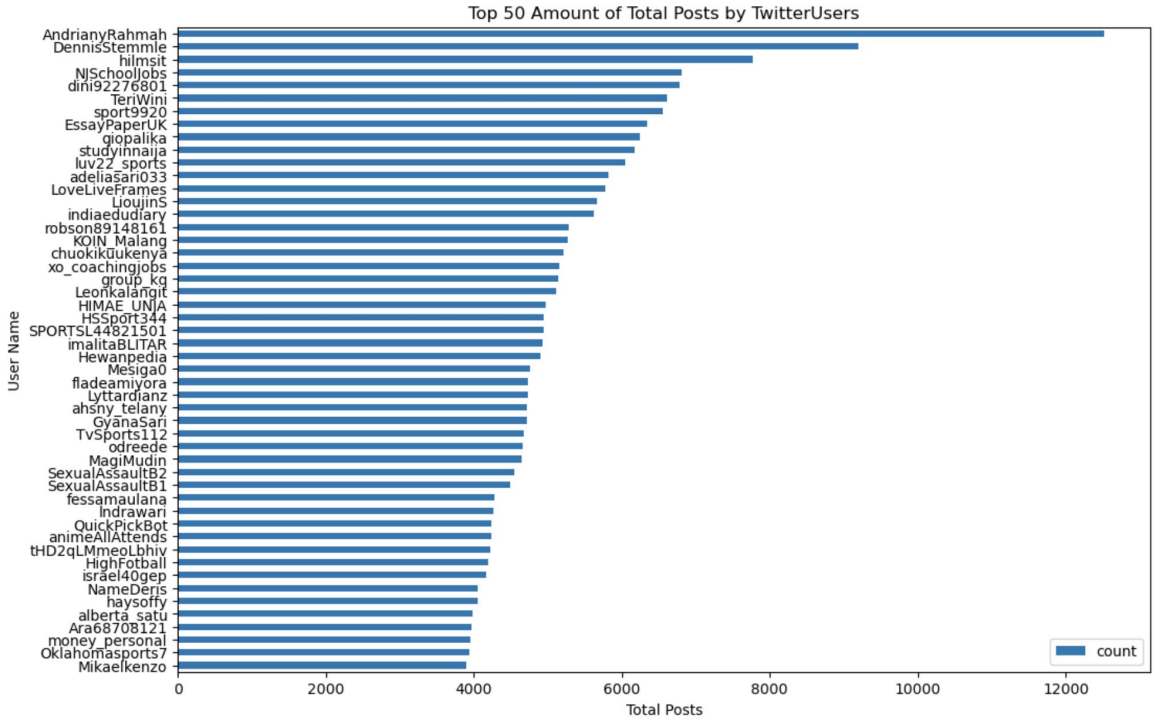
Filtering and Cleaning

To reduce the complexity of computation and clarity of analysis, I used the text column of the data and extract the most occurrence words besides stopwords to identify the keywords of interests in education.

school: 35321364	students: 4097465
college: 10222592	like: 3793172
high: 7962524	kids: 2894026
university: 7763989	professor: 2851083
schools: 5370633	people: 27344286

- After selection, I decided to keep 'college', 'high', 'university', 'students', 'public', 'private', 'secondary', 'primary', 'education', 'undergraduate', 'graduate' as my filtering keywords for the data to maintain the integrity of the data while also trimming down the unrelated data.
- After filtering, the dataset of interest has 30 million tweets.

Author Identification: Most Prolific Accounts



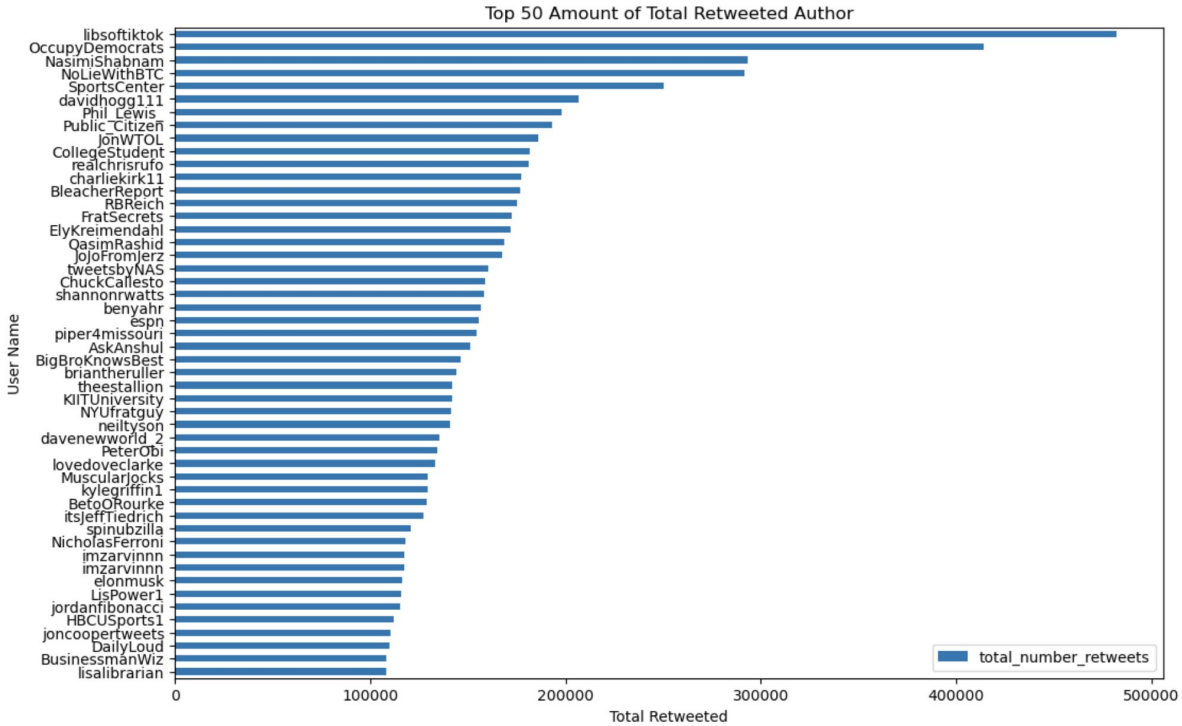
- Most of the top 50 accounts with the most original posts are related to sports and unverified and not reflective of education.
- Since college and high school sports are popular in the United States and highly related to the education system, my keywords for filtration left out many sports related tweets.
- None of the top 50 prolific accounts is verified, with the most followed account has 6000 followers.
- To avoid these posts, I decided on using number of retweets the account have as the metric to measure influence.

Author Identification: Most Retweeted Accounts

user_name	followers	Description	total_number_retweets	verified
libsoftiktok	1893691	Bringing you news you won't see anywhere else. All videos belong to their respective owners. 📧 libsoftiktok@gmail.com. DM submissions	482067	False
OccupyDemocrats	504212	Pro-Democrat political organization & news website. NY Times reported that our reach dominated Trump on Facebook before his ban. Founder: @OmarRiverosays	414295	True
NasimiShabnam	71862	Policy Special Advisor to the UK Minister of State for Refugees. Agent: info@theblairpartnership.com 🇬🇧 @TBP_agency	293392	True
NoLieWithBTC	338956	Podcast covering the top stories & interviews with the biggest names in politics. Hosted by @briantylercohen	291413	True
SportsCenter	41400369	Download the ESPN App ⬇️	250517	True
davidhogg111	1267103	Supposedly a "multi-millionaire, American spy, and paid actor" contactdavidhogg@gmail.com	206630	True
Phil_Lewis_	300076	detroit native. senior front page editor @huffpost. subscribe to my newsletter! 👇🏽	197817	True
Public_Citizen	535244	Public Citizen has been standing up to corporate power and holding government accountable for 50 years. We're people-powered and accept no corporate money.	193203	True
JonWTOL	2973	NW Ohio native. Reporter/Photographer for @WTOL11Toledo and @Go_419	185946	True
CollegeStudent	1876945	Contact: CollegeStudentofficial@gmail.com	181667	False
realchrisrufo	490037	Writer, filmmaker, activist. Manhattan Institute and City Journal. Sign up for my newsletter: http://christopherrufo.com/newsletter.	181261	True
charliekirk11	1912374	Founder & President: @TPUSA • Host: The Charlie Kirk Show • Click the link below to subscribe 🇺🇸	177580	True
BleacherReport	13501984	Become a B/R Insider and help shape the future of the app 👇	176982	True
RBReich	1585497	Berkeley professor, former Secretary of Labor. Co-founder, \n@InequalityMedia. Substack: http://robertreich.substack.com | Mastodon: @rbreich@masto.ai	175297	True
FratSecrets	381302	Being a part of the brotherhood means we keep each other's secrets 🤐	172417	False
ElyKreimendahl	147400	writer, comic, queer. i live tweeted the birth of my baby. @funnyordie @newrootsartists @humordarling mgmt: waldorf entertainment	171951	False

- Most of the accounts are social media personalities with big enough followers
- 60% of the top 50 most retweeted accounts are verified accounts, which is enough to assume their credibility.
- Based on description and research, there are 5 political entities, 5 education related workers, 10 sports/news media, 23 social media personalities, 7 others.

Author Identification: Most Retweeted Accounts



Limitations:

- User who joined the platform for longer time would more likely to have more retweets than those who joined later.
- Retweet count as a metric does not have to depend on the follower base of the account, retweet count does not have to be consistent for user to have high amount of retweets.
- The new verification requirement makes it easier for accounts to be verified by Twitter.

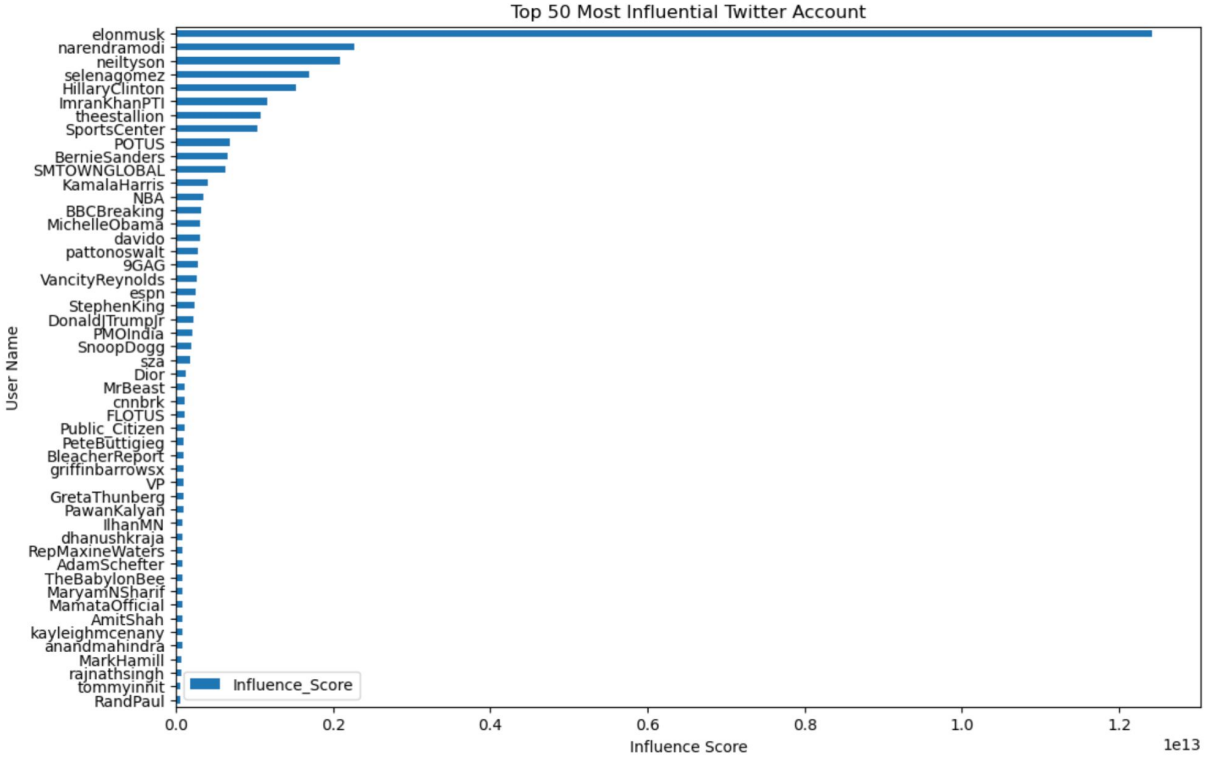
Author Identification: Influence Scores

To improve on the retweet count metric, I developed an influence scores to include the user's follower count as a credibility metric.
The metric is defined as: score = (total retweets/total posts) * the number of followers

user_name	followers	Description	total_number_retweets	verified	Average_Reach	Influence_Score
elonmusk	106476100	None	116620	True	116620.000000	1.241724e+13
narendramodi	84850082	Prime Minister of India	80295	True	26765.000000	2.271012e+12
neiltyson	14805242	Astrophysicist	140754	True	140754.000000	2.083897e+12
selenagomez	65889944	REVELACIÓN out now: http://smarturl.it/REVELACIONSG	25761	True	25761.000000	1.697391e+12
HillaryClinton	31426778	2016 Democratic Nominee, SecState, Senator, hair icon. Mom, Wife, Grandma x3, lawyer, advocate, fan of walks in the woods & standing up for our democracy.	48751	True	48751.000000	1.532087e+12
ImranKhanPTI	17789483	Chairman Pakistan Tehreek-e-Insaf & former Prime Minister of Islamic Republic of Pakistan	65561	True	65561.000000	1.166296e+12
theestallion	7600799	thee real Htown Hottie 😜	142205	True	142205.000000	1.080872e+12
SportsCenter	41400369	Download the ESPN App ⬇️	250517	True	25051.700000	1.037150e+12
POTUS	29370914	46th President of the United States, husband to @FLOTUS, proud dad & pop. Tweets may be archived: http://whitehouse.gov/privacy\nText me: (302) 404-0880	92863	True	23215.750000	6.818678e+11
BernieSanders	15731059	U.S. Senator for Vermont. Not me, us.	84233	True	42116.500000	6.625371e+11
SMTOWNGLOBAL	11294965	SMEntertainment Group Official Twitter	55483	True	55483.000000	6.266785e+11
KamalaHarris	19817342	Fighting for the people. Wife, Momala, Auntie. She/her. Official account is @VP.	40615	True	20307.500000	4.024407e+11
NBA	42052619	The 2022-23 NBA season continues Monday on NBA TV! 7:30pm/et: @Lakers/@BrooklynNets | 10pm/et: @ATLHawks/@trailblazers	24773	True	8257.666667	3.472565e+11
BBCBreaking	51629170	Breaking news alerts and updates from the BBC. For news, features, analysis follow @BBCWorld (international) or @BBCNews (UK). Latest sport news @BBCSport.	24352	True	6088.000000	3.143184e+11
MichelleObama	22224422	Girl from the South Side and former First Lady. Wife, mother, dog lover. Always hugger-in-chief. #TheLightWeCarry	13839	True	13839.000000	3.075638e+11

- 49 of the Top 50 influencers are verified accounts.
- There are 15 political entities from all over the world in the top 50 influential account, 12 celebrities, 1 education worker, 7 news/sports media, 10 social media personalities and 5 others.
- Political entities present as the most influential individuals in this scoring system suggests that government related entities have large influence on education under the current scoring system.

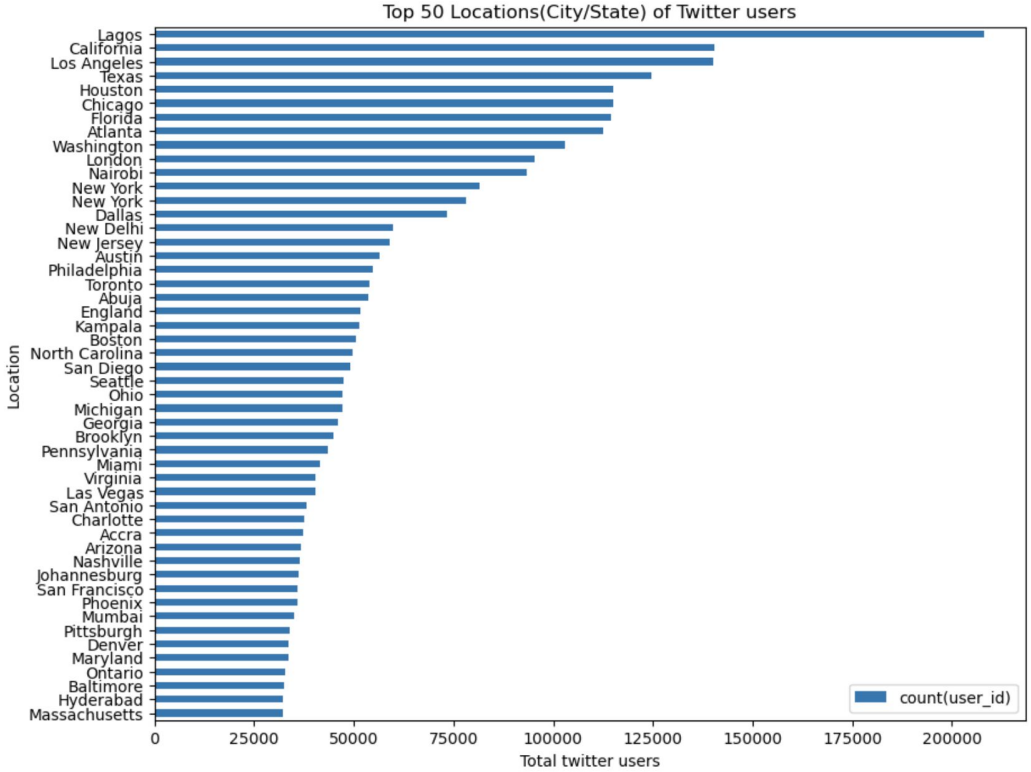
Author Identification: Influence Scores



Limitations:

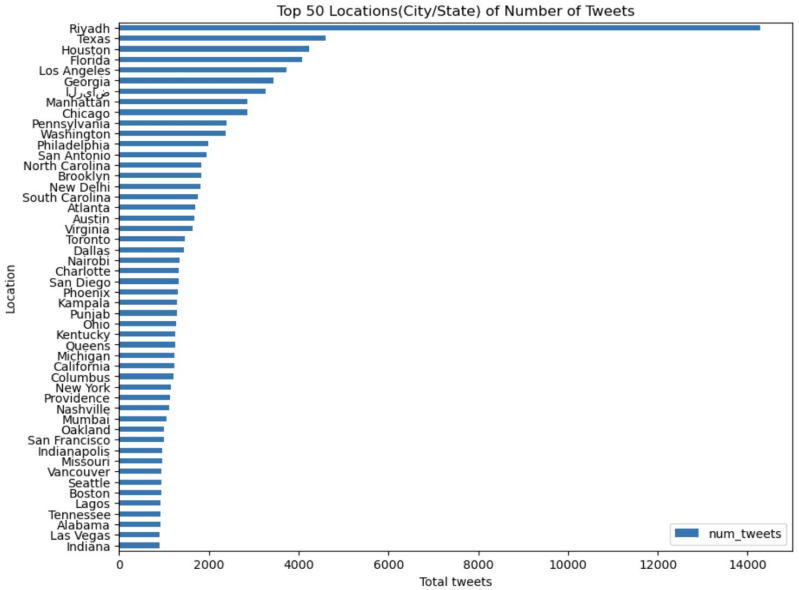
- Current scoring system is heavily influenced by the followers of the account, for instance Elon Musk only have one education related tweet but is the most influential.
- This system favors celebrities since they have both followers and retweets, but for education institutions such as universities, which usually have lower retweet counts and less significant followers.

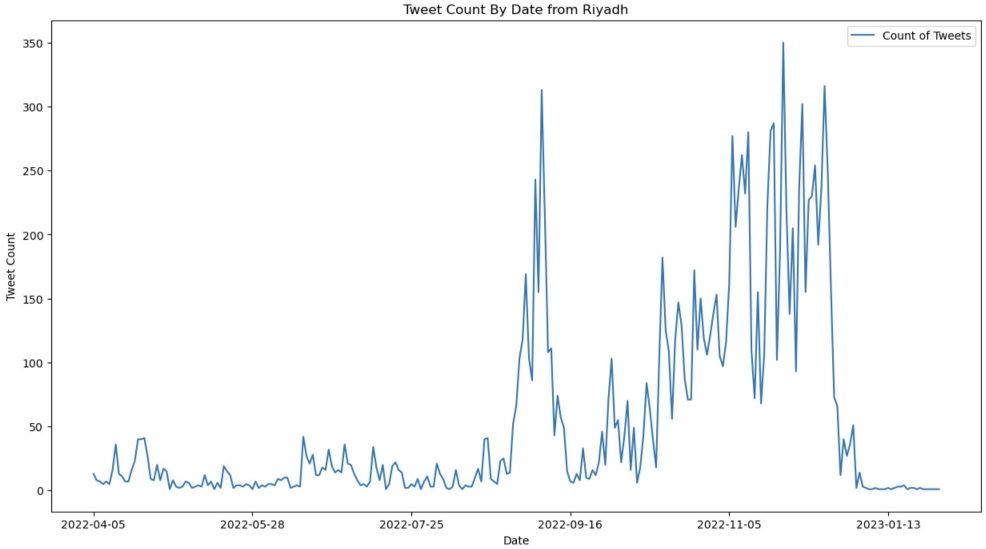
User Location Analysis



- Use the registered location information under ‘users’ to analyze the distribution.
- 40 of the top 50 locations of Twitter users are in North America, with most of the locations being english speaking.
- The user location data does not have uniform format, even after filtration, which caused the graph to have 2 New York entries, one is the state and one is the city, with different number.
- The locations/cities with the highest number of twitter users are mostly larger cities with the most population.
- Surprisingly, Lagos, Nigeria has the most number of Twitter users even though the country banned Twitter use for the public from June 2021 to January 2022.

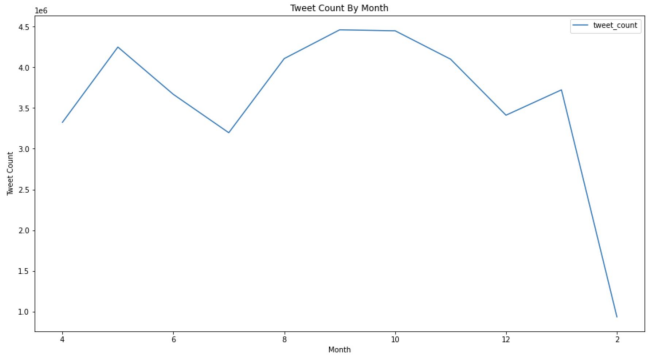
Tweets Location Analysis

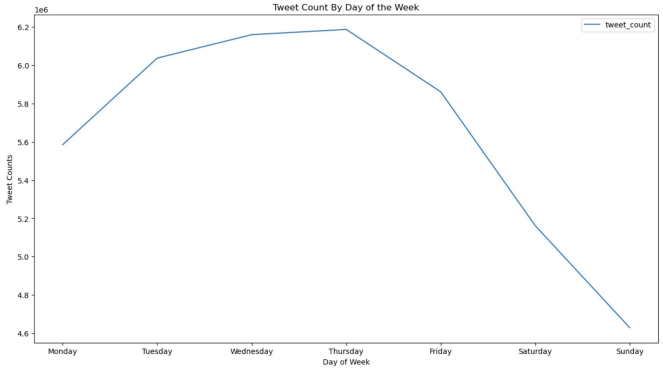


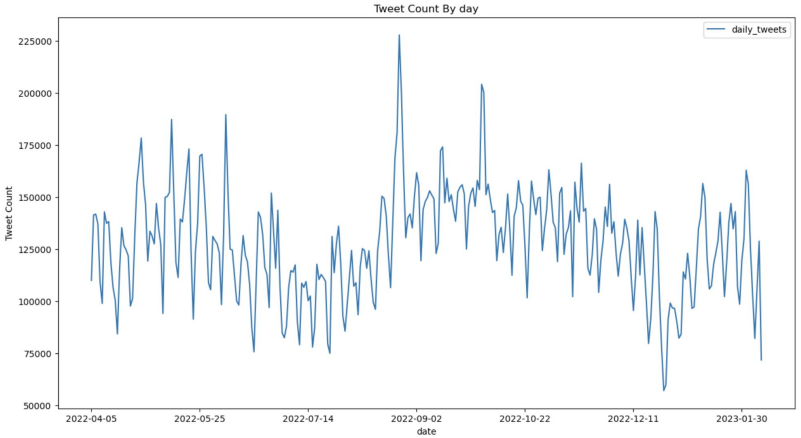


- Riyadh had the most number of tweets despite predominantly amount of tweets come from english speaking locations during the past 12 months.
- Peaks of daily tweets happened during July, August and December, which coincided with President Biden's Saudi Arabia visit and 2022 World Cup.
- The peaks of World Cup further demonstrate my filtration of choice left out significant amount of sports related tweets.

General Tweets Timeline



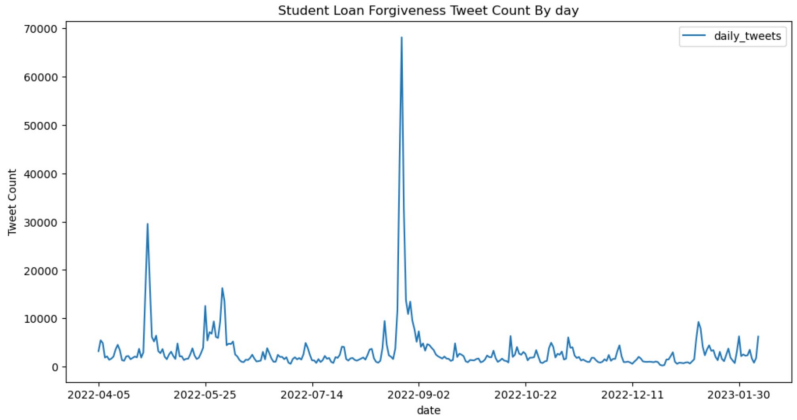


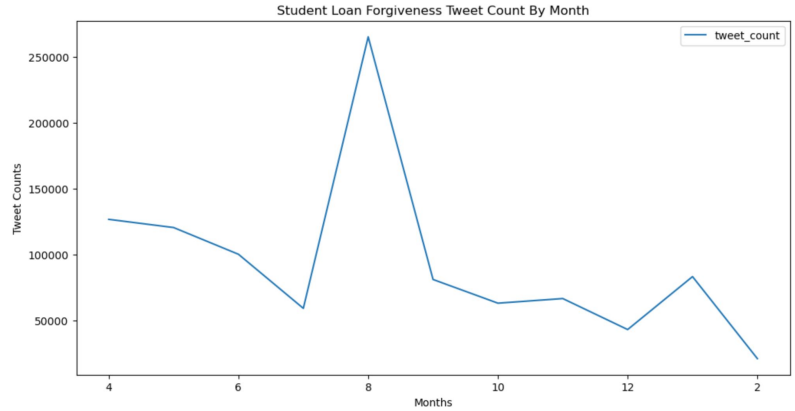


- Generally Speaking, Peaks of tweets happen during September when school resumes and the two valleys are July and December during vacations. For February 2023, the steep decline is caused by the lack of latest tweets data.
- Weekly, Monday to Thursday has a increasing trend, while tweets number decline during weekends.
- There are recognizable peaks during April and May last year, potentially caused by public events such as the Florida Math Book Ban.
- The most tweets about education in 2022 happened in late August, which coincided with President Biden's Student Loan Forgiveness.

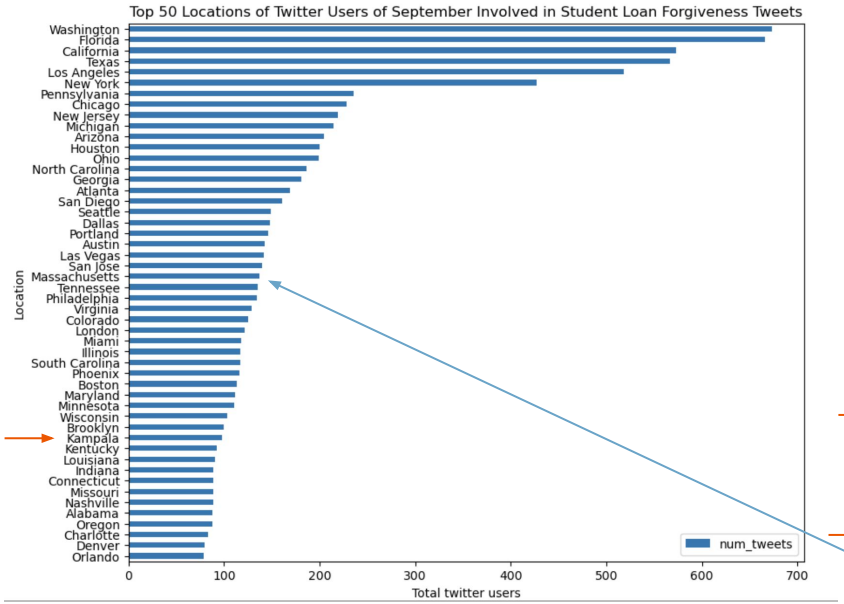
Timeline Peak: Student Loan Forgiveness

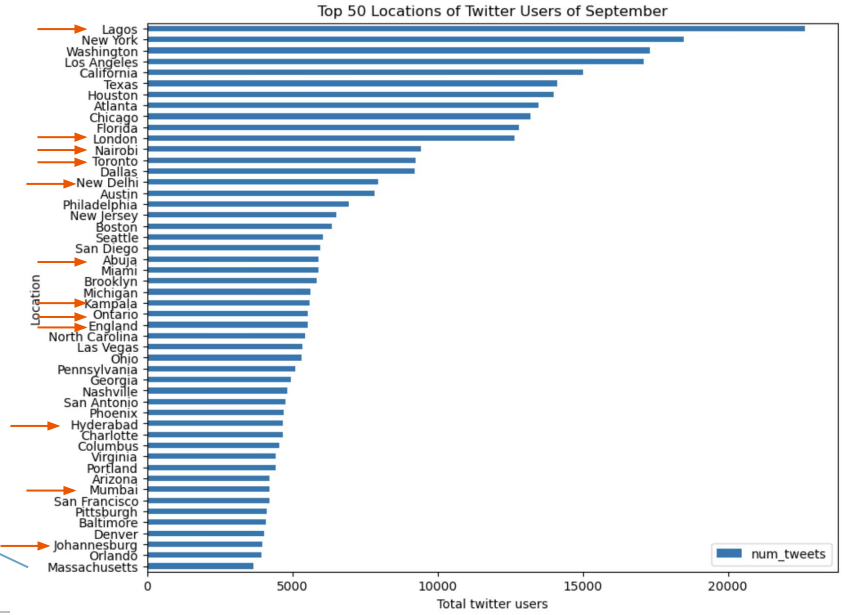
- After keywords filtration, there are 1 million tweets related to President Biden's Student Loan Forgiveness.
- Monthly tweets of the subject also peaked in August
- The peak happened at the end of August, at the same time of the announcement.
- The timeliness of the tweets volume suggests that Twitter as a platform can provide its users with the most recent updates in Education.





Timeline Peak: Student Loan Forgiveness





- To investigate the engagement of twitter users, I chose September's tweets in consideration of the time needed for public opinions to develop.
- Only 1 location of the top 50 locations of amount of users engaged in the public opinions of Student Loan Forgiveness is outside US compared to 12 of the top 50 locations are outside US when consider education topics in general.
- Locations such as Massachusetts that has numerous institutions with students carry student loan, experienced a noticeable increase in ranking.
- The change of location ranking suggests that Twitter user in the US were more involved in the topic.

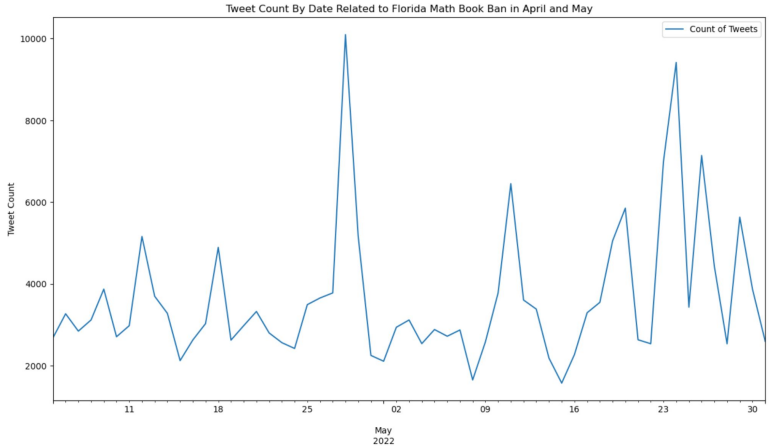
Florida Math Book Ban: Geo Engagement and Timeline

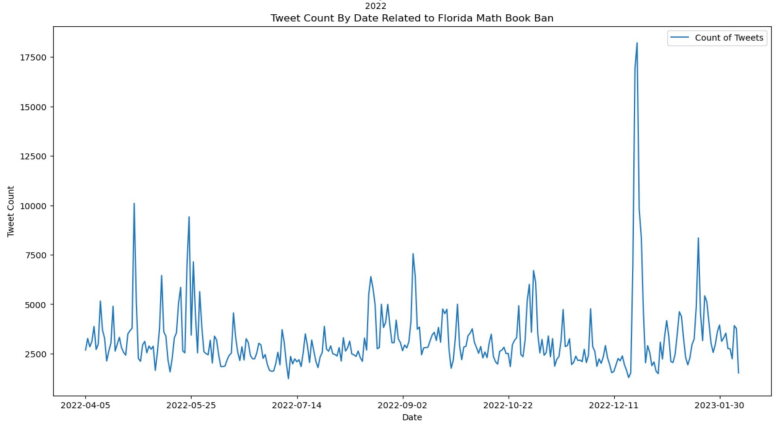
To understand the spread of public opinion on Twitter, I chose a regional public event such as the Florida Math Book Ban to understand how the engagement of twitter users change, and the timeliness of information on Twitter

1. I filtered the twitter data with keywords 'florida', 'math', 'book', 'ban' . After the filtering, there are 1,620,798 tweets related to the topic.
2. I analyzed the time of events and time of peak tweets volume to investigate the timeliness. From the timeline of tweets volume, the peak happened at the same time of the event, which suggests that Twitter provides its users with the most up-to-date information in education.
3. I compared the amount of tweets by user locations to understand the user engagement,the increase in ranking of Florida and the decrease of Lagos regarding this topic suggest that Twitter is the platform of choice for the public to obtain and express opinions.

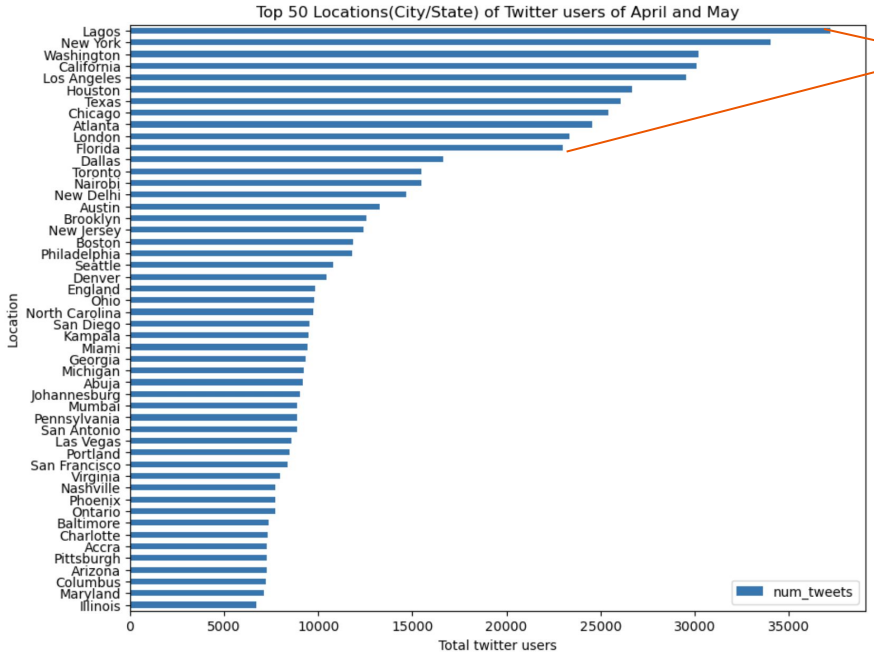
Topic Timeline

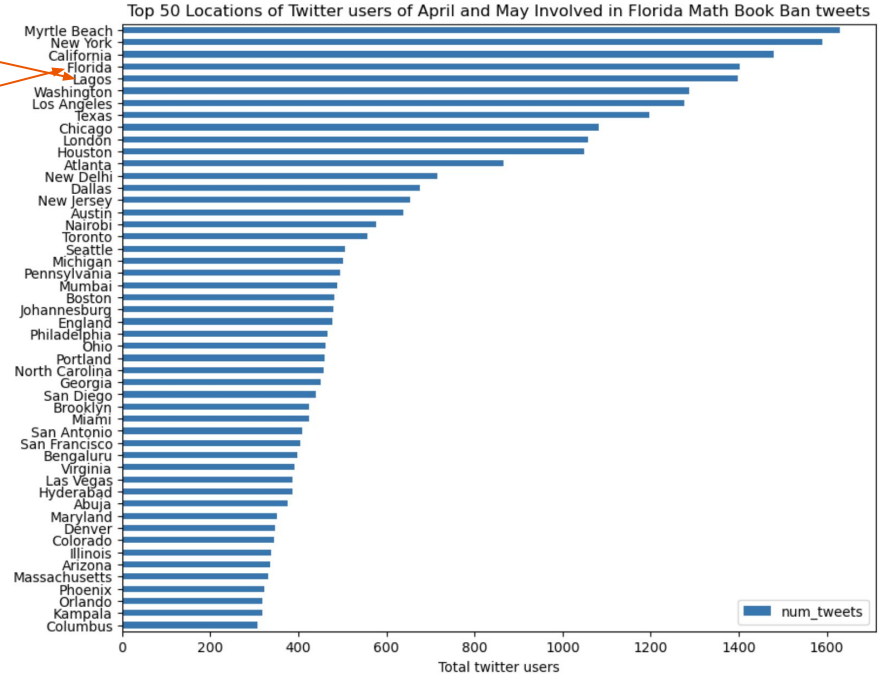
- From the trend plots, after filtering the dataset with keywords, there are significant peaks happen at the beginning and end of May, while also a significant peaks around Christmas.
- The anomaly during Christmas was excused due to the keyword of choice ‘florida’ is often associated with the holiday, so the anomaly is understandable.
- The trendline of April and May suggests that twitter as a public opinion platform responded quickly to the event since the first peaks appeared at the same time with the promulgation.
- Therefore I believe that as a public opinion platform, twitter provides timely information regarding public events.





Topic Geographical Engagement





The increase of user engagement of the users in Florida and the decrease in Lagos suggest Twitter is the platform of choice for the public, especially those who were closely related to the subject manner.

Tweets Uniqueness

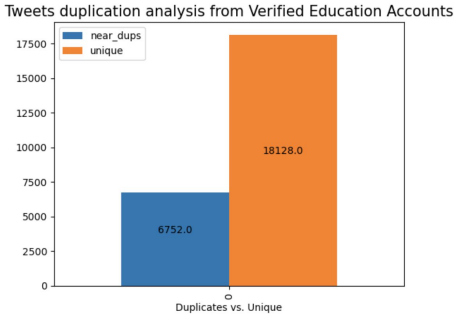
To understand the the uniqueness of tweets in different sectors, I roughly divided the data into 6 groups: News, Government, Education, Celebrities, Sports and Others with selection of keywords. For the uniqueness analysis, I mainly focused on the original tweets from verified Education, Government and News categories.

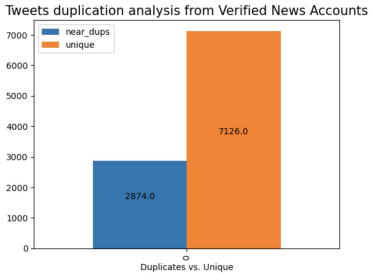
- I chose the Jaccard similarities studies for the analysis and set the threshold to be ½.
- The goal is to identify whether tweets about education are becoming homogeneous in different sector.

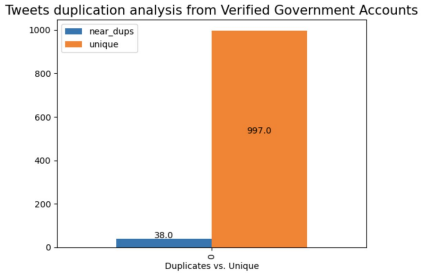
Limitations:

1. The classification of tweets into groups is highly subjective and premature, with high volume of tweets in group ‘Other’ since user name and description is not a reliable indicator to identify users.
2. Uniqueness does not suggest the credibility of tweets, news outlets report policies about education from the same source with duplications does not make them less credible.

Tweets Uniqueness







- From the Bar chart of verified education accounts, of total 24919 tweets, ¼ of the tweets are near duplicates, which could be a sign of them sharing the same source of information.
- After segmenting the original text data on government accounts, I only obtained 1035 tweets from verified account, which may not represent the whole picture.
- From the sample of 1000 tweets, I discovered that most of the tweets are unique with only few exception. Although the results may be fraudulent, it is possible to assume that government account such as senators and congressmen were using their twitter account as platforms to promote their policies to attract voters, that can explain the high amount of uniqueness.
- Finally, News outlet have the most significant proportion of duplicates possibly due to news accounts retweet and repost from the same source.

Conclusion

1. Author Identification
    a. Most of the prolific twitter accounts are sports related due to the limitation of keywords.
    b. Most retweeted accounts are mostly social media personalities.
    c. Influence score system identify the most number of verified accounts, especially political entities.
2. Location Analysis
    a. Riyadh has the most amount of tweets came out. Lagos has the most Twitter users.
    b. Most of the tweets and twitterers come from United States cities such as New York, Los Angeles, which is reflective of population.
    c. Relevant public are willing to use Twitter to engage in public opinion discussions as demonstrated by the student loan forgiveness and florida math book ban sample.
3. Timeline Analysis
    a. Weekly tweets peaks on Thursday and monthly tweets peak during Fall and Winter.
    b. The most daily tweets happen in late August caused by the Student Loan Forgiveness.
    c. Twitter can provide timely information about public events such as Florida Math Book Ban.
4. Uniqueness of Tweets
    a. Tweets are fairly unique in the government sector but not significantly unique in education.
    b. News outlets are similar in contents due to them getting the information from same sources.

Actionable Recommendations

Recommendations for Analytics:

1. To improve the analytical results, I would consider a better keywords selection since the words of choice this time left out a lot of sports related tweets in the dataset and inconclusive uniqueness test.
2. Spam accounts present big challenges for me during the analysis, there are accounts of different id but same name and description. I prematurely eliminated the duplicates. I should consider using created_at under the user column to extract the most recent user_name and ids.
3. Better score systems should be considered, since the current one can be heavily influenced by the number of followers that favors the celebrities. I should do more research on how to calculate the influence.
4. Utilize users and tweets location information can be uninformative when take into consideration of the use of VPNs, which may caused the large amount of users from Lagos, Nigeria where Twitter was banned last year.

Recommendations for Twitter:

1. Twitter should develop a better algorithm of recommendations based on a effective influence scoring system to provide users with accurate information.
2. Twitter should flag spam accounts based on the content they produce to avoid the dilution the spread of misinformation.
3. Increase the reach of government entities to help spread their political agendas to raise public awareness.