Agenda

- Executive Summary

- **Methodology and Source Data Overview**

- Tweet Clean-up and Filtering

- EDA and Extensive Usage of Available Variables

- Author Identification

- Location and Timeline Analysis

- Message Uniqueness Analysis

- Conclusions

- Actionable Recommendations

Executive Summary

This analysis aims to determine whether Twitter can be considered a reliable source of information for emerging trends and important topics in education. The most prolific Twitter users are news organizations, while celebrities are the most influential Twitter users by retweet volume. The top four countries with the highest number of Twitter users in education are the USA, Nigeria, India, and England, with the highest number of Twitter users in Asia. The timeline of tweets in Asia shows a pattern unique to India, which is related to college admission exams and enrollment season. Tweets about CRT and Hijab are concentrated in North America and India, respectively, and were influenced by recent events. The analysis also shows that messages are mostly unique, but selecting different Jaccard similarity for different organizations and topics may result in a different result. The high volume of tweets in September are most likely related to beginning of a new academic year in many countries rather than any new topic in education. Overall, Twitter can be a valuable tool for tracking emerging trends and important topics in education, but caution should be taken in interpreting the results.

Methodology and Source Data Overview

Objective: Determine whether Twitter can be considered a reliable source of information for emerging trends and important topics in education.

Data Source: Collection of Twitter data stored in Google Cloud Storage, containing approximately 100 million tweets related to education.

Methodology:
1. Discarded irrelevant tweets
2. Performed EDA to identify relevant variables for profiling Twitter users
3. Identified the most prolific and influential Twitter users by message volume and retweets and categorized them based on their organizations
4. visualized the distribution and location of Twitter users, detected data collection gaps, and analyzed message uniqueness and duplication.
5. Repeated the above steps for two specific topics, Critical Race Theory in the US and Hijab Row in India, to determine if there were any spikes in Twitter activity or shifts in geographical distribution of Twitter users related to these topics.

The data was processed on Spark using Python and relevant libraries, including Pandas, Matplotlib, and NLTK.

Tweet Clean-up and Filtering

- **Started with a 100M instances on Education**
- **Filtered out non-English tweets, which improves the accuracy of the analysis for English-speaking regions. Later, two specific topics were analyzed which were trending in the USA and India, regions where a significant portion of tweets are made in English**
- **Used a list of 68 key words given by ChatGPT to filter tweets related to K-12 & Higher Education**
- **Removed tweets that contain violent or inappropriate language to avoid skew in the results. Used 54 keywords given by chatGPT**
- **Ended up with 40M instances**

EDA and Extensive Usage of Available Variables

Performed a quality check by computing the percentage of missing values in each column and dropped most of the columns with missing values greater than 70%

Then looked at the rest of the variables and selected variables that are:

tweet-related:
coordinates
created_at
id_str
retweeted_status
tweet_text
text

user-related:
created_at
description
followers_count
id_str
name
screen_name
verified
location

retweet-related:
retweet_count
favorite_count
reply_count
quote_count

Author identification

Top 5 prolific Twitterers by message volume:
sportsthread
Student Finance England
Science Careers
TOI Cities
ExploreLearning

Top 5 influential Twitterers by message retweet:
Mrs N (user name is now changed to Melo's Mommy)
Kathy Jones-Hospod
Zeynep Arsel
Shabnam Nasimi
Tinu

Author identification

7 out of top 10 twitterers by original content and 5 out of top 10 twitterers by retweets are news organisations.

News organisations are the most prolific twitterers by tweet volume and celebrities are the most influential twitterers by retweet volume which makes sense as news organisations tweet often, and tweets of celebrities get highly retweeted by followers.
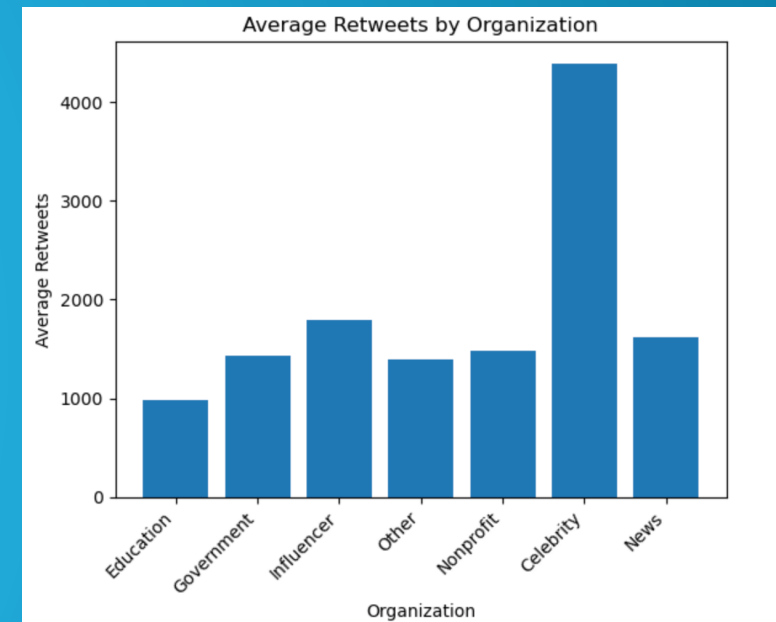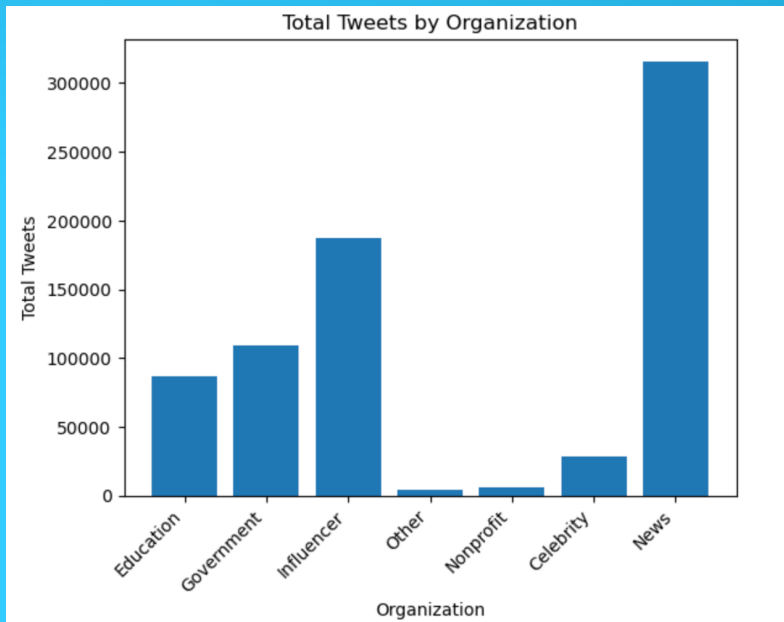
Location and Timeline Analysis

(Findings to begin with)

There is a weak positive correlation (r=0.555) between total tweets and tweet frequency of all users. However, there is no significant correlation (r=0.080) between tweet frequency of influencers and total tweets.


```
+-----------------+--------+
|    user_location|   count|
+-----------------+--------+
|             null|15637191|
|    United States|  418100|
|   Lagos, Nigeria|  258689|
|          Nigeria|  233503|
|            India|  208778|
|              USA|  148671|
|  Los Angeles, CA|  128698|
|  California, USA|  128250|
|  London, England|  118432|
|      Chicago, IL|  111863|
+-----------------+--------+
```


Timeline of Tweets

Distribution of twittererss across the globe

Top 4 countries with highest number of twitterers are USA, Nigeria, India and England.
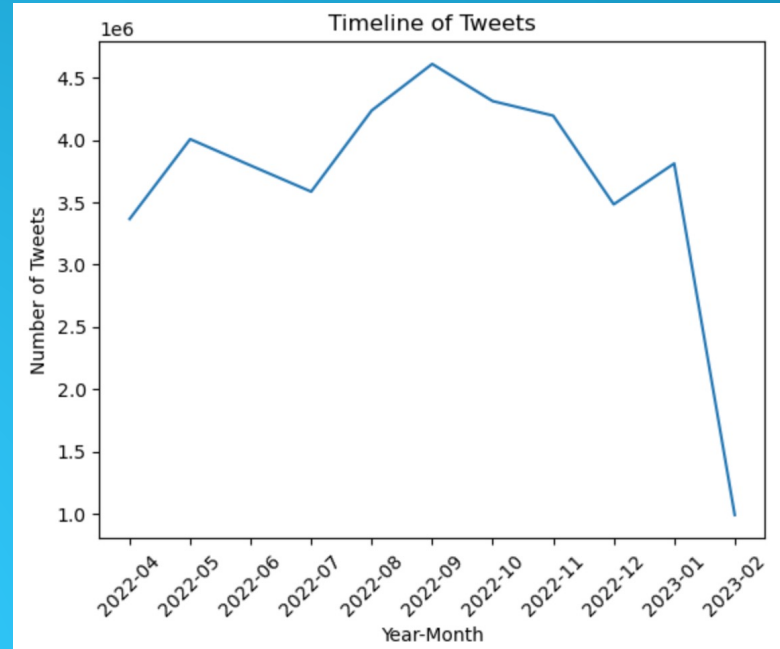
The peak of tweets in September could be due to the beginning of a new academic year in many countries. It is also the time when many students start applying for colleges and universities or prepare for standardized tests. Valleys in July and December are probably due to holidays
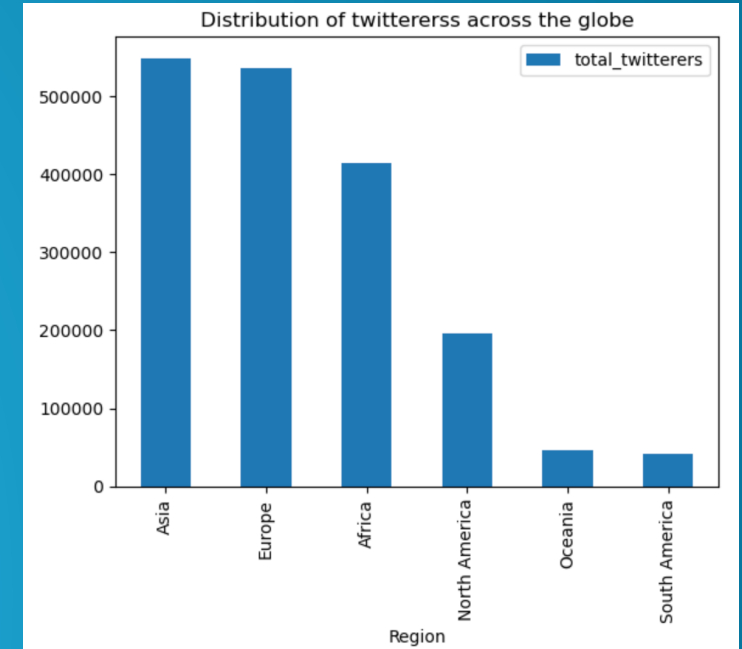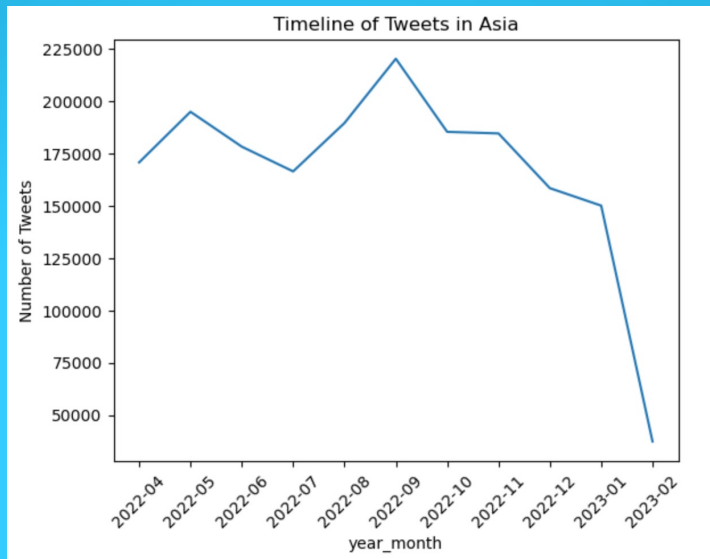
Highest number of twitterers are in Asia. Considering the fact that Asia has the highest number of internet users and social media users in the world, this analysis makes sense.
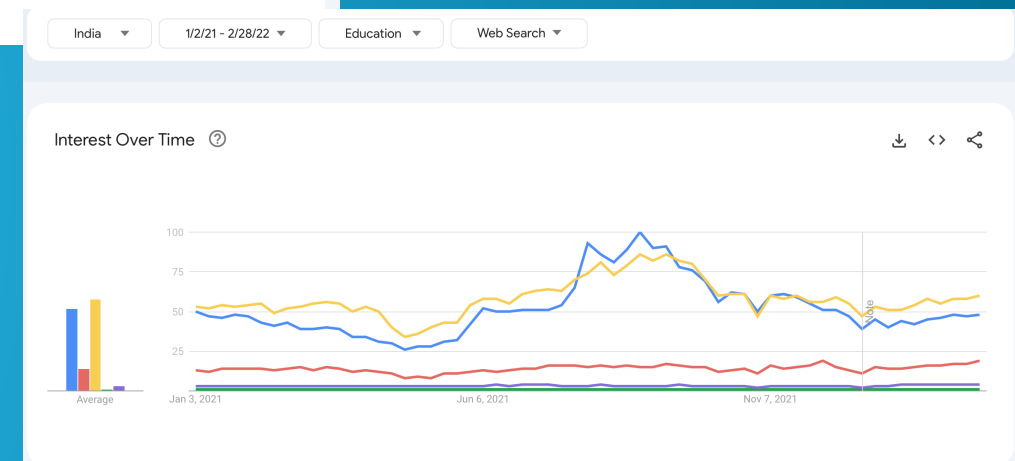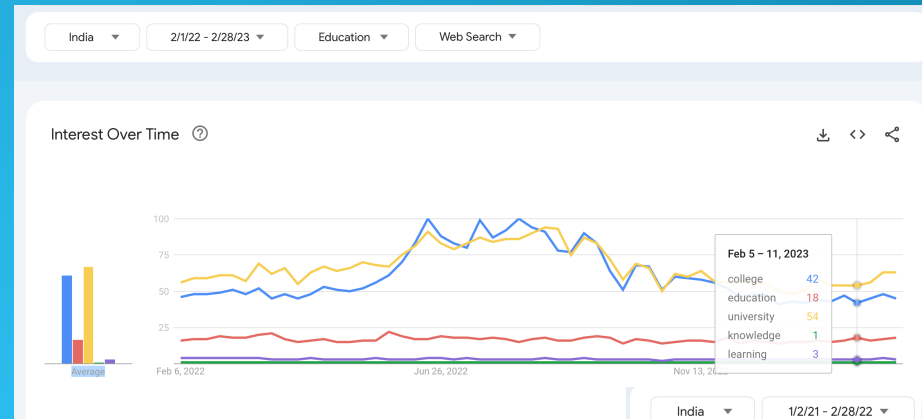
Location and Timeline Analysis

(Looking into Asia)

The timeline of tweets in Asia show an increasing trend from July 2022, a peak in September and a decreasing trend after that.

The following two graphs (with India as the selected region in 2021 and 2022) taken from google trends show similar behaviour.
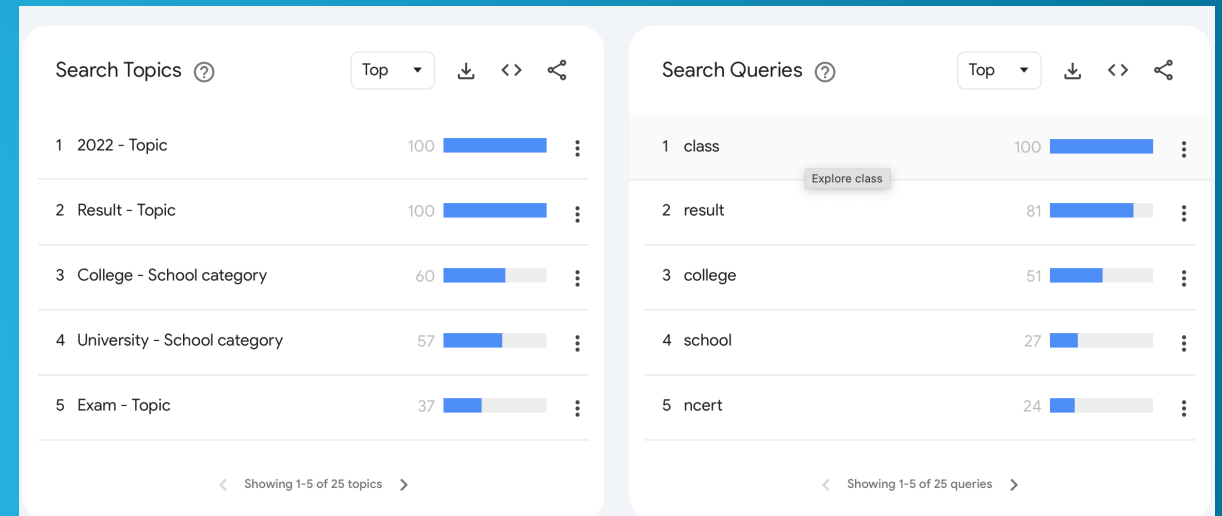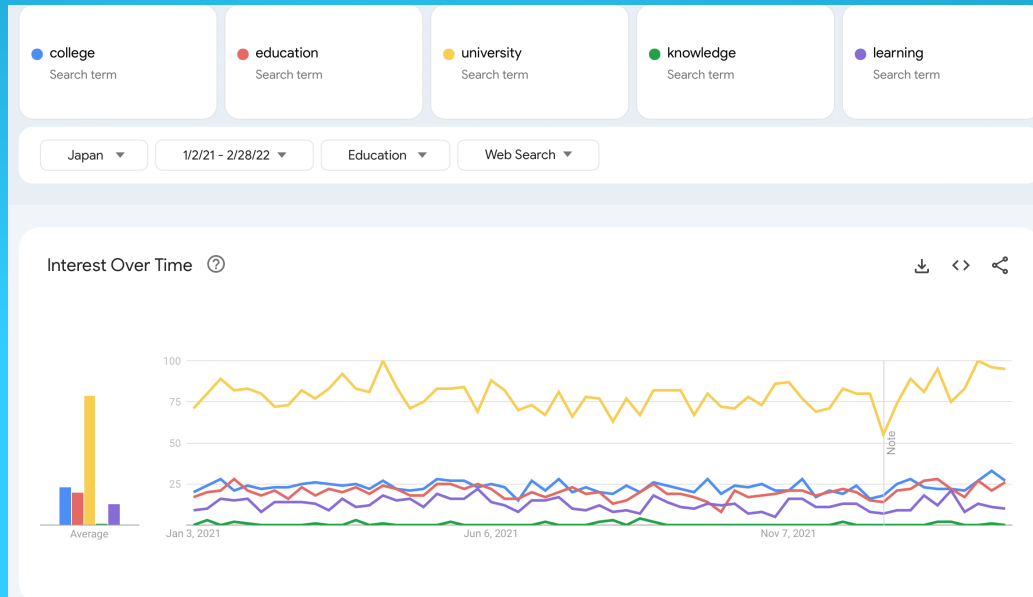
Location and Timeline Analysis
(Looking into Asia)

Japan which has the highest number of twitter users in Asia doesn't show this behaviour.

This pattern is unique to India which has the second highest number of twitter users in Asia. <u>The reason for this yearly pattern in India is its college admission exams and enrolment season.</u> Top search topics and queries from google trends back this up.

Location and Timeline Analysis
**(Critical Race Theory)**

CRT is an academic framework that examines how race and racism intersect with the law and social institutions, and how this intersection perpetuates systemic inequalities. In recent years, there has been a debate in the US on whether CRT should or should not be taught in schools or universities.
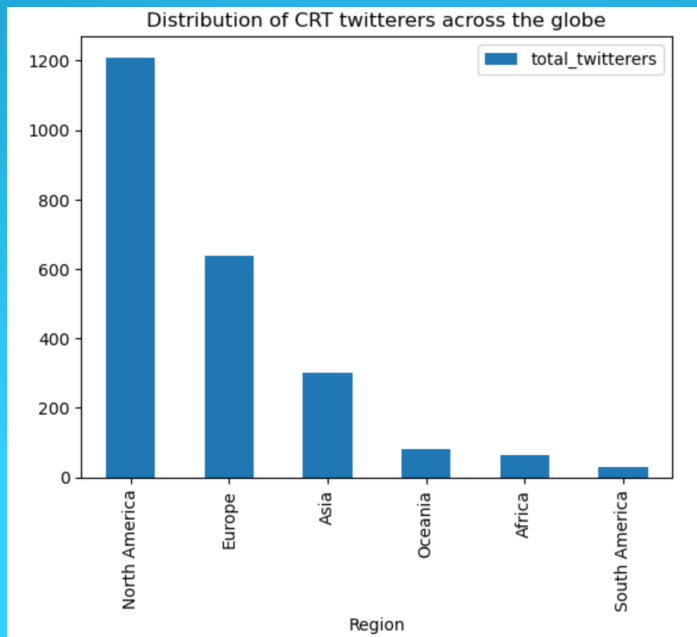
As one expects, highest number of tweets about CRT are from North America
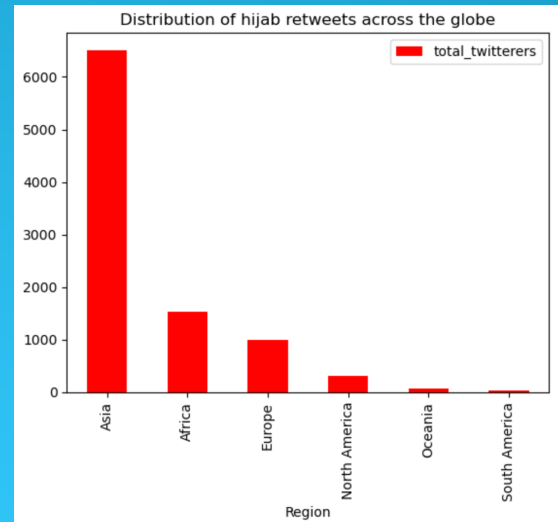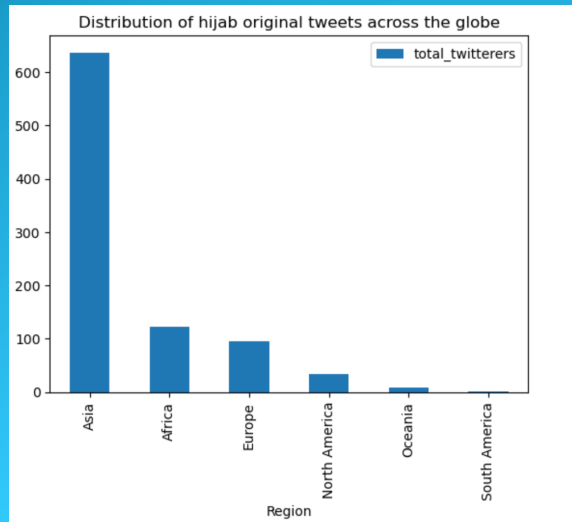
In January, Florida Governor DeSantis moved to ban CRT from Florida's Public Higher Education which might explain why Florida has a high tweet count and also the recent peak in timeline of tweets containing CRT.
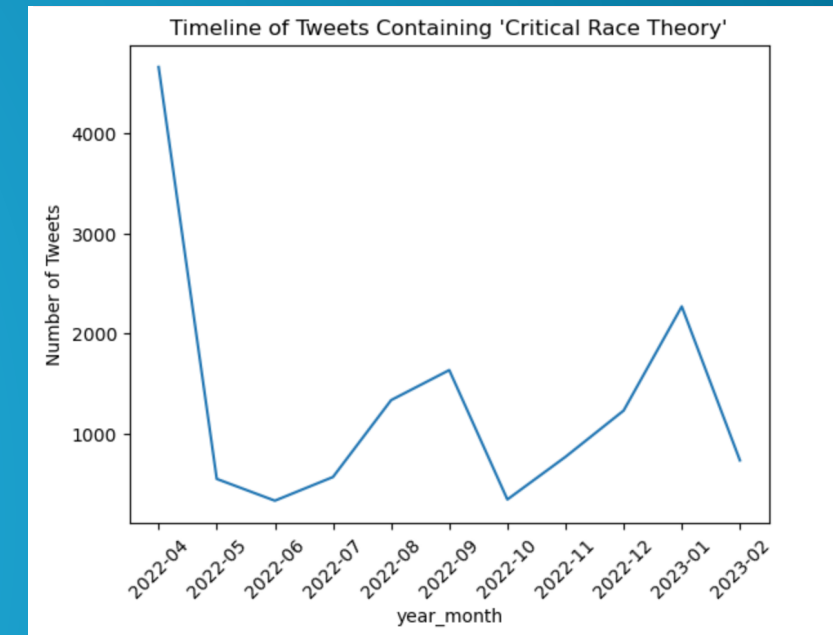
Location and Timeline Analysis
(#HijabRow)

Following the High Court interim order restraining all students from wearing any form of religious attire, schools and colleges across Karnataka State in India asked Muslim students to remove hijabs and burqas outside the school. This controversy led to heated debates and discussions on social media platforms like Twitter.

Distribution of hijab original tweets across the globe

Distribution of hijab retweets across the globe

```
+----------------+-----+
| user_location|count|
+----------------+-----+
|            null|16127|
|           India| 1365|
|New Delhi, India|  359|
|   Mumbai, India|  330|
|Bengaluru, India|  228|
|Hyderabad, India|  208|
|   United States|  173|
| London, England|  128|
|          Mumbai|  128|
|          Bharat|  121|
```


Timeline of hijab Tweets in Asia
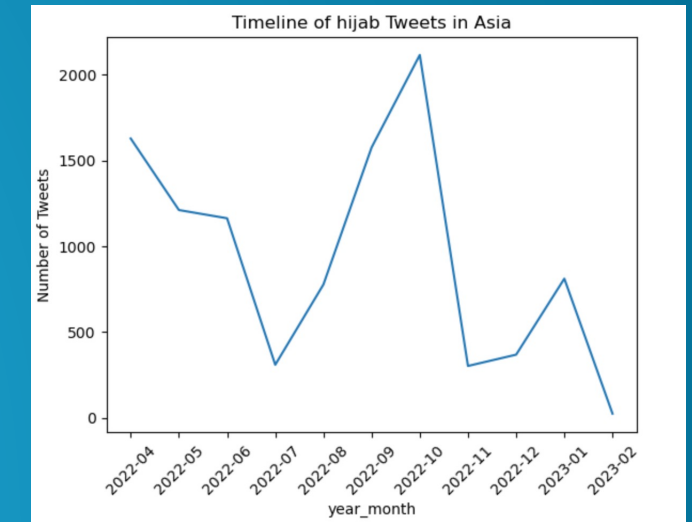
As one expects, highest number of tweets and retweets about hijab are from India in Asia

The Pamphlet ✔ @Pamphlet_in · 16 Sep 2022
Supreme Court says, "Educational Institutes have the right to prescribe uniforms"

#HijabRow #SupremeCourt

NewsDrum ✔ @thenewsdrum · 14 Oct 2022
#HijabRow: Hope in judiciary, says student who fought for hijab

#Hijab #HijabSplitVerdict #SupremeCourt #Udupi @Aliyassadi @sheik_hiba

In October, Supreme Court of India gave a verdict on the issue which explains the peak in tweets on twitter.

Message uniqueness analysis

Selected 0.5 Jaccard similarity from distribution of overall near-duplication in LSH plot to perform similarity analysis. For tweets, a Jaccard similarity of 0.5 may be appropriate because tweets are typically short and contain limited information. A Jaccard similarity of 0.5 indicates that two tweets share about half of their words, which may be sufficient to identify related or similar tweets.
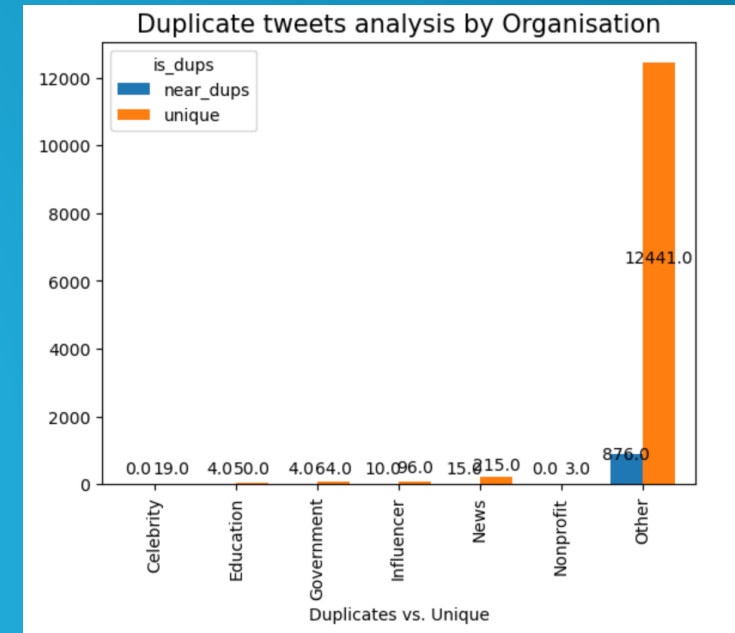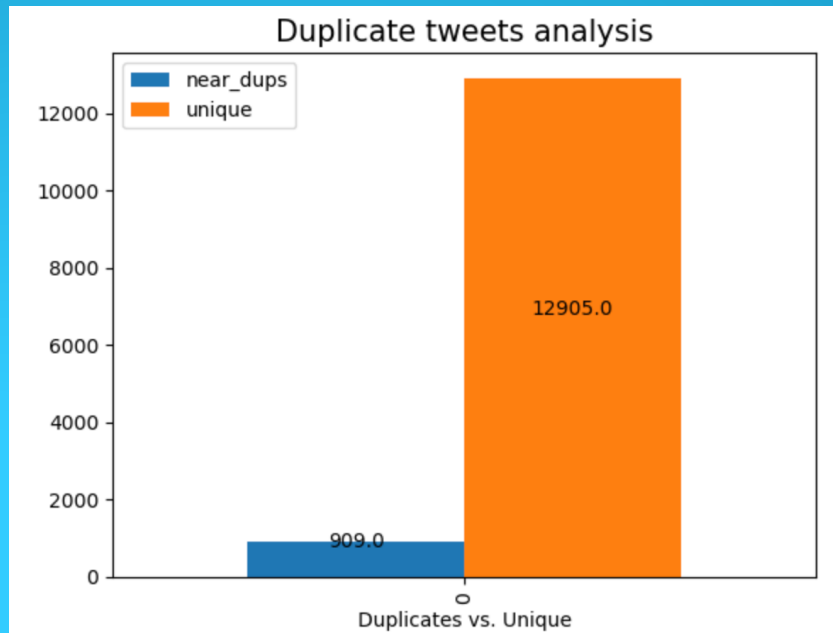
This analysis shows that messages are mostly unique. However, selecting different Jaccard's similarity for different organizations and topics may result in a different result

Conclusions

• Twitter can be a valuable tool for tracking emerging trends and important topics in education, but caution should be taken in interpreting the results.

• The high volume of tweets in September are most likely related to the beginning of a new academic year in many countries rather than any new topic in education.

• News organizations are the most prolific Twitter users, while celebrities are the most influential Twitter users by retweet volume. This could lead to a biased representation of the topics.

• The top four countries with the highest number of Twitter users in education are the USA, Nigeria, India, and England, with the highest number of Twitter users in Asia.

• The timeline of tweets in Asia shows a pattern unique to India, which is related to college admission exams and enrollment season.

• Tweets about CRT and Hijab are concentrated in North America and India, respectively, and were influenced by recent events.

• The analysis also shows that messages are mostly unique, but selecting different Jaccard similarity for different organizations and topics may result in a different result.

Actionable Recommendations

1.Provide tools (similar to google trends) for analyzing tweets: Twitter should provide tools for analyzing tweets to make it easier to identify trends and important topics. These tools should enable users to filter out irrelevant tweets and focus on relevant tweets. They should also enable users to identify influential Twitter users and categorize them based on their organizations.

2.Promote verified education-related accounts: Twitter should promote verified education-related accounts to increase their visibility and credibility. These accounts could include news organizations, universities, and education-related government agencies.

3.Provide a mechanism for flagging inappropriate content: Twitter should provide a mechanism for flagging inappropriate content. This will help ensure that are free from violent or inappropriate language.

4.Work with education organizations to increase engagement: Twitter should work with education organizations to increase engagement on the platform. This could include hosting Twitter chats on education-related topics or promoting education-related events on Twitter.

5.Highlight important topics in education: Twitter should highlight important topics in education to increase awareness and promote discussion. This could include promoting tweets related to important education-related events or featuring tweets from influential education-related Twitter users.