Educational Tweets Analysis

Mariam Adeyemo

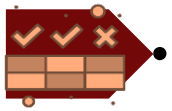
Big Data Final Project

March 9, 2023

Outline

- Executive Summary

- Methodology and Source Data Overview



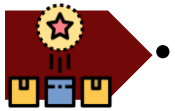
- Tweet Clean-up and Filtering

- Exploratory Data Analysis

- Author Identification

- Location Analysis

- Timeline Analysis



- Message Uniqueness Analysis

- Conclusions and Recommendations

Executive Summary

Executive Summary

Twitter can be considered a credible source of information, which reflects the emergence of important trends or topics in education because of the following reasons:

- Higher tweets can be attributed to emergent issues in education.
- Higher increase in the number of retweet facilitate the dissemination of new emerging trends and ideas in education.
- Verified organizations such as News Outlets, Educational Institutions, and Government Entities influences the conversations and discussions that takes place on Twitter. Though, the general public also help topics gain traction on Twitter by retweets from verified sources.

Objective

The aim is to identify whether Twitter can be considered a credible source of information, which reflects the emergence of important trends or ideas in education. The analysis is focused on the Twitterers identification, their geographical location, tweets timeline, and tweets uniqueness.

Methodology and Source Data Overview

Source Data Overview

The twitter data was gotten from the Google Cloud Storage, and it was in a JSON format. The tweets were related to both education and non-education topics.

Roughly **100 million Tweets (~500GB) with 39 columns** (excluding nested columns) was initially the size and shape of the data.

About **36.6 million Tweets with 26 columns** was finally used for the analysis after cleaning and filtering.

Methodology

- Google Cloud was used as the computing platform.
- PySpark was used for the data analysis, including Spark RDD, Spark Dataframe, and Spark SQL.
- Pandas was used for bar plots and graphs.
- Locality-Sensitive Hashing (LSH) was used to test for similarity amongst the tweets.

Tweet clean-up and filtering

As the tweets data contained topics related to education and non-education, the data (“text” column) was cleaned and tweets not relating to education was discarded using educational keywords.

Removing unwanted expression in the text

```
unwanted_expression = RegexTokenizer(inputCol="text", outputCol="words", pattern="\\W")

lower_case = udf(lambda x: x.lower())

tweets_raw = tweets_raw.withColumn("text", lower(regexp_replace("text", "[\$#,&%\".]", "")))
tweets_raw = unwanted_expression.transform(tweets_raw)
tweets_raw = tweets_raw.withColumn("text", concat_ws(" ", "words"))
tweets_raw = tweets_raw.drop("words")
```

Filtering the data to educational related tweets

```
pattern1 = "|".join(education_related_words)
pattern2 = "|".join(exclude_words)

#Discard irrelevant tweets using the filter words identified above
filtered_tweets = tweets_eng.filter(
    col("text").rlike(pattern1)
).filter(
    ~col("text").rlike(pattern2)
)
```

Educational keywords used for filtering

```
#creating a list of filter words for education
education_related_words = ['elementary school', 'middle school', 'high school', 'higher education', 'k-12', 'preschool',
                           'college', 'kindergarten', 'students', 'tuition', 'university', 'education', 'classroom',
                           'game based learning', 'teach', 'teacher', 'edu', 'digital education', 'education system',
                           'steam based learning']

exclude_words = ["died", "shoot", "kill", "killed", "deceased", "murder", "attack", "sex", "threesome", "horny",
           "shooting", "porn", "pornography", "shot", "gunned", "shootings", "gun", "guns", "uvalde", "football"]
```


Note: I excluded some words in the data as I realized these tweets were overshadowing the educational issues I was interested in.

Website used to identify educational keywords: https://www.teachthought.com/twitter-hashtags-for-teacher/

Exploratory Data Analysis

- The twitter data contained extensive list of root-level and nested-level attributes. The major root-level attributes were “user”, “tweet”, ‘entities” and “extended entities”. I used the twitter data dictionary and the “printSchema” function in PySpark to understand the various attributes, and this aided my features selection.

- The percentage of missing values was also calculated for each of the attributes. Features with greater than 70% missing values were discarded.

Raw Tweets Data

- ❑ Initial total tweets: **~100 million**
- ❑ Initial total columns(excluding nested-levels): **39 columns**
- ❑ % of poorly populated columns: **46%**
- ❑ % of features with little or no missing values: **54%**

Final Tweets Data

- ❑ Final total tweets: **36.6 million**
- ❑ Final total columns: **26 columns**

Feature Engineering

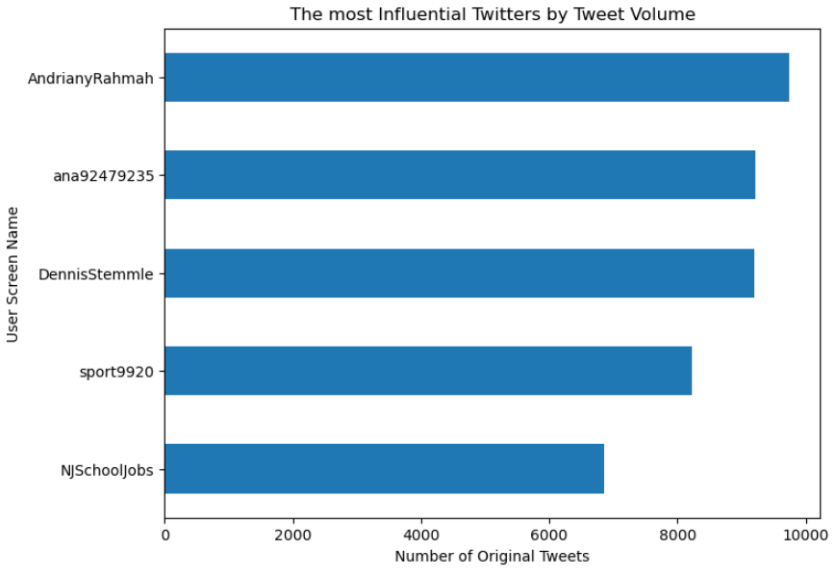
- ❑ **Country column** from user location
- ❑ **Region column** from Country column
- ❑ **Year, month, and days columns** from the tweet timestamp
- ❑ **Organization column** from the user screen name and user description

Author Identification — By Message Volume

The tweet data was filtered to only contain original tweets, and this was used for the analysis.

Table and plot showing the **top 5 most influential twitterers** by original tweet volume

user_screen_name	user_category	original_tweet_count
AndrianyRahmah	News Outlet	9741
ana92479235	University	9214
DennisStemmle	University	9194
sport9920	Someone Else	8221
NJSchoolJobs	University	6858



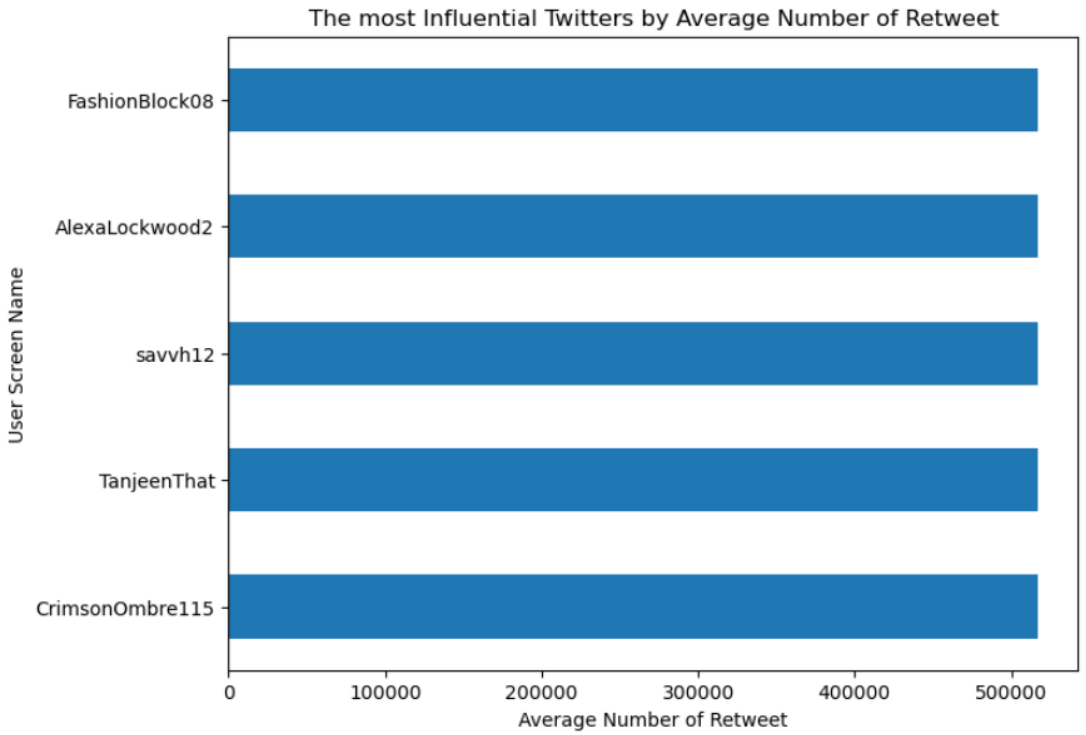
Example of tweets from AndrianyRahmah

Key Insights:

- **AndrianyRahmah** had the highest original tweet count, and **tweets about high school games**. This is not surprising that a News outlets is the most prolific twitters based on original tweet count because News outlets often post original content on Twitter, such as breaking news updates, exclusive interviews, and current events analysis.
- Though, **AndrianyRahmah has just 39 followers, his livestreams of the games had about 12,000 views** which is pretty huge.
- The **second most influential twitterer was ana92479235**, followed by DennisStemmle, sport9920, and NJSchoolJobs.

Author Identification — By Message Retweet

Table and plot showing the **top 5 most influential twitterers** by retweet volume



user_screen_name	user_category	Average_retweet
FashionBlock08	Someone Else	516855.0
AlexaLockwood2	Someone Else	516850.0
savvh12	Someone Else	516795.0
TanjeenThat	Someone Else	516791.0
CrimsonOmbre115	Someone Else	516779.0

Key Insights:

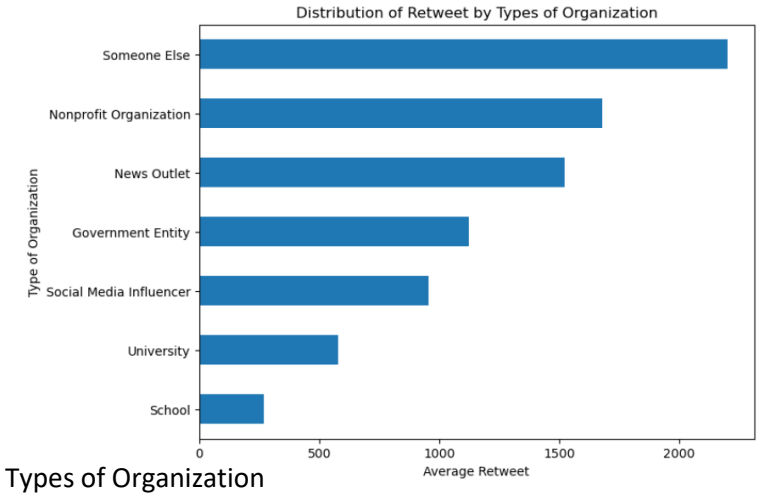
- FashionBlock08 is the most influential twitterer based on the average number of retweet while AlexaLockwood2 is the second most influential to retweet educational posts.

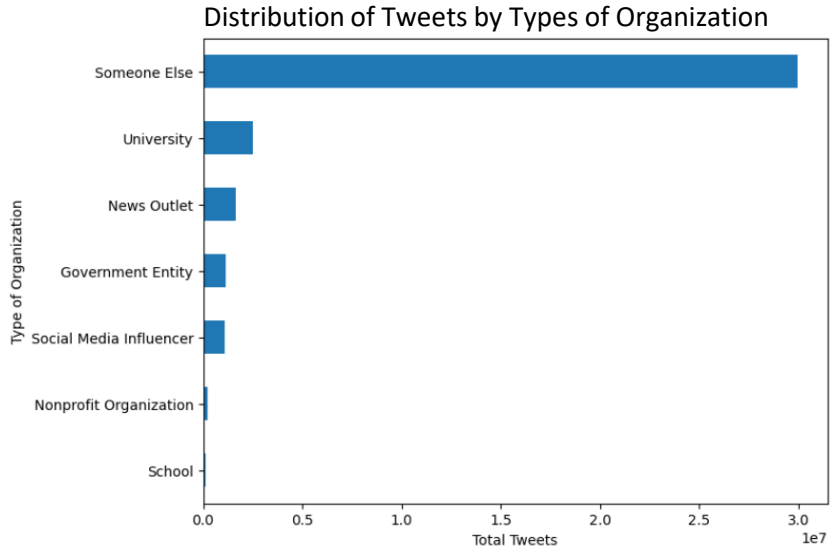
Author Identification — By Types of Organization

All the tweets was used for analysis to identify the most influential organization for disseminating educational posts.

Twitterers Identification based on Organization.

user_category	total_twitterers
Someone Else	9760162
University	479131
News Outlet	326442
Government Entity	231351
Social Media Influencer	150409
Nonprofit Organization	59008
School	16860





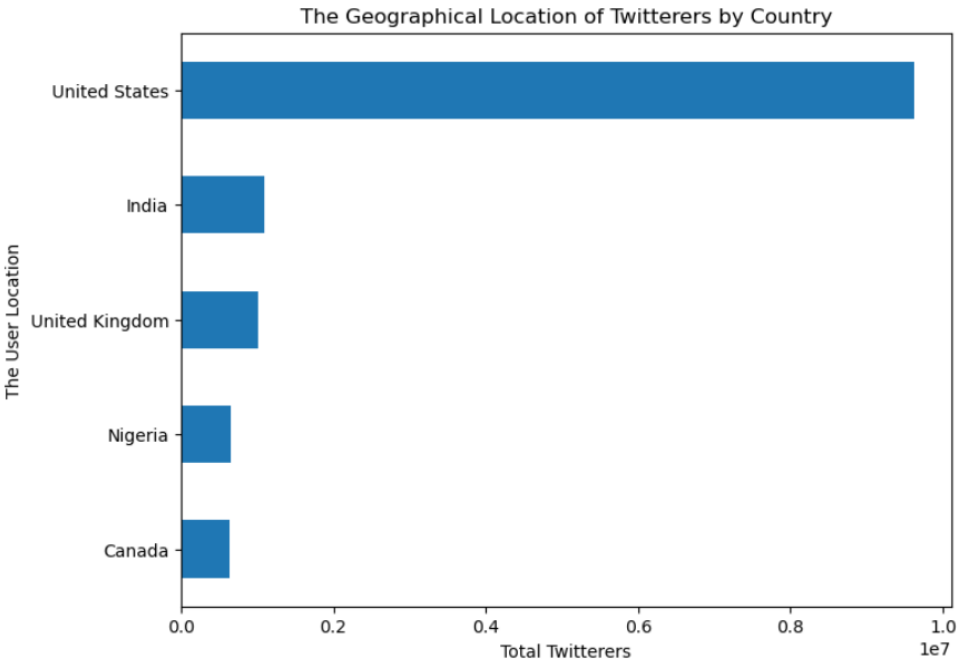
Key Insights:

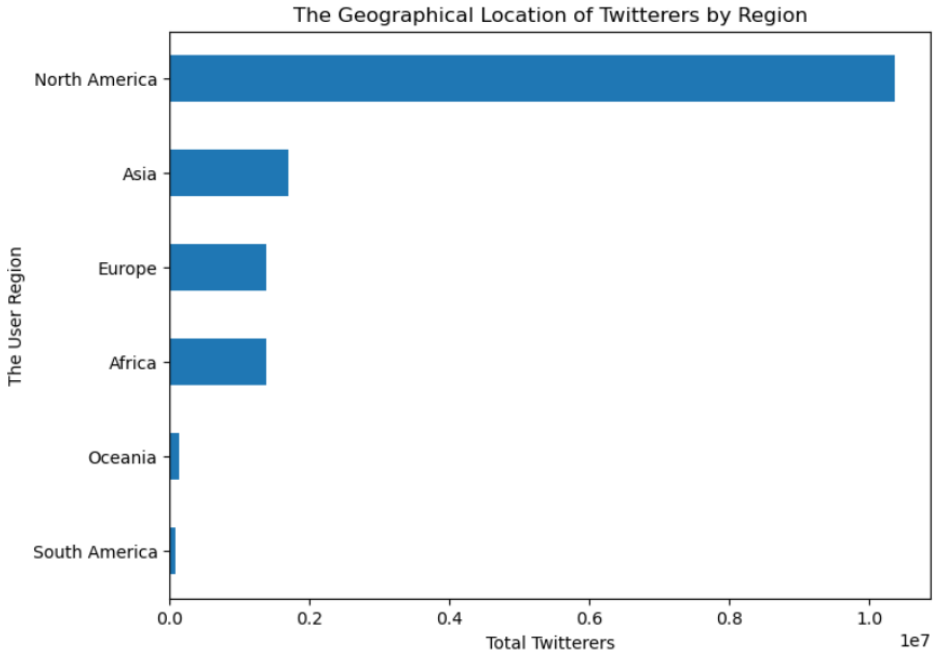
- Based on **the identification of twitterers by organization, "Someone Else" contributed most to the tweets about education.** This might be due to the fact that most individuals have social media presence compared to corporate organizations.
- Based on **the number of retweet, "Someone Else" also was the most prolific user** and this is similar to the findings in the previous slide.
- **Educational institutions had the least number of retweet** which is reasonable as they often have more formal tweets which might make it less engaging.
- **Someone else, University, and News Outlet generally have more tweets**

NB: Unique user id was used to identify who the twitterers are based on the organization.

Location Analysis — By Geographical Distribution

The user location data was used in this analysis to identify the location of the twitterers engaging in educational posts. I cleaned the user location data and derived two columns from it named “country” and “region”.





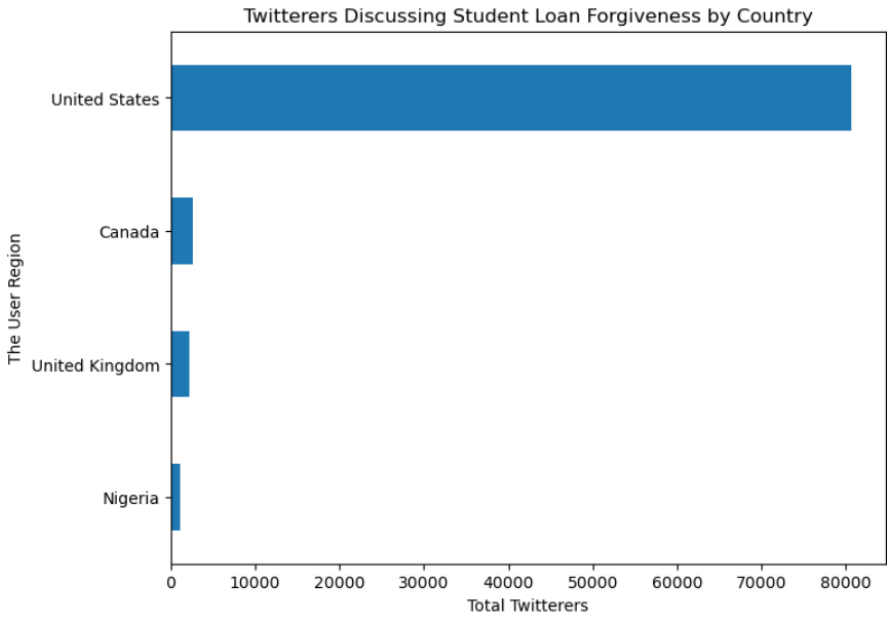
Key Insights:

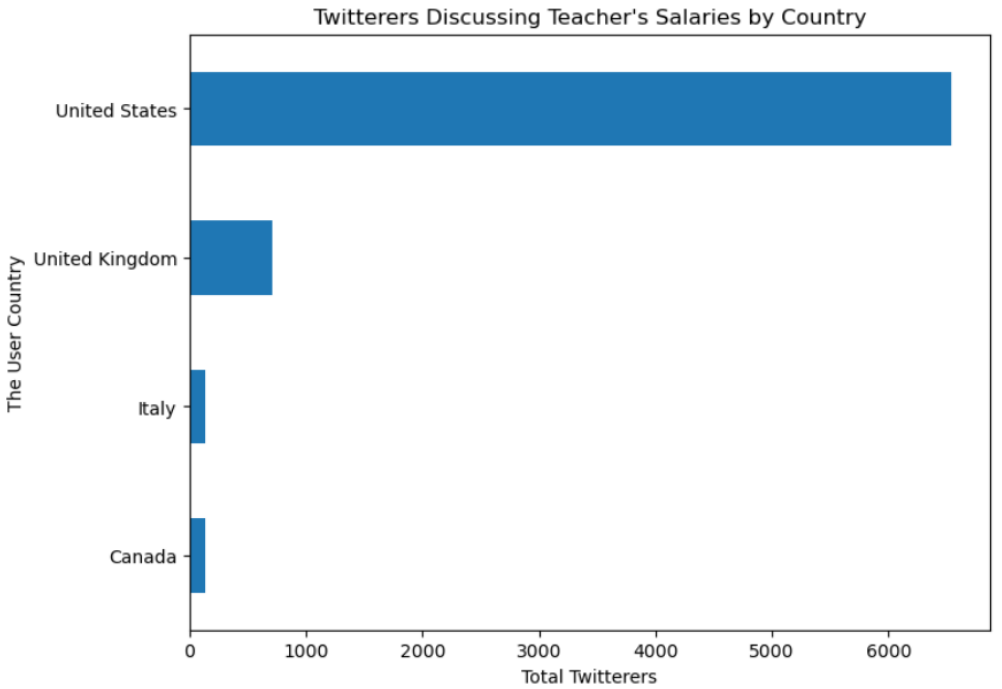
- For educational tweets**, most of the Twitterers are from the USA** (9.6 million users). This is similar to the findings of Statista where they observed that USA is the leading country with the highest number of Twitter users (~77 million users).
- This is due to the fact that Twitter is widely popular in the United States.
- The regional analysis also have similar distribution to the country analysis, **where North America had the highest number of Twitterers relating to educational tweets.**
- **India is the second highest country** to tweet about education.

https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/

Location Analysis — By Emergent Issues in Education

Based on research on the emergent issues in education, two issues were chosen for this analysis which are “**Student Loan Forgiveness**” and “**Teachers' Salaries**”





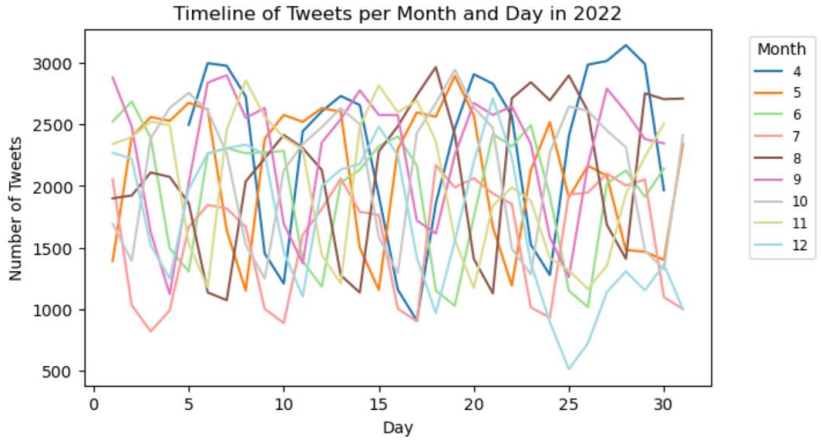
Key Insights:

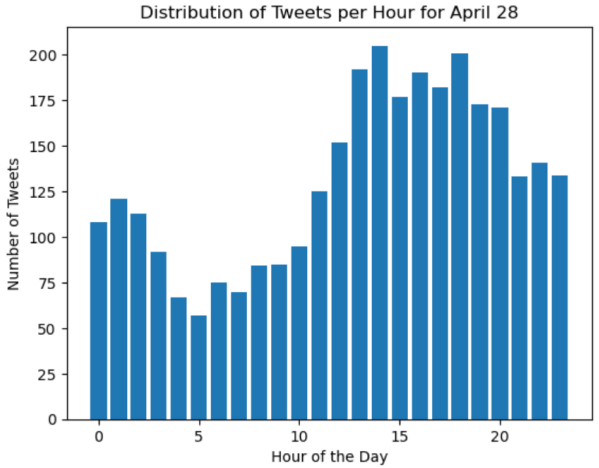
- **There is a slight relationship between the emergence of new issues in education and progression and locations of these Twitterers.**
- Emergent issues usually gain tractions based on the location of occurrence of the issue and the interest of the twitterers.

Timeline Analysis — By Verified Twitterers

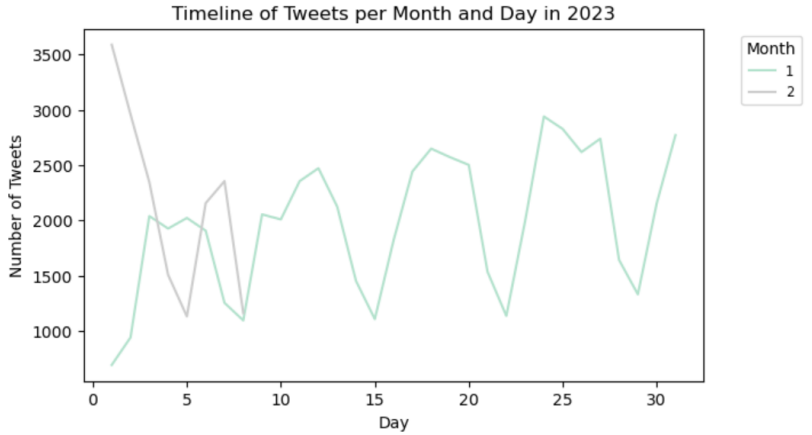
I compared the timeline of the tweets between verified and unverified twitterers to see if they both have significant peaks and valleys.

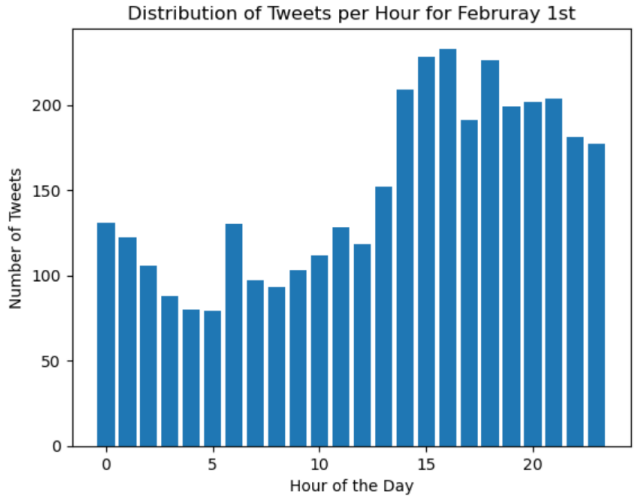
2022 Timeline Analysis for Verified users





2023 Timeline Analysis for Verified users





Key Insights:

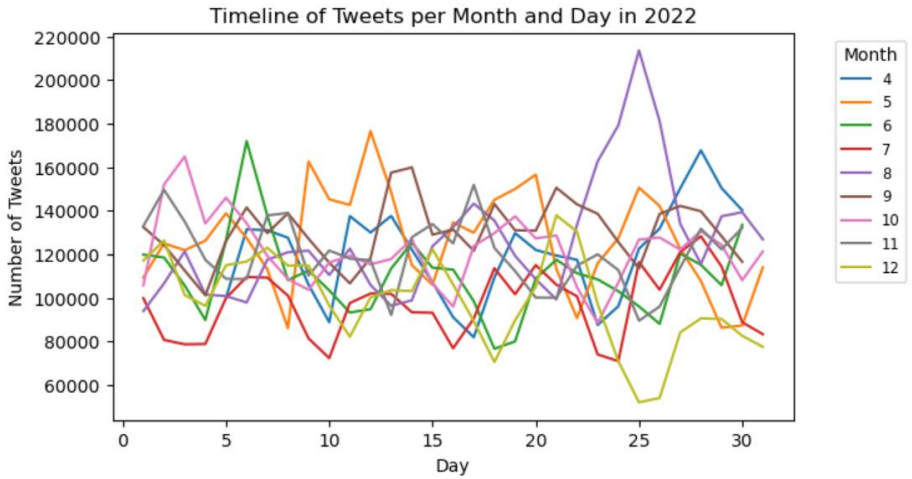
- For **2022**, various peaks was noticed throughout the month, however, **August 25 to 28 had the highest peaks**. The most predominant tweets during these period was about the students' loan forgiveness.
- There was a **significant drop in the number of tweets on December 25, 2022**. This might be due to the holiday season which led to less engagement.
- For **2023**, the **highest peak was noticed on February 1st**. One of the most predominant tweets was about the college board release of the revised Advanced Placement course in African American Studies.
- Most twitterers engaged more during the evening periods for both peaks.

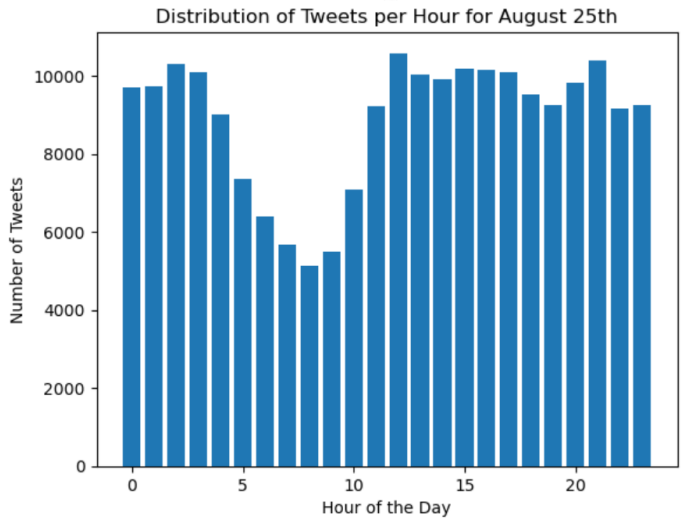
It was noticed that there was no data for January through March in 2022. Furthermore, the data for 2023 stopped at February 13th.

The verified twitterers are only 2% of the total twitterers analyzed (36.6 million)

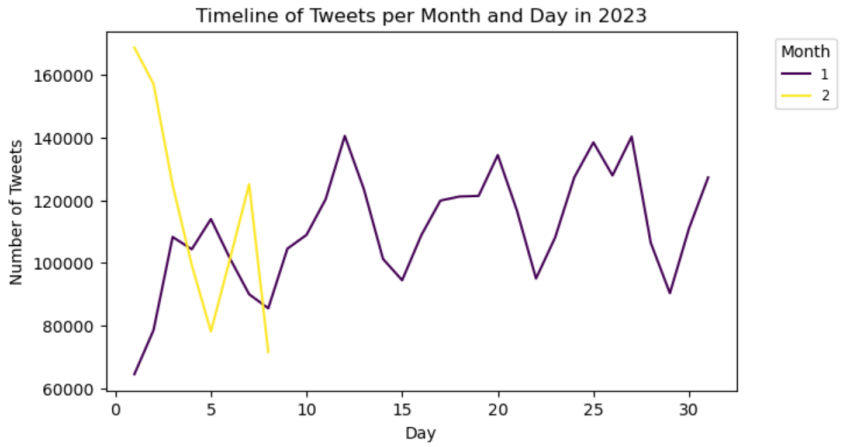
Timeline Analysis — By Unverified Twitterers

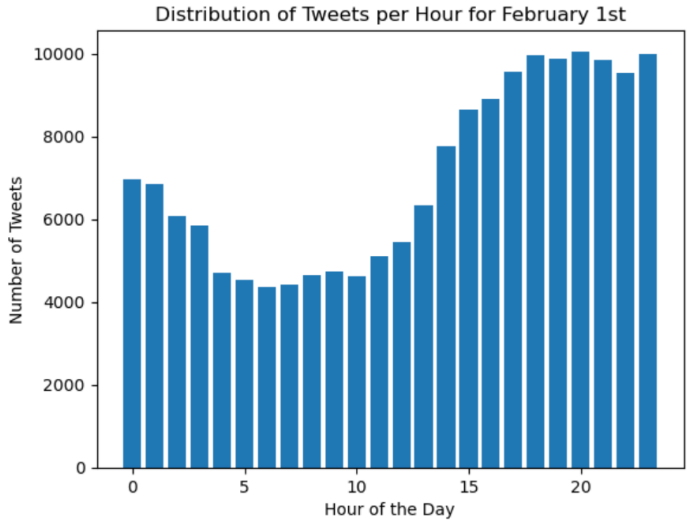
2022 Timeline Analysis for Unverified users





2023 Timeline Analysis for Unverified users





Key Insights:

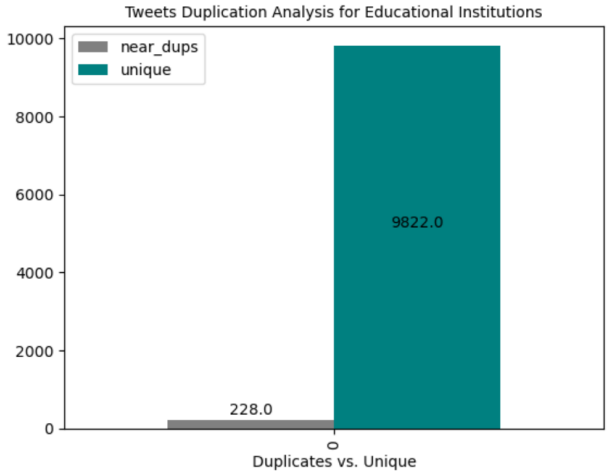
- For **2022**, various peaks was noticed throughout the month, however, **August 25 had the highest peaks.** The most predominant tweets during these period was about the students' loan forgiveness, and this is **similar to the result from the verified users.**
- There was also a **significant drop in the number of tweets on December 25, 2022.** This is also like the result from the verified users.
- For **2023**, the **highest peak was also noticed on February 1st**.
- Most twitterers engaged more during the evening periods for both peaks. However, during August 25, 2022, we see a high number of tweets too at the early hours of the day.

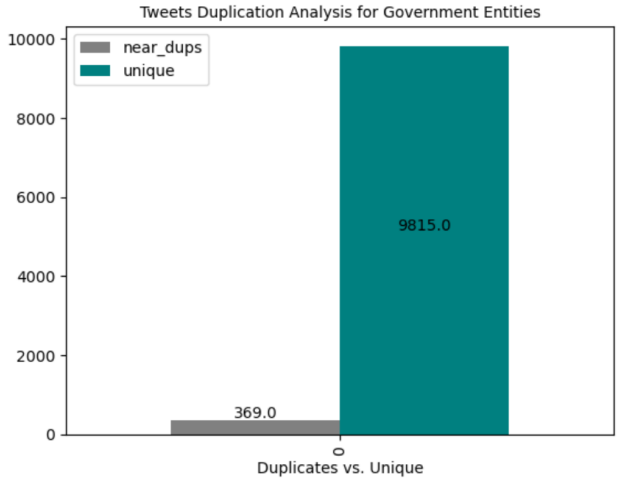
It was also noticed that there was no data for January through March in 2022. Furthermore, the data for 2023 stopped at February 13th.

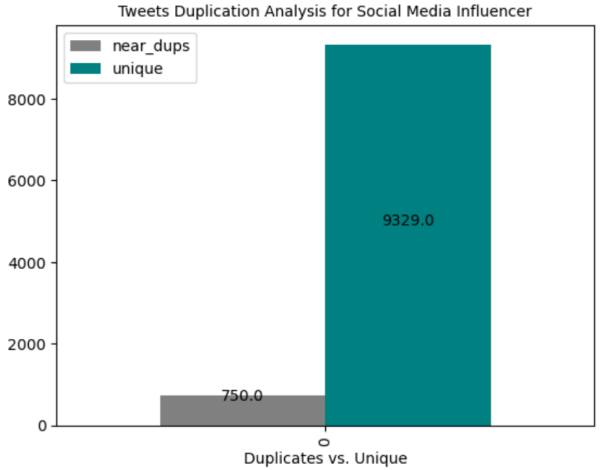
The unverified twitterers account for 98% of the total twitterers analyzed (36.6 million).

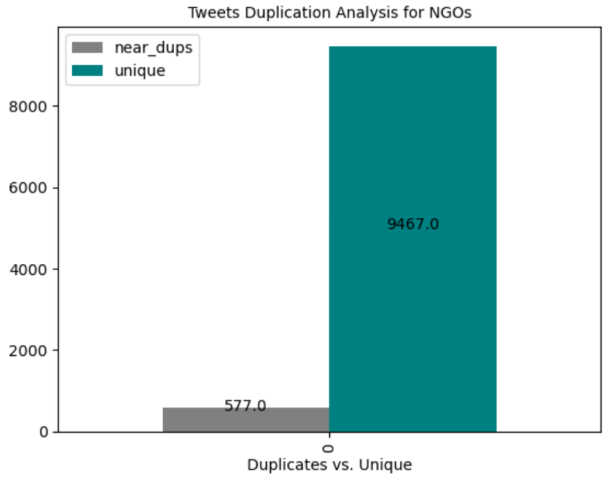
Message Uniqueness Analysis

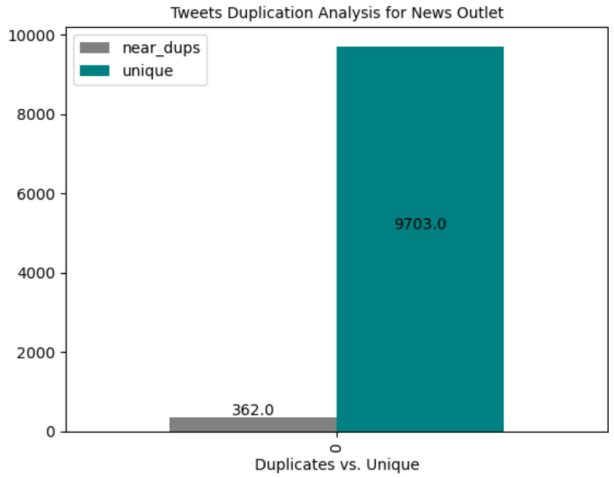
Initially, Jaccard threshold of 0.3, 0.5, and 0.7 was used to test for the similarities between the text to identify the right Jaccard similarity for the corpus. Based on the result of the evaluations, I selected Jaccard threshold 0.3 as the right threshold to be used for further analysis between each organizations.

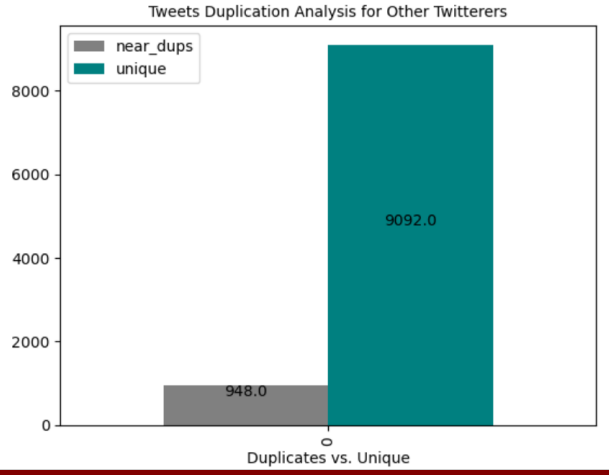












Key Insights:

- Based on the analysis of the similarities between tweets within each organization, **"Other twitterers and Social Media Influencer" had more duplicates compared** to other organizations.
- **Educational Institutions and News Outlets had the least duplicate tweets** among all the organizations analyzed.
- **In summary, most tweets are unique.**

Conclusions and Recommendations

Conclusions

- The different analysis carried out showed that Twitter can be considered a credible source of information, which reflects the emergence of important trends or topics in education.
- Higher tweets can be attributed to emergent issues in education.
- There is more tweets from unverified users compared to verified users because there are more of the unverified users in the Twitter data.
- The top 3 organizations that contributes more to educational tweets are Someone else, University, and News Outlet.
- Most of the twitterers are from the United States and most tweets are original contents.

Recommendations

- Most of the Twitter data was messy and contained lots of irrelevant features. Collection of only relevant features and regular cleaning and standardizing of the data will aid analysis.
- The Twitter data accounts for only 2% of the verified users on Educational content compared to 98% of unverified users. This result in an imbalance class and comparison analysis between the groups will be difficult, so an increase in the verified users would increase the reliability of information.

Thank You