Twitter as Credible Source of Educational Information

By: Sarmad Afzal

Agenda

01 Executive Summary with meaningful insights

02 Methodology and source data overview

03 Tweet clean-up and filtering

04 EDA and extensive usage of available variables

05 Author identification

06 Location analysis

07 Timeline analysis

08 Message uniqueness analysis

09 Conclusions and actionable recommendations

10 Appendix

Executive Summary

Social media platforms have become an important means for people to exercise their freedom of speech, allowing them to express their opinions and share their perspectives with a wider audience. Twitter is one of the most popular platform when it comes to sharing thoughts and reflection on events happening around the globe, for the sake of this project we will study educational activities happening on Twitter and identify if it can be considered as a credible source of educational information or not as some argue unchecked spread of misinformation on Twitter has a chilling effect on free expression and on there are bots too tweeting retweeting targeted message. Hence, to answer this question we will conduct following steps

- Exploring source data to find the quality and nosiness of tweets since random people can create accounts and start sharing their thoughts
- Identifying prolific authors of twitters by different tweet attributes such as message volume and their post engagements
- Visualizing the geographical distribution of twitterers to understand which country has most significant impact in field of education
- Conducting timeline analysis to see the trends in the series and what educational events are causing those trends in the data and either they are authentic or just random events
- Performing similarity analysis of tweets since people can retweet each other's post as well, this will help us understand how often people post their own thoughts on educational events in comparison to retweeting some influential stories

Methodology and Source Data Overview

Methodology

- Since the source data size is huge, distributed system on Google Cloud Platform is used
- PySpark is used for coding on Jupyter notebooks
- Python libraries such as Pandas and Numpy used for analyzing samples of data
- Matplotlib, Seaborn, GeoPandas used for plotting graphs and maps
- MinHashLSH is used for similarity analysis
- Project is accomplished in steps using different notebooks for different analysis and intermediate data was stored on Google Storage in Parquet format and some end results were stored in CSV files to be dealt with Pandas dataframe

Source Data Overview

- It's deeply nested JSON file with multiple nodes
  - For our analysis Root and Retweeted_Status nodes are of primary importance
- Size of data is ~500GB
- Number of tweets are ~ 100 Million
- 41 columns are there in total
- API objects twitter reference page: https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/object-model/tweet

Tweet Clean-up and Filtering

Filtering Educational Tweets

Out of 100M tweets, to finding the most relevant educational tweets, I took following 3 steps after understanding the schema of the JSON file using Twitter API website. *Tweet_text* and *text* columns which contains original tweet text and retweeted information respectively

Step 1

Based on tweet characteristics, I applied following filters:

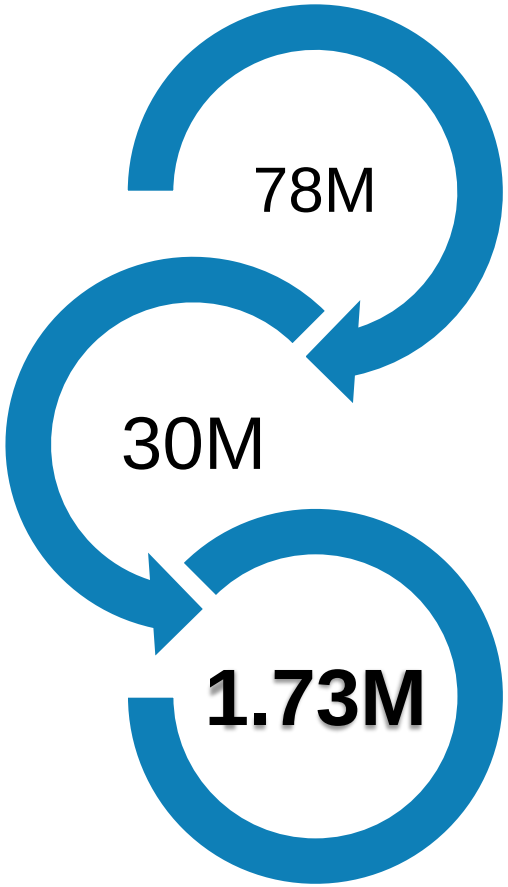
- Language to be English
- Removed sensitive tweets
- Chose tweets which are not truncated, as in not ending with '…' and have full content

Step 2

Matched tweet_text column based on keywords which are most frequent and common in the field of education. I used ChatGPT and Wikipedia to find these keywords

Step 3

Ensured text of the tweet as well as when it is retweeted with individual thoughts, contains education word atleast once in the content of text column and finally it gave me 1.73M Tweets



Data Filtering and EDA

Exploratory Data Analysis

Most of the API columns which I used for analysis comes from the root level and from retweeted status level

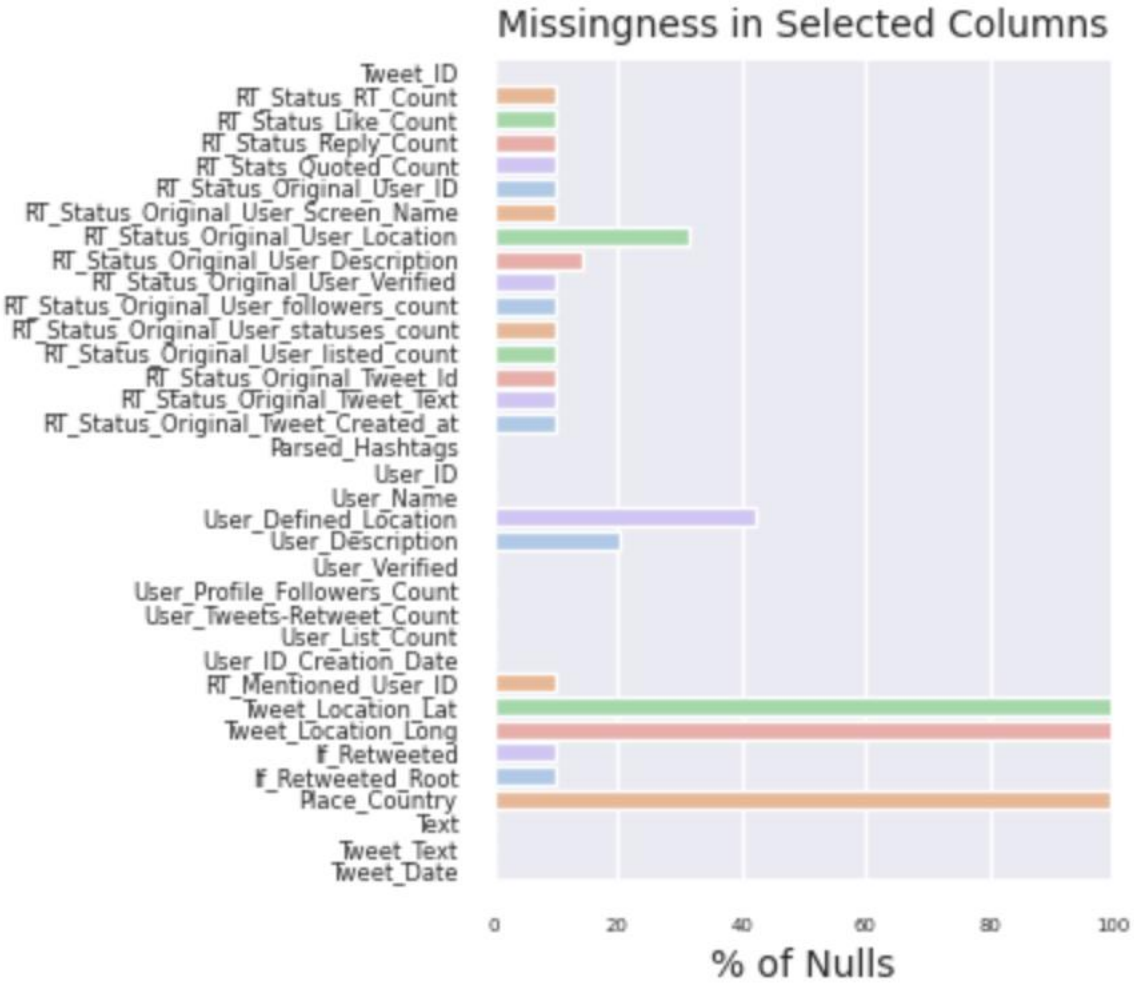
Retweeted status (RT_Status) contains the information when a tweet in the root level is actually a retweet, RT_Status keeps the original tweet details

Some columns are not directly related to analysis such as Parsed_Hashtags, Quoted_Count and etc but will be used to derive other factors such as engagement rate of a twitterer

Place_Country, Lat/Long from geo and coordinates API are 99% null hence we will stick to User_Defined_Location which is around 40% for location analysis

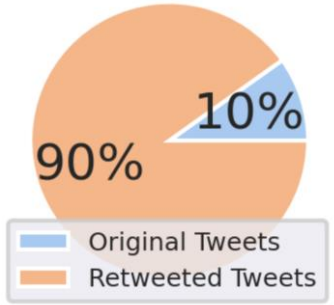
In RT_Status columns there is a constant baseline of nulls which represents the number of **original tweets that is 170K**

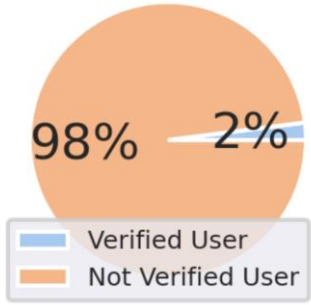
I have renamed the columns as per my convenience and a detailed dictionary is presented in the appendix



Author Identification

By Original Message Volume





To identify best authors by message volume, I filtered tweets which has retweeted column not equal to 'RT'

Only 10% of 1.73M tweets are original, mostly people retweet statuses of others

Out the top 5 profiles who tweet the most, none of them is verified, even though 2395 of 115292 users are verified

- Table 2 shows that verified twitterers are mostly News channels posting surprisingly less educational tweets than non-verified twitterers

To characterise them better I created a metric called **Activity Rate (AR)** which is the number of retweet and tweet by the profile / total followers on that profile

- *nj_education* appears to be best by activity rate as this profile tweets as well as retweets frequently related to education

The top profile *Shopyazanophu33* tweets in high volume but hardly receive any engagement which is evident by few followers as well as low AR whereas *getthatrightgtr* has a better profile in terms of followers with high AR

These twitterers also tend not to be part of many twitter lists

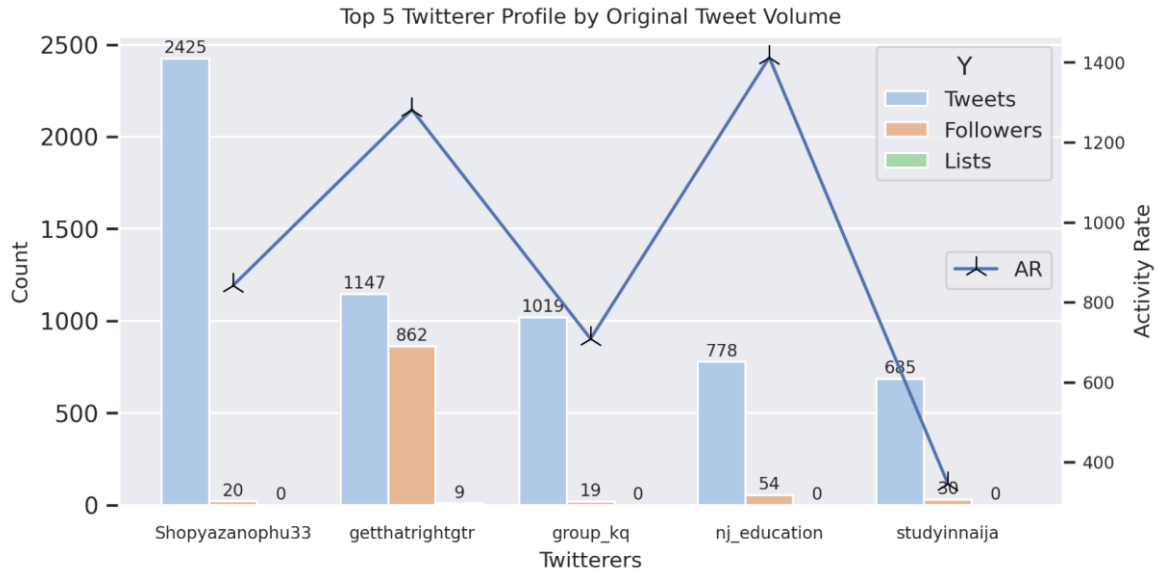
Number of tweets / message volume solely should not be used to identify an influential twitterer

Table A: Top 5 Profiles

User_Name	Shopyazanophu33	getthatrightgtr	group_kq	nj_education	studyinnaija
Tweet_Count	2425	1147	1019	778	685

Table B: Verified Top 5 Profiles

User_Name	TOICitiesNews	MiddleEastMnt	FoxNews	EdSurge	OUSDNews
Tweet_Count	30	29	28	27	27



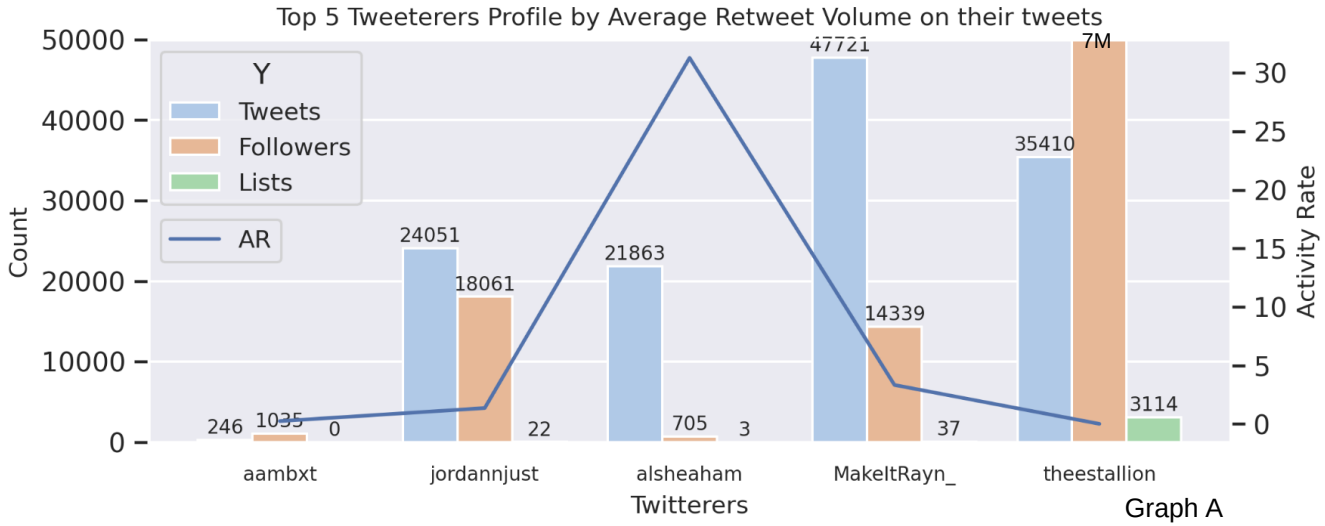
Author Identification

By Volume of Message Retweeted

To see influential authors by how often their messages are retweeted, I filtered tweets which have Retweeted as ‘RT’ and under RT_Status I used original user and average of the RT_count on their unique tweets

Table A: Top 5 Profiles

User_Name	aambxt	jordannjust	alsheaham	MakeItRayn_	theestallion
Avg Retweets / Tweet	230910.0	147265.3	106857.0	73903.0	66864.0

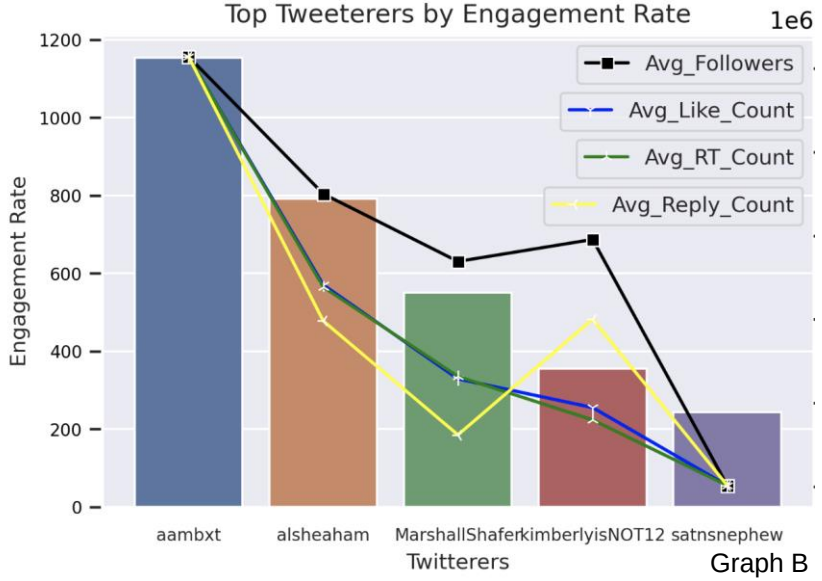


Graph A

Out of the top 5 only 1 profile is verified twitterer (*thestallion*) which is quite evident by the number of followers, see graph A

To further classify profiles as influential we created an engagement score which is sum of the avg likes, replies and retweets per tweet by the number of followers, all tweet attributes such as likes, replies follows identical trend, graph B

aambxt remains at top due to higher avg retweet and reply count per tweet



Graph B

**Line plots are not on the same scale as engagement rate, plotted here to see the trend*

Author Identification

By Entity

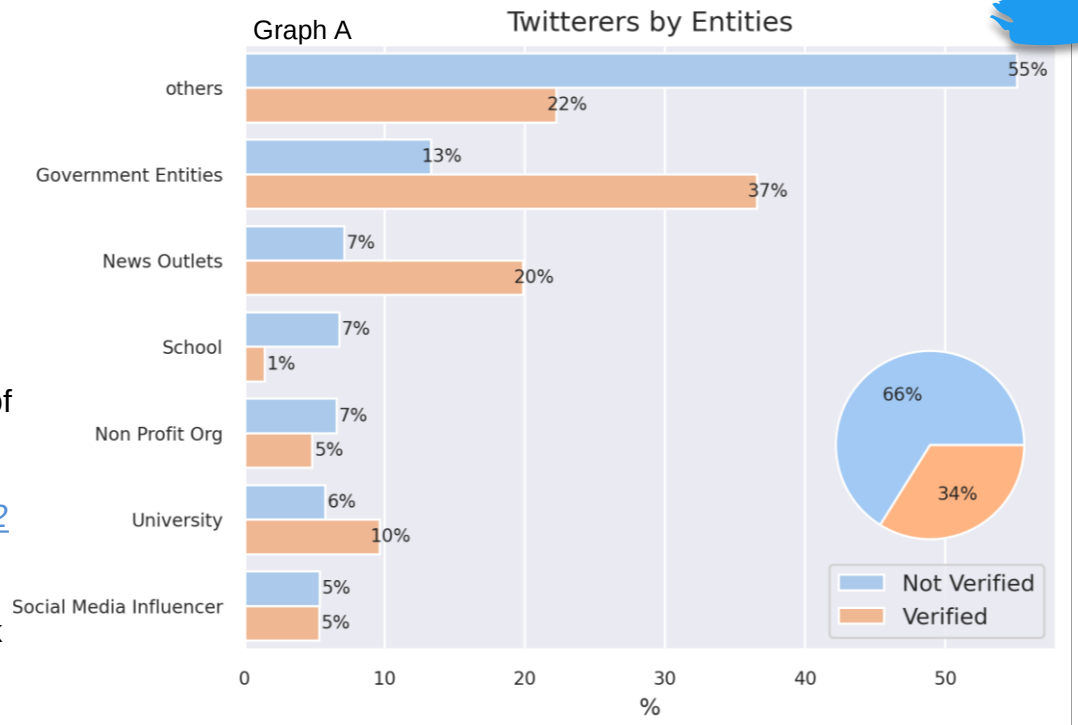
I filtered out tweets where user description is not null and then I used key words related to entities to characterize twitterer, these keywords were matched with the user description and user screen name columns.

37% of Verified twitterers belong to government entity, tweeting about education whereas, from not verified accounts almost 55% are random people, tweeting on educational topics

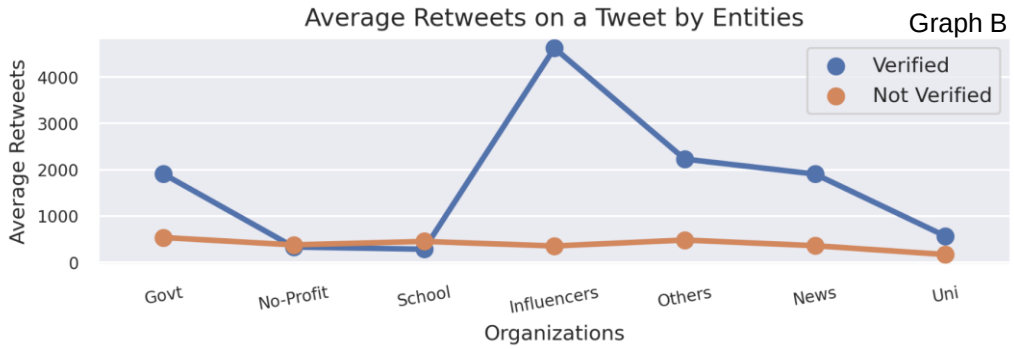
Universities and News and government entities have comparatively higher chuck of twitterers in verified accounts than not verified. On the contrary, schools despite of being not verified are speaking more on education than universities, upon further research, I came across *Higher Education Social Media. Engagement Report 2022* which reports some facts about universities which aids our analysis

- Median Engagement Rate for Higher Education Institutions: 0.096%
- Median Posting Frequency for Higher Education Institutions: 8.8 times per week

Graph A



Graph B



By analyzing the average number of retweets per tweet by entity, it is found that Social Media Influencers gets the highest number of engagement on their tweets which practically makes them the influencers

Overall trend in all entities is constant for not verified as they seldomly get retweets

Surprisingly verified school gets lower retweets on average than non verified which infers that most school are not actively using twitter

Location Analysis

Geographical Distribution of Twitterers

User location is used for location analysis after dropping nulls and cleaning the column, then only unique users were kept for this analysis to accurately identify counts per country with reduced rows to 556534

Analysis is performed on country level

Most of the twitterers posting regarding education are located densely in the US after that in India and then about the same number of twitterers are in UK and Nigeria

Since these are user defined locations and not geo tagged/mapped so it can contain random text as well can potentially reduce our analysis accuracy. However, since it has the least not null values, I stick to this column for location analysis

To support the geographical distribution, I plotted with Country inside Place column (1917 rows only) as well, this is associated with coordinates rather than randomly defined and it also appears to follow the similar distribution, showing top 6 in table B below where KSA comes on the board

Table A: Top 5 countries by Twitterers count

Countries	United States of America	India	UK	Nigeria	Canada
Count of Twitterer	144915	42121	28991	24580	17221

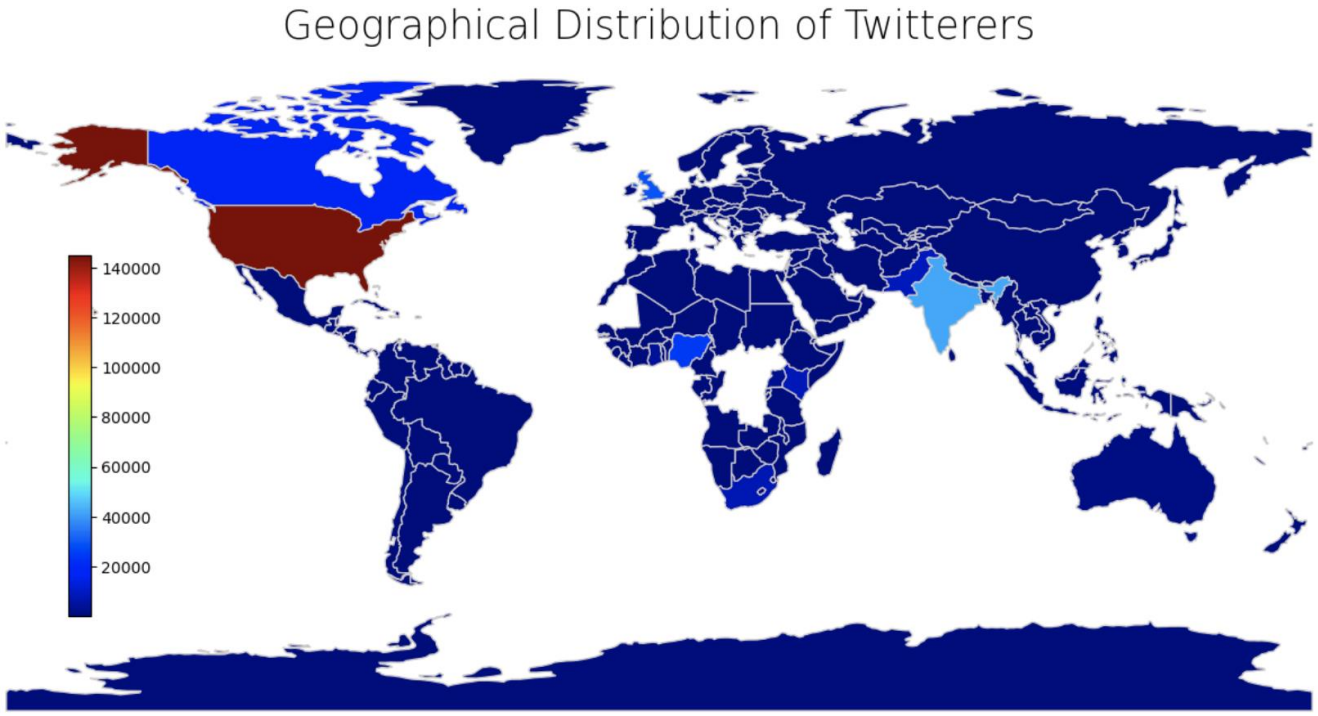


Table B: Top 5 countries pulled from coordinates API for Twitterers count

United States	United Kingdom	India	Kingdom of Saudi Arabia	Nigeria
899	195	144	144	100

Location Analysis

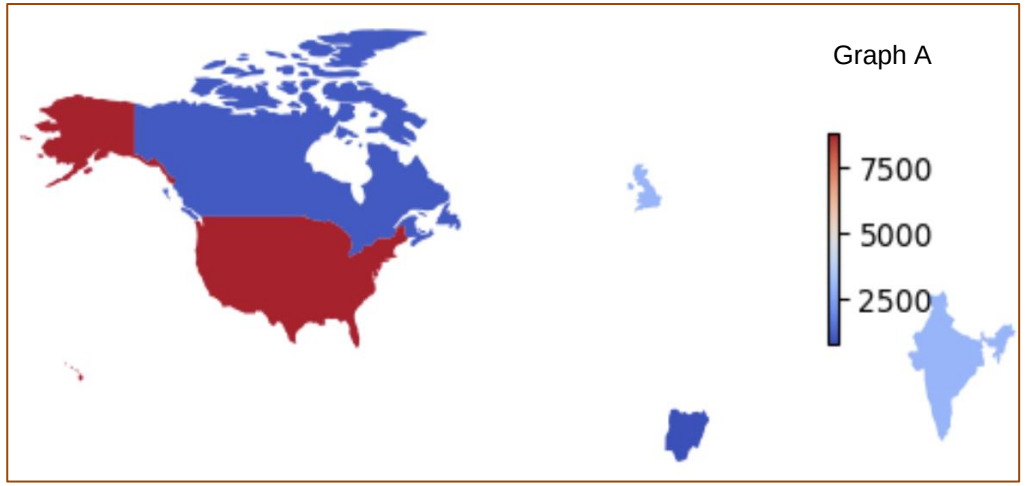
Topic Progression

I analyzed at progression of tweets which are related to word ‘ban’ in the field of education and found only around 40K tweets with distinct user who has defined location in their profile

It has been most talked around in the same top 5 countries we found by count of twitterers

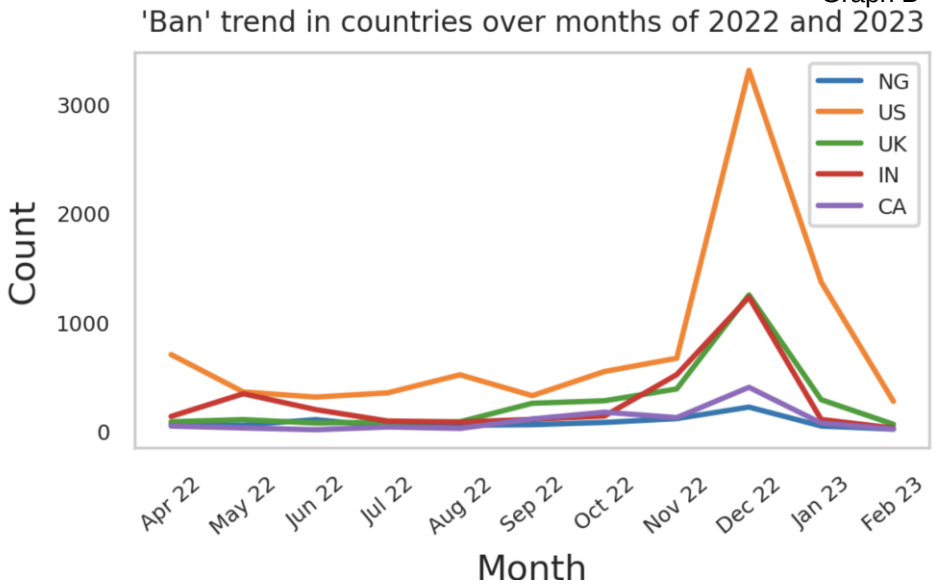
More people in Canada than Nigeria despite of having lesser educational twitterers are raising their voices against educational ban

Graph A



United States of America	India	United Kingdom	Canada	Nigeria
8737	2920	2912	1033	827

Graph B

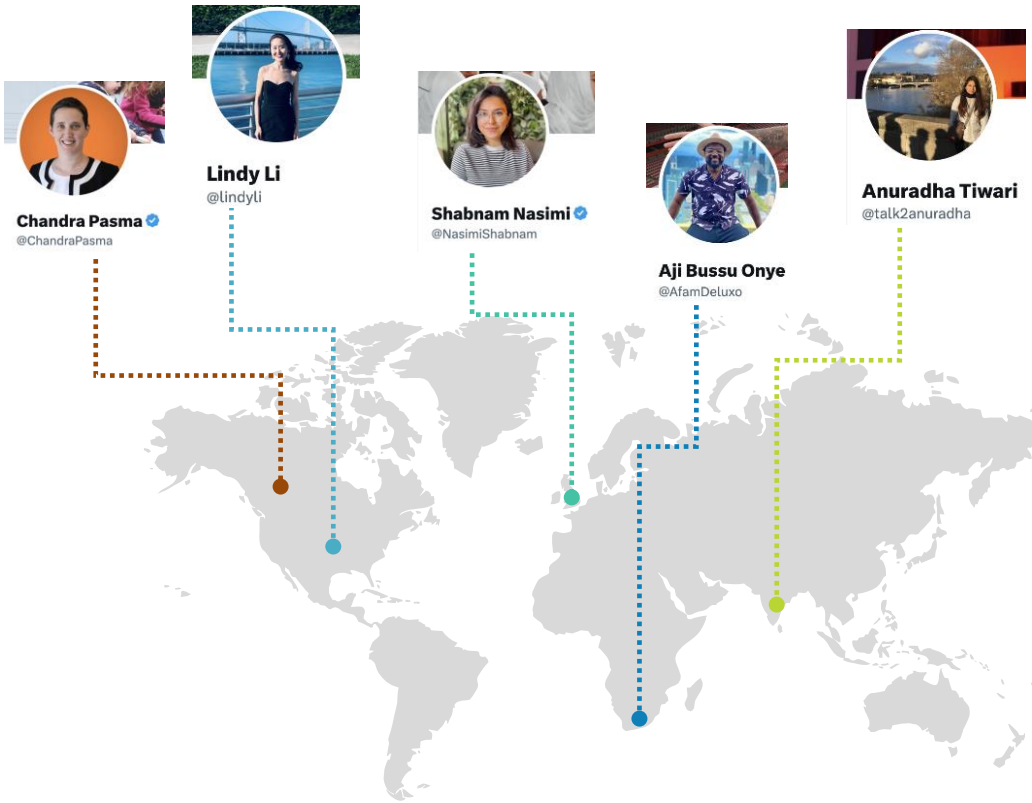


In the year 2022, tweets about banning education-related topics were posted randomly throughout the year. However, in December 2022, this topic gained significant attention worldwide as shown by the peak in all countries on graph B.
Moving forward, let's discuss the event that caused this peak in trend.

Location Analysis

Topic Progression – Peak in Trend

In the top 5 countries by twitterers count, following are profiles speaking volume about ban related to education on twitter



Profile URL are linked on profile pictures
Tweet URL is linked on tweet snapshot

Since the user defined location is not geo tagged and randomly added by the users, so for Canada, UK and USA I found *Shabnam Nasimi* to be the most influential twitter for educational ban tweets but upon exploring the place country column I found her residing in UK and then I selected 2nd most influential twitterers for USA and Canada. Top 5 twitterers in this topic in the top 5 countries are listed in the appendix

Below is the snap of the tweet with highest engagement in this topic causing peak in the trend

This tweet is regarding educational ban for women in Afghanistan, posted by a UK based twitterer who works in Government Entity

As expected, people are posting tweets about the events which are happening all over the world and not just only in their locality

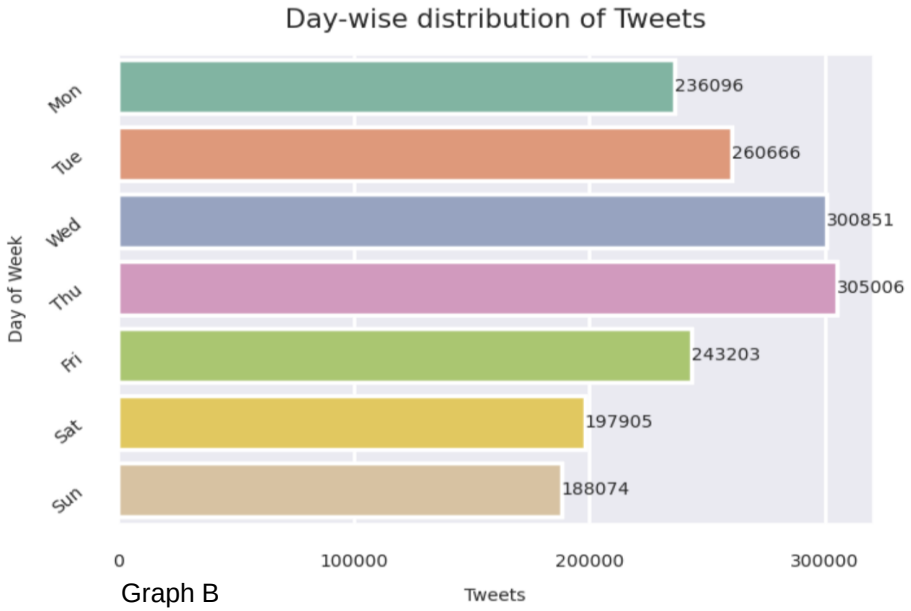
Educational issues are related to topic progression and people from across continents tend to highlight and condemn such activities

Timeline Analysis

Time Series Distribution of Tweets

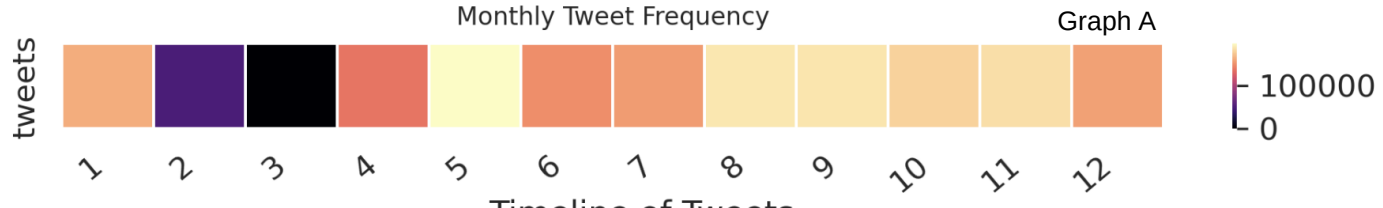
From graph B, mostly people tweet about education during Wednesdays and Thursdays and relatively few wants to spend their weekends thinking about education instead of having fun

From graph C, Verified twitterers sets the trends and then random users retweet such post as can seen by the synchronized behavior of tweet counts by random and verified accounts

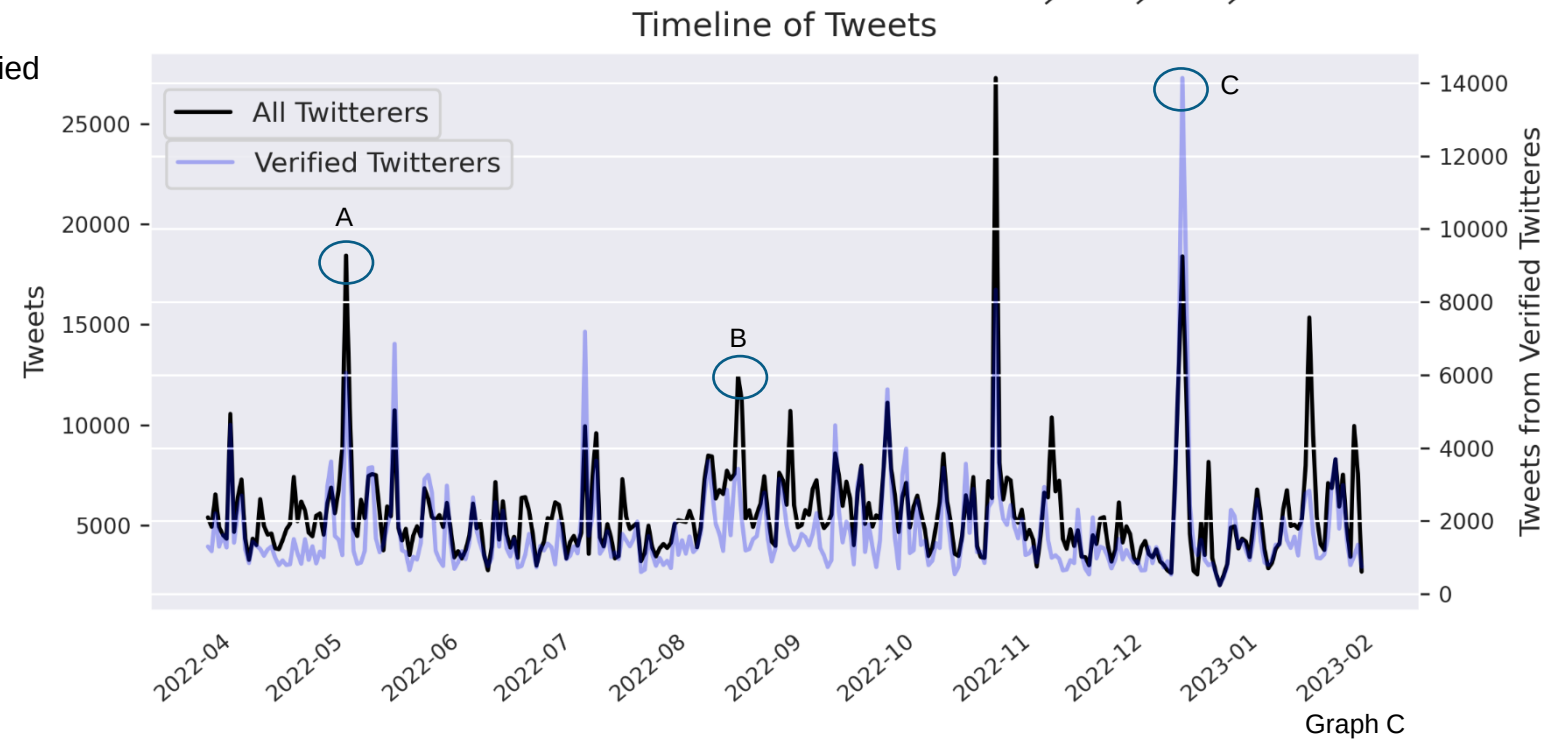


Graph B

In graph A, the cumulative performance of tweets for each month of all years shows a gap in the data for the month of March. The month of June has the highest number of tweets, which could be due to the start of the summer break and people being relieved of academic stress.



Graph A



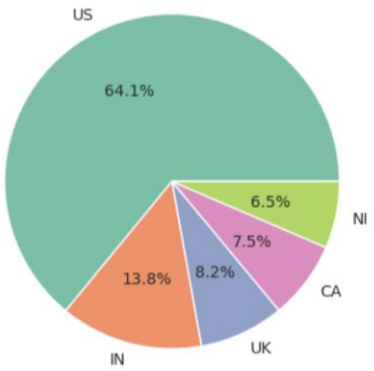
Graph C

There are some spikes observed in the trend and will be discussed next

Timeline Analysis

Spike A – Dates 1-10th May 2022 - Total tweets 55K

Distribution by country



Best Tweet

Leah Remini
@LeahRemini

Going to college for the first time at 51 after living most of my life in a cult and only having an 8th grade education hasn't been easy.

But I took my first NYU final exams this week and I'm glad I decided to educate myself.

It's never too late. It's tough, but it's worth it.

2:38 PM · May 7, 2022

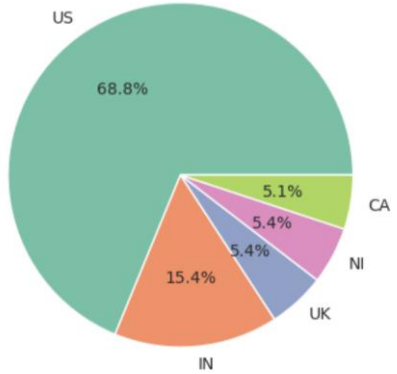
15.1K Retweets **1,600** Quote Tweets **390.4K** Likes

Top Hashtags

Hashtag	count
education	937
ChildrenOfTigray	631
Education	547
savesoil	293
MPTETVARG3_51000	278

Spike B – Dates 25 - 31st Aug 2022 - Total tweets 51K

Distribution by country



Best Tweet

AlwaysAshley
@AshleyAlready

Okay, so I rearranged my whole life to avoid student loan debt, prepaid for my kids' college educations since their births, and now everyone's getting $10,000 forgiven?

I think that's great 😊 Congrats, guys. A rising tide lifts all boats.

4:04 pm · 24 Aug 2022

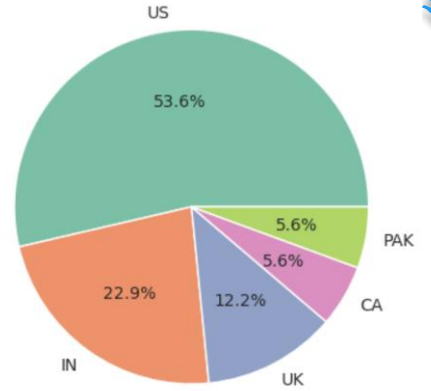
21.4K Retweets **1,385** Quote Tweets **325.9K** Likes

Top Hashtags

Hashtag	count
education	761
90percentscholarship	600
Tigray	317
KCR	265
Education	240

Spike C – Dates 15 - 31st Dec 2022 - Total tweets 100K

Distribution by country



Best Tweet

Shabnam Nasimi
@NasimiShabnam

A day after of the Taliban BANNED female university education, women & girls have come out on the streets of Kabul protesting against the decree.

They chant —"Either for everyone or for no one. One for all, all for one"

Amplify their voices.

Top Hashtags

Hashtag	count
Tigray	3215
education	1389
Education	932
Taliban	783
Afghanistan	736

Tweet URL is linked on tweet snapshot

Message Uniqueness Analysis

Using MinHash LSH Method

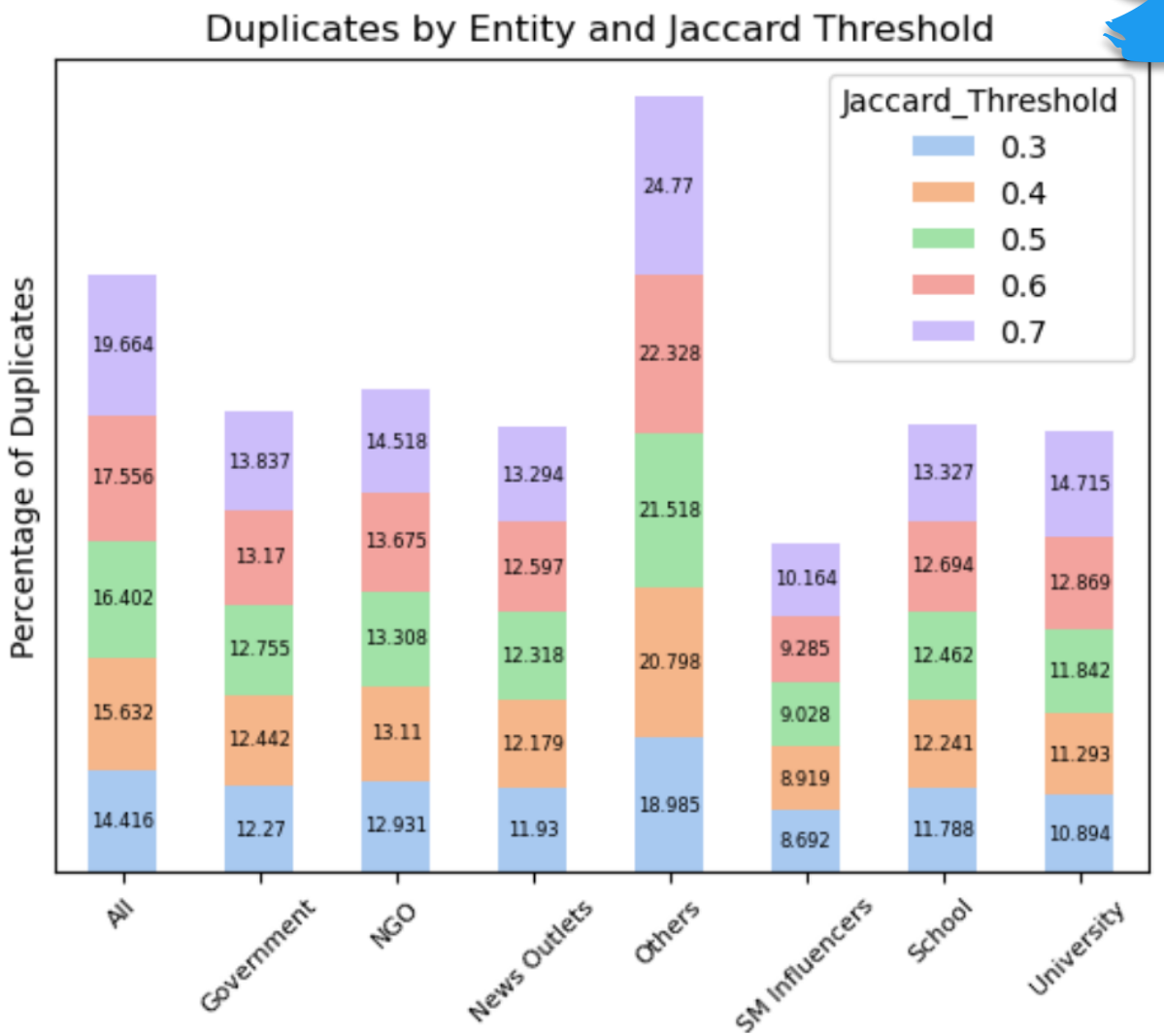
Similarity analysis is conducted using a sample of 10K tweets in all records as well 10k samples in each of the entities

Most of the messages appears to be unique as even at Jaccard threshold of 0.7 there are only around 20% of duplicates, with an exception in others category which has around 25% duplicates as these are random people frequently retweeting others tweets

Out of all entities, Social Media Influencers tweet new and unique messages

For all entities except others, a good Jaccard threshold value is 0.5 as it have fairly equal percentage of duplicates

For other, 0.4 will be a reasonable threshold as 'others' category includes random individuals



Insights

I analysed a dataset of 1.73 million tweets relevant to education, of which only 10% were original tweets. Most users had random profile attributes. When analysing influential authors, only 1 out of the top 5 Twitter users was verified. To address this, I developed an engagement rate metric to identify influential twitterers based on their profile attributes.

Random user do not belong to any entity while verified Twitter users were mostly affiliated with government organizations instead of educational institutions. This was unexpected and was further explained in the Higher Education Social Media Engagement Report, which showed a low engagement rate of 0.08%.

User-defined locations were used for location analysis due to missing data in geotagged locations. Most of the 1.73 million tweets analysed were from the US, India, UK, Canada, and Nigeria. The most influential authors were found to be based in these countries, with the US being the top-ranked country.

'Ban' in education was a hot topic in December 2022 due to the ban on women from attending universities in Afghanistan. Even though the event occurred in Afghanistan, the top 5 countries tweeted about it the most, indicating global concern and awareness about current events.

Spikes in tweet counts related to educational bans, student loans, and returning to college were identified by examining the available data timeline.

Overall these tweets are unique messages but among the various entities, Social Media Influencers were found to have the highest number of unique messages, indicating that their content is diverse and not repetitive.

Conclusion and Recommendations

Conclusion

Based on my analysis, I found that Twitter is a credible source for educational topics that are politicized, but not for new research in education. This is because there are very few official university accounts on Twitter, and out of the top educational Twitter users, only one was verified. Moreover, with the changes in management, verified Twitter accounts are becoming less reliable and can be obtained for a small fee.

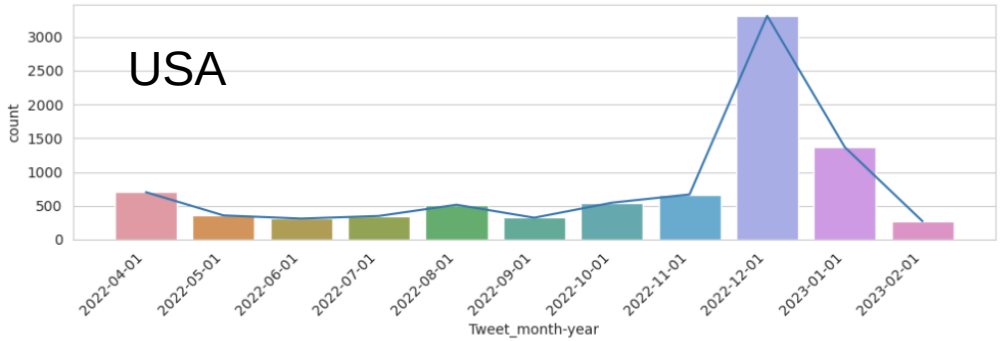
Recommendations

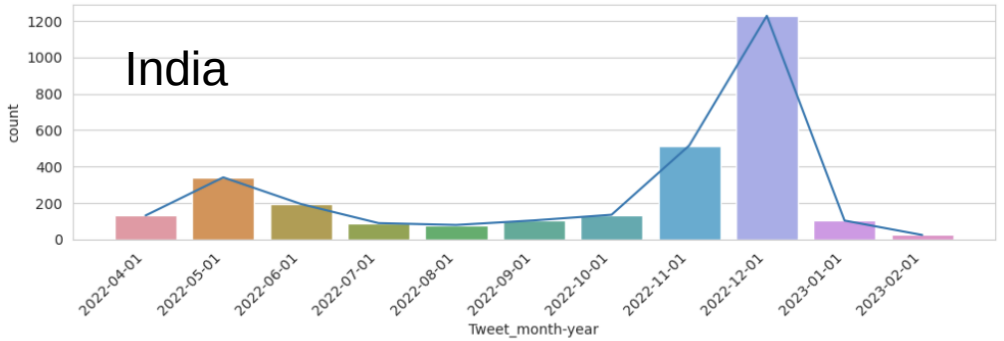
- Employing sentiment analysis can enhance the categorization of educational tweets based on their nature.
- Getting better location data with geotagging to accurately locate the user
- Having impression count of a tweet would enable us to determine a more precise engagement rate, facilitating a more accurate classification of influential users.
- Verified Twitter accounts should not be regarded as a credible source of information since it can be bought cheaply

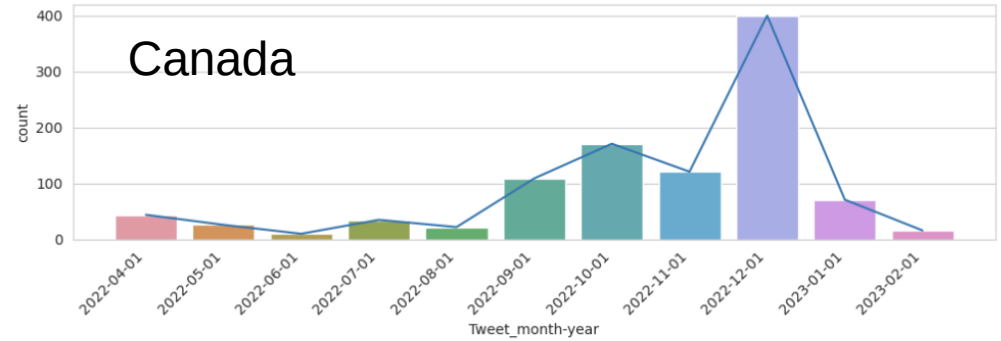
Appendix

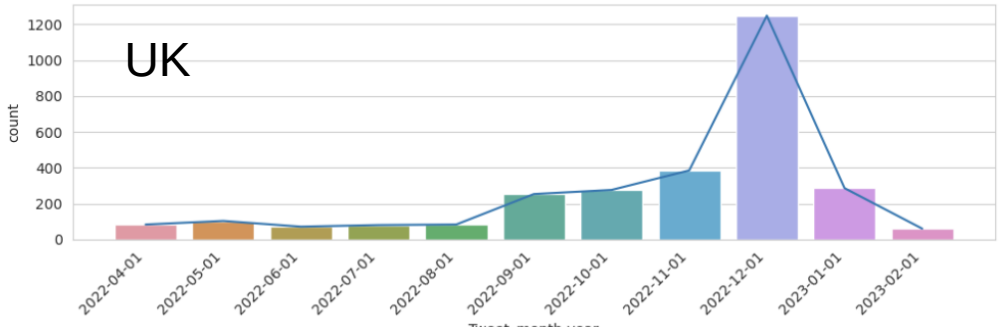
Appendix

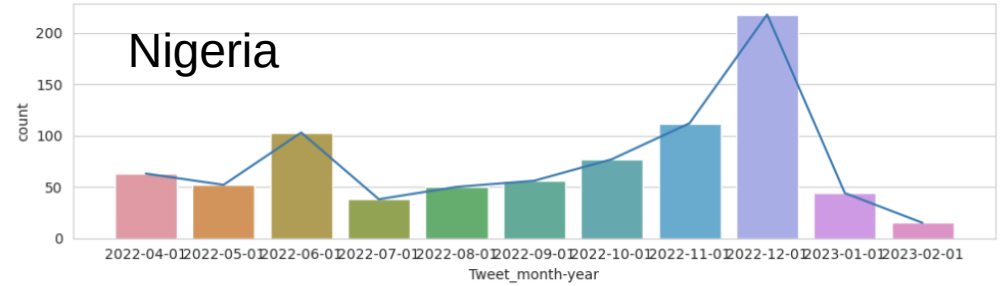
Topic Progression for Bans related to Education by countries











Appendix

Top Profiles talking about Bans related to Education in top 5 countries

User_Defined_Location	RT_Status_Original_User_Screen_Name	RT_Status_Original_User_ID	Tweet_Counts	Avg_RT
United States of ...	NasimiShabnam	3317348164	3110	14762.346623794212
United States of ...	lindyli	270132611	762	3378.3937007874015
United States of ...	null	null	630	null
United States of ...	AfricanArchives	808806102	456	2811.252192982456
United States of ...	Lancegooden	1029094268542099457	416	2194.1875

User_Defined_Location	RT_Status_Original_User_Screen_Name	RT_Status_Original_User_ID	Tweet_Counts	Avg_RT
Canada	NasimiShabnam	3317348164	478	12383.640167364018
Canada	ChandraPasma	984232620975247361	164	478.6219512195122
Canada	BhutilaKarpoche	867462325049208833	153	562.562091503268
Canada	Jenncun05871935	1417428142977962007	111	778.972972972973
Canada	BBCYaldaHakim	27831488	92	188.3586956521739

User_Defined_Location	RT_Status_Original_User_Screen_Name	RT_Status_Original_User_ID	Tweet_Counts	Avg_RT
UK	NasimiShabnam	3317348164	1516	8946.482189973614
UK	paul__johnson	35720019	307	1808.5765472312703
UK	BBCYaldaHakim	27831488	294	176.7312925170068
UK	DanielJMath1	1397127641874976768	287	2166.5644599303137
UK	educationgovuk	143039548	157	800.1783439490446

User_Defined_Location	RT_Status_Original_User_Screen_Name	RT_Status_Original_User_ID	Tweet_Counts	Avg_RT
Nigeria	AfamDeluxo	159638969	155	310.81290322580645
Nigeria	henryshield	1000955520	151	627.1059602649007
Nigeria	FS_Yusuf_	468685040	130	584.0
Nigeria	firstladyship	201237617	111	406.1081081081081
Nigeria	NasimiShabnam	3317348164	95	13333.884210526316

User_Defined_Location	RT_Status_Original_User_Screen_Name	RT_Status_Original_User_ID	Tweet_Counts	Avg_RT
India	talk2anuradha	1095547021	527	2616.4440227703985
India	NasimiShabnam	3317348164	467	8171.471092077088
India	RishiSunak	1168968080690749441	457	3411.205689277899
India	ashoswai	70355674	214	864.2803738317757
India	ANI	355989081	186	369.5860215053763

Appendix

Data Dictionary for API columns equivalent in local Data Frame

	Twitter JSON	Local DataFrame
0	id	Tweet_ID
1	entities.hashtags.text	Parsed_Hashtags
2	user.id	User_ID
3	user.screen_name	User_Name
4	user.location	User_Defined_Location
5	user.description	User_Description
6	user.verified	User_Verified
7	user.followers_count	User_Profile_Followers_Count
8	user.listed_count	User_List_Count
10	geo.coordinates	Tweet_Location_Lat-Long
11	retweeted	If_Retweeted_Root
12	place.country	Place_Country
13	text	Parsed_Hashtags
14	text	Text
15	tweet_text	Tweet_Text

	Twitter JSON	Local DataFrame
16	created_at	Tweet_Date
17	retweeted_status.entities.user_mentions.id	RT_Mentioned_User_ID
18	retweeted_status.retweet_count	RT_Status_RT_Count
19	retweeted_status.favorite_count	RT_Status_Like_Count
20	retweeted_status.reply_count	RT_Status_Reply_Count
21	retweeted_status.quote_count	RT_Stats_Quoted_Count
22	retweeted_status.user.id	RT_Status_Original_User_ID
23	retweeted_status.user.screen_name	RT_Status_Original_User_Screen_Name
24	tweeretweeted_status.user.locationt_text	RT_Status_Original_User_Location
25	retweeted_status.user.description	RT_Status_Original_User_Description
26	retweeted_status.user.verified	RT_Status_Original_User_Verified
27	retweeted_status.user.followers_coun	RT_Status_Original_User_followers_count
28	retweeted_status.user.statuses_count	RT_Status_Original_User_statuses_count
29	retweeted_status.user.listed_count	RT_Status_Original_User_listed_count
30	retweeted_status.id	RT_Status_Original_Tweet_Id
31	retweeted_status.text	RT_Status_Original_Tweet_Text