Twitter Analysis

BDP Final Project

Kaitong Hu
December 2022

Agenda

- Executive Summary
- Methodology & Source Data
- EDA & Available Variables
- Tweet Clean-Up & Filtering
- Author Identification
- Location Analysis
- Timeline Analysis
- Message Uniqueness Analysis
- Conclusions & Recommendations

Executive Summary

Given Twitter’s huge user base and high volume of tweets posted each day, many people are using Twitter as a main tool to gather information on the topics they are interested in. However, do all the tweets on this platform provide meaningful insights on a specific topic?

This project aims to assess whether Twitter can be considered a credible source of information, which reflects the emergence of important trends or topics in education, through four dimensions:

- Who posts these tweets?
- Where is a tweet published from?
- When is a tweet posted?
- How unique is a tweet?

After investigation, I found that Twitter could be a great source of information. However, further steps are needed to determine if it could be considered as a credible source.

Methodology & Source Data

Data Source

- ~100 million tweets related to education
- Location:

  gs://msca-bdp-tweets/final_project

File Types

- The original dataset is stored in **JSON** files
- The processed dataset is stored in **Parquet**

Visualization

- Matplotlib
- Seaborn
- Chart types used: **line charts & bar charts**

Platform & Tools

- **Google Cloud Platform**: a cloud computing service
- **PySpark:** primarily used Spark DF and RDD to analyze the data

Methods & Functions

.filter()
.withColumn()
.select()
.groupby()
.agg()
.limit()
.rlike()
.contains()

Jaccard Similarity

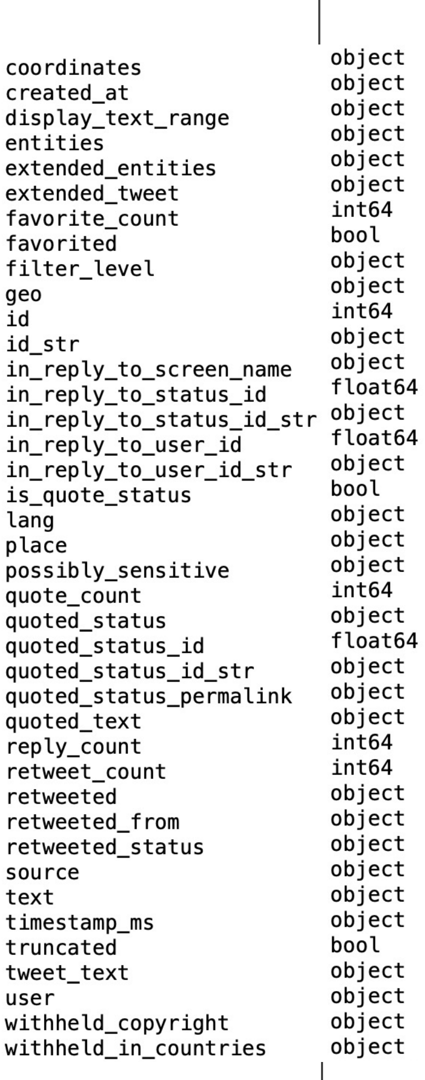
Packages used:
pyspark.ml.feature
nltk.corpus
Functions used:
MinHashLSH
CountVectorizer
stopwords

Source Data Overview

- File type: JSON file
- File size: ~100 million rows, where each row contains relevant information about a tweet
- Features: 40 variables in total, many of them are json objects

Data Type	Number of Columns
boolean	3
float64	3
int64	5
object	29

- Details of data type for all columns can be seen on the right



```
coordinates                  object
created_at                   object
display_text_range           object
entities                     object
extended_entities            object
extended_tweet               object
favorite_count                int64
favorited                      bool
filter_level                 object
geo                          object
id                            int64
id_str                       object
in_reply_to_screen_name      object
in_reply_to_status_id       float64
in_reply_to_status_id_str    object
in_reply_to_user_id         float64
in_reply_to_user_id_str      object
is_quote_status                bool
lang                         object
place                        object
possibly_sensitive           object
quote_count                   int64
quoted_status                object
quoted_status_id            float64
quoted_status_id_str         object
quoted_status_permalink      object
quoted_text                  object
reply_count                   int64
retweet_count                 int64
retweeted                    object
retweeted_from               object
retweeted_status             object
source                       object
text                         object
timestamp_ms                 object
truncated                      bool
tweet_text                   object
user                         object
withheld_copyright           object
withheld_in_countries        object
```

Tweet Clean-up & Filtering

- **Topic:** "Critical Race Theory"
- **URL:** https://chicago.chalkbeat.org/2022/3/1/22957083/illinois-legislation-curriculum-transparency-critical-race-theory-bill

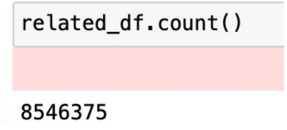
Define Key Words to Filter Out Irrelevant Tweets

critical race theory
racist
racism
sexism
sexist
sex
inequality
teacher
classroom
bills
legislation
anti-racism

Tweets that contain any of these key words are kept

Final Dataset

Approximately **8.5%** of the rows in the original dataset are kept



```
related_df.count()
```

```
8546375
```

Only **12 cleaned variables** will be used in the analysis

(details of these variables are discussed in next few pages)

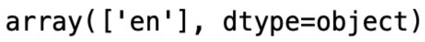
Exploratory Data Analysis (EDA)

01 created_at

All the tweet are created in the year of 2022

02 Lang (excluded)

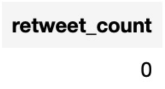
This column contains only 'en', thus it is meaningless to my analysis



```
array(['en'], dtype=object)
```

03 retweet_count (excluded)

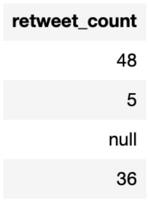
The retweet_count column contains only zero, not useful at all



retweet_count
0

04 retweet_status.retweet_count

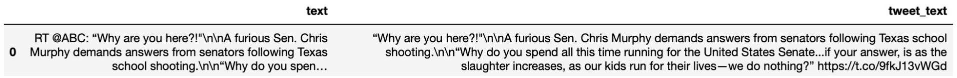
Within the retweet_status json object, I found a retweet_count variable that could be used



retweet_count
48
5
null
36

05 text vs. tweet_text (use tweet_text)

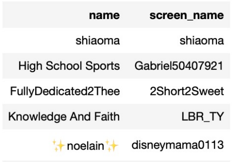
The text column and tweet_text column are pretty much the same, except text also contains a username



	text	tweet_text
0	RT @ABC: “Why are you here?!"\n\nA furious Sen. Chris Murphy demands answers from senators following Texas school shooting.\n\n“Why do you spen…	“Why are you here?!"\n\nA furious Sen. Chris Murphy demands answers from senators following Texas school shooting.\n\n“Why do you spend all this time running for the United States Senate...if your answer, is as the slaughter increases, as our kids run for their lives—we do nothing?” https://t.co/9fkJ13vWGd

06 name vs. screen_name (use screen_name)

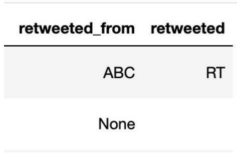
The screen_name column does not have any emoji, and it is unique



name	screen_name
shiaoma	shiaoma
High School Sports	Gabriel50407921
FullyDedicated2Thee	2Short2Sweet
Knowledge And Faith	LBR_TY
✨noelain✨	disneymama0113

07 retweeted_from & retweeted

retweeted_from is a better variable we can use to identify original content



retweeted_from	retweeted
ABC	RT
None	

08 coordinates (exclude)

Only 1% of the rows have values in this column

Feature Selection & Engineering

12 variables used in total

8 Original Features

created_at

tweet_id

tweet_text

user_screen_name

user_location

user_description

retweeted_status.retweet_count

retweeted_from

These three variables are extracted from the **user** json object

4 New Features

Organization

- Key words for each organization type are first specified
- Every tweet is classified into an organization type **based on the user_description**

cleaned_location

- Derived from **user_location** column
- The terms before the ',' in user_location are used as the value for this column

date

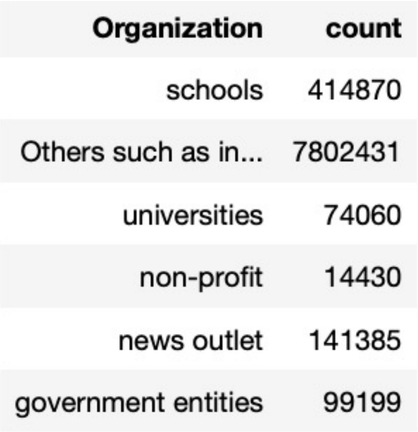
- Derived from **created_at**
- Extracted the year, month and day value from created_at column

month_year

- Derived from **created_at**
- Extracted the year and month value from created_at column

Author Identification Analysis

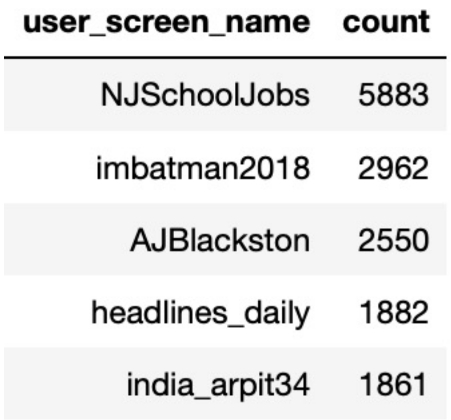
# of tweets by organization types:



Organization	count
schools	414870
Others such as in...	7802431
universities	74060
non-profit	14430
news outlet	141385
government entities	99199

Among 8546375 tweets, about **91%** of them are posted by users in **“Others”** group

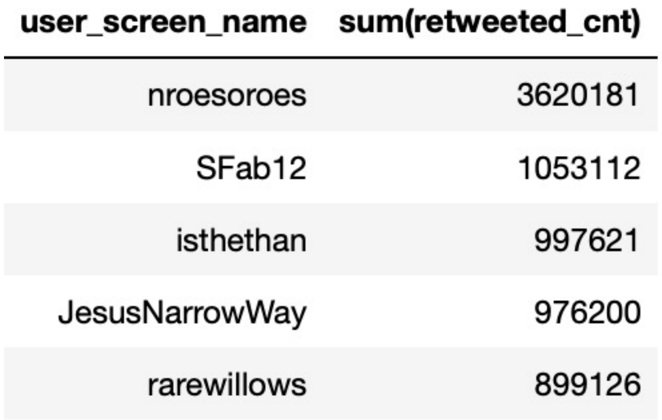
Top five twitterers by # of original tweets:



user_screen_name	count
NJSchoolJobs	5883
imbatman2018	2962
AJBlackston	2550
headlines_daily	1882
india_arpit34	1861

The user **NJSchoolJobs** has posted **the highest** number of original content related to the topic chosen

Top five twitters by # of retweets by other users:

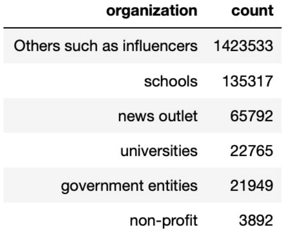


user_screen_name	sum(retweeted_cnt)
nroesoroes	3620181
SFab12	1053112
isthethan	997621
JesusNarrowWay	976200
rarewillows	899126

Users with the highest number of retweets by others are **different** from users who posted the most. **nroesoroes** received the highest number of retweets by other users

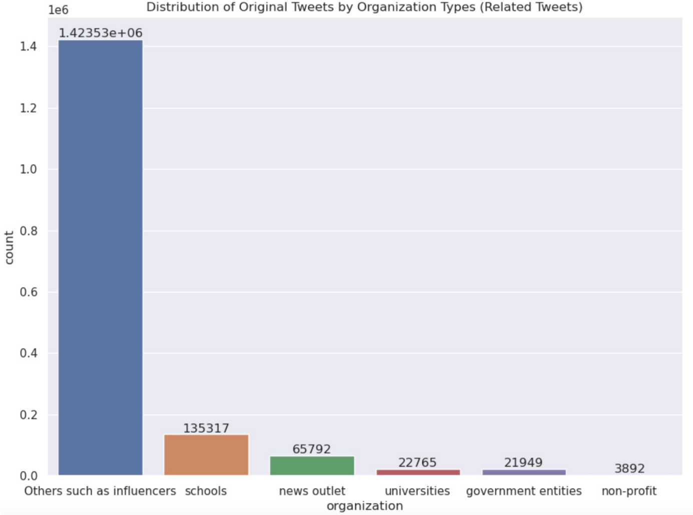
Distribution of Tweet/ Retweet Volume by Organization Type

Original Tweets by Organization Types

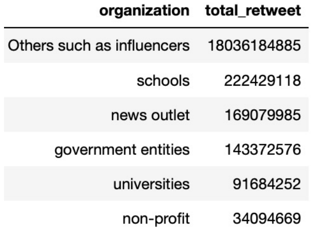


organization	count
Others such as influencers	1423533
schools	135317
news outlet	65792
universities	22765
government entities	21949
non-profit	3892

Since the "Others such as influencers" group has the highest number of tweets, it is reasonable to expect that this group also has the highest number of original tweets. Groups that followed are "schools" and "news outlet"

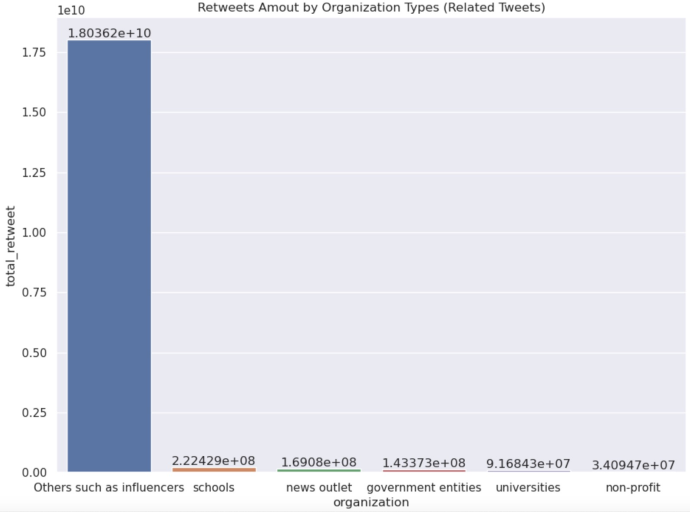


Number of Retweets by Organization Types



organization	total_retweet
Others such as influencers	18036184885
schools	222429118
news outlet	169079985
government entities	143372576
universities	91684252
non-profit	34094669

Likewise, we can also expect that "Others such as influencers" group has the highest number of retweets by others. "schools" and "news outlet" are ranked 2nd and 3rd, respectively. The "non-profit" group has the fewest number of retweets

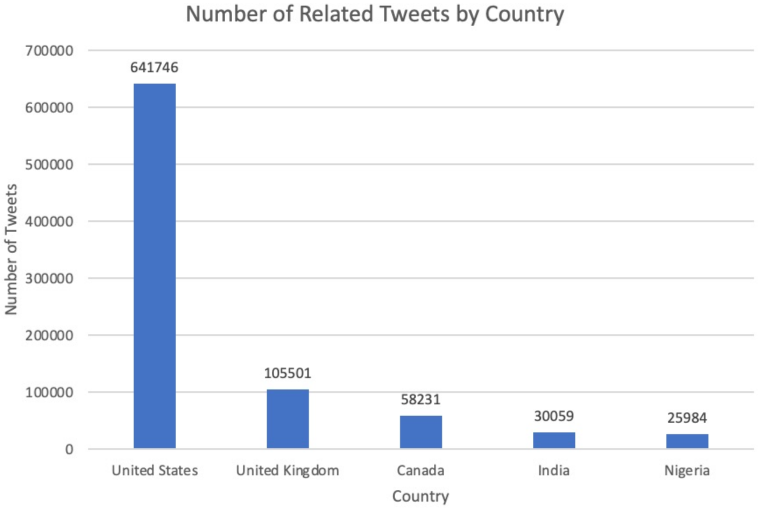


Location Analysis

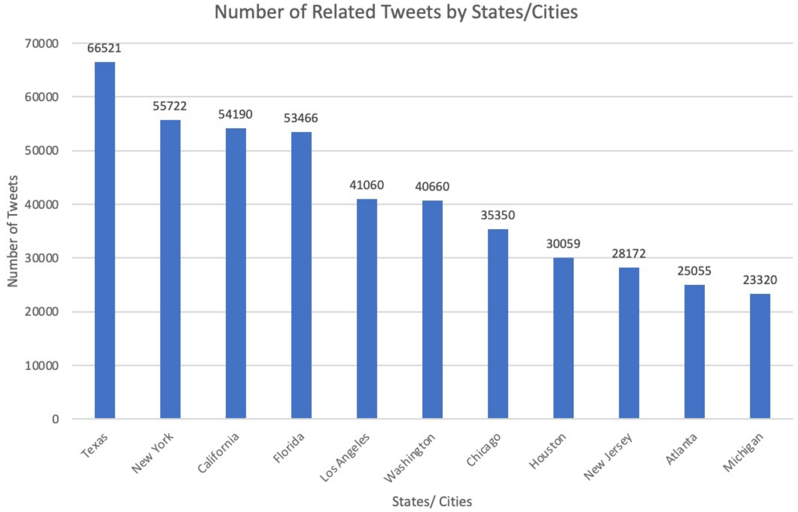
about 10% of the filtered dataset (~860,000 rows) are used

In this sample dataset, about **75%** of the tweets are posted by users in the **United States**, while approximately 12% of the tweets are from users in the UK

The below bar chart displays the states/cities with the highest number of tweets related to the topic chosen. One reason for this distribution could be that these locations have **higher numbers of education institutions**

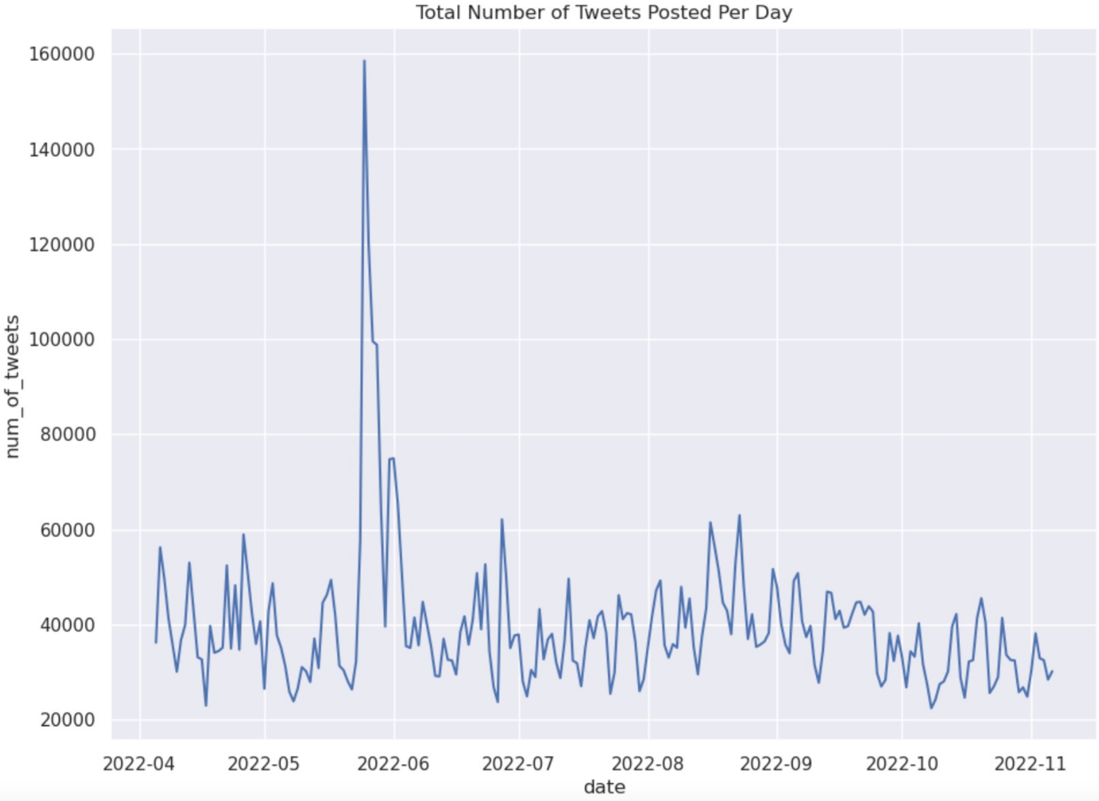


For United States specifically



*Note: because some users only specified the state in which they are in, the counts for states **do not** include the counts for cities that belong to that specific states

Timeline Analysis



The line graph on the right represents the distribution of tweets posted from April 2022 to November 2022. There was a **surge** of tweets at the **end of May**, which coincided with the time at which news about whether the "critical race theory" should be banned in Illinois schools came out

There are **obvious gaps** between different dates on the amounts of related tweets posted

Since **mid-August**, the number of tweets related to the "critical race theory" is in a **decreasing trend**

Message Uniqueness Analysis

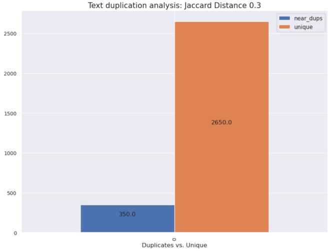
Jaccard similarity and MinHashLSH are used to assess uniqueness

For a sample of 3000, regardless of the organization type

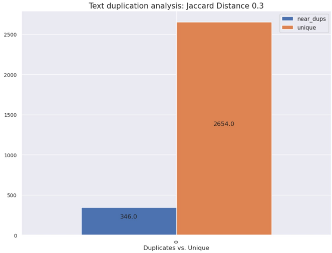
Jaccard Distance	Near Duplicate	Unique
0.3	406	2594
0.5	419	2581
0.7	600	2400

After comparing the actual text itself, I found a **Jaccard Distance of 0.3** best captures the uniqueness of tweets. Using a Jaccard Distance of 0.3, about 13.5% can be considered as near-duplicate, while 86.5% are unique

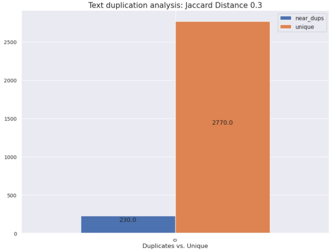
government entities



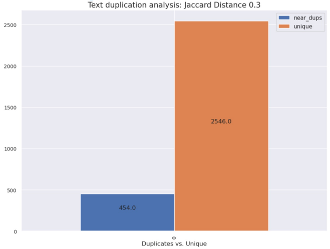
universities



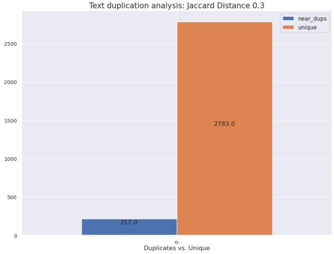
schools



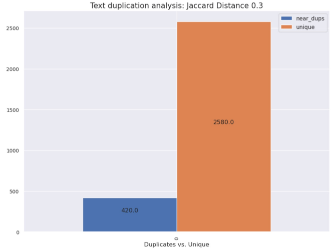
non-profit



news outlet



Others such as influencers



From the graphs on the left, we can see that users from groups, such as **"schools"** and **"news outlet"**, tend to post **more unique** tweets than users classified in other groups

Conclusion

After thoroughly analyzing the tweets as discussed in this presentation, it is reasonable to conclude that Twitter **can be a useful source of information** that reflects the emergence of important trends or topics in education. We can use this platform to better understand the public's opinion regarding a topic we are particularly interested in. **However**, tweets on Twitter **should not be** considered as **credible source** of information for the following reason:

- Most of the tweets are posted by individuals who **do not** belong to any credible institutions
- The consistency between surge of tweets and the emergence of new hot topics **could be a coincidence**, further analysis is needed to confirm the relationship

Recommendation

In order **to assess the credibility** of messages posted on Twitter, we need to:

- Identify ways to accurately classify twitterers into different groups
- Determine if the differences in the number of tweets posted before & after a specific event is statistically significant (hypothesis testing may be needed)
- Determine if these related tweets indeed form meaningful opinions on a topic