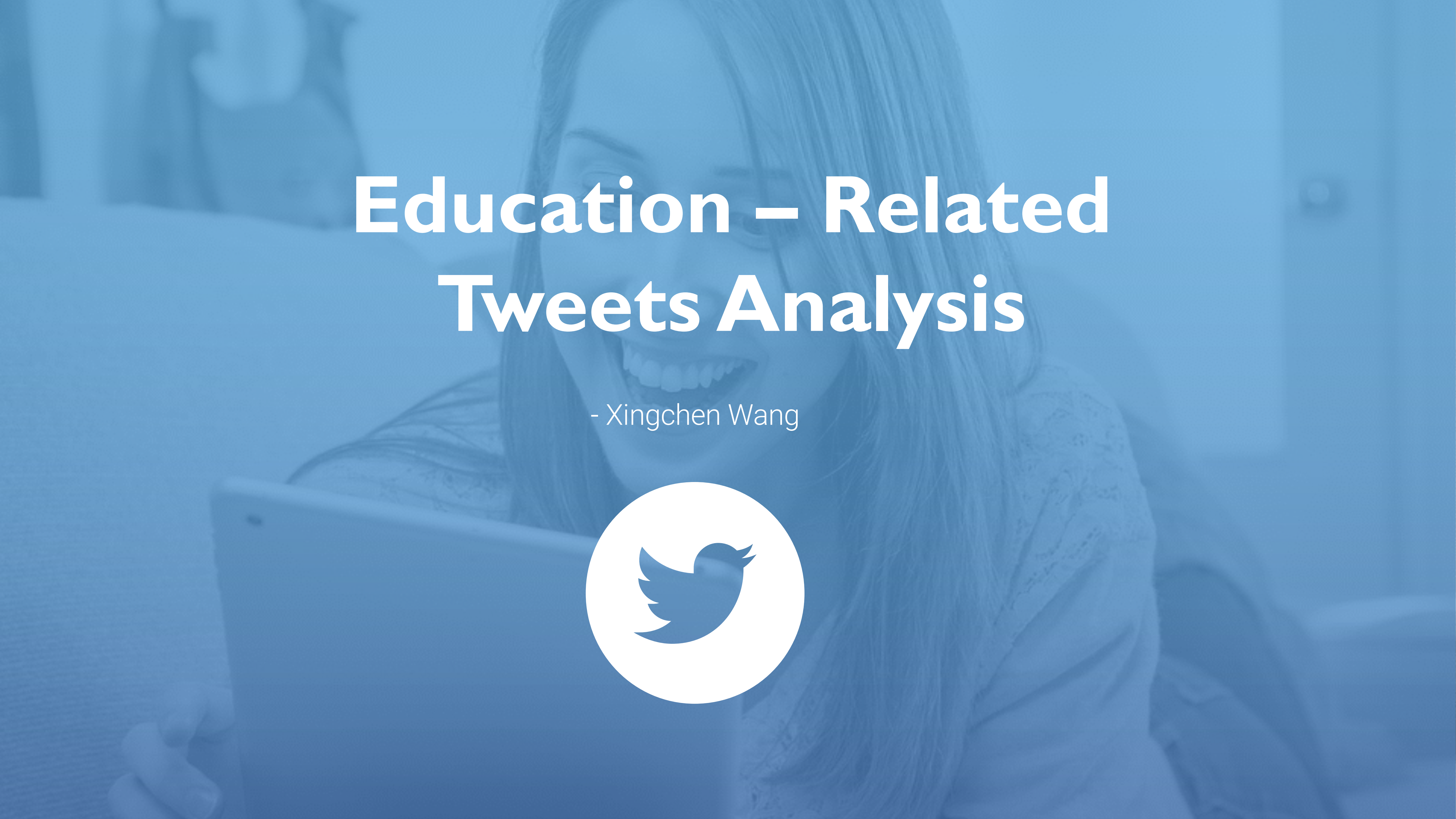


# Education – Related Tweets Analysis

- Xingchen Wang

# EXECUTIVE SUMMARY

1. Spike in Twitter activity will generally reflect the emergence of new hot topics in education, like the Uvalde school shooting.
2. Sports and school related accounts and have the most original tweets and tweets from news outlets and social media influencers are retweeted most.
3. Most of the tweets from verified users are original since they all have similarity below 15% regardless of the type of organization.
4. If there is an emergence of new issues in a certain country, the Twitter users in that country will have more activity.

# METHODOLY & SOURCE DATA OVERVIEW

Methodology

- Pyspark for coding and analyzing
- Pandas and matplotlib for plotting and visualizing
- LSH for text similarity analysis

Data Overview

- 39067882 rows of data
- Stored in google cloud
- 18 columns

| Column | % NA |
|---|---|
| Created_at | 0 |
| User_id | 0 |
| lang | 0 |
| text | 0 |
| Verified_user | 0 |
| description | 17 |

| Column | % NA |
|---|---|
| Retweeted_user_id | 34 |
| Retweeted_user_screen_name | 34 |
| Retweeted_user_description | 36 |
| Retweeted_status | 34 |
| retweeted | 0 |
| Retweeted_followers_count | 34 |

| Column | % NA |
|---|---|
| User_screen_name | 0 |
| Quoted_status | 93 |
| place | 98 |
| coordinates | 99 |
| Followers_ count | 0 |
| category | 0 |

# TWEET CLEAN-UP AND FILTERING

- Select text containing words that are most related to education.

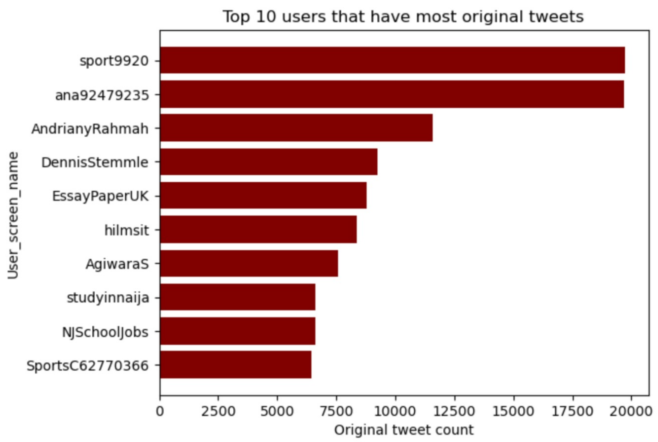
( teach, primary school, instruction, curriculum, learning, college, university, kindergarten, secondary school, high school, tuition, course, textbook, undergrad, instructor, primary education, secondary education,student,literacy, math)

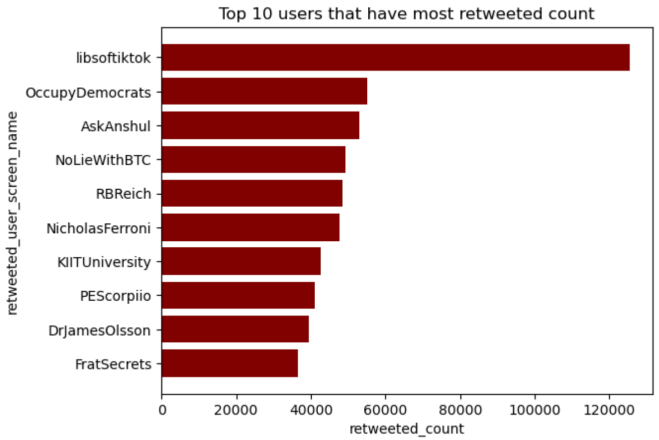
- Filter text language to English.
- For the text similarity analysis, filter verified users since they are more reliable.

# EDA AND EXTENSIVE USAGE OF AVAILABLE VARIABLES

- Use column retweeted status to obtain the retweeted user information.
- Coordinate column has more NA values than place, so use coordinates in place column to plot geographical distribution of twitters.
- Determine social media influencer by using column follower count (> 50000)  and keyword (influencer, tiktok, youtube, facebook, vlog, blog) involved in description and user screen name.
- Use keywords (news, gov, school, university, ngo) in description and user screen name to categorize government entities, school, university, news outlet, non-profit organization.
- Use column verified to determine whether user completes identity verification and filter government entities based on column verified (true).
- Use column retweeted to filter retweets and use retweeted status and quotus status to filter original tweets.

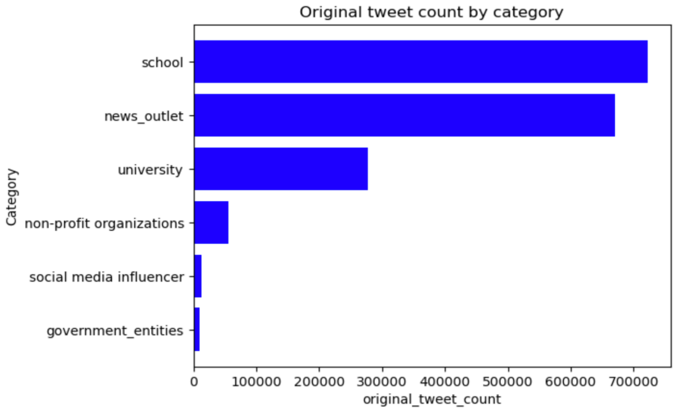
# AUTHOR IDENTIFICATION

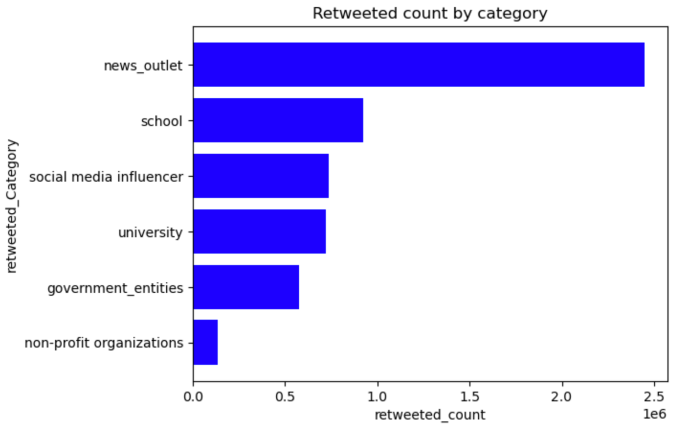




From user screen name:

- Sports related accounts have most original tweets
- Social media influencer have most retweeted count.





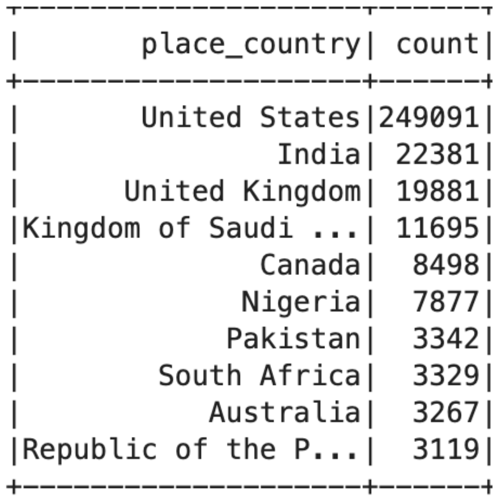
From type of organization:

- School related accounts generate most original tweets
- Tweets from news outlet are retweeted by most users.

Later two graphs ignore someone else for better visualization

# LOCATION ANALYSIS

- United States, India, Australia and european countries have a lot of twitter users who generate or retweet tweets related to education.



```
+--------------------+------+
|       place_country| count|
+--------------------+------+
|       United States|249091|
|               India| 22381|
|      United Kingdom| 19881|
|Kingdom of Saudi ...| 11695|
|              Canada|  8498|
|             Nigeria|  7877|
|            Pakistan|  3342|
|        South Africa|  3329|
|           Australia|  3267|
|Republic of the P...|  3119|
+--------------------+------+
```

Fig1: TOP 10 country with most tweet counts

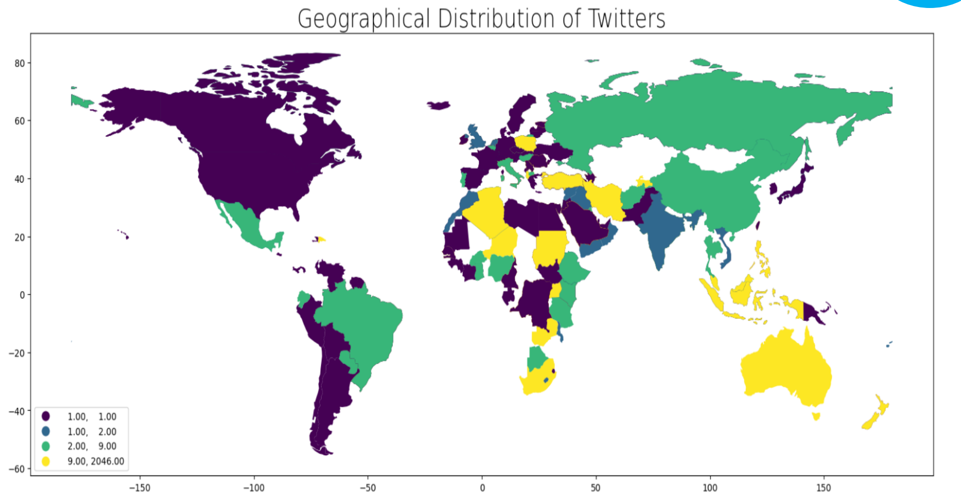


Fig2: Geographical distribution using non-duplciated country name

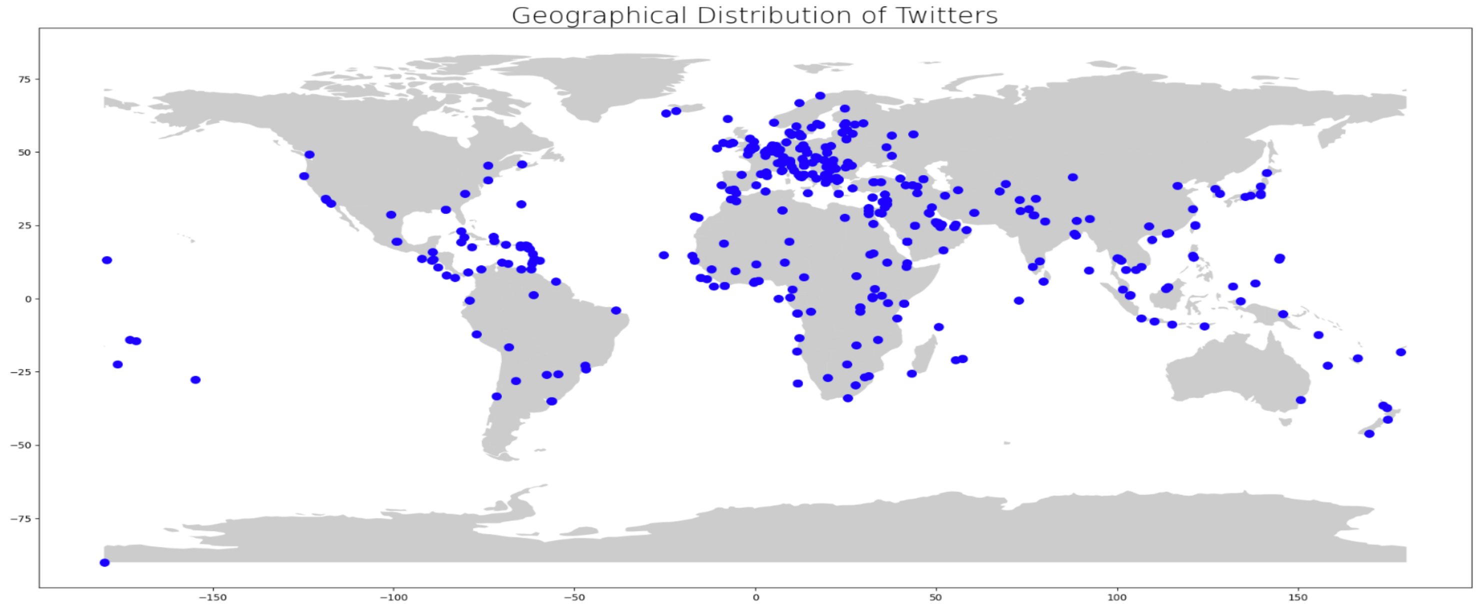
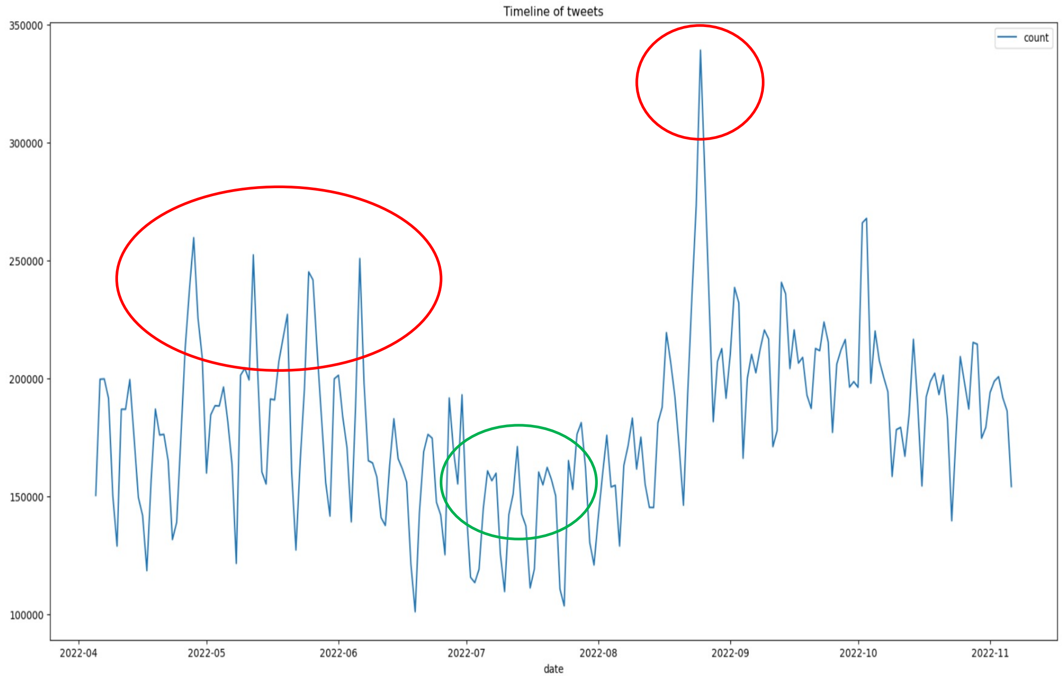
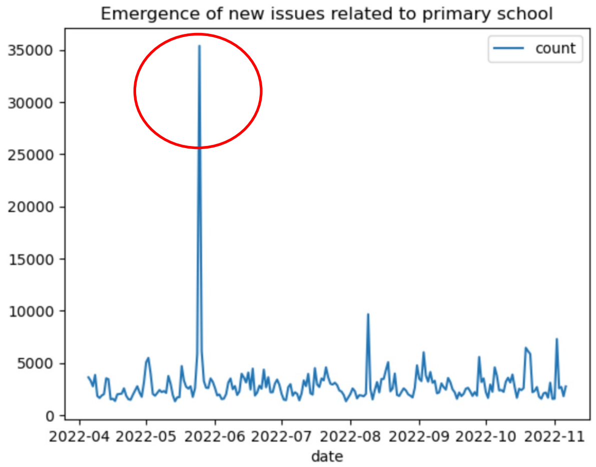


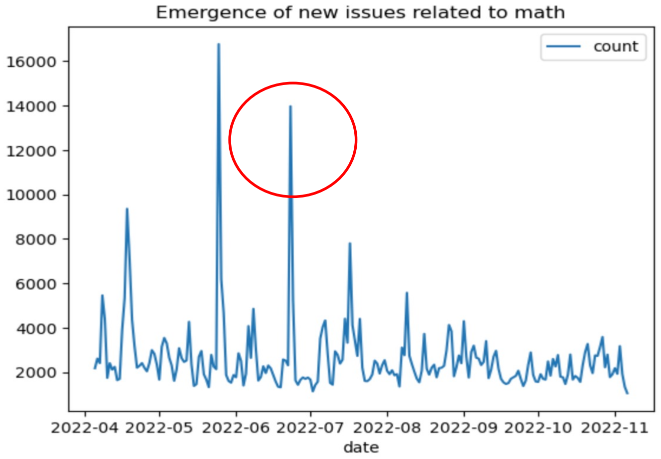
Fig3: Geographical distribution using non-duplicated longitude and latitude

# TIMELINE ANALYSIS



- There is a peak on 2022-05 and 2022-08, and 2022-07 may have some data collection gaps.





- The peak on 05-25 shows an issue related to primary school emerges, which is Uvalde school shooting.

- The peak around 06-23 shows an issue related to sex education, and the author uses math as an allusion.

# MESSAGE DUPLICATION ANALYSIS

| category<br>jaccard \ distance | Government entities | School | University | Non-profit organization | Social Media influencer | News outlet | Others |
|---|---|---|---|---|---|---|---|
| 0.3 | 8% | 3% | 2% | 8% | 4% | 5% | 2% |
| 0.5 | 9% | 5% | 4% | 10% | 6% | 7% | 3% |
| 0.7 | 15% | 15% | 7% | 14% | 13% | 13% | 9% |

- In the table, every cell represents the percentage of duplication for each category at different jaccard distance. It is obvious to see that most tweets are original content, rather than copies of the original tweets since the similarity is kind of low for each category. The possible reason is that we filter verified users in this part, and they generate more original tweets rather than copies of other tweets. Tweets from non-profit organizations and government entities have relatively higher duplications than other categories.

# RECOMMENDATIONS

- 1. If people want to look at important trends or topics in education, they can pay more attention to tweets from social media influencers, news outlets and schools.
- 2. Verified users have most original tweets, so people who look for real original education-related tweets, they should choose verified users. Twitter should encourage more users to complete identity verification to avoid the spread of fake news. Verified accounts only take up 1.7% right now, which is a quite low proportion.
- 3. Twitter should also try to collect more data about where twitters locate to better visualize the geographical distribution since most of them are null in our dataset.
- 4. Twitter dataset should update the user screen name for the same user id regularly since outdated user screen name will cause difficulty in analyzing related data.

# THANK YOU !