Analyzing Educational Tweets with PySpark: A Professional Approach

By Kishor Kumar Reddy Mannur

MSCA 31013 IP02 Big Data Platforms

Agenda

1. Executive Summary with data-driven insights
2. Methodology
3. Source Data Overview
4. Tweet clean-up and filtering & Perform EDA (Exploratory Data Analysis)
5. Author Identification (Prolific/ Influential)
6. Perform Location Analysis (Geographical Significance)
7. Plot Timeline Analysis (Trends in **Tweeting**)
8. Message uniqueness analysis (Credibility Analysis)
9. Conclusion (with Actionable Recommendations and Suggestions)
10. Appendix

Executive Summary

Twitter is a widely used social media platform that enables users to express their thoughts, views, news, and other information through brief text messages known as 'tweets.' The site has become an **essential source of information** for many individuals and organizations worldwide. Nevertheless, owing to the simplicity of accessibility, the data on Twitter may be disorderly and occasionally unreliable.

Some of the key questions that this text analysis project on Twitter can help to answer include:

Who are the most prolific/ influential Twitterers?

What are the most popular topics being discussed?

Where are the Twitter users located?

When are the tweets most likely to be posted? Any significant peaks and valleys?

And, finally, how unique are the messages being shared, and are Twitter users copying and pasting each other's tweets?

Methodology

- Using PySpark, I read and analyzed the tweet data from the GCP bucket.
- SparkSQL, Pandas, and other libraries for plotting and analyzing small chunks of data.
- Saving Parquet files as checkpoints after each stage of analysis to avoid blockages.
- To create geographical, bar, and line plots, I used Seaborn and Matplotlib.
- I utilized MinHash, LSH, and Jaccard similarity to check for uniqueness and perform credibility analysis.

Source Data Overview

- Data Source:

  - Google Cloud Storage Bucket (gs://msca-bdp-tweets/final_project)
  - Total Number of Records/ Tweets (**~ 100 Million =** 99,994,342**)**
  - Total Features/ Columns (**39**)

```
#Checking for the total number of records
total_recs = tweets_df.count()
total_recs
```

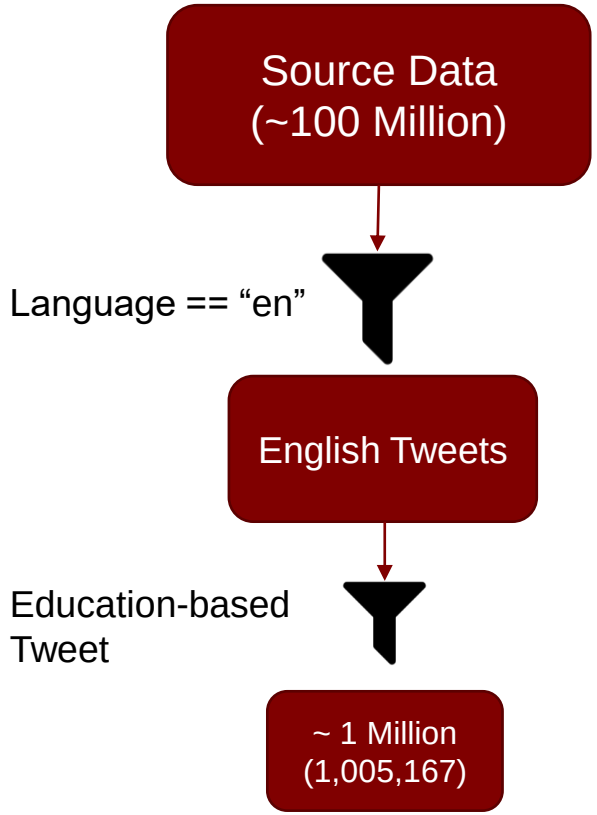
```
99994342
```

```
#Checking how many columns are present to proceed with EDA
total_features = len(tweets_df.columns)
total_features
```

```
39
```

Data Filtration and Cleaning

- Starting with data clean-up, I used some top education-related keywords to get relevant tweets.
- Also, following a reverse approach, I got rid of the irrelevant tags by excluding them from the dataset.
- Using language filter, took English language tweets only for analysis.
- Upon filtration, I arrived at relevant records of around ~1 Million tweets



Exploratory Data Analysis

- Based on the initial analyses, I observed there are many fields as per Tweet API that are empty and some basically were mostly **null**.

- After filtering the data based on the need, I selected the columns that are required for my analyses and discarded the rest.

	created_at	id	geo_coordinates	user_name	followers_count	verified_user	user_location	user_description	reply_count	retweet_count	retweeted_status	tweet_text	text
0	0	0	997584	0	0	0	451915	231810	352608	352608	352608	0	0

- The columns/ fields of interest are **user name**, **date of tweet creation**, **the location of the user**, **user description, retweeted status** (is a retweet or not), and the **tweet text**

```
twitter_df1.printSchema()

root
 |-- is_retweeted: string (nullable = true)
 |-- rt_original_id: long (nullable = true)
 |-- rt_original_user: string (nullable = true)
 |-- handle_name: string (nullable = true)
 |-- is_verified: boolean (nullable = true)
 |-- followers: long (nullable = true)
 |-- total_tweets: long (nullable = true)
 |-- tweet_type: string (nullable = false)
```

Author Identification Analysis

```
#To get the original tweets, we filter the records where the retweet status is null
count_overall = df.count()
origina_tw_cnt = df.filter('retweeted_status is null').count()


print('Count of total tweets before filtering:', count_overall)
print('Count of the original tweets:', origina_tw_cnt)
```

```
Count of total tweets before filtering: 1005167
Count of the original tweets: 352608
```

By checking for the original vs overall content in the tweets, it is fascinating to observe how the original tweets are only **30%** of the whole

Based on the number of **tweets**, followers count and verified status, I could observe the most prolific and influential Twitterers are **Times Now**, DNA followed by Rising Kashmir

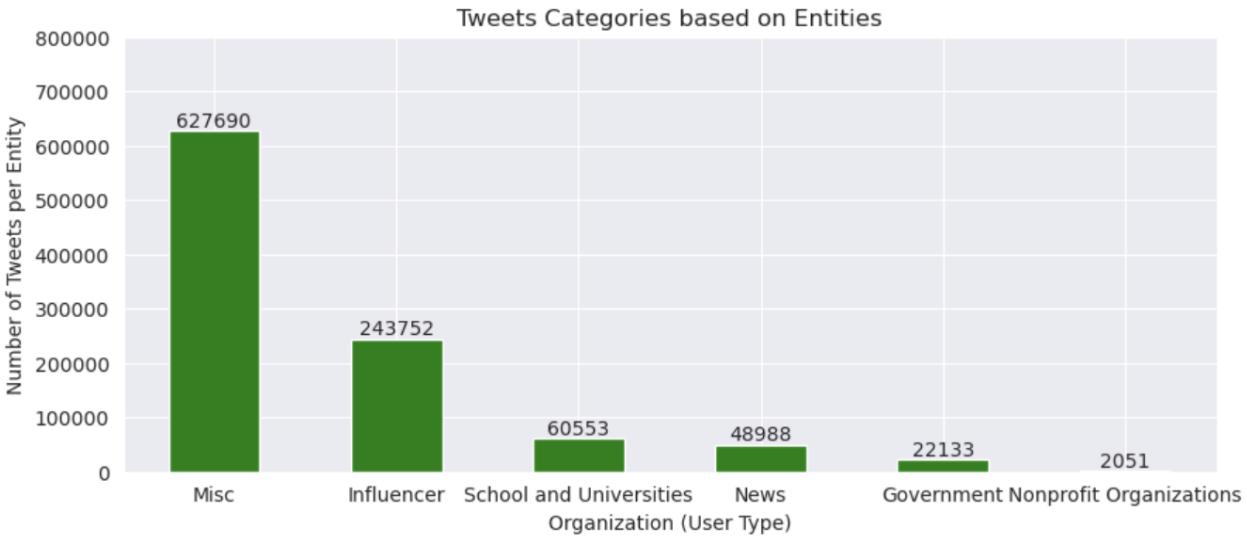
user_name	number_of_tweets	followers_count
TIMES NOW	68	698318444
DNA	72	164936270
Rising Kashmir	115	25537739
U.S. News Education	67	22136817
The Tribune	132	21082137

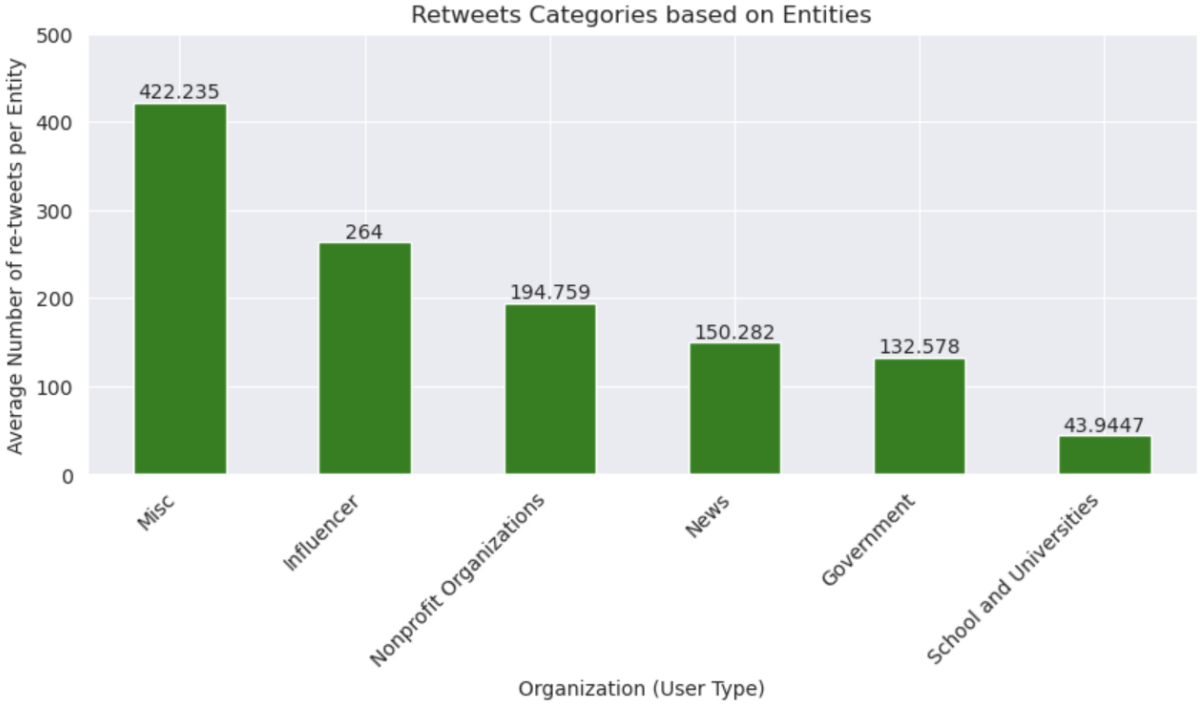
Based on the number of **retweets**, I calculated the retweet rate as (tweets count * average retweets) which helps us get the most influential twitterers

- **Brendan Schneider**
- Lori Lite
- National Education Union

user_name	tweet_rt_rate
Brendan Schneider	524.8
Lori Lite	308.41
National Educatio...	180.0
TIMES NOW	136.0
The Tribune	132.0

Entity Based Classification



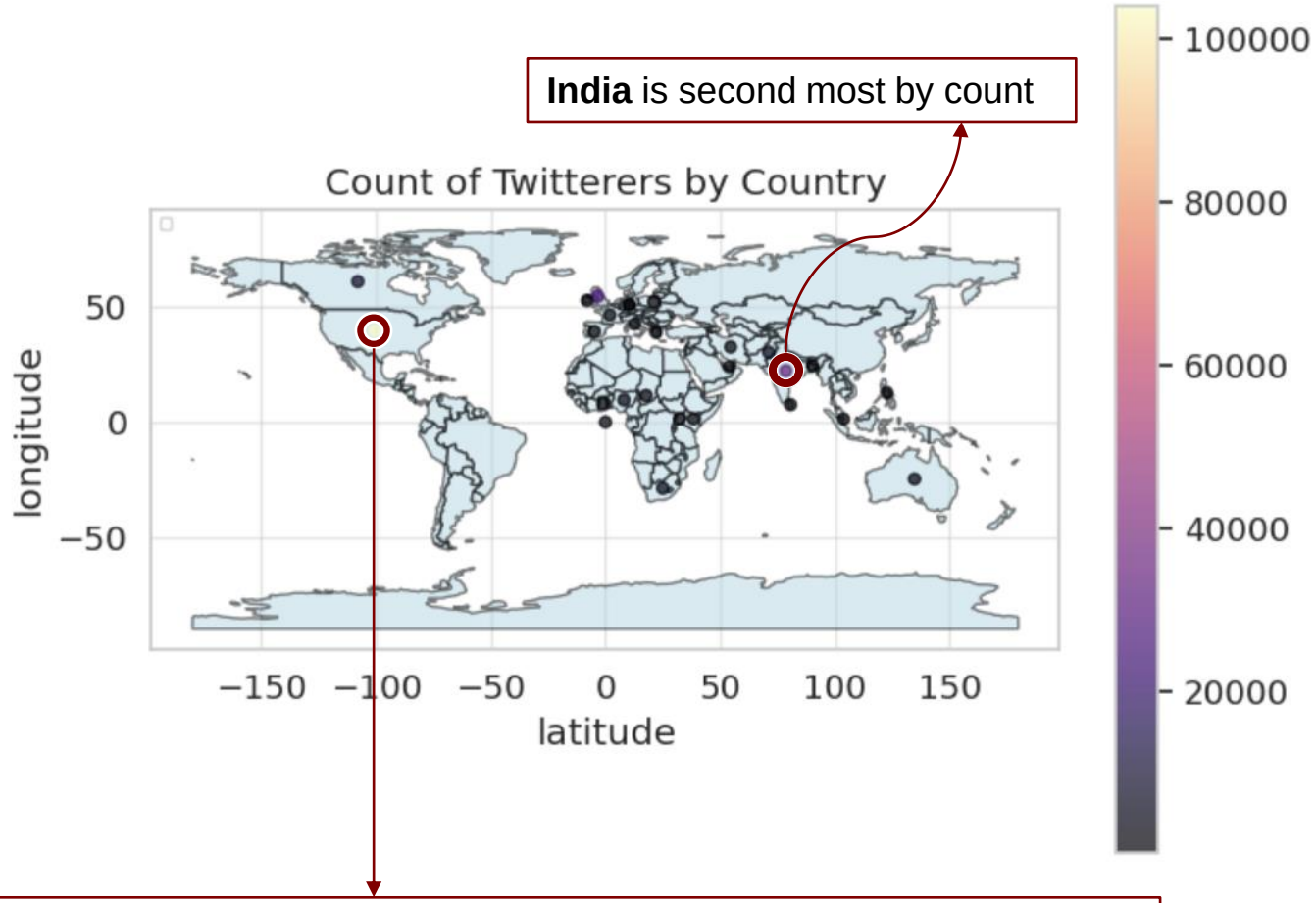


It can be seen that the majority of the tweets (original) come from the **Influencer** entity, who in this new-age world, has more influential power, followed by the Schools and Universities

Since individual twitterers like us are many, the group of **Misc** (miscellaneous) shows the general public does share their opinions on education.

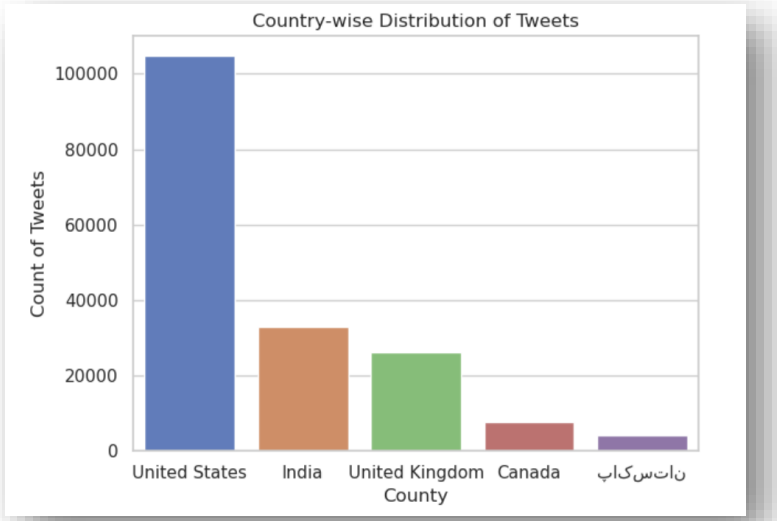
The average retweets are more from **Influencers** than any other entities. However, it was interesting to observe how the **Non-profit organizations** about education retweet more than tweeting

Geographical Analysis



India is second most by count

We can observe the majority of the tweets come from the **US**



- By using the user.location field and mapping them onto the location coordinates (using **geocode**), I plotted the heatmap as well distribution of the Twitter users from around the world.
- It can be observed that most come from the United States, followed by India, and the UK stands third in this category.
- Heat map helped me best visualize the concentration of twitterers

Topic Progression By Geographical Location

	date	user_location	count(1)
0	2022-09-18	India	647
1	2022-05-16	South Africa	305
2	2022-11-03	United Sates	259
3	2022-05-16	Johannesburg, South Africa	244
4	2022-09-14	Patna, India	208

Based on the frequency of tweets,

- I could see that there was trending news in India and South Africa on 2022/09/18 and 2022/05/16 respectively
- Digging deeper, these are the topics that were trending on the respective days.

Leaked videos of women bathing, an alleged suicide, and protests: Chandigarh hostel MMS scandal explained

FP Explainers, September 19, 2022 09:44:24 IST

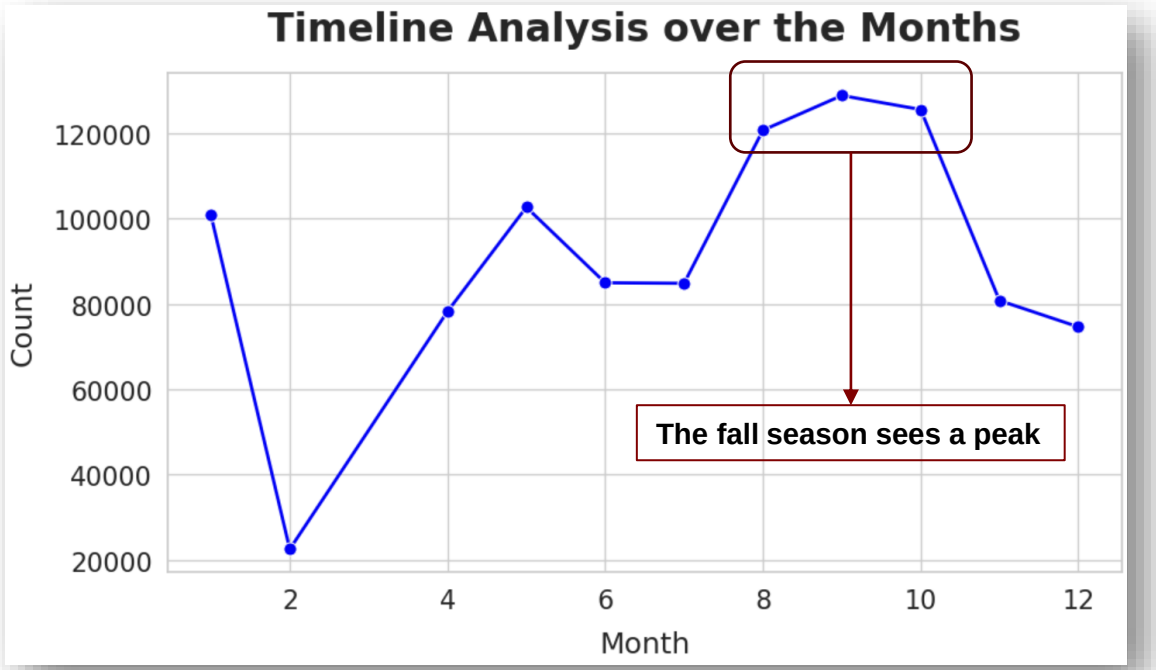
SA: Uni student who urinated on books and the laptop of a black student expelled

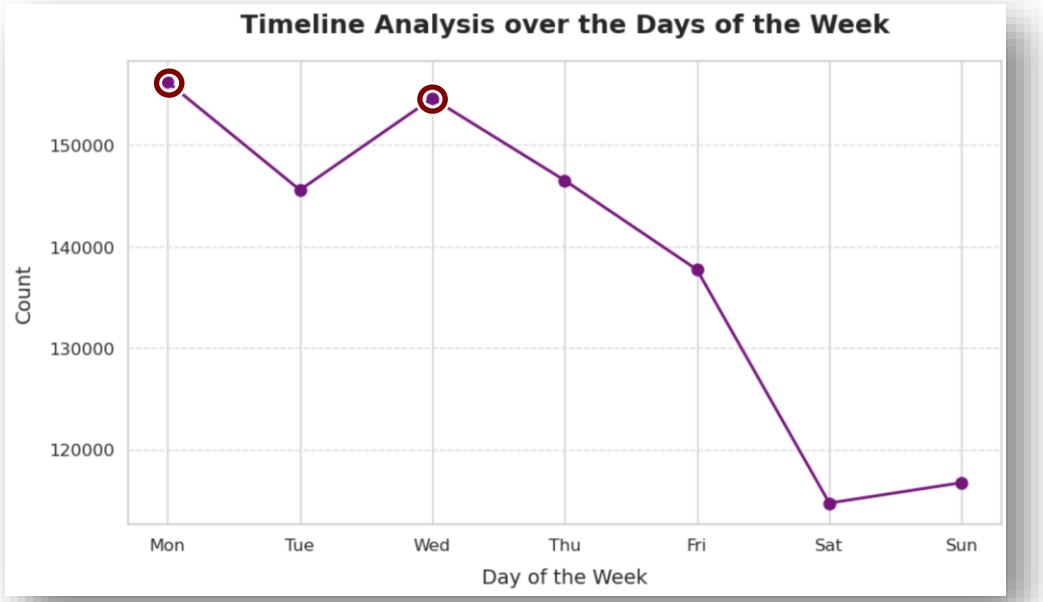
By Rédaction Africanews with AFP
Last updated: 22/07 - 16:34

Timeline-based Analysis

The majority of the tweets are from 2022 and some are from 2023.

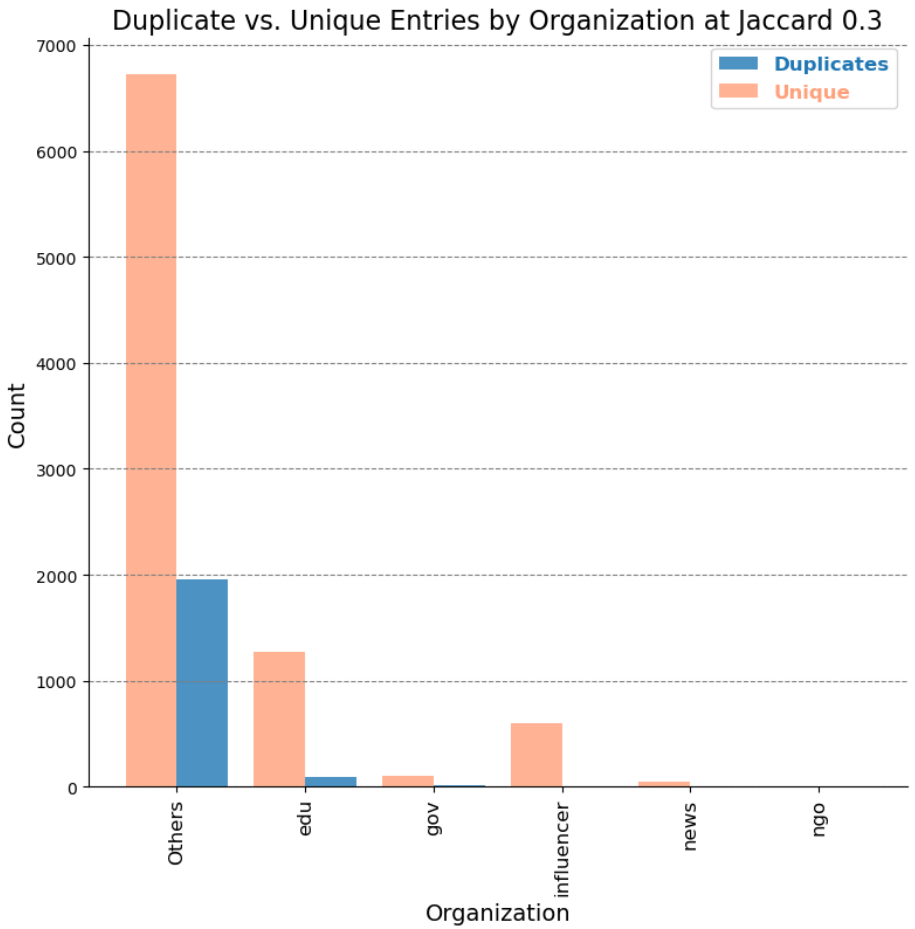
```
year
2022    851826
2023    120240
Name: count_id, dtype: int64
```





It is evident that education tweets are at a high number during the fall season owing to the beginning of a new school/ college year. Likewise, Mondays and Wednesdays see a rise in tweets at the start of the week for schools and then decline toward weekends (holidays)

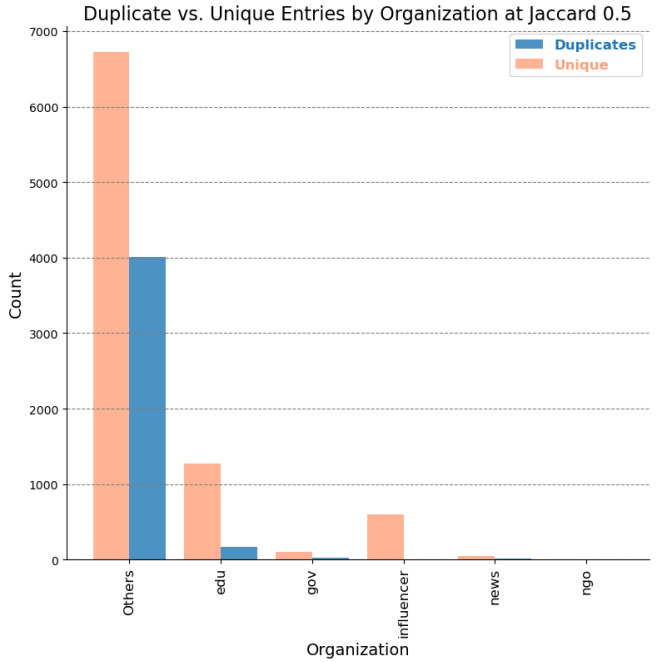
(Similarity Analysis) Twitter as a Credible Source



Analyzing the uniqueness of tweets on a sample of 10k records using the **MinHashLSH** algorithm, I could fairly distinguish between unique and duplicate tweets.

Government, Influencers, and News are the entities that have the least duplicates

At a **Jaccard** similarity of **0.3**, I could observe that the duplicates are at **~ 2,000 which is 20%** of the sample.



At Jaccard distance 0.5

Conclusion

- Based on the analyses conducted, Twitter seems to be a credible source of information but at times can be misleading. However, it is always better to cross-check for credibility against news from top articles and other reliable sources.

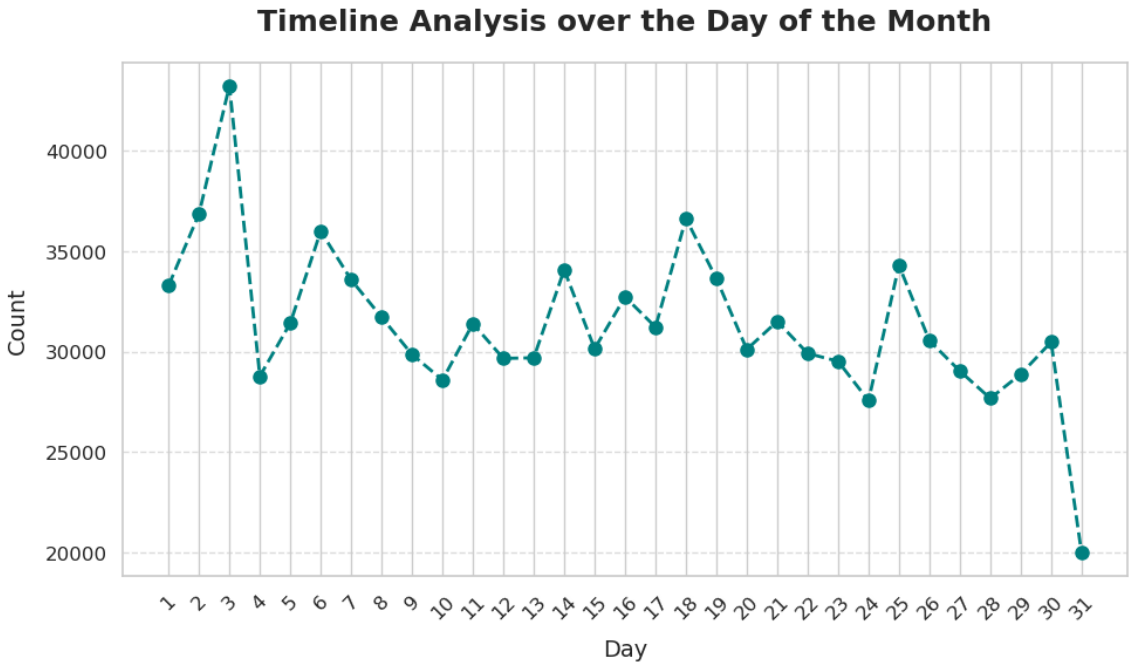
- Out of the ~100M records, after filtering based on requirement and reaching ~1M records, the observations are:

- Influential/ Most Prolific Twitterers – **Time Now** based on original content and **Brendan Schneider** according to the retweet count
- Geographical and Time analyses helped us understand that the **US** is where the major activity happens and that the **fall season** and the **start of the week** are where the most tweets are posted.
- Out of the relevant ~1M records, drawing a sample of ~10,000 records, I could observe that **~20%** of the tweets are **duplicates**. This goes to show that Twitter, after all, is a good source of information

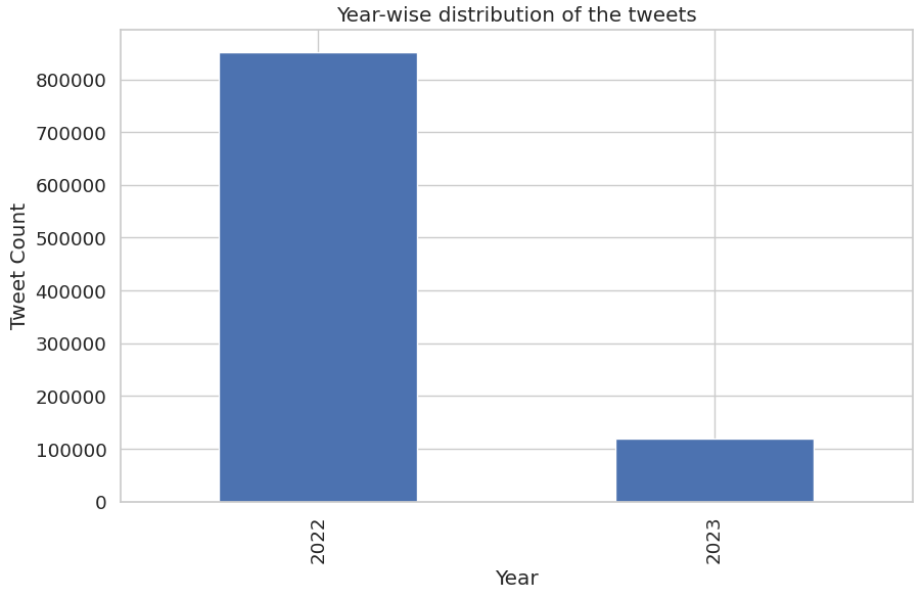
Future Scope & Recommendations

- The imbalance in the type of users on Twitter, the uncertainty of the verified status, and bot activity make Twitter lose its credibility as a reliable source of information
- Twitter can be treated as a secondary source of information or go-to if there is no other way.
- Raw data that we procured is noisy and has a lot of imbalance which could lead to bias in the analysis. Better to confirm with other sources.
- However, these challenges can be overcome by certain data-driven actions:
  - Be aware of the Twitter handle the information is coming from, the follower count, the type of audience, the reply count, and other factors before trusting the piece of information.
  - Impersonation on Twitter is a major concern and a serious issue in the current AI world. Bot activity adds to this inconvenience.

Appendix



Over the month, we can see the behavior of the average twitterer regarding education



Source data spans over two years