Twitter Data on Education

Linh Le

07th December 2022

Executive Summary

The purpose of this report is to examine the possibility of using Twitter as a credible source to identify the emergence of trends or topics in education

Project Background

- Twitter is one of the most popular social media platforms today, where users broadcast short posts known as tweets
- It has 100 million daily active users and 500 million tweets sent daily
- Twitter is often used to receive news, follow celebrities, or stay in-touch with friends
- For this project, 500GB of tweets is analyzed to determine whether Twitter can be considered a credible source of information, specifically in education

Key questions to tackle…

- ❑ Who are the most influential Twitterers that are talking about topics in education?
- ❑ Where are these Twitterers located?
- ❑ What are the timelines of these tweets talking about topics in education?
- ❑ How unique are these tweets, or are they mostly duplicates?
- ❑ What are the conclusions and recommendations based on the data?

Methodology and Source Data Overview

Various platform and tools were deployed for this analysis

Methodology

Google Cloud Platform	Cloud computing service to store and process tweets data
PySpark	Used for large-scale data processing
Pandas & Matplotlib	Data visualization
Locality Sensitive Hashing	Text-similarity analysis for original tweets and all tweets
Analysis Methodology	Tweets are separated and labelled as original and retweet for calculation

Source Data Overview

100 million records, 500 GB, stored in nested JSON format

Contains Tweet, User, Geo, Entities objects

Tweet objects have 22 attributes, User objects have 22 attributes, Geo objects have 11 attributes, and Entities objects have 6 attributes

Tweet Clean-up and Filtering

After cleaned and filtered, the final dataset to be analysed includes 32 million records and 24 columns

Tweet Filtering

- Since we want to focus on educational topics, a list of words relating to education was used to filter out irrelevant tweets
- The following words were used to filter desired tweets:

higher education	**k-12**	**teachers**	**primary school**	**high school**	**curriculum**	**secondary school**	**university**
tuition	**student**	**homework**	**classroom**	**research**	**educational**	**college**	

Tweet Clean-up

- We look at the number of missing values in our chosen features to ensure that they are well populated. However, we do not have sufficient tweet's geo data in our dataset, so we settled for a field ("coordinates") that has a lot of missing data

Number of Missing Data in Relevant Columns

created_at	0	**favorite_count**	0	**retweeted_status_retweet_count**	27,151,650	**user_created_at**	0	**user_favourites_count**	0	**user_location**	28,388,235
id	0	**retweet_count**	0	**tweet_longitude**	74,864,424	**user_id**	0	**user_followers_count**	0	**user_screen_name**	0
text	0	**reply_count**	0	**tweet_latitude**	74,864,424	**user_description**	12,593,959	**user_friends_count**	0	**user_verified**	0

Data run on 32M records

Exploratory Data Analysis

Flatten Dataset

The dataset comes in deeply nested JSON format. Each record has multiple attributes with STRUCTURE data type, which required flattening and re-labelling.

Discard Irrelevant and Poorly Populated Columns

The dataset has 61 attributes, but not every column is needed for the analysis or is well populated.
Dropping irrelevant columns and poorly populated columns was conducted.
The timeframe of the dataset is from April 2022 to November 2022.

Describe Numeric Columns in Dataset

summary	retweeted_status_ retweet_count	user_favourites_ count	user_followers_ count	user_friends_ count
count	19,771,692	30,334,243	30,334,243	30,334,243
mean	2,226	44,489	9,311	1548
stddev	7,338	91,717	279,286	5692
min	0	-1	0	0
max	516,928	3,407,404	82,199,011	2,218,970

Author Identification

Data run on 32M records

Volume of Original Tweet by Twitterer

Username	User Type	Number of Original Tweets
sport9920	Other	19726
ana92479235	Other	19703
AndrianyRahmah	News Outlet	11597
DennisStemmle	University	9236
EssayPaperUK	School	8788
jaeyunowins	School	8413
hilmsit	School	8352
AgiwaraS	School	7567
adeliasari033	School	7322
studyinnaija	University	6607

Volume of Retweeted Times by Twitterer

Username	User Type	Number of Retweeted Times
PEScorpiio*	Other	1826531158
NicholasFerroni	Social Media Influencer	1761595304
ChrChristensen	University	1033252600
brndxq	Social Media Influencer	776258895
mattxiv	Social Media Influencer	709330484
Mr_JCE	News Outlet	704632195
polevaultpower	Social Media Influencer	652023001
CathyMarksKrpan	School	584476648
Ernie_Zuniga	News Outlet	577185975
KianSharifi	News Outlet	533426243
MichaelWarbur17	Social Media Influencer	483372535

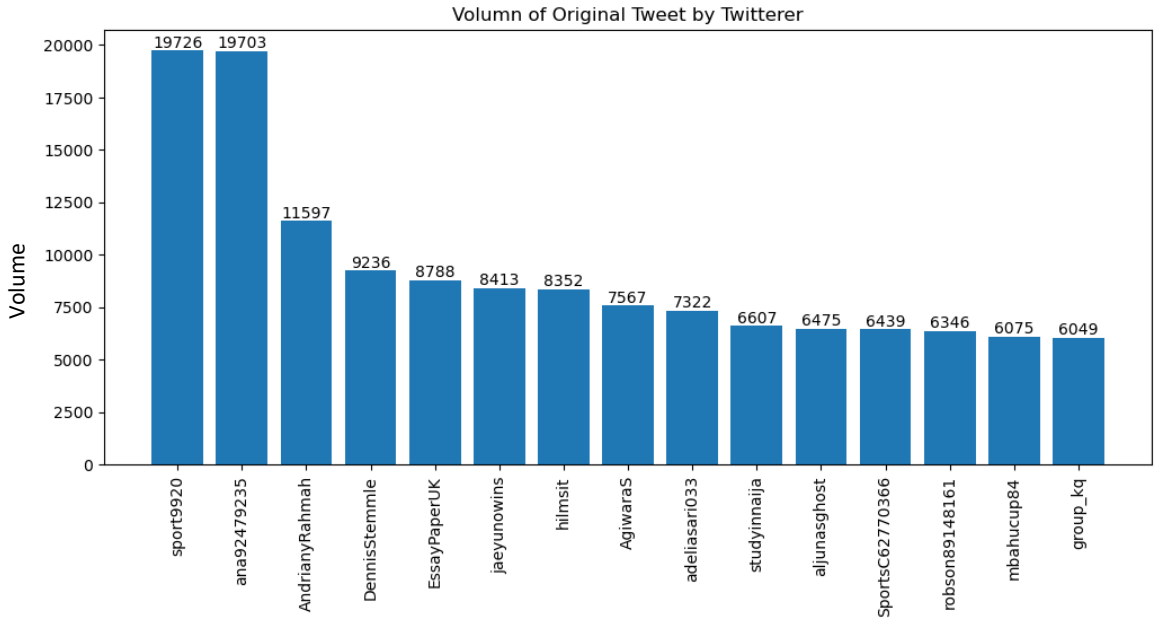
**Note: Upon investigation, user PEScorpiio has one tweet that contributed to his 1 billion+ retweets, which is an unusual event and thus not as meaningful to be included in our analysis, where we are trying to look for patterns*

Data run on 32M records

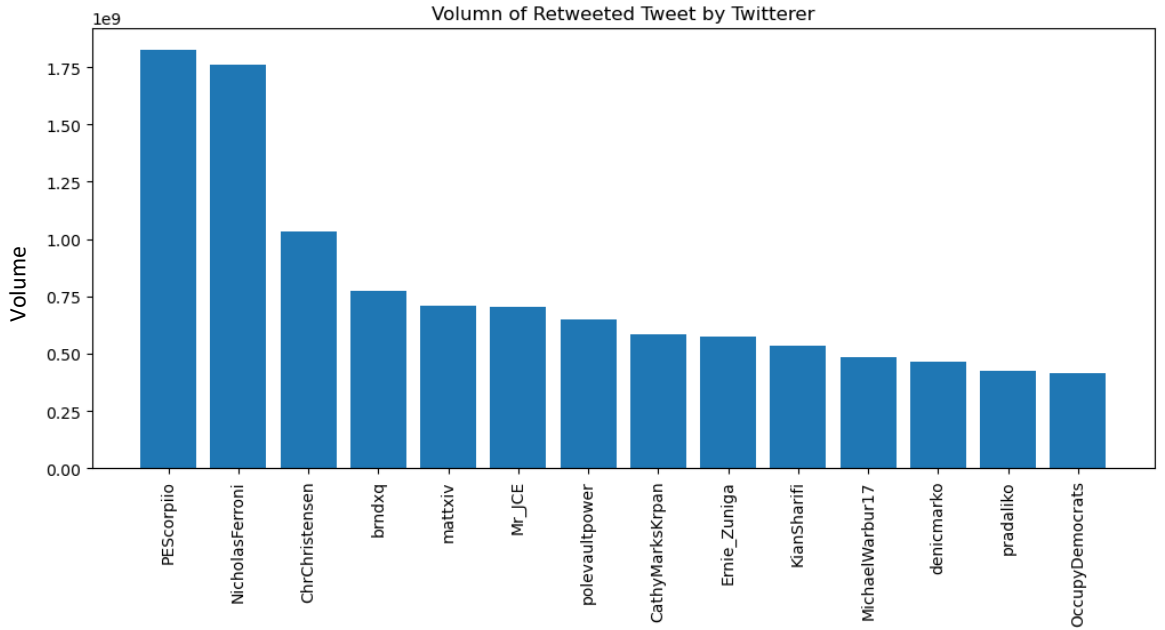
Author Identification - Visualization

Out of the top 10 most prolific twitterers in producing original tweets, more than half are associated with a school or university. However, the users in top 10 most retweeted are social media influencers and news outlets

Volume of Original Tweet by Twitterer



Volume of Retweeted Times by Twitterer

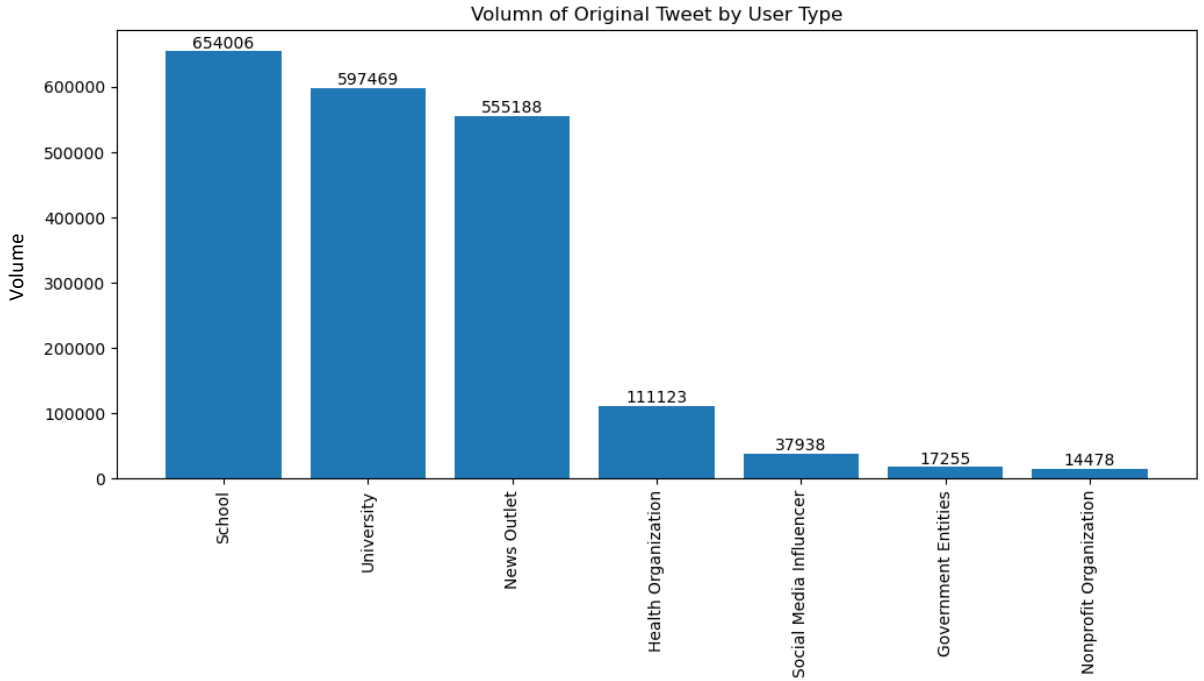


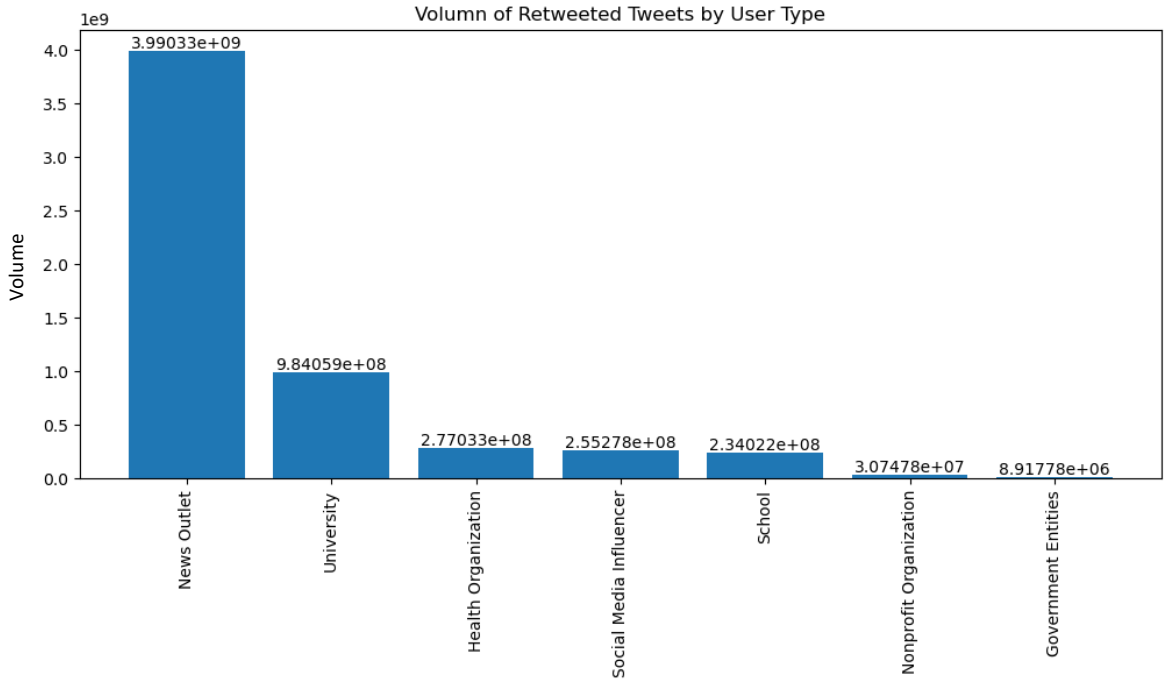
Data run on 32M records

Author Identification - Visualization

Unsurprisingly, schools and universities produce the most original tweets, whereas news outlet get retweeted the most

A lack of Social Media Influencers in our top list of user types getting most retweeted might be misleading due to the difficulty in identifying social media influencers. Therefore, the social media influencers' numbers might be understated.





Data run on 32M records

Location Analysis

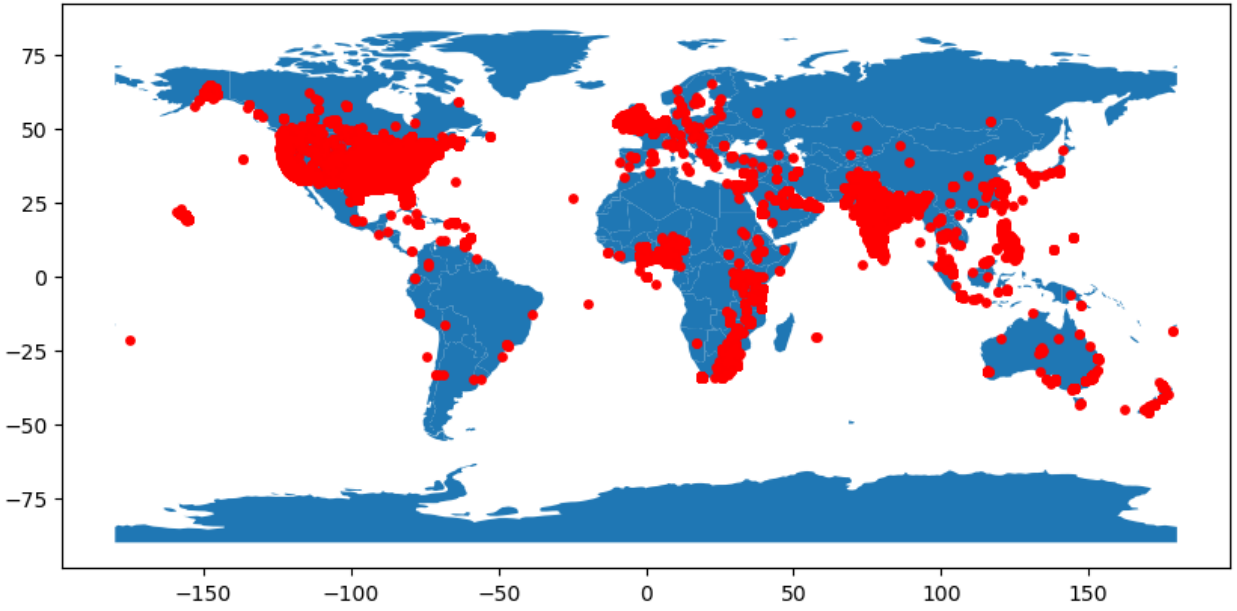
Most tweets are identified as being in the United States

It is worth noting that the filter words choice and the language (English) we used to clean our data in data cleaning contribute to the high concentration in education tweets in the United States

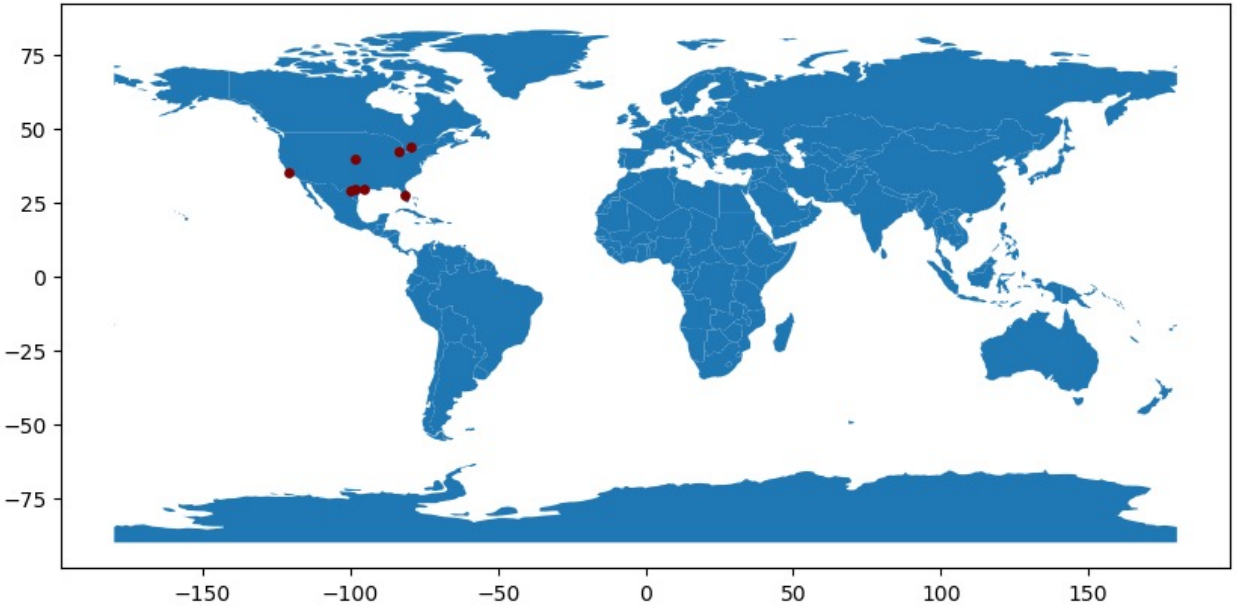
There is no relationship between the emergence of new issues in education and progression and locations of Twitterers tweeting about education

It is most likely because education is such a broad, localized topic, thus it is harder to see an effect that an emergence of a new topic has on the twitterers

Map: Location of Twitterers tweeting about education



Map: Location of tweets on topics in education such as **book ban, school shootings, standardized test, student loan** and **teacher tenure**



Location Analysis

Data run on 32M records

As expected, the majority twitterers in our dataset reside in the United States

- Again, this could well be due to the bias introduced by the filter words list, and the language of the tweets used. Also, Twitter is the most popular in the United States.
- United Kingdom and India follows in the second and third rank

Country	Number of Twitterers
United States	98,303
United Kingdom	10,946
India	10,125
Nigeria	4,544
Canada	4,293
South Africa	1,957
Pakistan	1,820
Australia	1,644
Kingdom of Saudi	1,625
Kenya	1,562
Uganda	1,268
Ireland	1,247

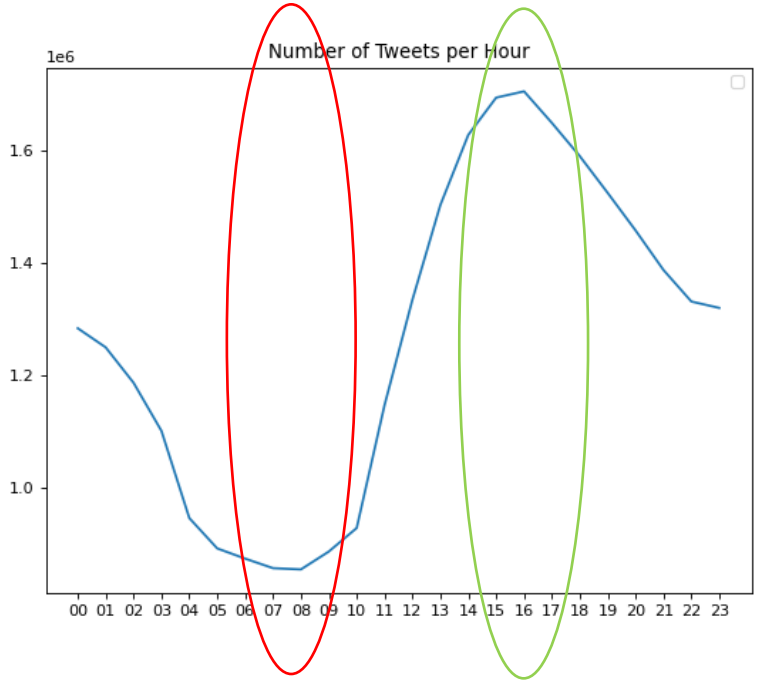
Data run on 32M records

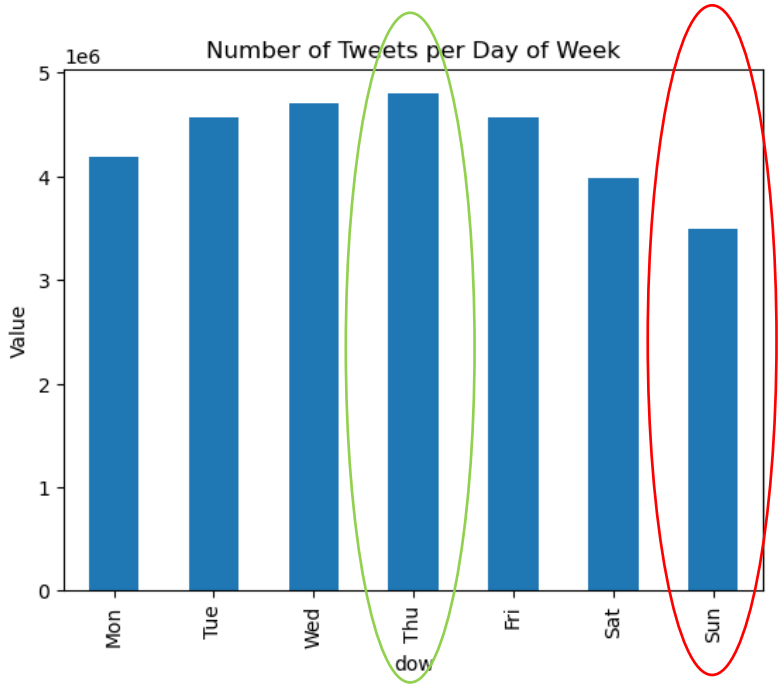
Timeline Analysis

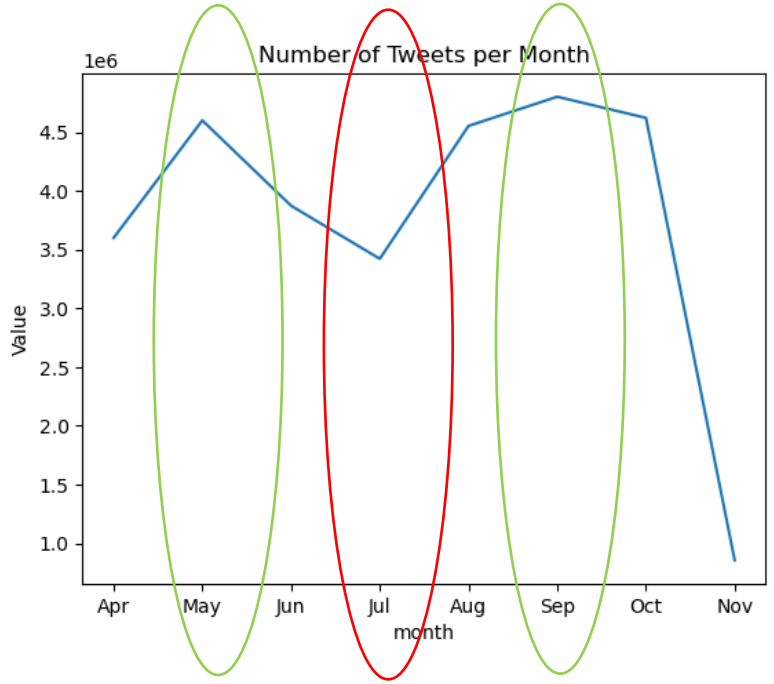
There is missing data in January, February, March and December of 2022, thus affecting our ability to analyze a yearly tweeting trend

There are some apparent trends in the timeline in which education tweets are generated.

- They are more often generated in the afternoon (3-4PM), and the least frequent around 7-8AM
- They are more often generated in weekday (Wed-Thu), and the least frequent in Sun-Mon, consistent with the days of school schedule
- There are most often generated in September and May, and the least frequent in the summer, consistent with the months of school schedule







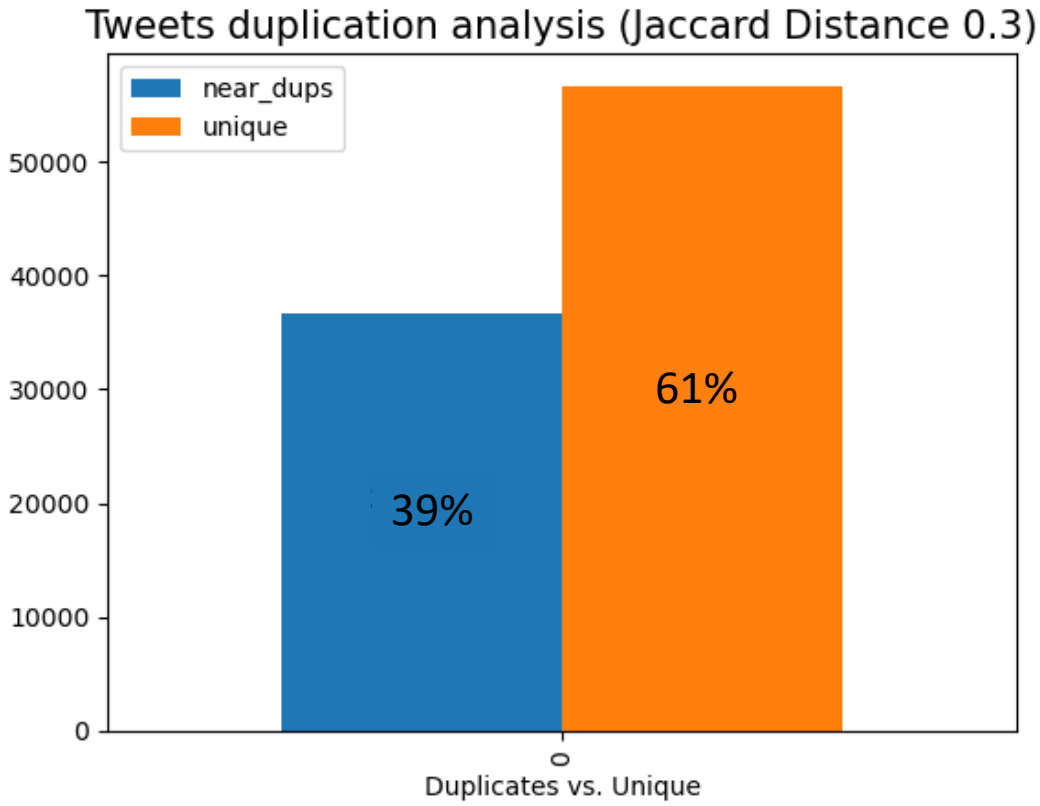
Data run on 93K records

Message Uniqueness Analysis – All Tweets

A sample of 93,329 tweets were run to test for their uniqueness as a proxy for the uniqueness of whole dataset

More than one-third of tweets are near-duplicates

- Since the sample data that is used to run the similarity analysis on is random, and is composed by many different types of twitterers (not just verified twitterers), we see a high rate of near-duplicates of 39% (with a Jaccard distance of 0.3)
- This high rate of near-duplicates includes retweets and very similar tweets, because retweeting does broadcast the messages of the original tweets, and such should be included in the total pool of tweets in uniqueness consideration
- A Jaccard distance of 0.3 was chosen for this corpus of text due to the short nature of each tweet. It is easier for a short word document to have similar words, thus the threshold for it to be identified as near-duplicate should be higher



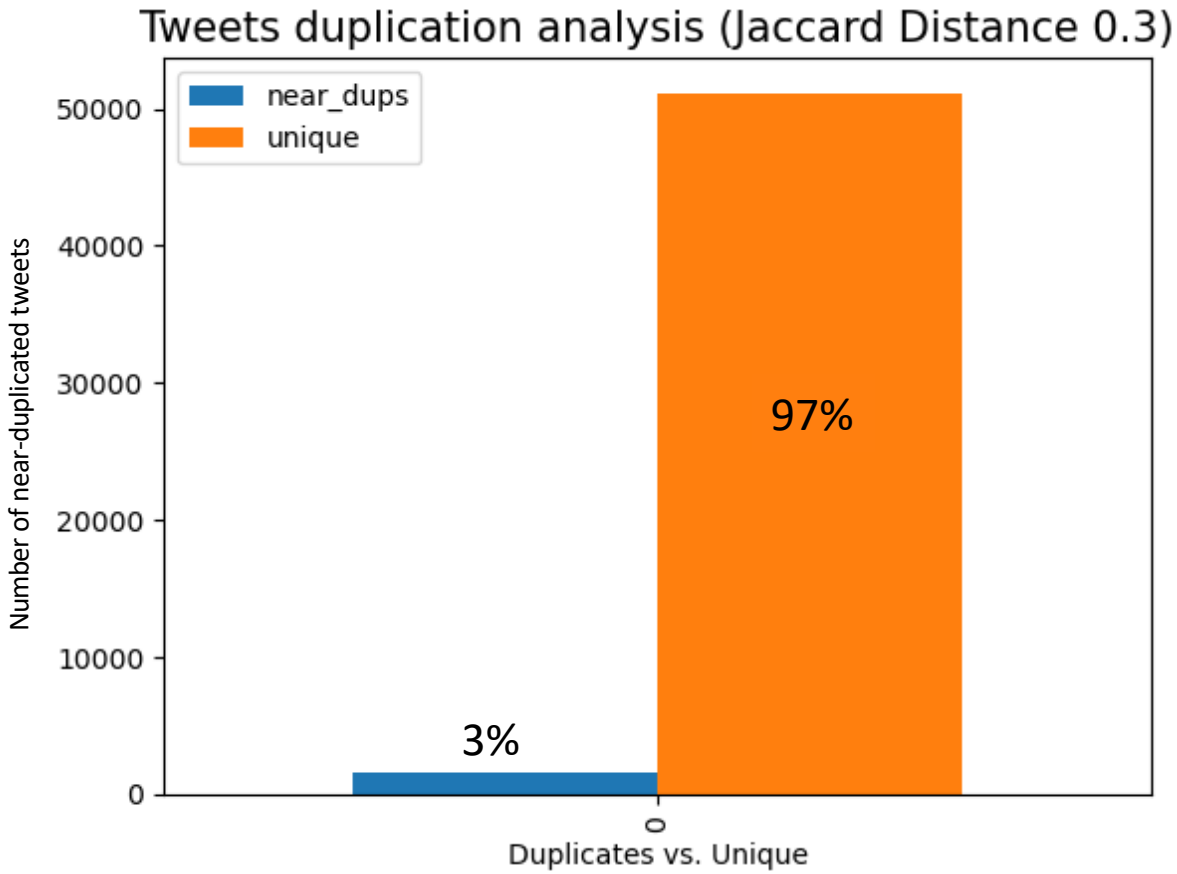
Data run on 53K records

Message Uniqueness Analysis – Original Tweets only

If limiting to only original tweets (not retweets), holding the same Jaccard distance of 0.3, the near-duplicates rate decreases significantly

Original tweets are mostly different from one another

- A different sample data, with only original tweets (not retweets) was run to determine the similarity of these tweets
- Using the same Jaccard distance of 0.3, the high rate of near-duplicates observed in the last similarity test was no longer apparent, having a drastic drop from 39% to 3%
- The original tweets seem to be extremely unique, and thus the high rate of similarity observed in the last test was contributed largely by the retweets, which goes to show the ease and power of knowledge spreading just by a click to retweet

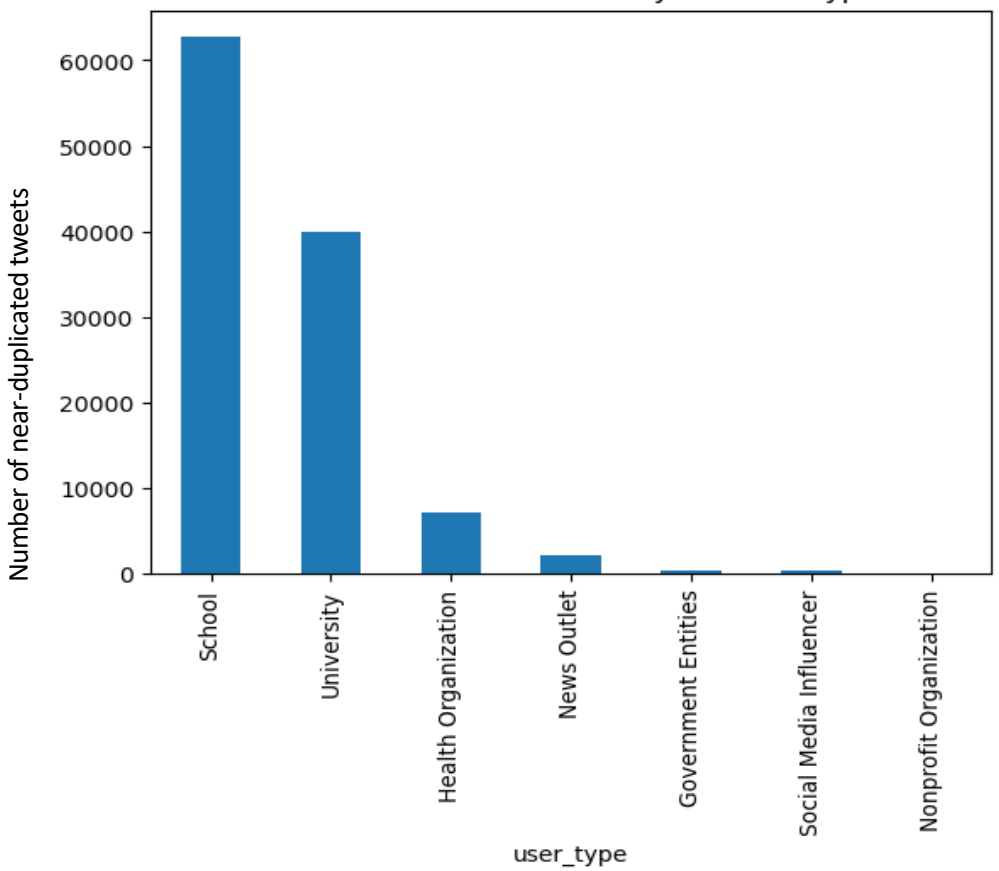


Message Uniqueness Analysis – By User Type

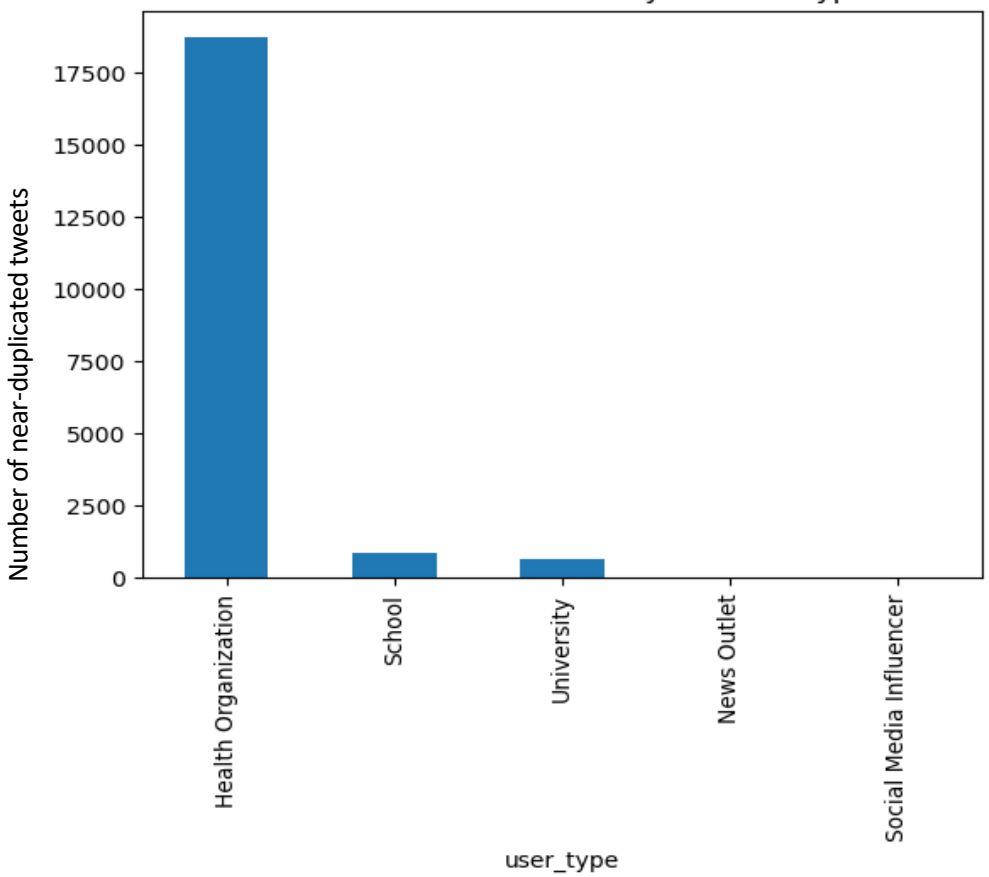
School seems to have the most near-duplicated tweets when retweets are included, but Health Organization has the most near-duplicated tweets when there are only original tweets

- As we have established before, the high rates of near-duplicate tweets in schools and universities are most likely due to the number of retweets, whereas Nonprofit organizations have the fewest similar tweets.

Analysis on Dataset including retweets



Analysis on Dataset excluding retweets



Conclusions and Recommendation

Conclusion	Recommendation
Majority of Twitter's tweets are retweets	Twitter should be more conscious of their retweet function, since this could be used to spread information fast and powerfully and could easily be abused. On the other hand, if the government, or policy makers want to measure the impact of a policy has on the population, they could measure the retweeting rate of such policy
The majority of tweets do not mention education topics, but rather just regular social media posts (photos in school, experiences in college, etc.). A large number of tweets belonging to education on Twitter are talking about school-shootings.	School shooting continues to be an aching problem that attract much attention and conversation in Twitter. This issue could overshadow other more educational topics when doing analysis, thus resulting in inaccurate insights. Controlling for school shooting tweets might be a good way to improve our analysis. In addition, controlling for tweets that are more social media related, instead of news-related, will also improve the analysis. This will require a more sophisticated method to label relevant/irrelevant tweets than the current text mapping method
There are clear trends in time and date that users are more active when tweeting about education	Education tweets have a clear trend of occurrence in accordance to the school year, and weekdays. Knowing this insight, we can know when to start collecting data and when we can possibly stop to minimize the efforts to collect data for analysis