Twitter Identity In Education Analysis

Big Data Platforms – Winter 2023

Final Project Education

Minh Vo

Agenda

- Executive Summary
- Data Overview
- Methodology
- EDA
- Author Identification
- Location Analysis
- Timeline Analysis
- Message uniqueness analysis
- Conclusions & Recommendations

Executive Summary

Twitter is currently one of the primary social media platforms in the world, where a huge amount of information is posted daily by millions of individuals, including influencers, as well as news and organizations.

This project focuses on analyzing users tweeting about education and identify whether Twitter can be considered a credible source of information for emerging topics in education.

Key Initiatives:

- Who are the influential Twitterers regarding education messages?
- Where are these Twitterers located?
- Are there any relationships between Twitterers' location and emergence of hot issues in education?
- What are the timelines of these education-related tweets?
- Are these tweets unique or copied from each other?

Data Overview & Methodology

Data Overview

- Twitter Data is retrieved from Twitter API
- 100 million tweets (500GB), under JSON format.
- Consists of Tweet, User, Geo, and Entities objects
- Fundamental attributes: *user_id, created_at, text, user, entities,* and *place*

Methodology

- Data processing and storage: Google Cloud Platforms
- Data Analysis: PySpark
- Visualization: Seaborn, Matplotlib
- Text Similarity Analysis: MinHash & Jaccard

Exploratory Data Analysis

Data Cleaning:

- 10 education-related key words were selected to filtered out irrelevant tweets are ***education, K-12, teachers, professors, students, university, college, schools, curriculum***

  After filtering: **99,9M rows** → **24,7M rows**

Feature Selection:

- There are more than 60 attributes with child object in the dataset, yet not every attributes will be used.
- Thus, only 15 features are selected for analysis. ***Text***, ***User***, ***Retweet***, **and *Place* are important features**

text	lang		
retweeted	retweeted_from	retweeted_status	retweeted_status.*retweet_count*
place.*country*	place.bounding_box.*coordiates*		
user.*id_str*	user.*screen_name*	user.*location*	user.*description*
user.*followers_count*	user.*created_at*	user.*verified*	

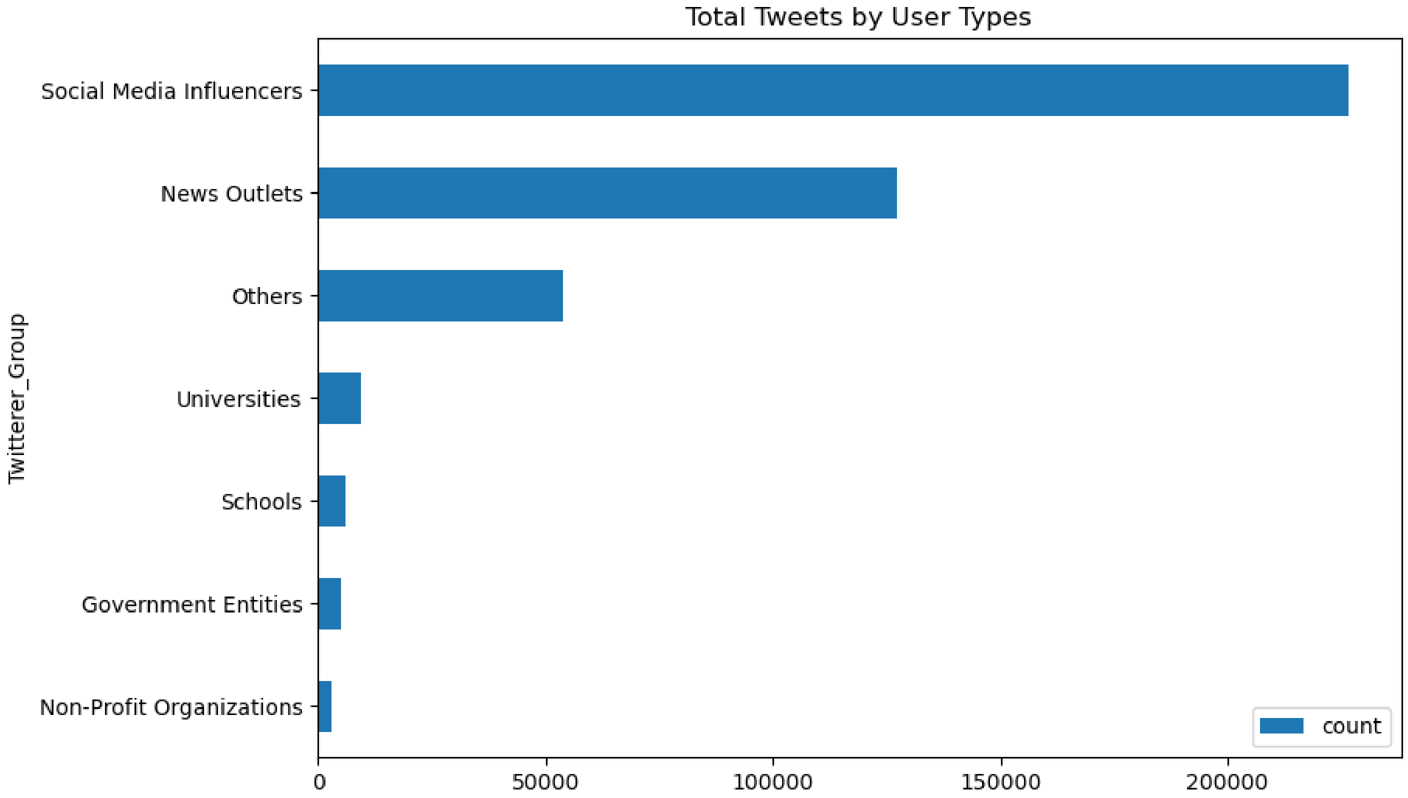
Data Filtering:

- To enhance the accuracy of analysis, missing values are also filtered out from *country, coordinates, location,* and *description.*

Author Identification

Tweet Messages are mostly from the educational social media influencers

- Number of verified Twitterers accounts for 2% of the data.
- Verified Twitterers are divided into 7 groups.
- Majority of Twitterers are **social media influencers**, followed by the **news outlet**.

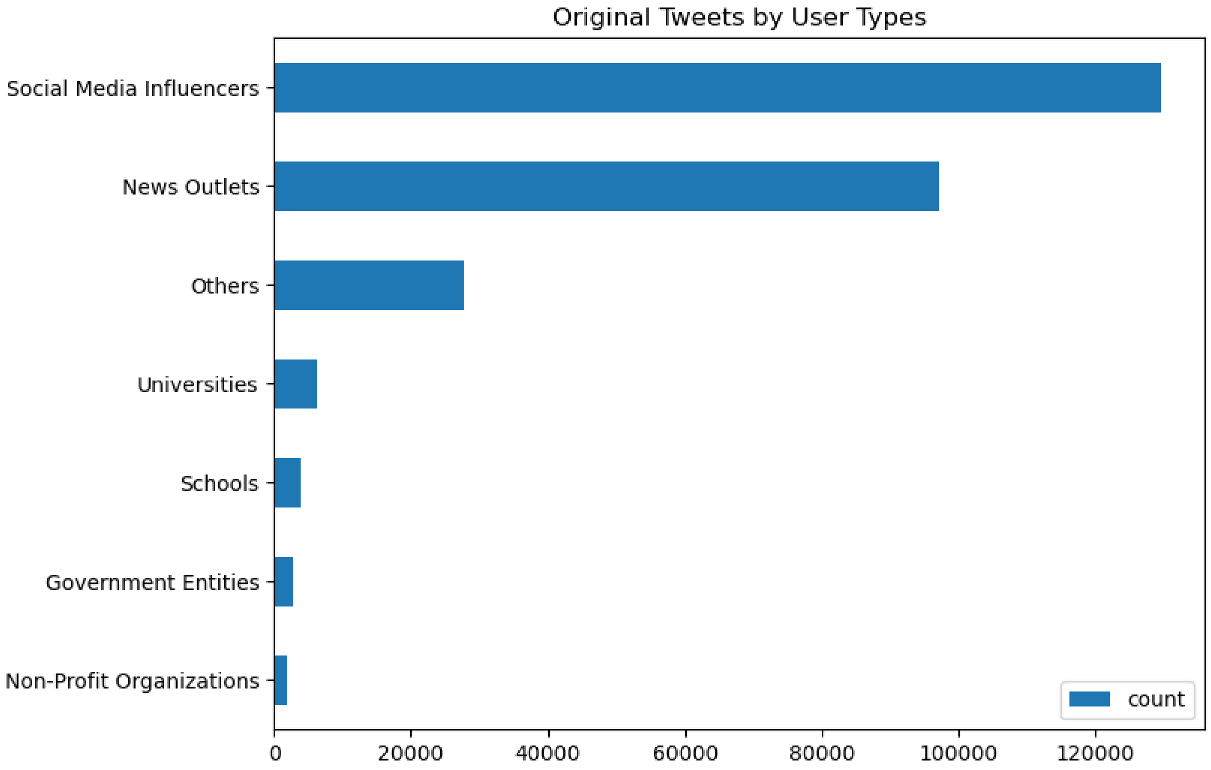


User Groups	Total Tweets
Social Media Influencers	226,766
News Outlets	127,342
Others	53,943
Universities	9641
Schools	6209
Government Entities	5067
Non-Profit Organizations	2928

Author Identification

Top Verified Users By Original Messages are from Education Influencers & News

- There are approximately 8 millions of original tweets among 24 millions of tweets. Majority of them are posted by the social media influencers with 226,766 tweets.
- Apart from influencers, news outlets also generate a number of education-related tweets (127,342 tweets).

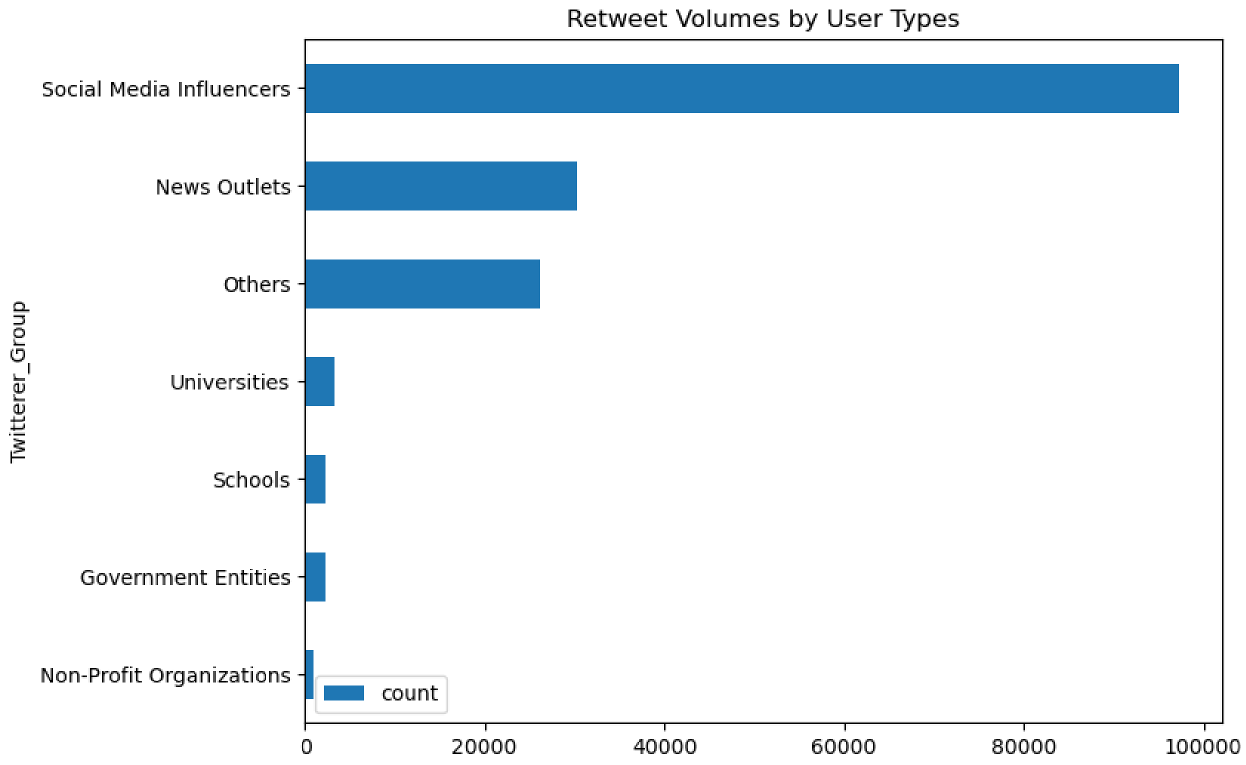


User Name	User Group	Original Tweets	Followers
sportsthread	Social Media Influencers	3148	206,502
USNewsEducation	News Outlets	764	332,591
timeshighered	News Outlets	726	317,637
SF_England	Social Media Influencers	677	127,004
TOICitiesNews	Social Media Influencers	636	27,844
ExploreLearning	Social Media Influencers	576	9,477
WashTimes	News Outlets	514	447,359
SportsBookWire	Others	471	3,847
tes	News Outlets	439	362,347
DeAngelisCorey	Social Media Influencers	430	129,470

Author Identification

Education-related messages are mostly retweeted from Social Media Influencers

- Compared to the original tweets, the retweet volumes are much higher.
- There are not many tweets by News Outlets as well as organizations that are retweeted by Twitterers. In other words, Twitterers tend to retweet education posts generated by influencers.

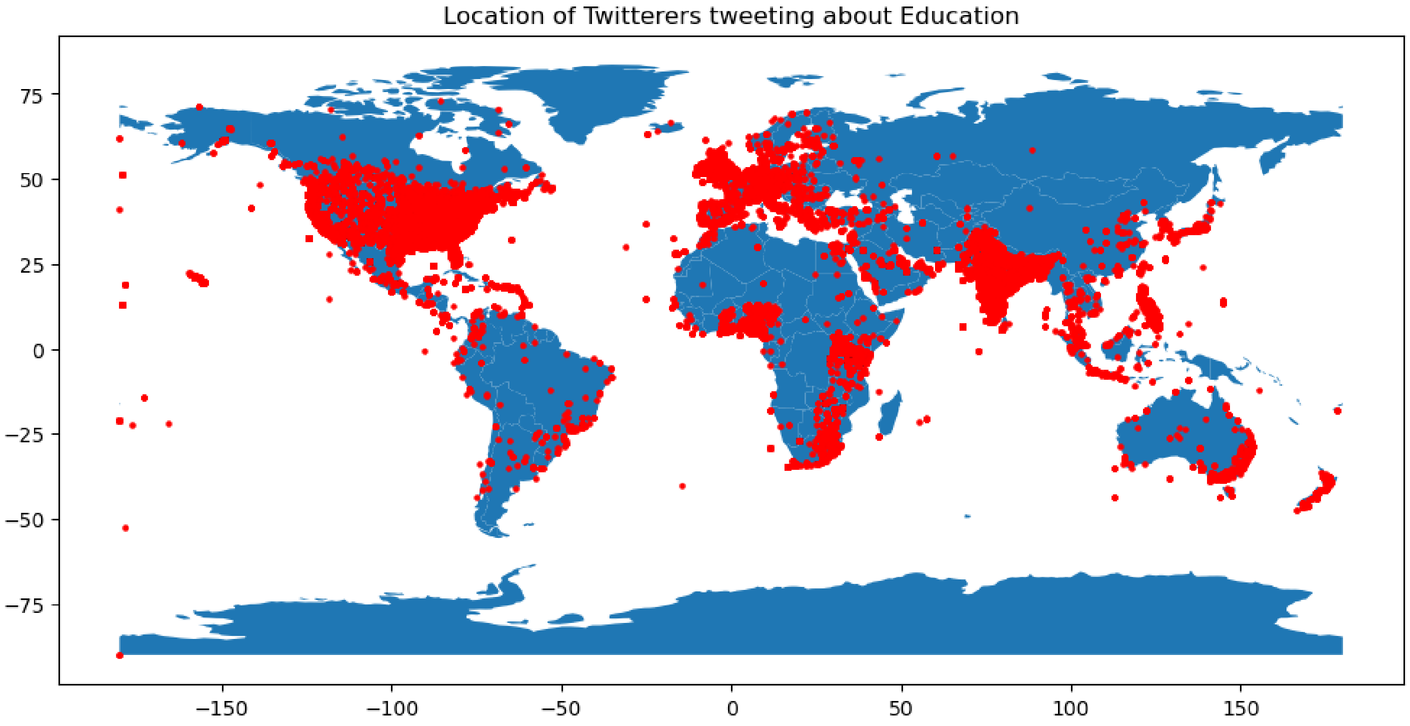


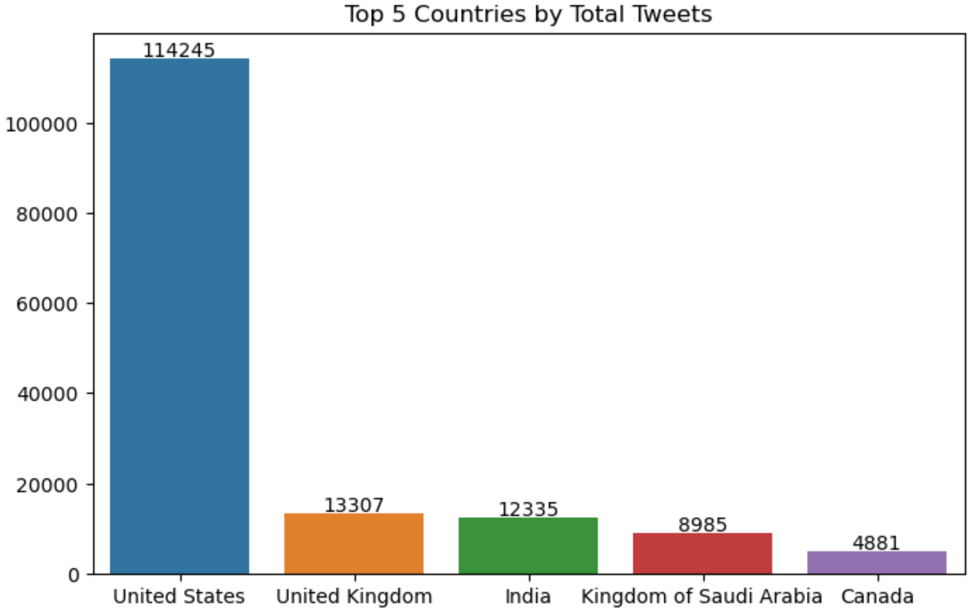
User Name	User Group	Retweets	Followers
NasimiShabnam	Social Media Influencers	440,753	175,063
ValaAfshar	Social Media Influencers	373,676	976,616
LeratoMannya	Social Media Influencers	317,801	78,164
otepofficial	Social Media Influencers	302,430	46,726
JonesHospodTX	Social Media Influencers	190,364	19,399
jwomack	Social Media Influencers	170,683	5,422
FarrahFazal	Social Media Influencers	151,149	10,703
MarthaKelly3	Social Media Influencers	144,308	30,711
2tall4u2	Social Media Influencers	143,002	52,239
juscohen	Social Media Influencers	135,084	7,330

Location Analysis

Majority of Twitterers come from the U.S.

- Most of the Twitterers are in the U.S., especially in the East Coast.
- Despite a huge gap in tweet volumes between countries, based on the map, it can be seen that Twitter is also a common social media platform in United Kingdom and India.



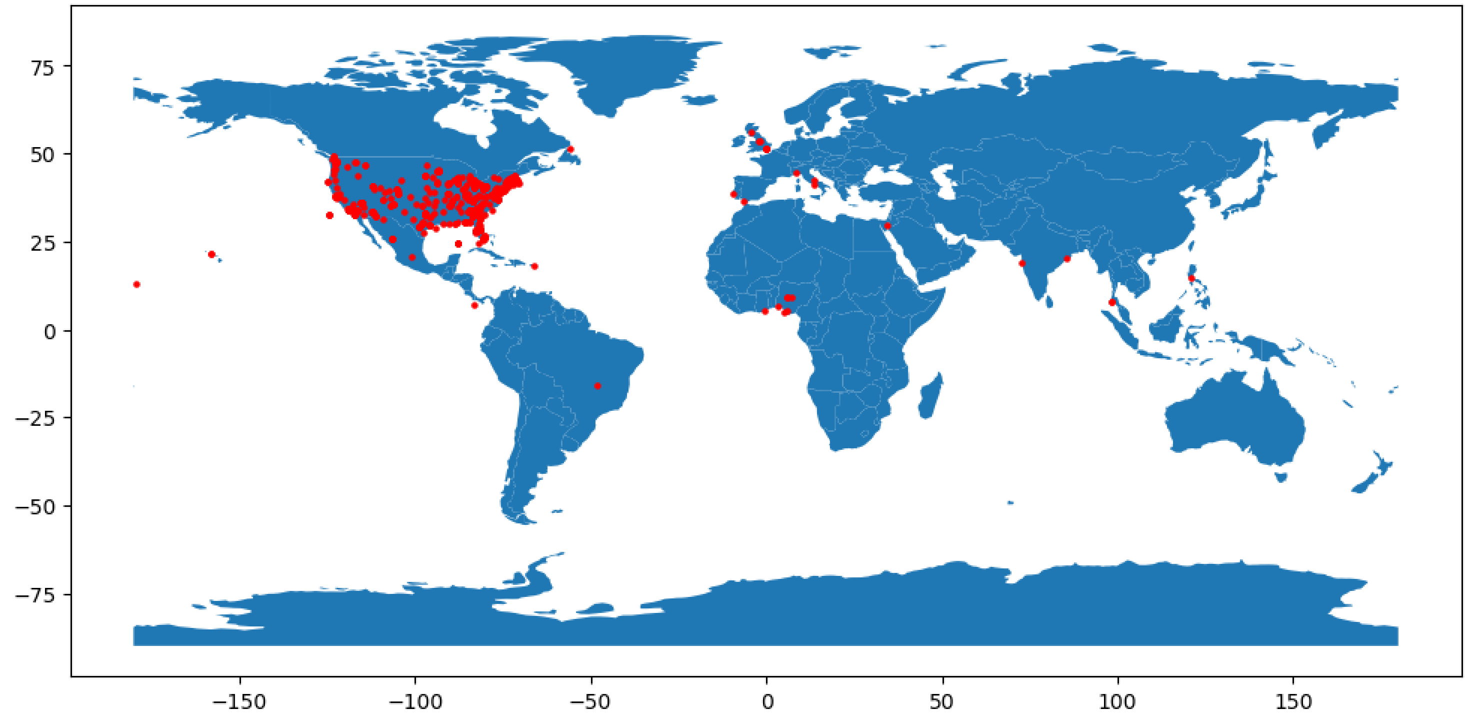


Location Analysis

“Book ban” and “Student loans” topics tend to be more common among U.S. Twitters.

- Majority of Twitterers who tweet about *‘book ban’* and *‘student loans’* issues are in the U.S compared to the remaining countries in the top 5.
- This is obvious since U.S. Twitterers account for the highest number. Hence, there seems to be no relationship between the emergence of new issues in education and these Twitterers’ locations because of the bias.

Location of Twitterers tweeting about hot topics in education



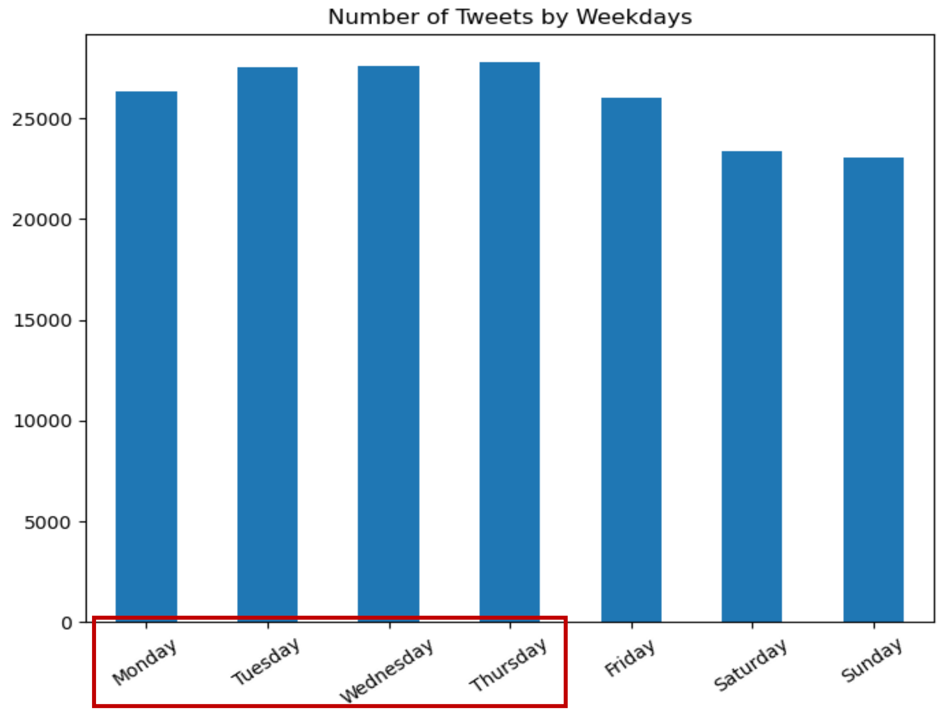
Top 5 countries with the highest number of Twitterers tweeting about **‘book ban’** & **‘student loans’**

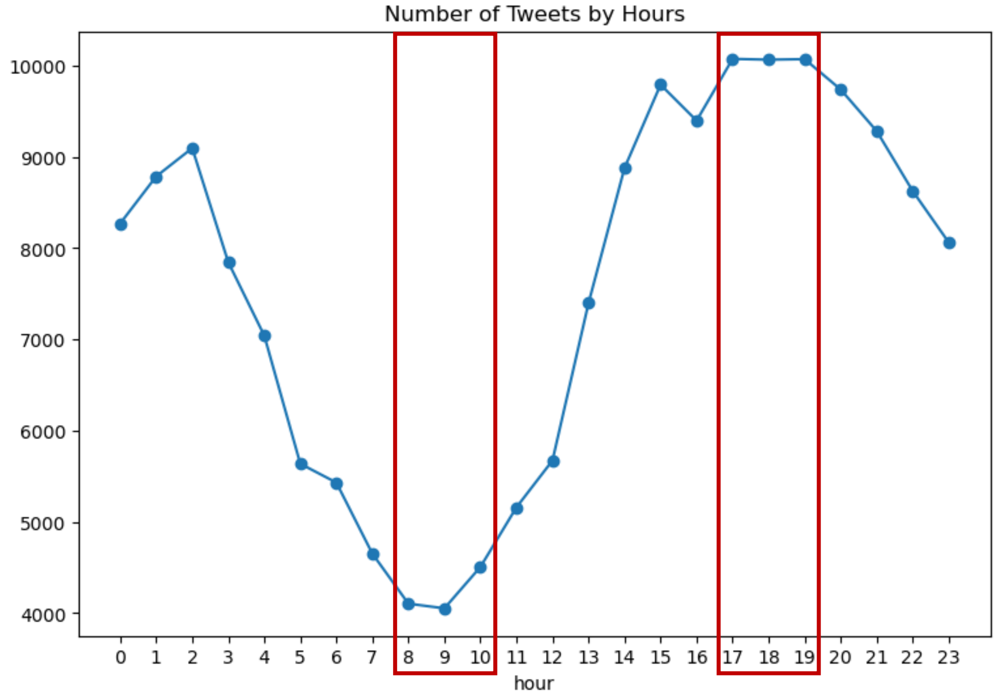
Country	Total Tweets
United States	476
United Kingdom	8
Nigeria	6
Canada	4
Thailand	2

Timeline Analysis

Education-related tweet volumes tend to be higher on weekdays and after 5 PM during a day

- During the week, there are more tweets generated during **weekdays**. Particularly, **Tuesday** to **Thursday** recorded the most tweet volumes.
- However, this number decreases from Friday. Therefore, the number of tweets on weekends are lower than on weekdays.
- One of the reasons could be because schools as well as news outlets and organizations tend to post and update latest news and information on weekdays to keep users updated faster and easier.
- Throughout a day, the highest volume of tweets occurs after working hours, which is from **5 – 7 PM**; while the fewest tweets are posted from 8 – 10 AM, when most people go to schools and work.

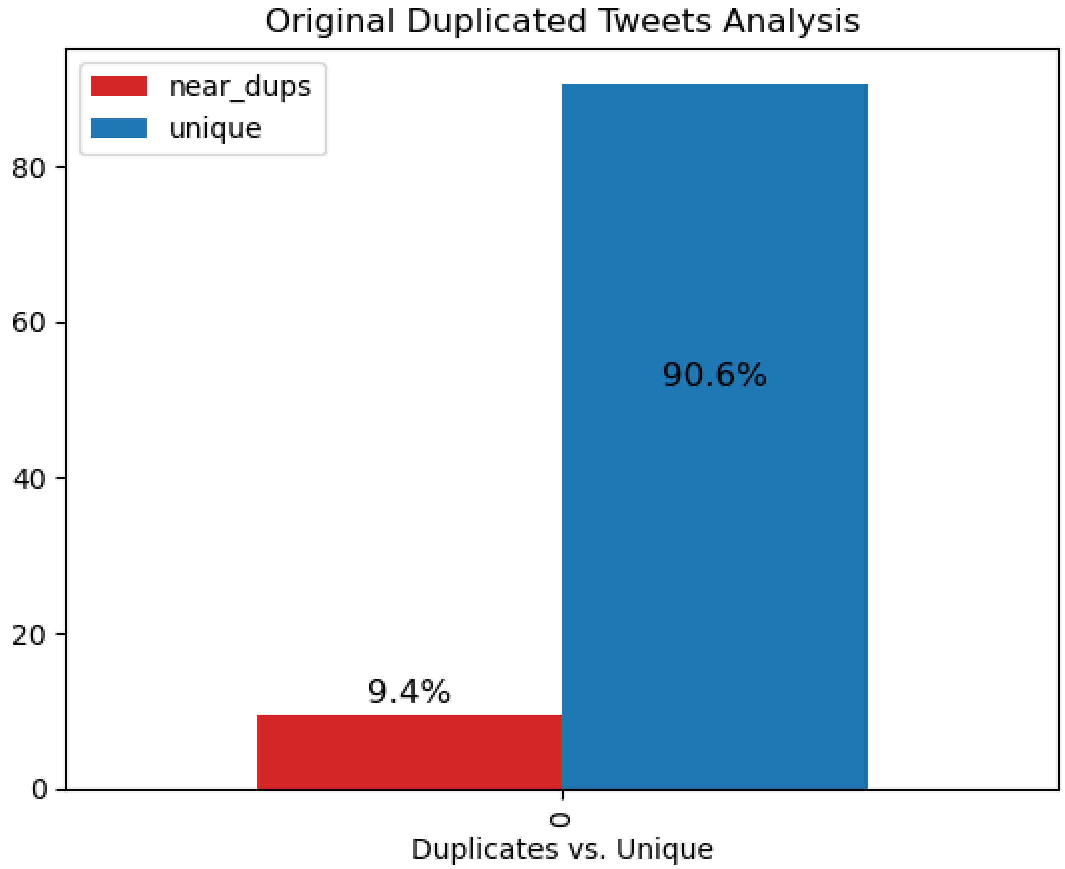




Message Uniqueness Analysis

Given around 8 millions of original tweets and Jaccard distance of 0.3, over 90% of the tweet volumes are unique.

- Only original tweets were used as the retweet messages can cause the duplicated tweet volumes to be high.
- MinHash and Jaccard similarity (with distance = 0.3) was applied to measure tweets' uniqueness.
- Unexpectedly, the number of unique tweets account for up to 91%. Since these tweets are originally generated by verified Twitterers, the possibility of having high duplicated tweets is relatively low.



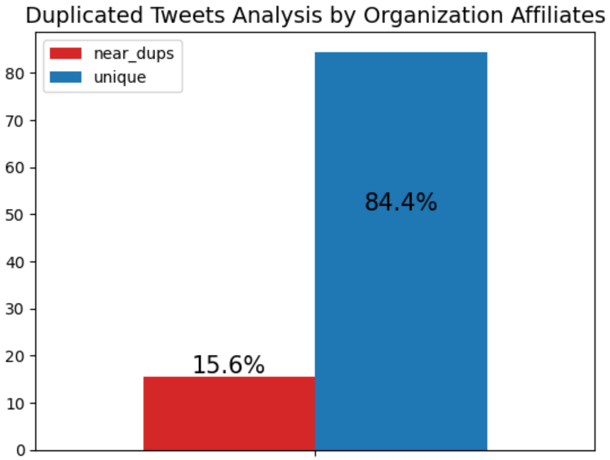
**Because of the huge data volume, 1% of data was randomly selected for analysis.*

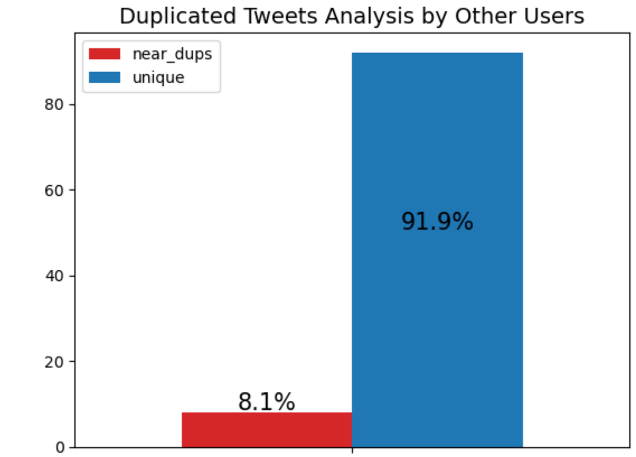
Message Uniqueness Analysis

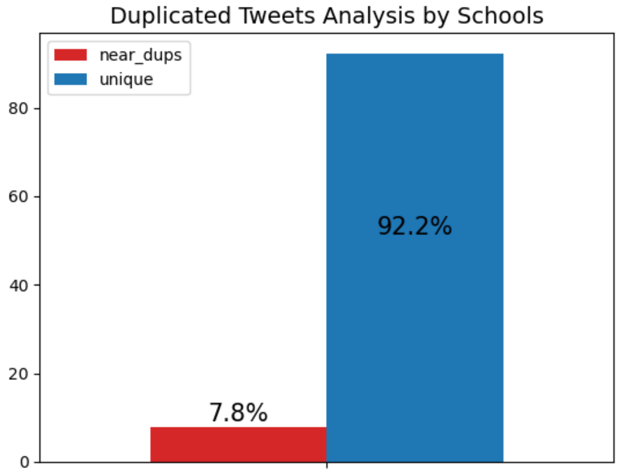
All types of user groups have high percentage of unique original tweets

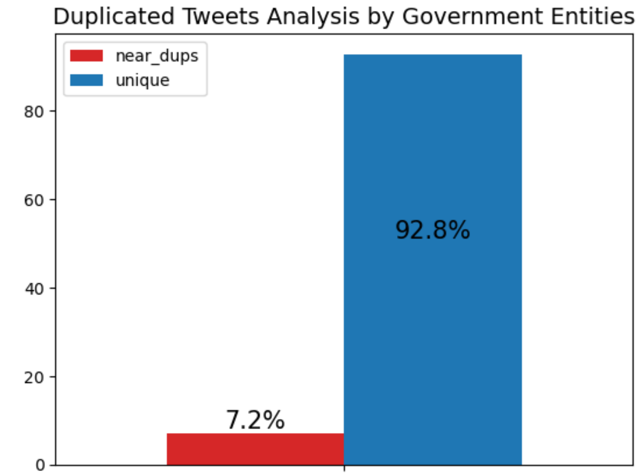
- Given the Jaccard distance equal to 0.3, it is shown that organization affiliates tend to have higher proportion of “near duplicates” tweets (15.6%).
- Remaining user groups have low duplicated tweet volumes (under 10%).

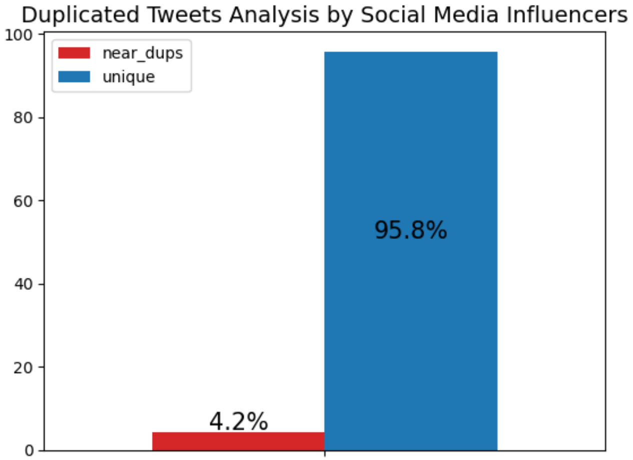
**10% of data from News Outlets and Influencers, and 5% of data from Other Users were taken as samples for analysis due to the high tweet volumes.*

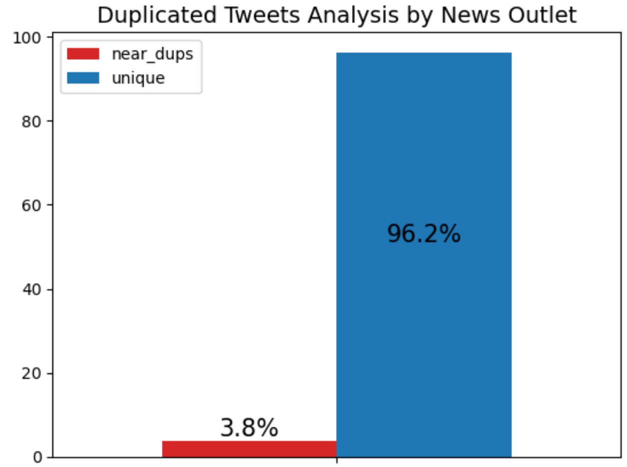

Duplicated Tweets Analysis by Organization Affiliates


Duplicated Tweets Analysis by Other Users


Duplicated Tweets Analysis by Schools


Duplicated Tweets Analysis by Government Entities


Duplicated Tweets Analysis by Social Media Influencers


Duplicated Tweets Analysis by News Outlet

Conclusions & Recommendations

Conclusions

- Social media influencers and News Outlets have a great contribution to the credible source of education-related information on Twitter. Verified influencers also have higher number of retweeted message volumes compared to other user groups.
- However, since majority of Twitterers are not influencers or particular organizations, their tweets are mostly statuses regarding schools, studying, etc., rather than post about a specific education topic.
- There is no strong relationship between the emergence of new issues in education and these Twitterers' locations.

Recommendations

- Focusing on social media influencers in education field. There are many influencers who usually tweet meaningful and impactful messages but have not been widely recognized, probably because they are not verified Twitterers yet.
- Measuring the influence of Twitters given their total followers, total liked tweets.
- Knowing which education topics are becoming hot issues as well as when and where they occur is crucial.
- Understanding when Twitterers tend to tweet based on timeline can be an effective way to help increase tweeting interactions and identify emergence of important trends in education.

THANK YOU!