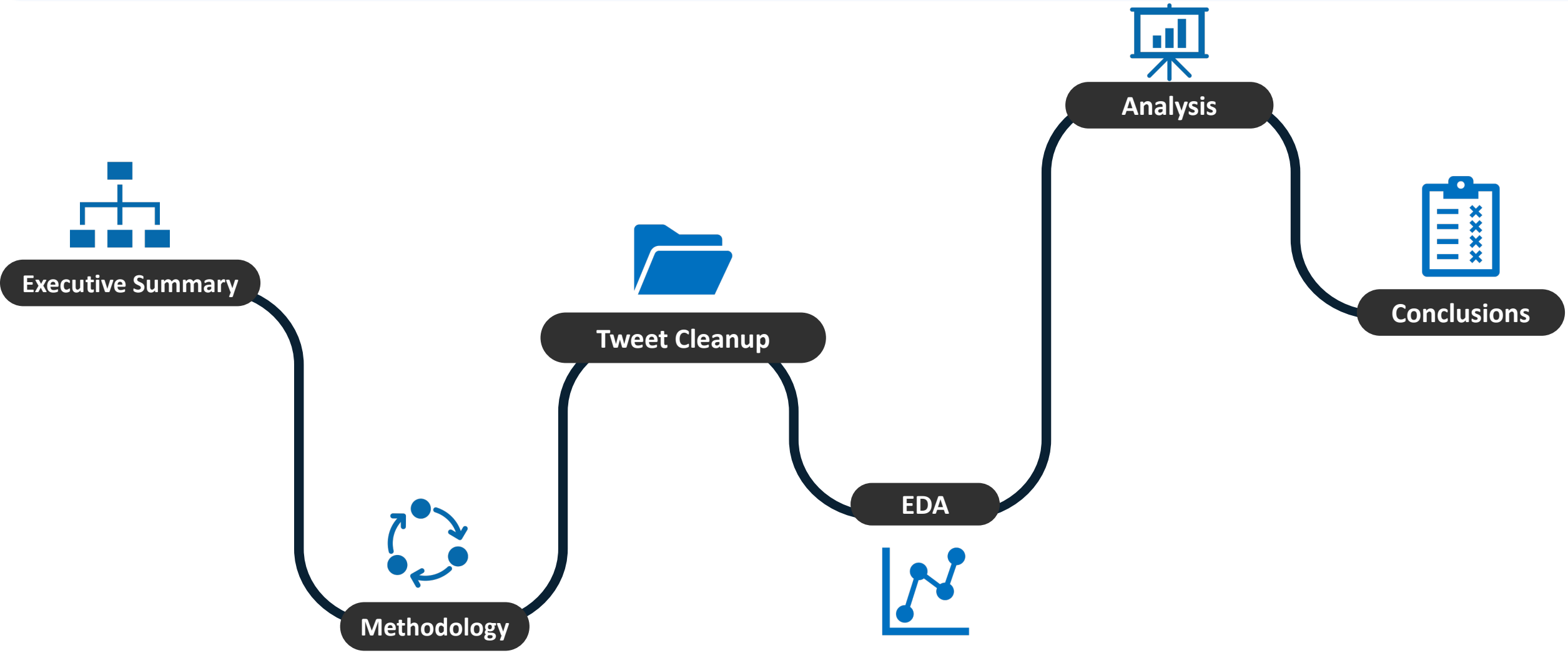
TWITTER: CREDIBLE INFORMATION SOURCE

Presenter: Snigda Gedala

Dec 08, 2022

Agenda


Executive Summary
Methodology
Tweet Cleanup
EDA
Analysis
Conclusions

EXECUTIVE SUMMARY

Background:

Twitter is a social networking service on which users post and interact with messages known as "tweets". Registered users can post, like, and retweet tweets, while unregistered users only have the ability to read public tweets. Twitter provides valuable data to be able to analyse and draw insightful patterns and hence making decisions accordingly.

Objective:

In this project, the objective is to identify whether Twitter can be considered a credible source of information, which reflects the emergence of important trends or topics in education. In this process, we will be performing twitterer identification, location analysis, timeline analysis and tweets uniqueness.

METHODOLOGY AND SOURCE DATA OVERVIEW

Background

- Overall Analysis : PySpark
- Data Visualisations : Matplotlib, Pandas
- Text Similarity Analysis : LSH

Data Overview

- Number of rows : ~ 100M
- Number of columns : 40
- File Type : JSON

Columns that had null values greater than 70%

```
['coordinates',
 'display_text_range',
 'extended_entities',
 'extended_tweet',
 'geo',
 'in_reply_to_screen_name',
 'in_reply_to_status_id',
 'in_reply_to_status_id_str',
 'in_reply_to_user_id',
 'in_reply_to_user_id_str',
 'place',
 'possibly_sensitive',
 'quoted_status',
 'quoted_status_id',
 'quoted_status_id_str',
 'quoted_status_permalink',
 'quoted_text',
 'withheld_copyright',
 'withheld_in_countries']
```

TWEET CLEANUP AND FILTERING

Step 1 : Choosing tweets who are written only in English using the column “lang”

```
twitter_df=twitter_df.filter(twitter_df.lang=='en')
```

Step 2 : Filtering tweets related to primary, secondary and higher education

Selecting all the tweets related to k-12 by using highest frequency of words that are related to this topic – education.

```
tweets_key=twitter_df.filter(lower(col('text')).contains('primary education')|\
                             lower(col('text')).contains('secondary education')|\
                            lower(col('text')).contains('primary school')|\
                            lower(col('text')).contains('elementary education')|\
                            lower(col('text')).contains('normal school')|\
                            lower(col('text')).contains('elementary school')|\
                            lower(col('text')).contains('compulsory education')|\
                            lower(col('text')).contains('school')|\
                            lower(col('text')).contains('middle school')|\
                            lower(col('text')).contains('junior school')|\
                            lower(col('text')).contains('public school')|\
                            lower(col('text')).contains('education')|\
                            lower(col('text')).contains('educate')|\
                            lower(col('text')).contains('teacher')|\
                            lower(col('text')).contains('student')|\
                            lower(col('text')).contains('schoolmate')|\
                            lower(col('text')).contains('university')|\
                            lower(col('text')).contains('college')|\
                            lower(col('text')).contains('interschool')|\
                            lower(col('text')).contains('textbook')|\
                            lower(col('text')).contains('higher education')|\
                            lower(col('text')).contains('high school')|\
                            lower(col('text')).contains('students')|\
                            lower(col('text')).contains('schools')|\
                            lower(col('text')).contains('curriculum')|\
                            lower(col('text')).contains('undergraduate')|\
                            lower(col('text')).contains('grades')|\
                            lower(col('text')).contains('cgpa'))
```

Number of rows after cleanup:

~75.7M

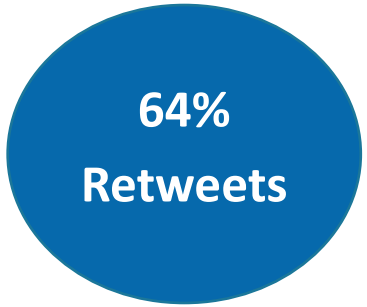
EDA AND EXTENSIVE USAGE OF AVAILABLE VARIABLES

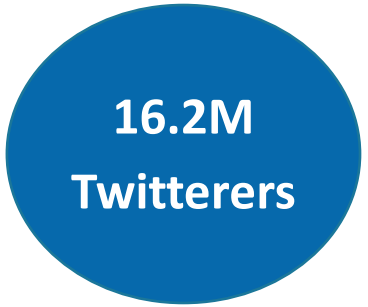
Step 1 : Null Value Check

Observed that there are columns where more than 70% pf the data is nulls. Hence, removed them from the dataset

Step 2 : Understanding the variables avaivable

Looked into the variables and finalized which variables are useful for the analysis

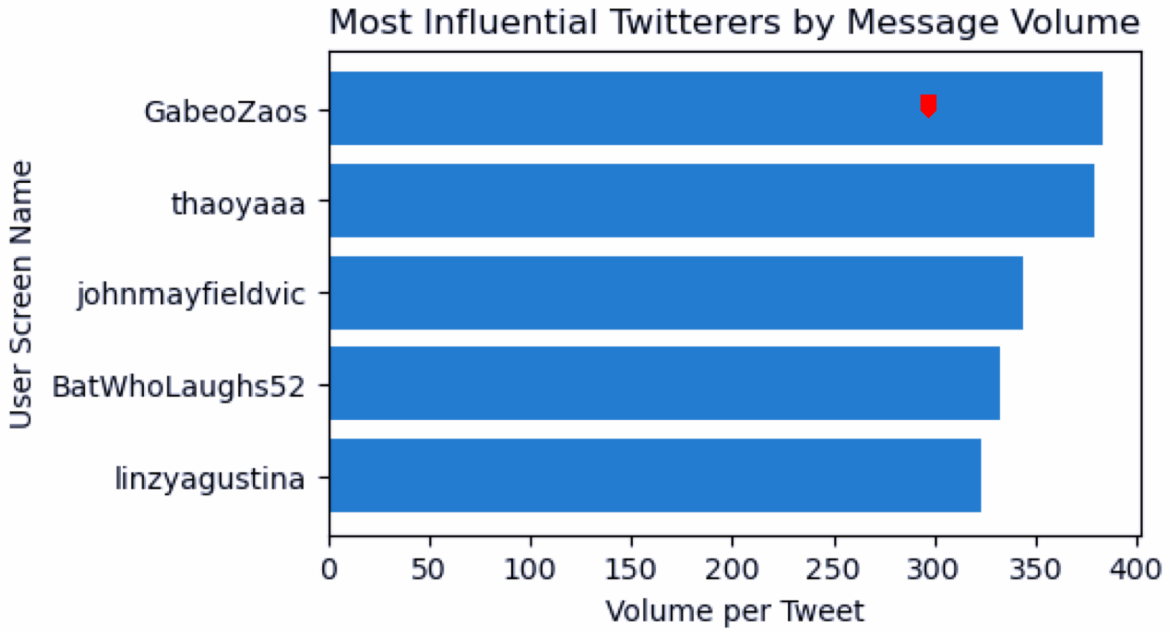




AUTHOR IDENTIFICATION

- By Message Volume

- GabeoZaos has posted highest volume of tweets about education.
- Thaoyaaa has posted second highest volume of tweets about education

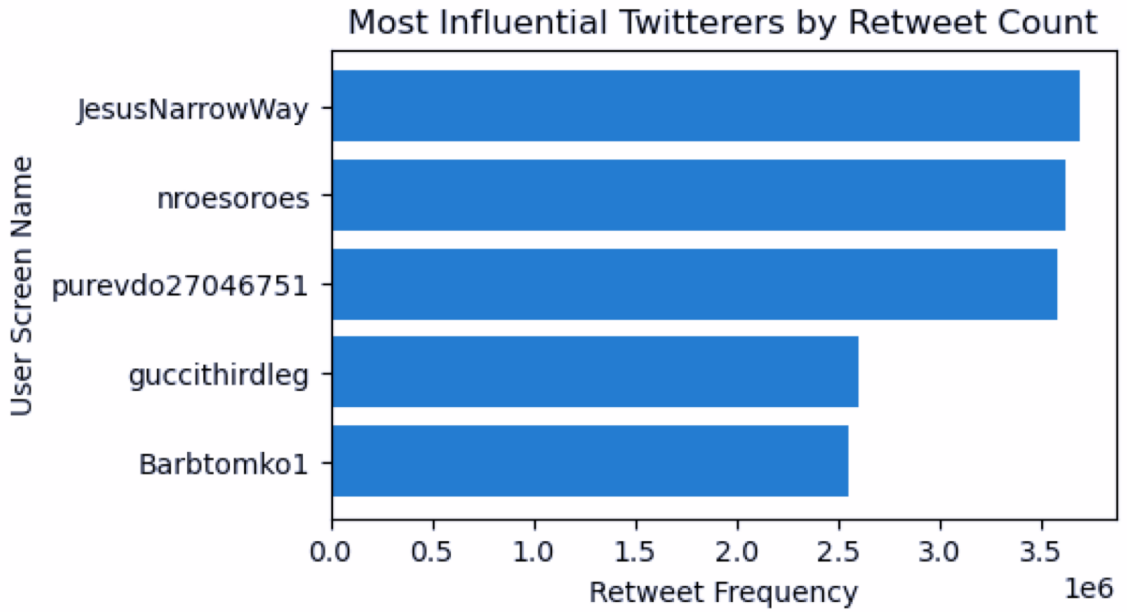


	id_str	screen_name	Volume
0	1458935628264095745	GabeoZaos	383.0
1	1273615219168653312	thaoyaaa	379.0
2	1547675376523939842	johnmayfieldvic	344.0
3	1155297287465328641	BatWhoLaughs52	332.0
4	1718044848	linzyagustina	323.0

AUTHOR IDENTIFICATION

- By Message Retweet

- JesusNarrowWay tweets(about education) have highest number of retweets(~3.69M).
- nroesroes tweets(about education) have second highest number of retweets(~3.62M).

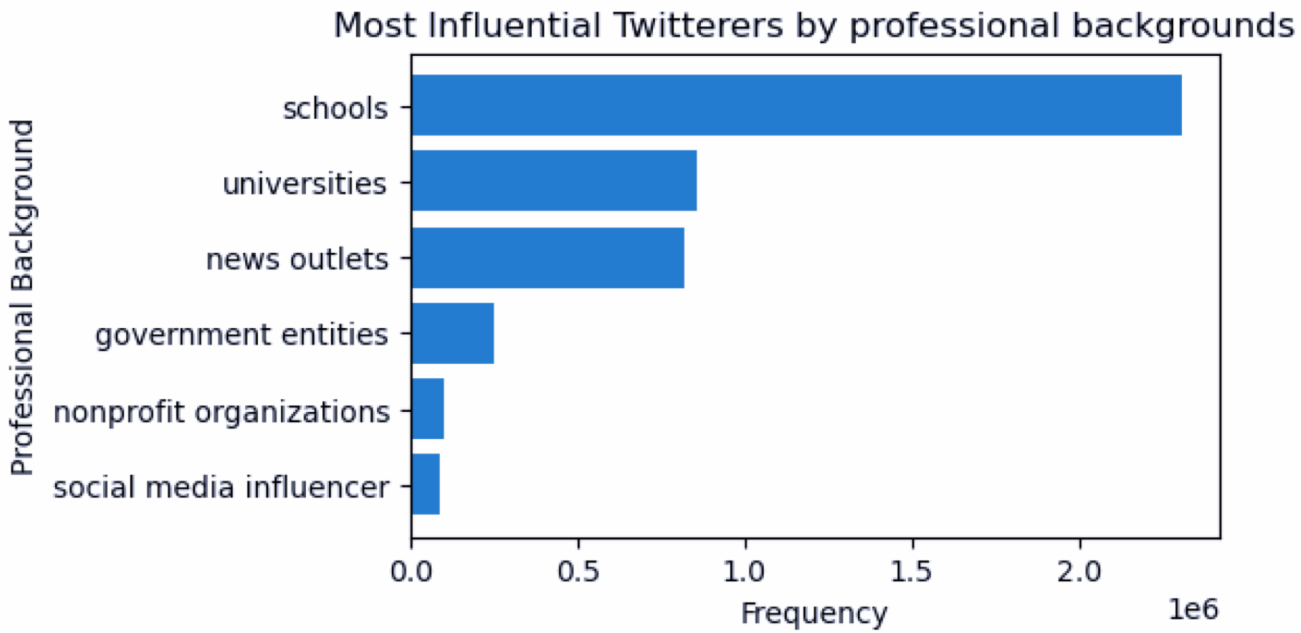


	id_str	screen_name	Total_Retweet_Count
0	362084030	JesusNarrowWay	3688172
1	1516373316033863690	nroesoroes	3623418
2	1552239312195817474	purevdo27046751	3578311
3	1501043135639695367	guccithirdleg	2602455
4	804430039	Barbtomko1	2552806

AUTHOR IDENTIFICATION

- By Professional Background

- Twitterers belonging to schools have highest number of tweets about education(~2.3M).
- Twitterers belonging to universities have second highest number of tweets about education(~0.85M).

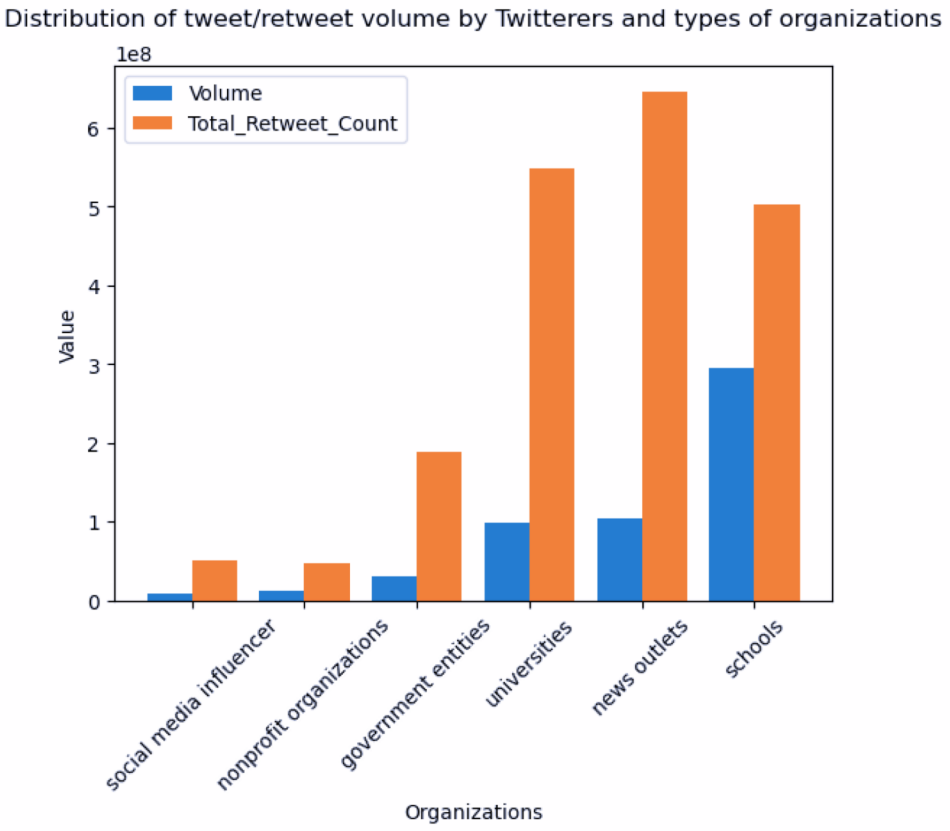


background	count
social media influencer	89703
nonprofit organizations	100866
government entities	248966
news outlets	818864
universities	858565
schools	2308094

AUTHOR IDENTIFICATION

-Distribution of tweet/retweet volume by Twitterers and types of organizations

- Twitterers belonging to schools have highest volume of tweets about education.
- Twitterers belonging to news outlets have highest number of retweets about education.

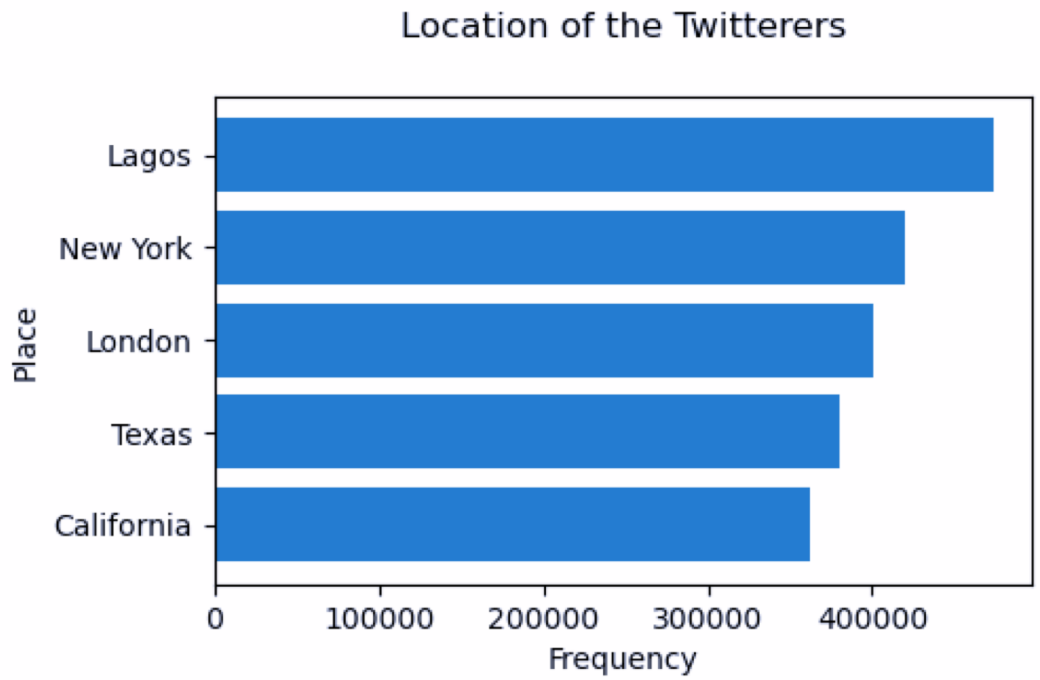


pf_bg	Total_Retweet_Count	Volume
schools	503120263	294083309
news outlets	646457982	103633968
universities	549272378	98080781
government entities	188454315	29777504
nonprofit organizations	46703835	13001086
social media influencer	49871848	7888800

LOCATION ANALYSIS

- Geographical Distribution

- Highest number of twitterers belong to Lagos.
- Second highest number of twitterers belong to New York.
- It is observed that three places in the top 5 are in United States itself, while the top one belongs to Africa and top 3 belongs to United Kingdom.

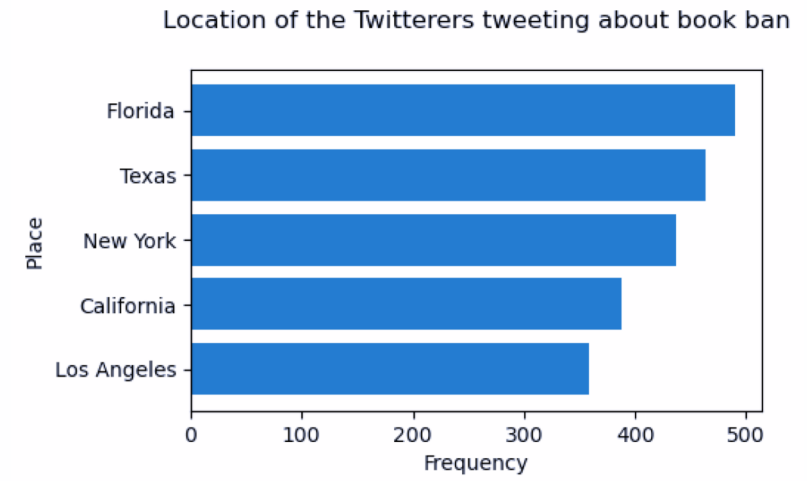


user_place	count
Lagos	473387
New York	419528
London	400238
Texas	380219
California	361585

LOCATION ANALYSIS

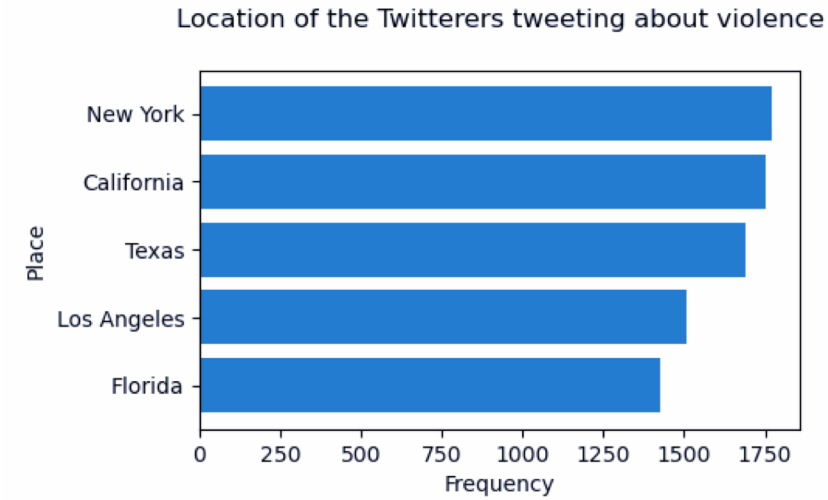
- Emergence of new issues in education and locations

- Highest number of twiterrers of Florida tweeted about book ban.



	user_place	count
0	Florida	490
1	Texas	464
2	New York	437
3	California	388
4	Los Angeles	358

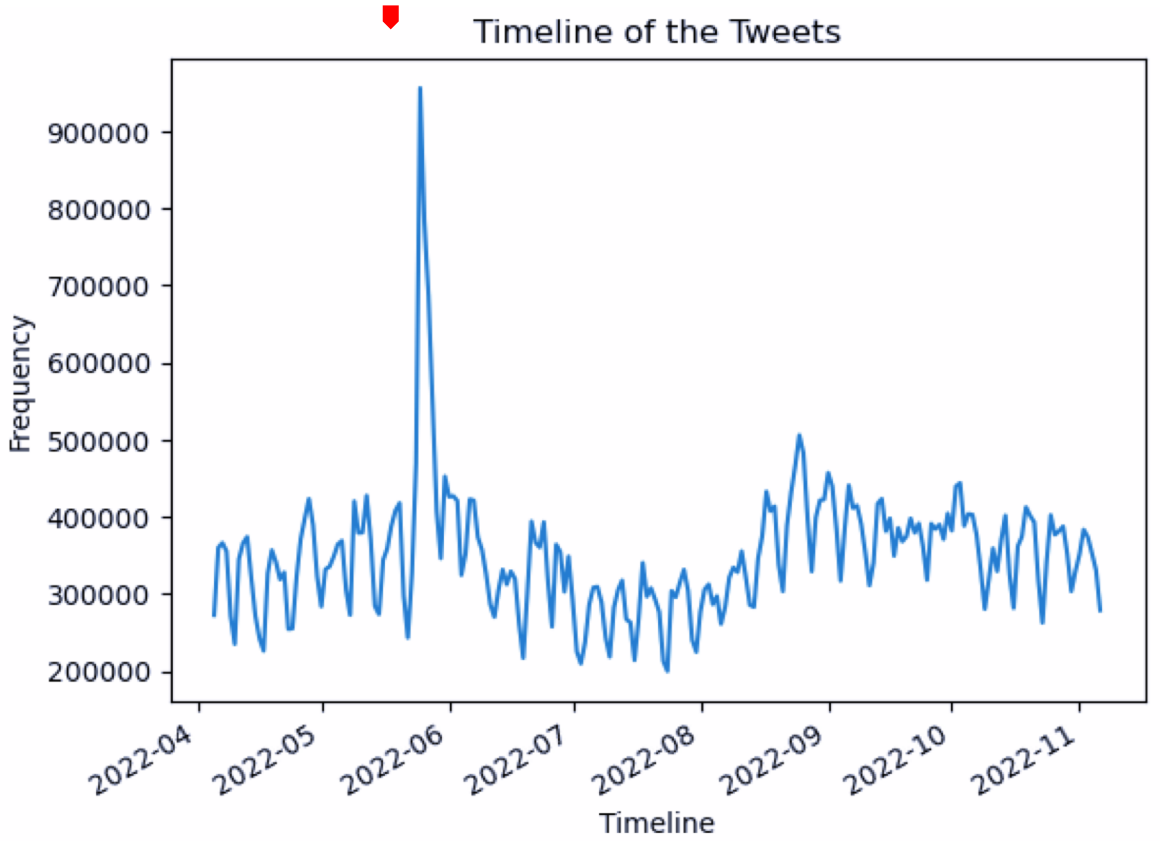
- Highest number of twiterrers of New York tweeted something related violence and education.



	user_place	count
1	New York	1771
2	California	1750
3	Texas	1688
4	Los Angeles	1507
5	Florida	1424

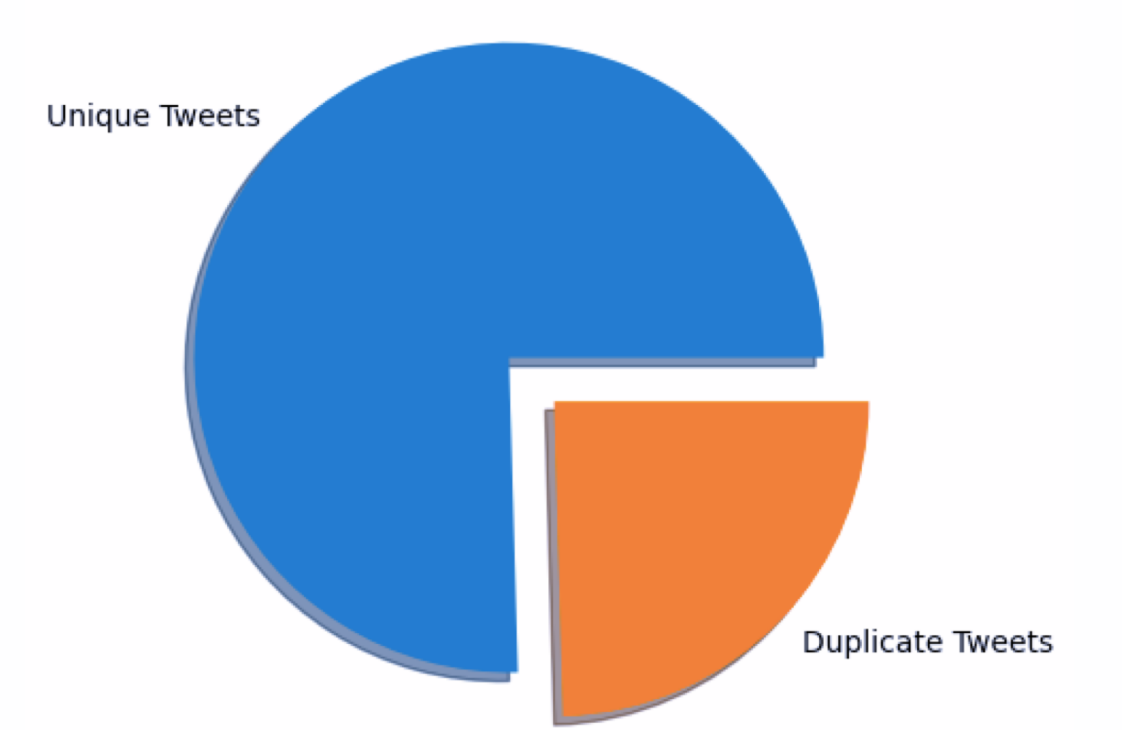
TIMELINE ANALYSIS

- It is observed that there is a very significant sudden peak around the end of June of 2022.
- Rest of the months seem to have similar pattern of tweets



TWEETS UNIQUENESS

- 75.45% of the tweets were unique tweets and the rest 24.55% tweets are duplicate.
- Jaccard distance used : 0.5



CONCLUSIONS AND ACTIONABLE RECOMMENDATIONS

Conclusions:

- The Twitter API offers a wealth of information about public opinion and is extremely helpful in data mining applications.
- The most educational tweets have been tweeted by GabeoZaos.
- Tweets by JesusNarrowWay (about education) have received the most retweets (3.69M).
- Twitter users from colleges and universities are the ones who tweet about education the most (2.3 million).
- Although Twitter users are highest among schools and institutions, news organizations receive the most retweets.
- The majority of tweets about education are from the United States.
- It is observed that education trends are usually being reflected the most in the big cities of USA like New York and California.

Recommendation:

- The current data belong to only 2022. Instead, data should be taken from a wider range of timeline to be able to better understand if the tweets reflects the emergence of important trends or topics in education.

THANK YOU