TWEETING ABOUT EDUCATION

CREDIBILITY AND TRENDS

Submitted by: Sakshi Shende

AGENDA

- Executive Summary
- Methodology and Source Overview
- Tweets Clean-Up & Feature Selection
- Exploratory Data Analysis
- Author Identification
- Location Analysis
- Timeline Analysis
- Tweets Uniqueness Analysis
- Conclusions
- Actionable Recommendations

EXECUTIVE SUMMARY

twitter is one of the most popular social media platforms that enable real-time communication among **~100M users** worldwide and **~500M tweets** sent daily.

Twitter has become an important source of real-time news and information related to education, news, sports, health, etc.

The project aims to identify whether Twitter is a credible source of information for important trends and topics in education by answering the following questions:

- Who are the most influential Twitterers?
- What type of Twitter users contribute to the platform's education-related discussion?
- Are there any spikes in activity and shifts in geographical distribution to Twitterers?
- Are the tweets original content or just copies of existing tweets and retweets?

METHODOLOGY AND SOURCE OVERVIEW

Methodology

Google Cloud Platform: Cloud computing platform to store Twitter data

Google Dataproc: cloud service for big data processing and storage

PySpark: used for large scale data processing

Pandas: used for data cleaning, manipulation and data analysis

Matplotlib: used for data visualization

SimHashLSH: used to analyze tweet similarity

Source Data Overview

~100 M tweets, ~500GB of data

Tweets are spread across **50K** JSON files in a nested format.

Tweets are collected on topics of education.

Contains Tweet objects, User objects, Geo objects, Entities objects, and Extended entities objects

Data from **April 2022** to **February 2023** is present.

TWEETS CLEAN-UP & FEATURE SELECTION

The final dataset used for analysis has 58 million records and 17 features.

- Only English-language tweets are taken into account for this project.
- Tweets containing the following keywords were kept and the ones that are unrelated to education were eliminated:

kindergarten k-12 textbook marksheet preschool tuition academic classwork

student curriculum syllabus degree assessment learning knowledge education

- Features that weren't needed for the analysis were discarded.
- By examining their missing values, we eliminated the sparsely populated features. Even though there were more than 90% missing values, the features we kept for analysis are shown in the table below. Extreme data loss would result if these features were dropped.

Feature Name	User Description	User Country	User_Location_Coords	Retweet_Count	Retweet_status	quoted_status
Missing Value %age	16.9%	99.2%	99.2%	37.38%	37.38%	90.27%

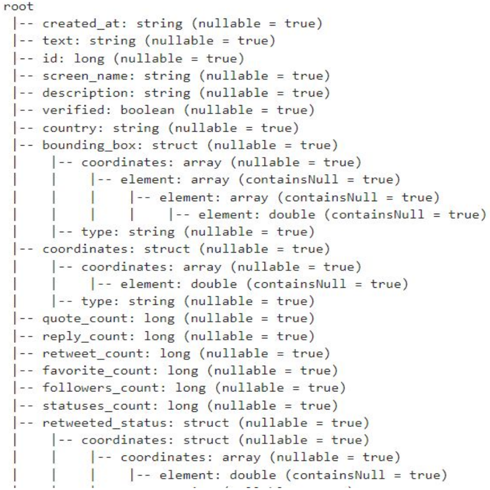
EXPLORATORY DATA ANALYSIS

- **17 features** and **58 million records** make up the dataset for analysis.

- Since the data is in nested JSON format, only the relevant features were chosen to be kept in the dataset for analysis.

- After data cleaning, filtering and maintaining only the relevant features, The dataset still contains null values in several of the columns because eliminating them would result in data loss.

- Data from **April 2022 to February 2023** is currently available.



```
root
 |-- created_at: string (nullable = true)
 |-- text: string (nullable = true)
 |-- id: long (nullable = true)
 |-- screen_name: string (nullable = true)
 |-- description: string (nullable = true)
 |-- verified: boolean (nullable = true)
 |-- country: string (nullable = true)
 |-- bounding_box: struct (nullable = true)
 |    |-- coordinates: array (nullable = true)
 |    |    |-- element: array (containsNull = true)
 |    |    |    |-- element: array (containsNull = true)
 |    |    |    |    |-- element: double (containsNull = true)
 |    |-- type: string (nullable = true)
 |-- coordinates: struct (nullable = true)
 |    |-- coordinates: array (nullable = true)
 |    |    |-- element: double (containsNull = true)
 |    |-- type: string (nullable = true)
 |-- quote_count: long (nullable = true)
 |-- reply_count: long (nullable = true)
 |-- retweet_count: long (nullable = true)
 |-- favorite_count: long (nullable = true)
 |-- followers_count: long (nullable = true)
 |-- statuses_count: long (nullable = true)
 |-- retweeted_status: struct (nullable = true)
 |    |-- coordinates: struct (nullable = true)
 |    |    |-- coordinates: array (nullable = true)
 |    |    |    |-- element: double (containsNull = true)
```

- Describe the numerical features:

Summary	Quote_count	Reply_count	Retweet_count	Favorite_count	Followers_count	Tweets_count
Count	58482831	58482831	36618068	58482831	58482831	58482831
mean	0.0	0.0	2,777.66	0.0	6,797.45	42,855.67
stddev	0.0	0.0	9,012.63	0.0	245,088.00	108,955.45
Min	0	0	0	0	-1	-1
max	0	0	516855	0	132031077	132031077

AUTHOR IDENTIFICATION

Twitter is used by diverse range of individuals and organizations.

Top 7 Twitterers by Message Tweets

user_name	user_type	total_tweets
Sportsthread	Education Organization	3,467 million
KenyaPower_Care	News Outlet	2,457 million
TOICitiesNews	News Outlet	1,009 million
htTweets	News Outlet	711 million
IndianExpress	News Outlet	581 million
Independent	News Outlet	534 million
Thehill	Government Entity	529 million

These users are primarily news outlets, with exception such as sportsthread, which is an education organization. These users have significant number of tweets related to education, suggesting that they may be **valuable sources of information**.

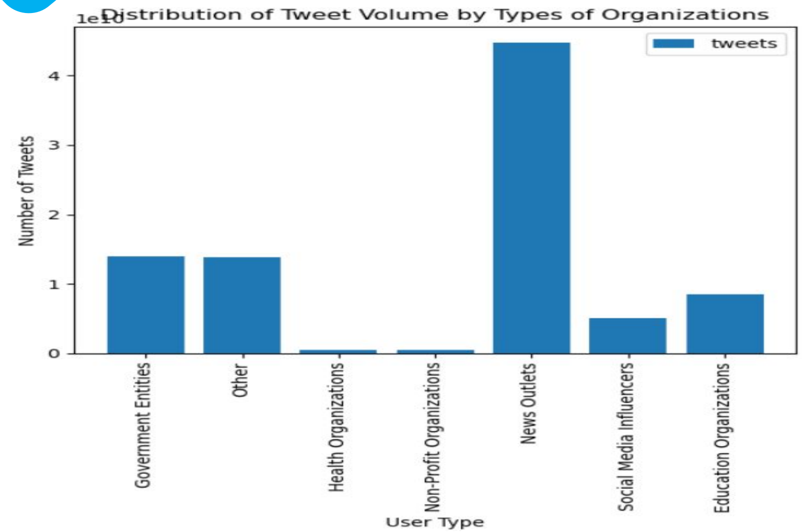
Top 7 Twitterers by Message Retweets

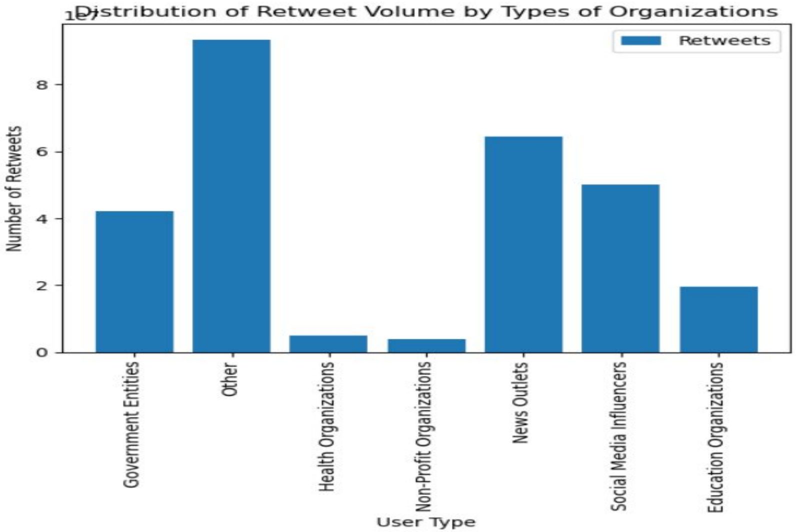
user_name	user_type	total_retweets
ValaAfshar	Social Media Influencer	1.06 million
JonesHospodTX	SociaGovernment Entity	0.48 million
*ReignOfApril	Other	0.42 million
LeratoMannya	News Outlets	0.41 million
* AnsisEgle	Other	0.402 million
Bruno_J_Navarro	News Outlets	0.39 million
se4realhinton	Social Media Influencer	0.38 million

These users have significant number of retweets related to education, indicating that their content is **influential** and widely shared in education community on Twitter.

*Note: Due to a lack of keywords in their descriptions, several users were incorrectly classified as Other. They could be news outlet and social media celebs

AUTHOR IDENTIFICATION





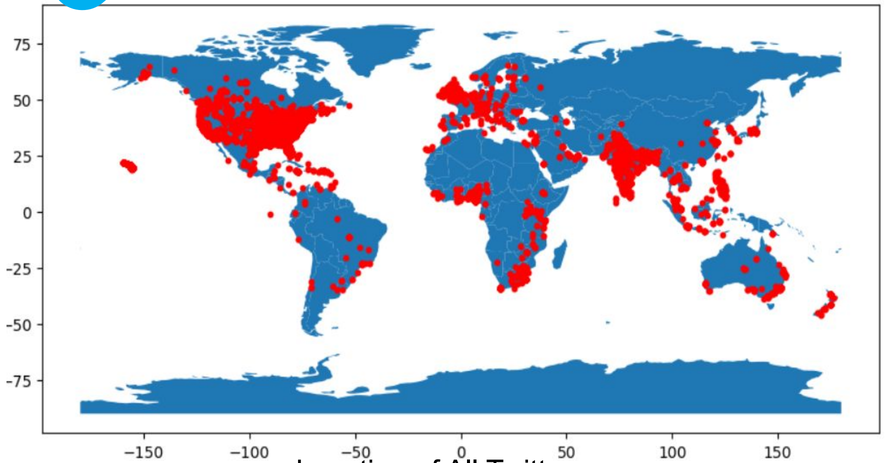
After news outlets and government entities, education groups accounted for 8.5 billion of all tweets. This demonstrates how aggressively Twitter is being used by organizations in the education sector to provide information and interact with their audience.

The least amount of tweets, however, are from health organizations. It's possible that they prioritize healthcare-related issues above educational ones because of this.

Random Twitter users had the highest number of retweets, followed by news outlets and social media influencers in the context of education-related tweets.

It is important to note that this outcome may have occurred because our dataset's user description is not properly populated.

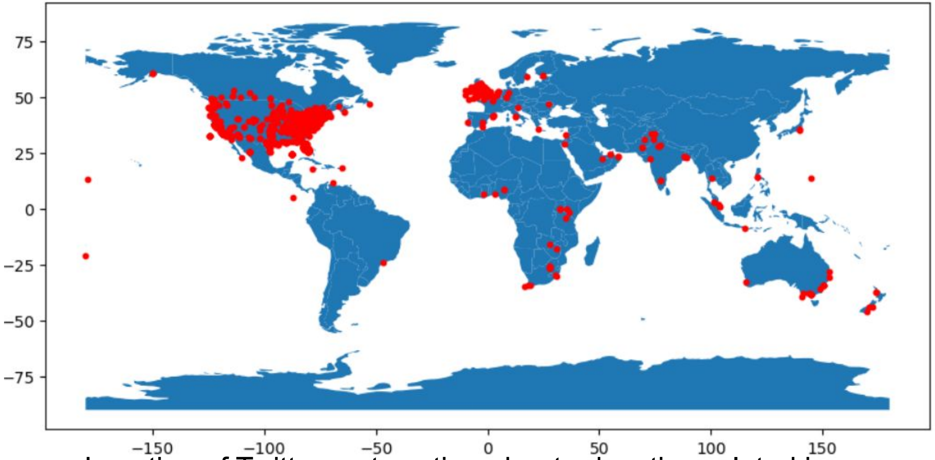
LOCATION ANALYSIS



Location of All Twitterers

Identified the most number of twitterers in the United States of America, followed by the United Kingdom and India.

Since we only used English keywords to clean our data and the majority of people in these countries speak and understand English, this is probably why there are so many tweets on education in these countries.

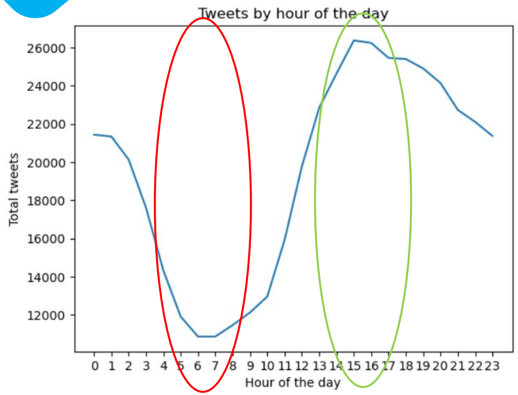


Location of Twitterers tweeting about education-related issues

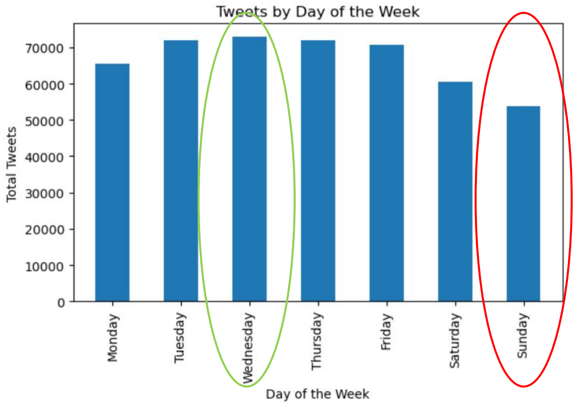
Identified no relationship between the emergence of new issues in education and the progression and locations of these Twitterers.

It is likely because issues like bullying and school safety are complex that involve kids, teachers, and parents. The emergence of new issues may be driven by a variety of factors, including political and social factors, which can make it difficult to establish a clear relationship.

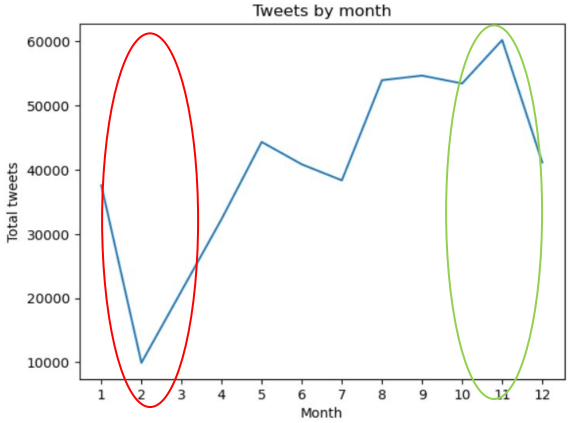
TIMELINE ANALYSIS



- **Peak activity hours:**13:00 pm till 17:00 pm. This time is generally considered a lunchtime for many people and after school hours for students. Hence, more people could check their accounts on social media.
- **Off-peak hours**: 5:00 am to 7:00 am. During this time, a lot of people probably sleep. So, fewer users may be online, which would reduce the number of tweets.



- **Most popular day**: Wednesday. It could be due to the fact that many companies and social media influencers schedule their content to go live. Also, Wednesdays are often associated with popular hashtags, which could encourage users to tweet more.
- **Least popular day**: Sunday is typically a day off for many school, business, and government organizations. Thus, there might be fewer tweets on this day.



- **Spike observed** from October to November. It could be due to the fact that many schools and universities begin their academic year. Also, these are the months when college applications are due in many countries.
- **Valley observed** from January to February. This is the time of winter break and the beginning of the year, so people may be focused on New Year's resolutions or other personal goals.

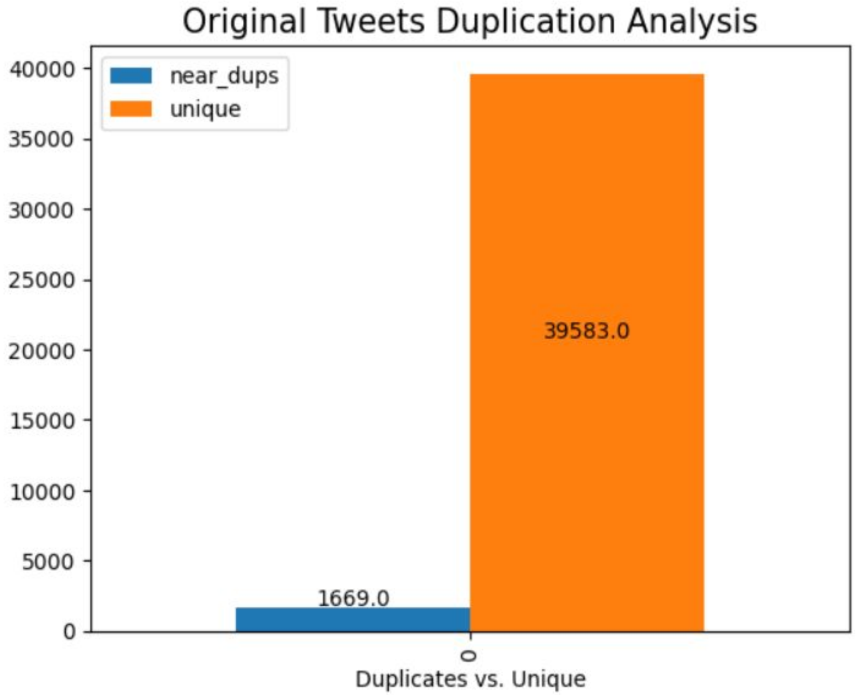
We calculated the difference between the consecutive days and observed **no data collection gaps**.

TWEET UNIQUENESS ANALYSIS

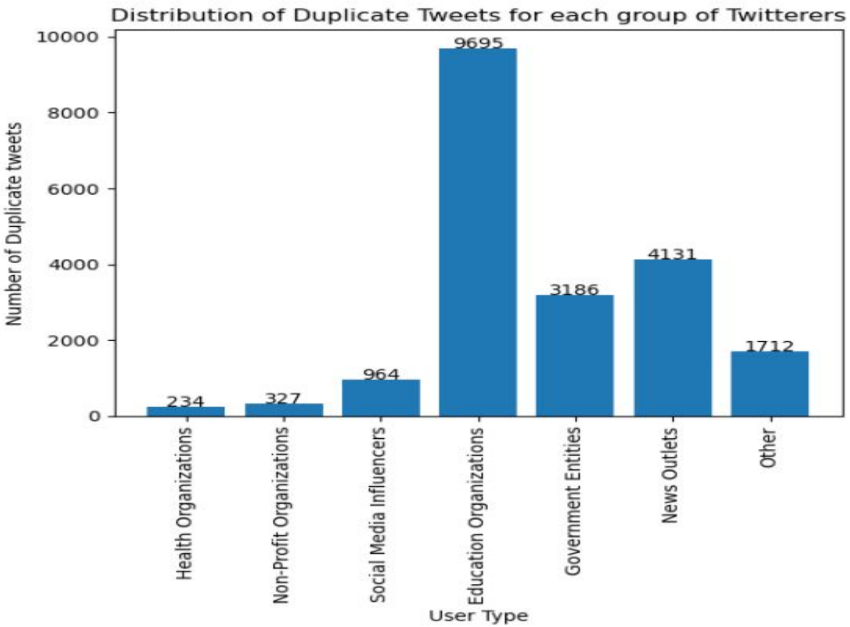
A random sample of **41K records** was taken to test the uniqueness of original tweets.

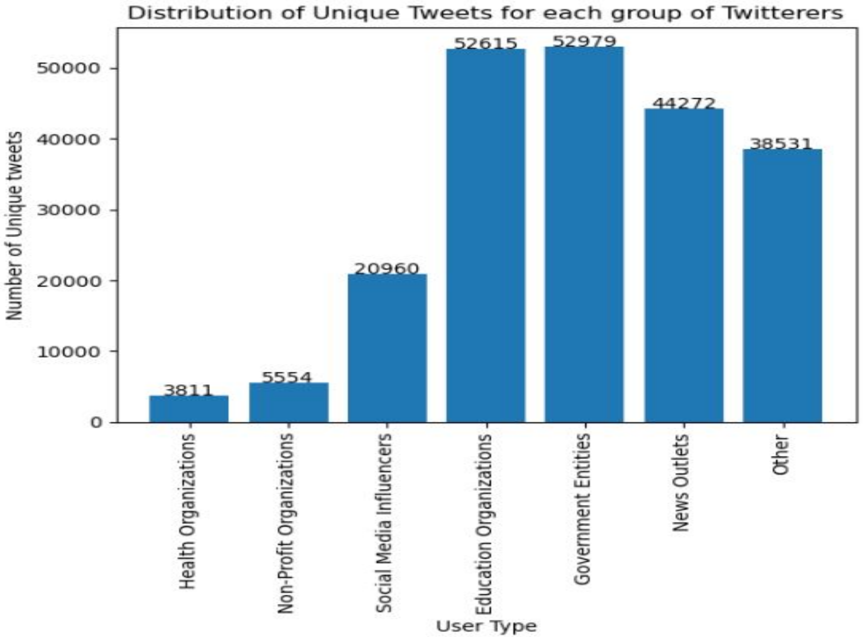
More than 91% of tweets that are original are unique.

- **41K records** were randomly selected as a sample from the entire dataset. So, this analysis does not include all of the tweets in the dataset.
- A **Jaccard distance threshold of 0.3** is chosen for tweets because they are short in nature and they tend to have higher degree of variability in terms of word choice and syntax.
- 91% of tweets are found to be original while only 8% of tweets are found to be duplicates.
- It implies that the majority of Twitter users are creating original content rather than just retweeting or copying other people's tweets.



TWEET UNIQUENESS ANALYSIS





Education Organizations tweeted the most near duplicate tweets and unique tweets (after government entities) related to education. It could be that they may retweet the same message as a way to amplify a particular message or campaign that could increase the chances of their message being seen by a wider audience.

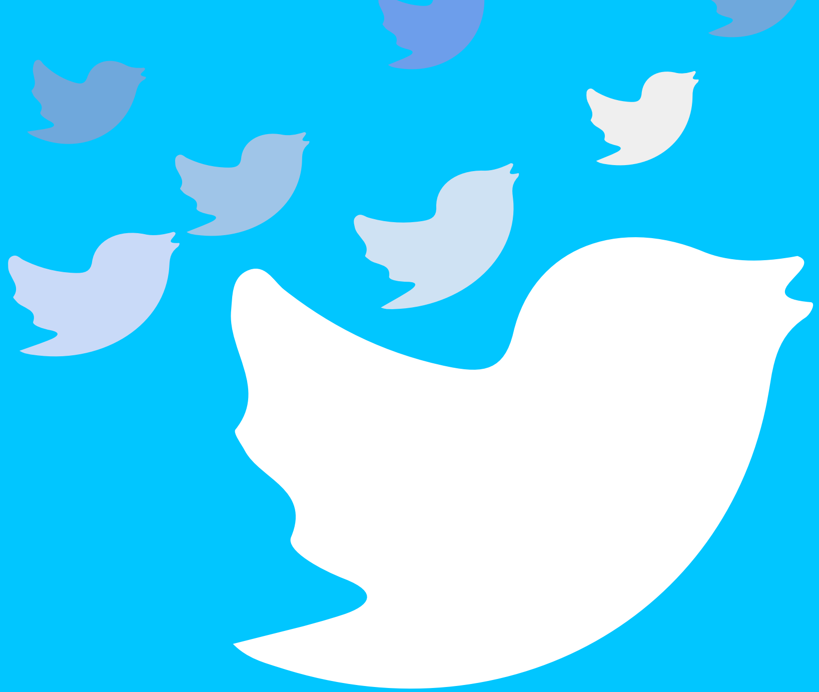
The health organizations tweeted the least number of near duplicate tweets as well as unique tweets.

CONCLUSION

- We are left with about 58 million tweets after eliminating irrelevant ones, of which just 853K are from verified accounts.
- The majority of tweets come from news organizations, which makes them a reliable source of information.
- The majority of twitter users are based in the US. It can be as a result of our decision to filter the data to only include tweets in English.
- No data collecting gaps were observed when we looked at the timeline of the the tweets.
- Weekly, Thursdays saw the highest volume of tweets. It could be because many businesses and social media influencers plan to publish their material.
- Only 8% of tweets sent on Twitter are duplicates; the majority (91%) are original tweets. The most duplicate tweets were created by groups working in the field of education to spread a certain message.
- Due to the fact that the study is conducted on randomly selected data, the result may differ if the same analysis is carried out on the complete dataset.

ACTIONABLE RECOMMENDATIONS

- Through cleaning process should be conducted to eliminate irrelevant tweets and ensure accurate analysis of education-related tweets.
- Sentiment analysis could help understand the overall mood of the public regarding specific education-related topics and identify emerging trends.
- K-Means clustering could be used to select clusters of tweets for analysis, rather than relying on word-based filtering.
- Develop a better strategy to measure the relevance and popularity of education-related tweets beyond just retweet/message volume.
- To learn more about how education-related topics are viewed and discussed locally, carry out a more detailed regional analysis.
- Use the results of analysis to identify key education-relation trends and develop targeted strategies for engaging with audience.



THANK YOU!!