Uncovering Education on Twitter -

A data-driven analysis on educational tweets

Presented by Sirivanth Paladugu

Table of Contents

Executive Summary with meaningful insights

- Twitter is a social media platform where users can send and receive short messages known as "tweets," which can be used for socializing, news gathering, and marketing purposes.
- The platform is also been used to spread false or misleading information, conspiracy theories, and propaganda, often through the use of bots and fake accounts.
- The objective is to analyze credibility of Twitter as a source of information and profile its users to understand their location and behavior, including whether they copy or create unique tweets. By conducting various analyses on education-related tweets, I aim to provide valuable insights into the opinions of Twitter users.
- The analysis I followed is as follows:
  - EDA
  - Author Identification
  - Location and Timeline Analysis
  - Message Uniqueness Analysis

Methodology

- For coding, I have used Pyspark in JupyterLab
- EDA involves using pandas to create a data frame, which helps organize and manipulate data, and visualizations to identify patterns and relationships in the dataset. Used seaborn and matplotlib for visualizations.
- To identify the active users, original and retweeted tweets are divided. This helps to identify patterns and improve the accuracy of author identification.
- To identify behavioral patterns and trends among users, the location is determined using geopandas and frequency of their tweets is analysed by plotting time series plots.
- To identify the similarity in tweets, I used MinHash LSH.
- For faster and more efficient analysis, intermediate data is stored in parquet files to improve data management.

Source Data Overview

- Format - **Nested JSON**
- Total size of the data - **498.7 GB**
- Number of tweets before filtering - **~100 million**
- Number of features - **40 Columns**
- The data is stored in **Google Cloud**
- Useful links/References:
  - https://www.merriam-webster.com/dictionary/twitterer
  - https://en.wikipedia.org/wiki/K%E2%80%9312
  - https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/object-model/tweet

Tweet clean-up and filtering

- Initially, I implemented a sequence of filters and transformations to eliminate irrelevant data, including tweets lacking user location or follower count, as well as those not in English. (Count is **46,698,507**)
- 50 education-related keywords were selected from the initial filtered Twitter data based on their relevance to the topic, with the aim of gaining insights into users' opinions and behaviors regarding education.
- Additionally, **explicit words** that were deemed irrelevant to the analysis were identified and removed from the dataset. This step ensures that the analysis is focused on the relevant information.
- The resultant tally of tweets following the process of filtration is **24,074,764**.

EDA - Data Sampling and Variable Selection

- The dataset contains multiple null values, hence I dropped the redundant columns with highest count of nulls which do not serve any importance for the analysis.
- I performed EDA on a sample size of 1,034,879 and kept columns with null values, like **coordinates**, even though they had over 95% nulls. I did this while keeping in mind that some users might not want to share their location for a variety of reasons.
- Some important columns I selected are **tweet_text, in_reply_to_status_id, retweeted_status.user.screen_name, user.screen_name, user.location, user.verified, user.followers_count** and **user.statuses_count.**

EDA - Analysis on Selected Variables

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1034879 entries, 0 to 1034878
Data columns (total 15 columns):
 #   Column          Non-Null Count    Dtype
---  ------          --------------    -----
 0   created_at      1034879 non-null  object
 1   tweet_text      1034879 non-null  object
 2   replied         147488 non-null   float64
 3   coordinates     1617 non-null     object
 4   location        9752 non-null     object
 5   is_quote        1034879 non-null  bool
 6   is_retweeted    1034879 non-null  object
 7   Retweeted_id    659167 non-null   float64
 8   Retweeted_user  659167 non-null   object
 9   Profile         1034879 non-null  object
 10  user_location   635672 non-null   object
 11  user_verified   1034879 non-null  bool
 12  followers       1034879 non-null  int64
 13  total_tweets    1034879 non-null  int64
 14  hashtags        1034879 non-null  object
dtypes: bool(2), float64(2), int64(2), object(9)
memory usage: 104.6+ MB
```

Fig - 1

I have selected 14 columns which would better suit for my analysis.

- Retweeted_id, Retweeted_user and replied contains nulls because the nulls are the original tweets.

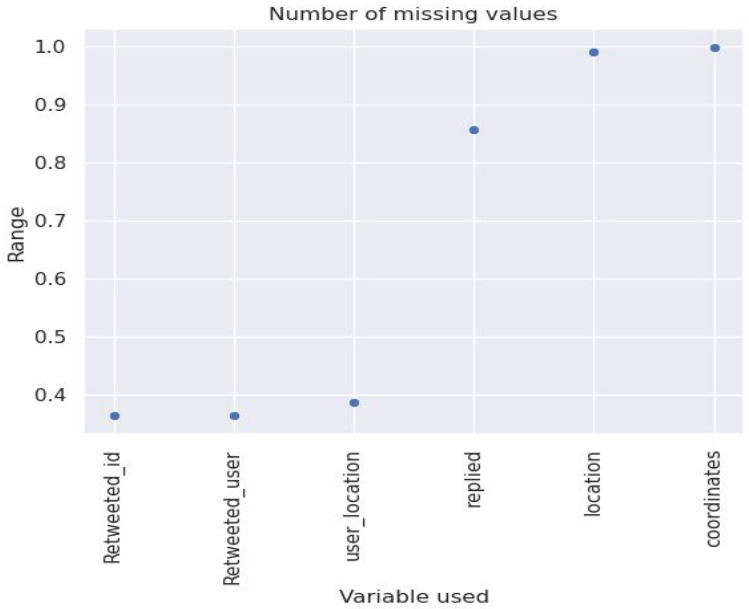


Fig - 2

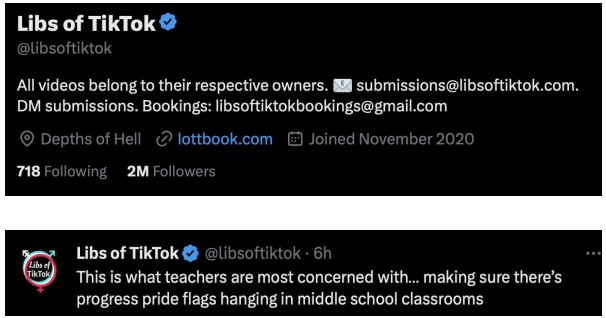
I have plotted the nulls for the selected columns from range of 0 to 1.

Author Identification - Prolific/Influential Twitterers

	Profile	total_content
0	la_patilla	6445536
1	laxradar	4572875
2	tropicalisimany	4362559
3	dogandwinelover	4286192
4	PulpNews	4187042

By analyzing the data, we can note that la_patilla has the highest number of content overall, while EssayPaperUK has the highest count of original content.

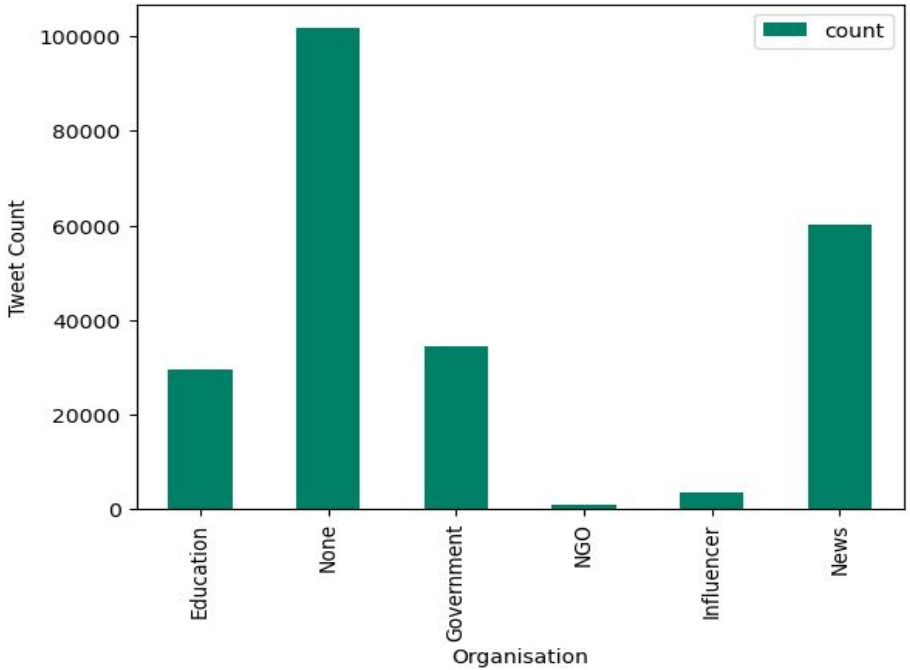
	user	original_content
0	EssayPaperUK	6264
1	LoveLiveFrames	5774
2	indiaedudiary	4804
3	AndrianyRahmah	4541
4	group_kq	4541



	user_name	avg_retweet
0	libsoftiktok	126443
1	NasimiShabnam	55607
2	OccupyDemocrats	44228
3	NoLieWithBTC	37775
4	StephenKing	36392
5	AskAnshul	33412

The table displays the frequency at which the tweet has been retweeted, with libsoftiktok appearing at the top of the list.

Author Identification - Influential Twitterers



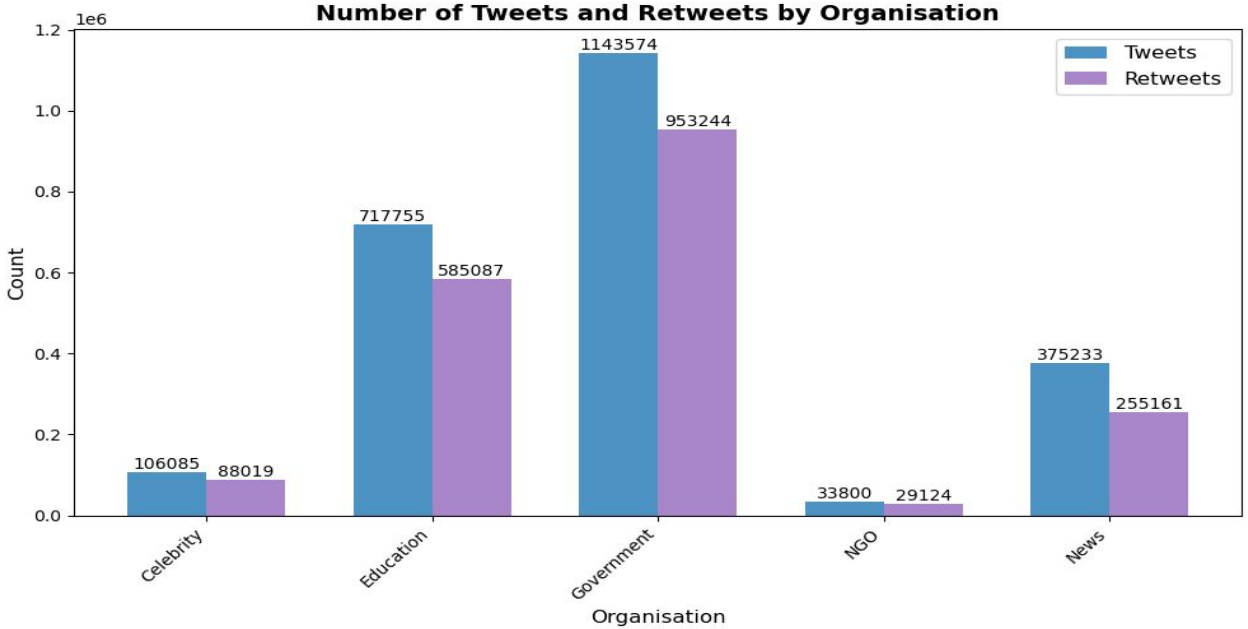
To group each user, a set of keywords relevant to a specific organization were selected.

There are five categories into which the users are grouped: Education, Government, NGO, Influencer, and News.

The "None" entity includes other user accounts that do not fit into any of the five categorized organizations mentioned.

- When excluding the "None" entity, the **news organization has the most influential Twitterers**, while the NGO has the least influential Twitterers out of all the categories.

Author Identification - Tweet/Retweet distribution by organisation

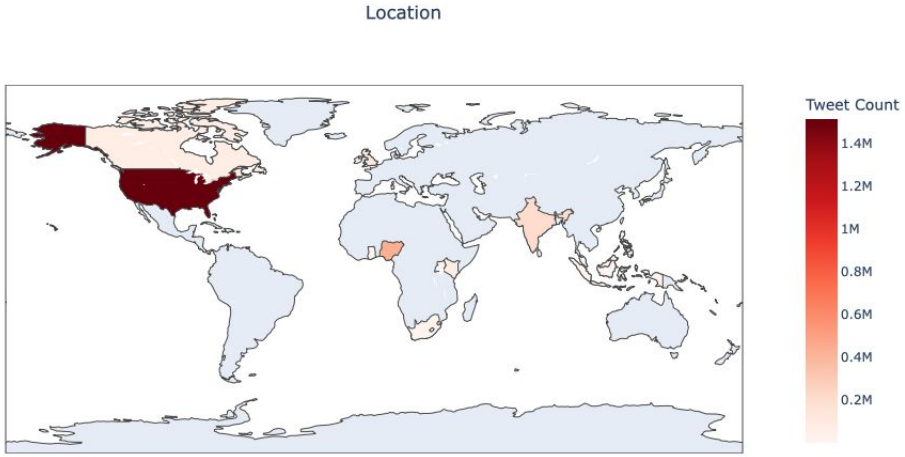


I have categorized all other individuals who have a significant following and influence in various industries as celebrities.

I am here considering 5 entities to see the tweet/retweet distribution.

- Despite the fact that NGOs and celebrities tweet less frequently about educational topics, their tweet-to-retweet ratios are similar.
- However, it is surprising to note that education and news entities do not have the expected distribution of retweets.
- Amongst all the organizations, the government entity has the highest count for both tweets and retweets.

Location Analysis - Where are these twitterers located?



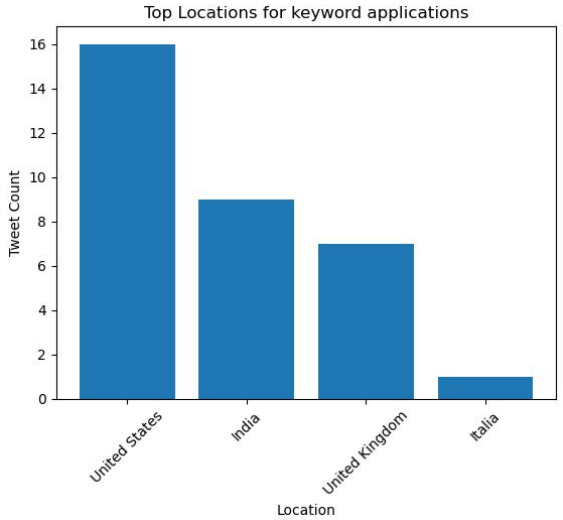
Graph - 1

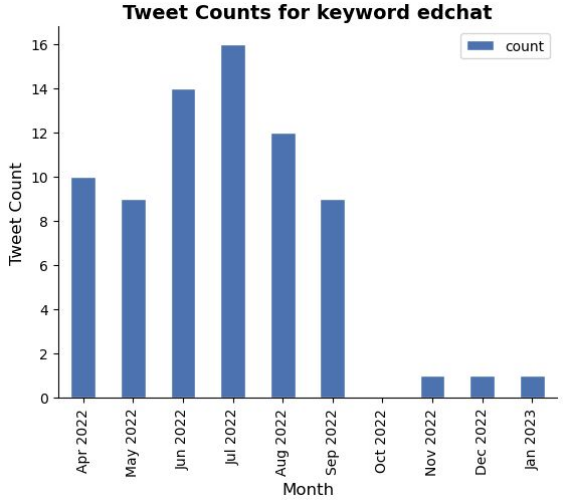
	country	count(1)
0	United States	1514122
1	Nigeria	431609
2	India	206892
3	United Kingdom	109658
4	Canada	66945

Fig - 1

The graph has been plotted taking into consideration the location of the users, and upon analysis, it is discernible that the United States has recorded the highest number of tweets, with Nigeria and India following suit in second and third place, respectively. This data may serve as a valuable insight into the online activities of individuals across the globe, highlighting the prevalence of twitter usage in different regions.

Location Analysis - Emergence of important trends



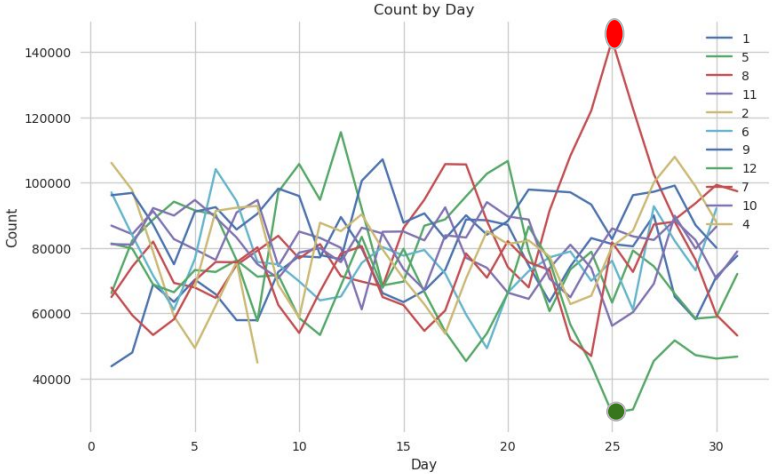


The keyword "**edchat**" was used to analyze the trend and users tweeting about educational topics. The United States had most tweets. This keyword is trended in July, June, and August when students have summer break, thus share their knowledge on educational topics.

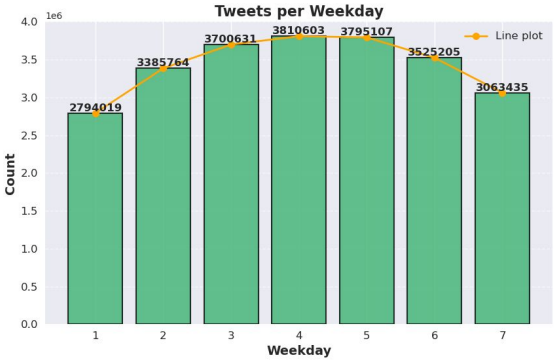
Benjamin K. Woodcock (He/Him) @bnjmn_k_wdcck · Jun 21, 2022
A 🧵: Looking into the education stances for each of the @MIGOP gubernatorial candidates as of today, **June** 21, **2022**. #miched **#edchat**
1 1

Mrs. Monty's Class @MrsMonty5thGrd · Jun 7, 2022
Just registered for the @Kahoot EDU Summit **2022** happen next week - **June** 14 and 15! Looks like two days filled with some incredible PD! Interested in joining @imkoulbanis, @BarbaraGruener, @AnneLeBlanc2? #edtech **#edchat** #Kahoot
4 5

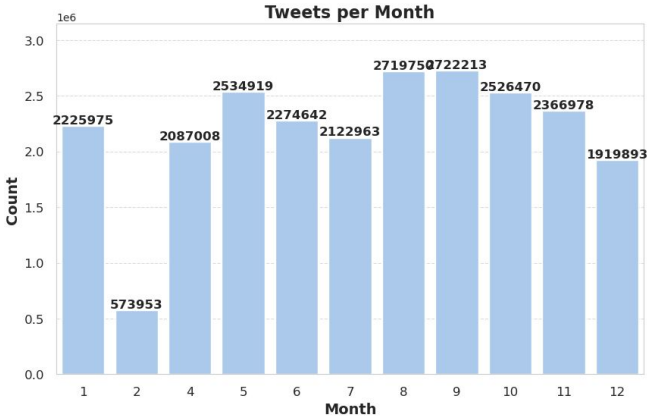
Timeline Analysis



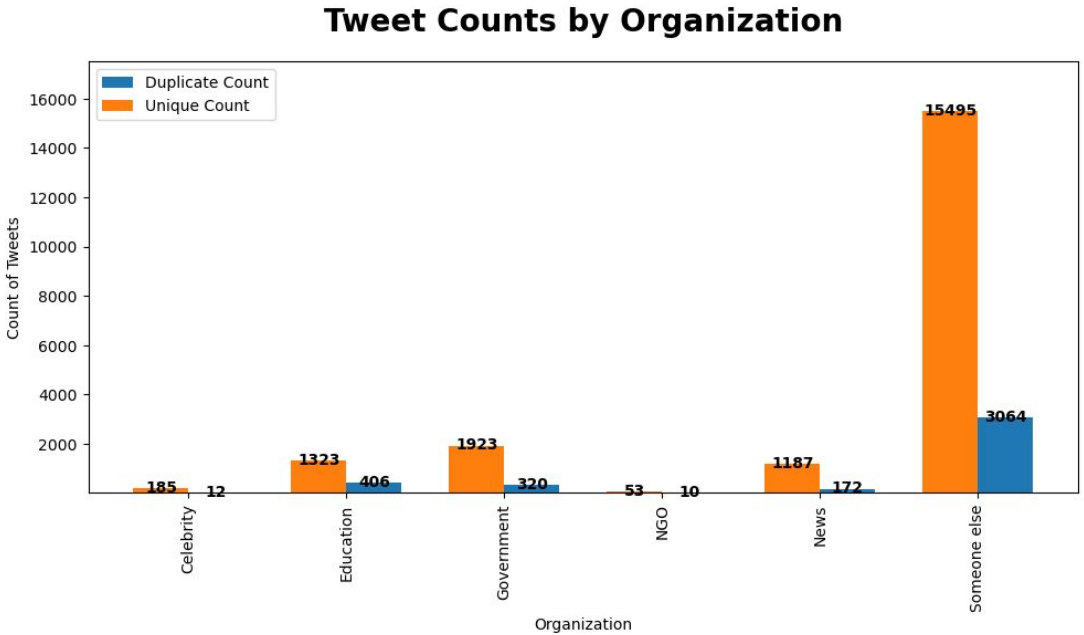
Based on the plotted graph, it can be observed that the day with the **highest** number of tweets was on **August 25th, 2022**, while the day with the **lowest** number of tweets was on **December 25th, 2022**. However, it is important to note that there are missing data points, as the available data only covers the period from **April 5th, 2022 to February 8th, 2023**, and does not include any data for the month of **March**.



Thursday and **Friday** recorded the highest number of tweets, while **September** and **August** were the months with the highest tweet counts. This is because many educational institutions commence their fall semesters during this period.



Message Uniqueness Analysis



A Jaccard distance threshold of 0.5 is used to measure the level of similarity between sets. Out of a sample of 20,166 records, only 1005 of them are duplicates, accounting for roughly 5% of the total records. We can say that the **tweets are mostly unique**.

- It can be observed that there are significantly fewer duplicate vs unique count for the categories of Celebrity and NGO, while the category labeled as "Education" has the highest count of unique vs duplicate.

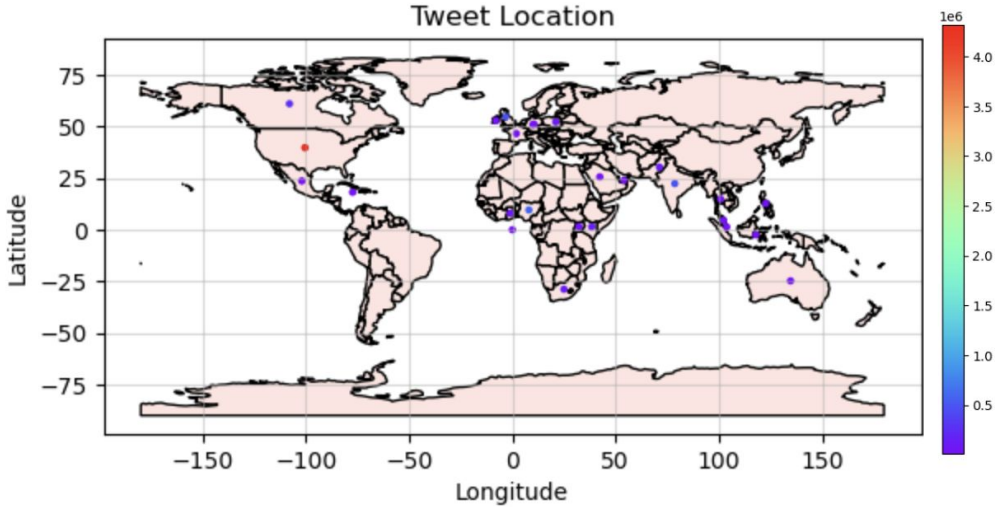
Conclusions

- Based on my analysis, I have come to the conclusion that Twitter can be a **Credible source of information**. However, it is advisable to verify the trends on popular news articles as there are some non-verified users on Twitter who may mislead the trends and provide incorrect information.
- By applying filtering criteria based on requirements, I was able to significantly reduce the amount of data from **100 million** tweets to **24 million** tweets.
- Compared to other organizations, the **news organization** has a higher number of influential Twitterers.
- **United States** has a prominent position in the emergence of important **trends** related to education, and many educational tweets come from this region.
- The reason for the high tweet count during the fall season is that many educational institutions begin their academic year in **August or September**. Additionally, **Thursdays** have a higher tweet count compared to other days.
- Out of the sample of 20,166 records, only a small percentage of 5% are duplicates, which amounts to 1005 records. Therefore, the majority of the tweets are **unique**.

Actionable Recommendations

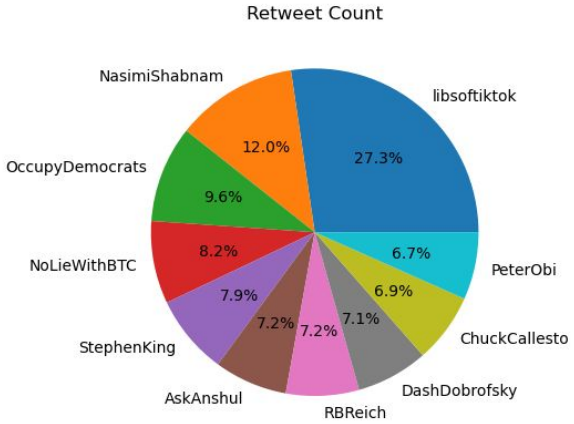
- Blocking **non-verified** users with no follower count and restricting the **use of bots** on Twitter is a suggestion that could potentially lead to more valuable and accurate information on the platform.
- To improve the accuracy of location analysis, it would be beneficial to have a good amount of information regarding coordinate data.
- Though the emerging issues on the twitter often receive significant attention, it is recommended to know how much of what you read there is **actually true**.
- To maintain the **credibility** of the data, it is important to **verify the reliability** of the source where data is gathered.
- As the part of project's future scope, doing **sentiment analysis** on the tweets could assist identify if they reflect opposition or support towards the emerging issues in the educational sector.

Appendix

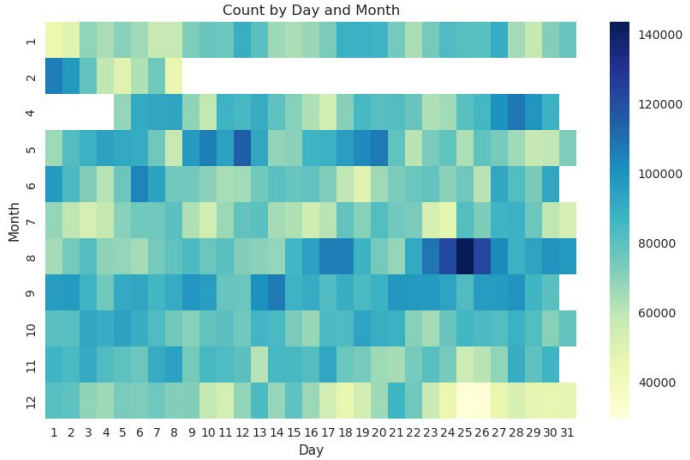


Geomap is plotted based on the limited coordinates provided.

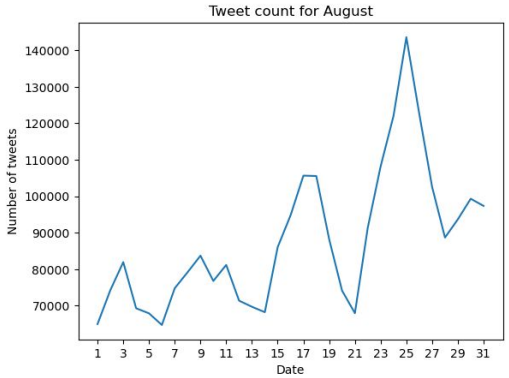
For retweet count I saw how often the tweet is getting retweeted in percentage.



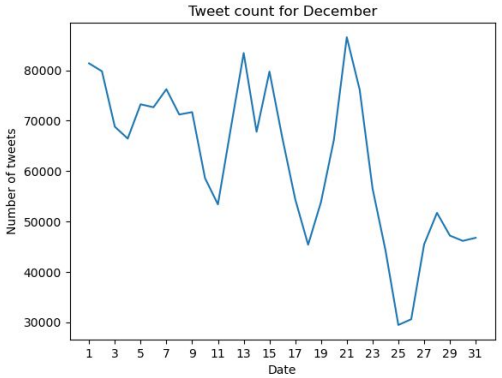
Appendix



The heatmap is plotted to see if there is any data gap in the given timeline. There is no data from 9th february to month end. Complete data for march month is missing.



Highest tweet count on 25th August.



Lowest tweet count on 25th December.