Twitter Analysis & Report

MSCA 31013 BIG DATA PLATFORMS
FINAL PROJECT
~GARIMA SOHI

Agenda

01 Executive Summary

Objective

To identify whether Twitter can be considered a credible source of information, which reflects the emergence of important trends or topics in education.

01

Background

Twitter is a platform for users to share information about anything, which sometimes makes it difficult to identify the credible source of information regarding any emerging trend. This brings in the need to evaluate the authenticity of the tweets to rely on Twitter as a trusted source.

02

Process

Twitterers profile will be evaluated with respect to the content posted. For example, are user tweeting about k-12 or higher education, is there any geographical distribution associated with high volume of tweets, are these users mostly government institutions, universities and credible non-profit organizations, or just random users, tweeting about their schools, teachers, etc.

03

03

02 Methodology

Implementation of the project took place in following modules:

01 DATASET SELECTION & EDA

02 MOST PROFILIC USERS ANALYSIS

03 GEOGRAPHICAL ANALYSIS

04 TIMELINE ANALYSIS

05 TWEET UNIQUENESS ANALYSIS

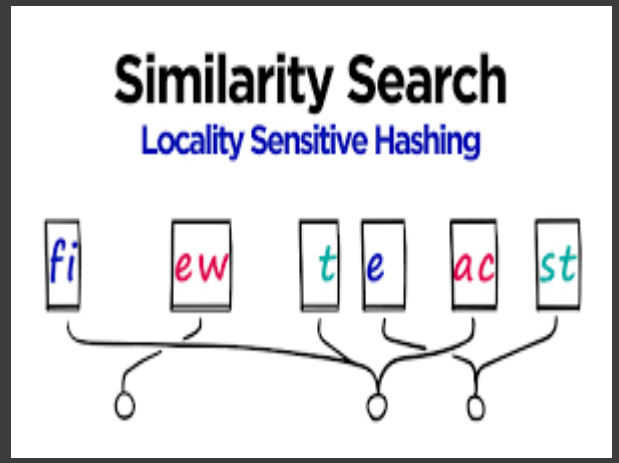
Tools/Packages/Files Used:

PySpark

pandas

seaborn

matplotlib



Parquet

Source Data Overview

Google Cloud Platform

Data is stored on Google Cloud Storage.

Big Data

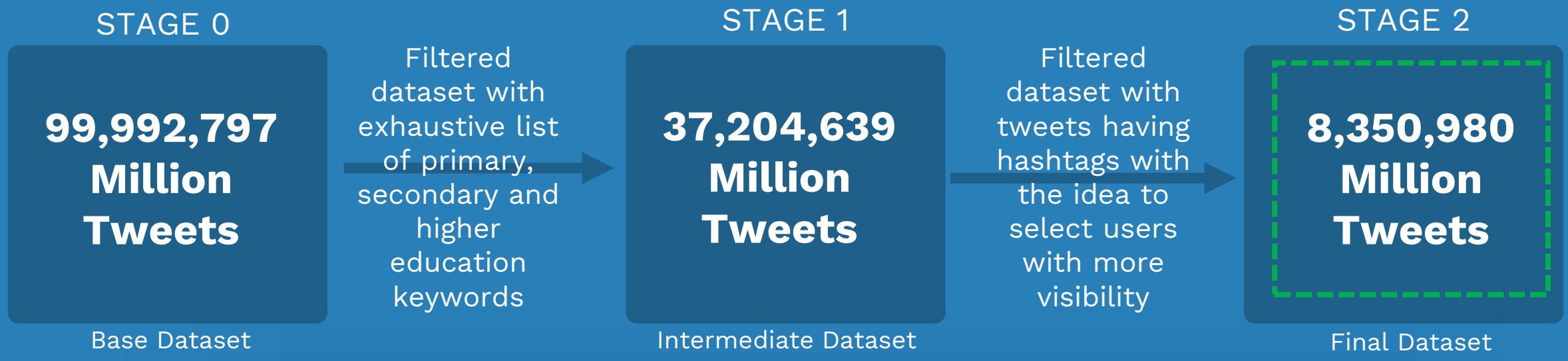
1. Dataset is a combination of multiple JSON files, with ~500 GB size.
2. Rows = ~100 Million (99,992,797)
3. Columns = 40

Tweets Description

Tweets are collected on the topics of education, schools, universities, learning, knowledge sharing, etc.

03 Tweets Data Cleaning & Pre-processing

Stages of Tweets clean-up & filtering:



Identifying and selecting relevant columns for project findings:

User Details

- Unique ID
- Name
- Location
- Description

Geographical

- Country Wise tweets distribution
- User's location wise tweets distribution

Tweets/Retweets

- Original tweets
- Retweets
- Total tweets

Timelines of tweets

- Timeline at which tweets were created

04 EDA – Exploratory Data Analysis

- Final Dataset Details: ~8.3 rows and 14 columns
- Tweets were created from 05th April 2022 to 06th November 2022 in the dataset.

```
Tweets Timeline:
2022-04-05 04:21:38
2022-11-06 22:53:10
```

- Out of the selected variables, following columns are having some null values basis respective reasons:

➢ retweet_user_description / retweet_user_name / retweet_user_id are only populated for retweets, rest rows are null.

➢ Some users didn't mention description and location so it can be seen that user_location / user_description are having some null values.

➢ place_country showing the geographical distribution is having most null values.

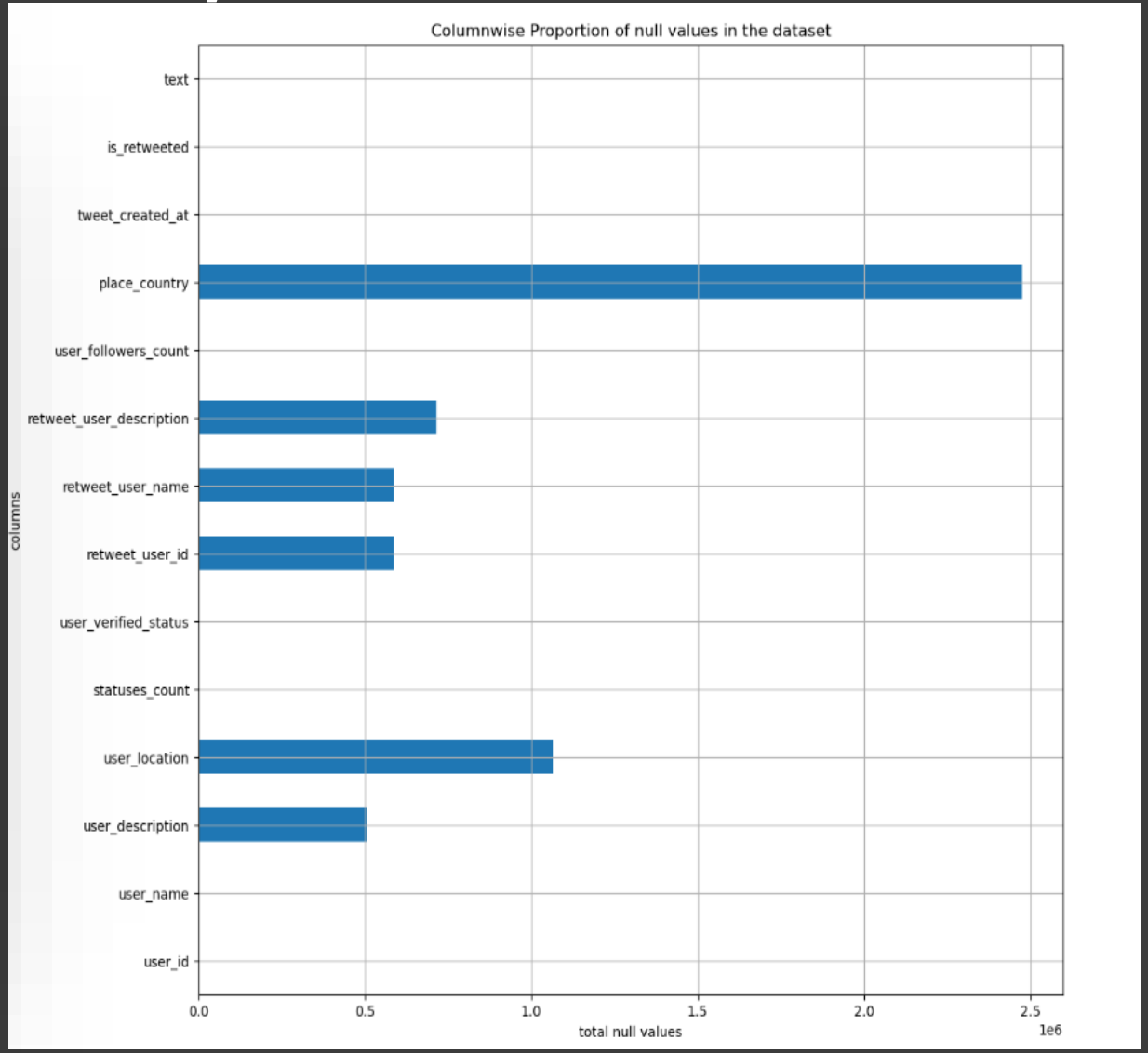


Fig. 1.1
Proportion of null values in the final selected dataset

05 Author Identification – by original tweets

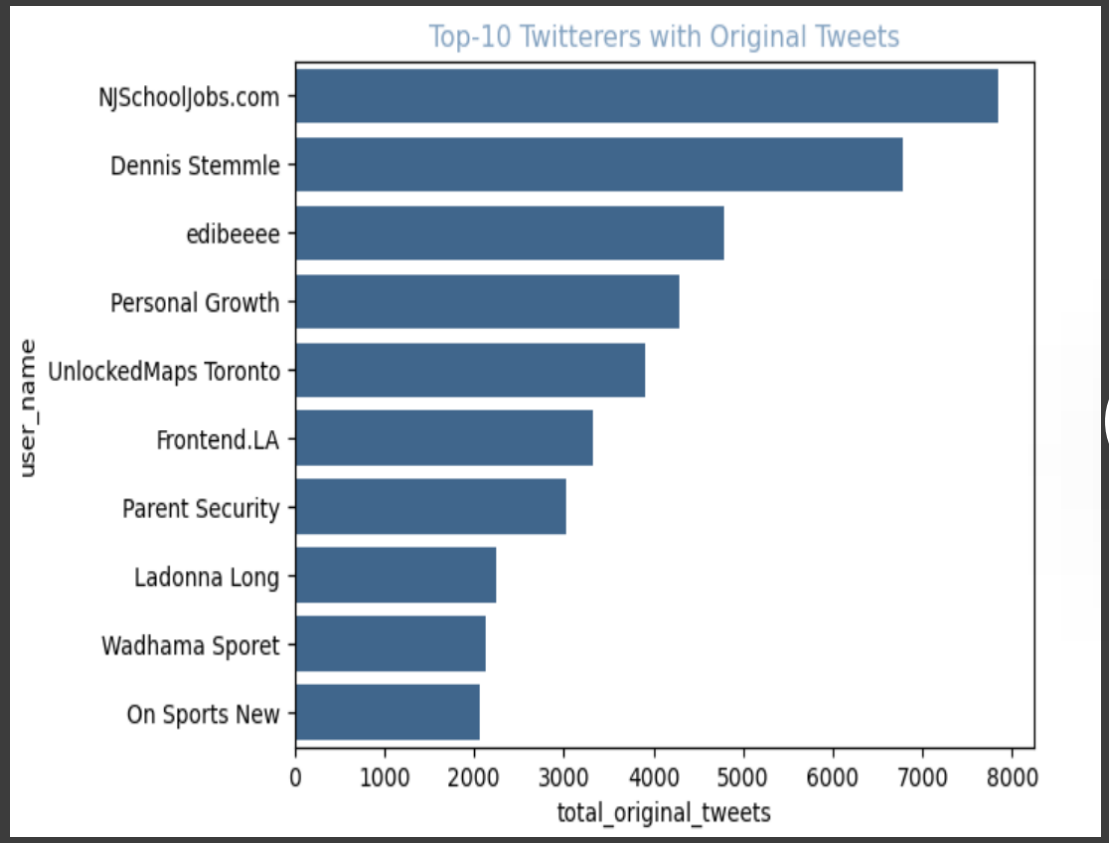


Fig. 1.2
Top-10 users with most original tweets volume

vs

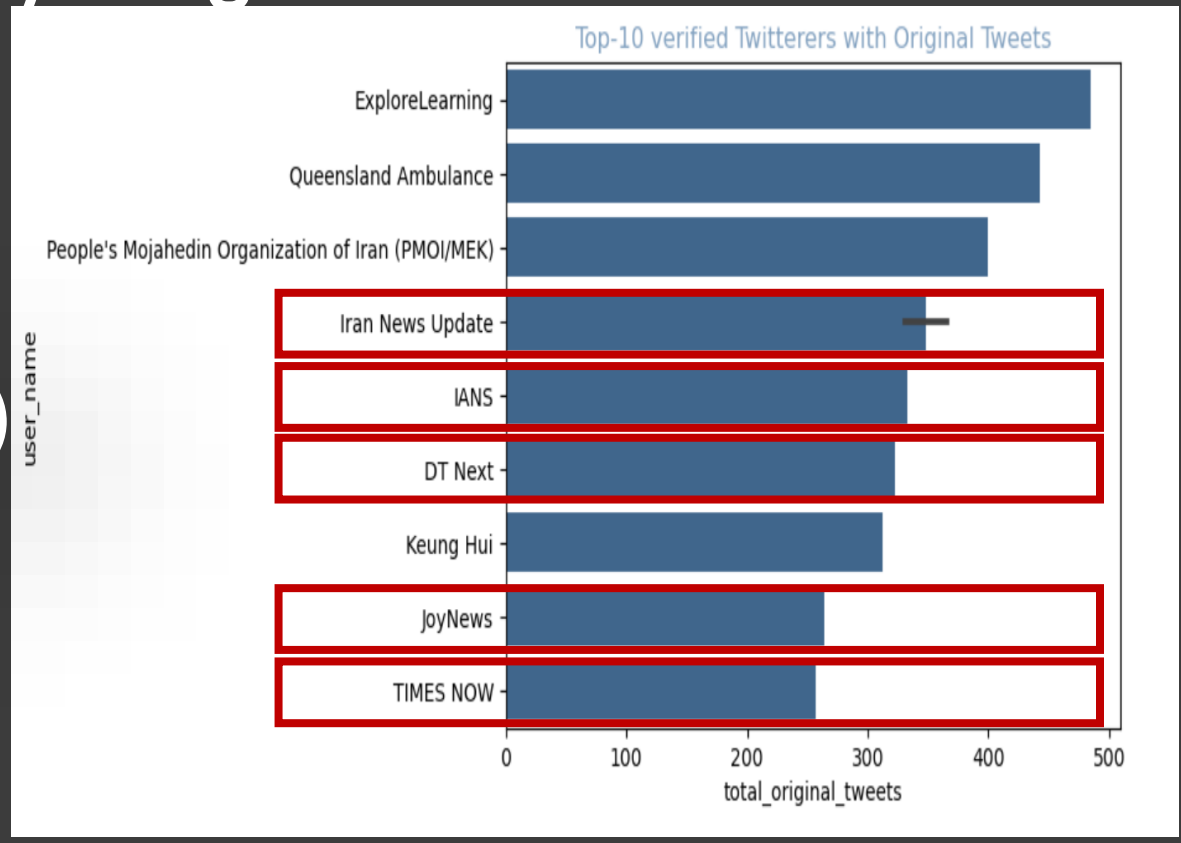


Fig. 1.3
Top-10 verified users with most original tweets volume

- Overall Users: In Fig. 1.2, there are majority random users with original messages. Top being the NjSchoolJobs, which is the leading advertiser of education jobs for New Jersey's Schools.
- Verified Users: In Fig. 1.3, 50% are news organizations, namely Iran News Update, IANS, DT Next, Joy News & TIMES NOW.

05 Author Identification – by retweets

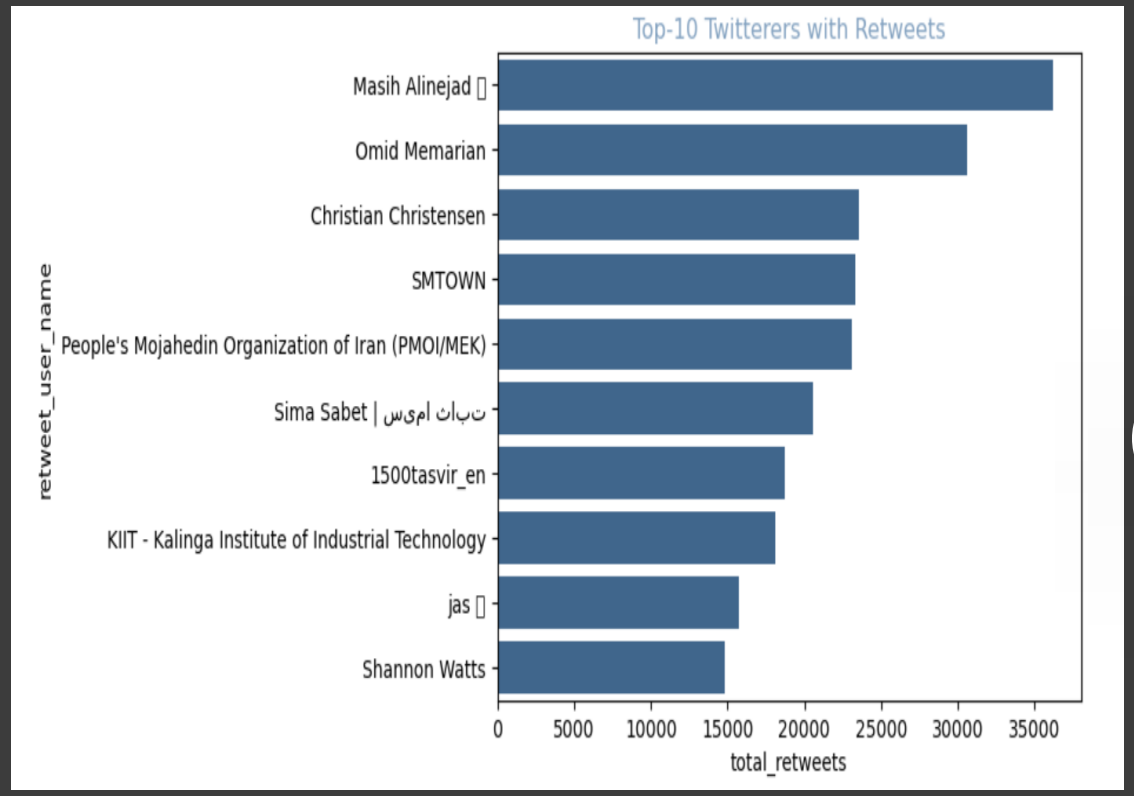


Fig. 1.4
Top-10 users with most retweets volume

vs

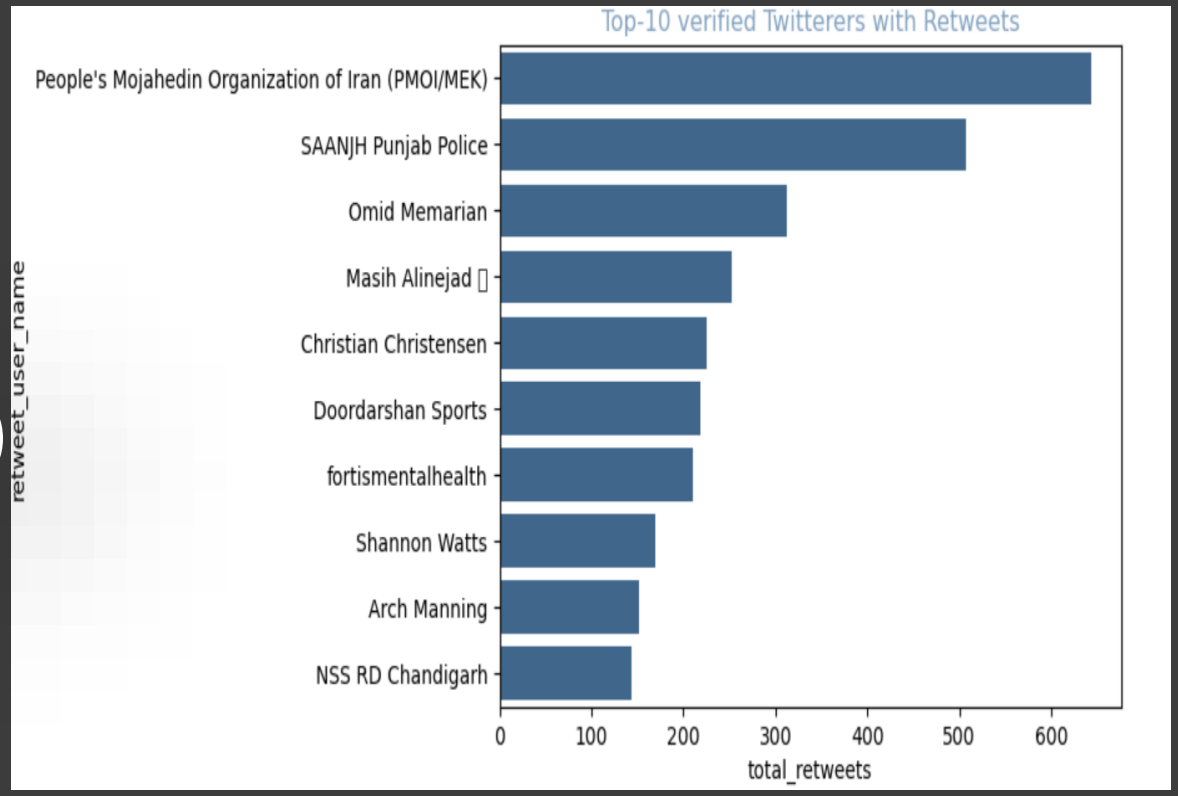


Fig. 1.5
Top-10 verified users with most retweets volume

- Overall Users: In Fig. 1.4, the top three users are into journalism with most retweets, followed by the combination of random users, one organization and educational institutions.
- Verified Users: In Fig. 1.5, miscellaneous groups can be seen from organizations, journalists, sports channel, health organization.

05 Author Identification – by types of orgs

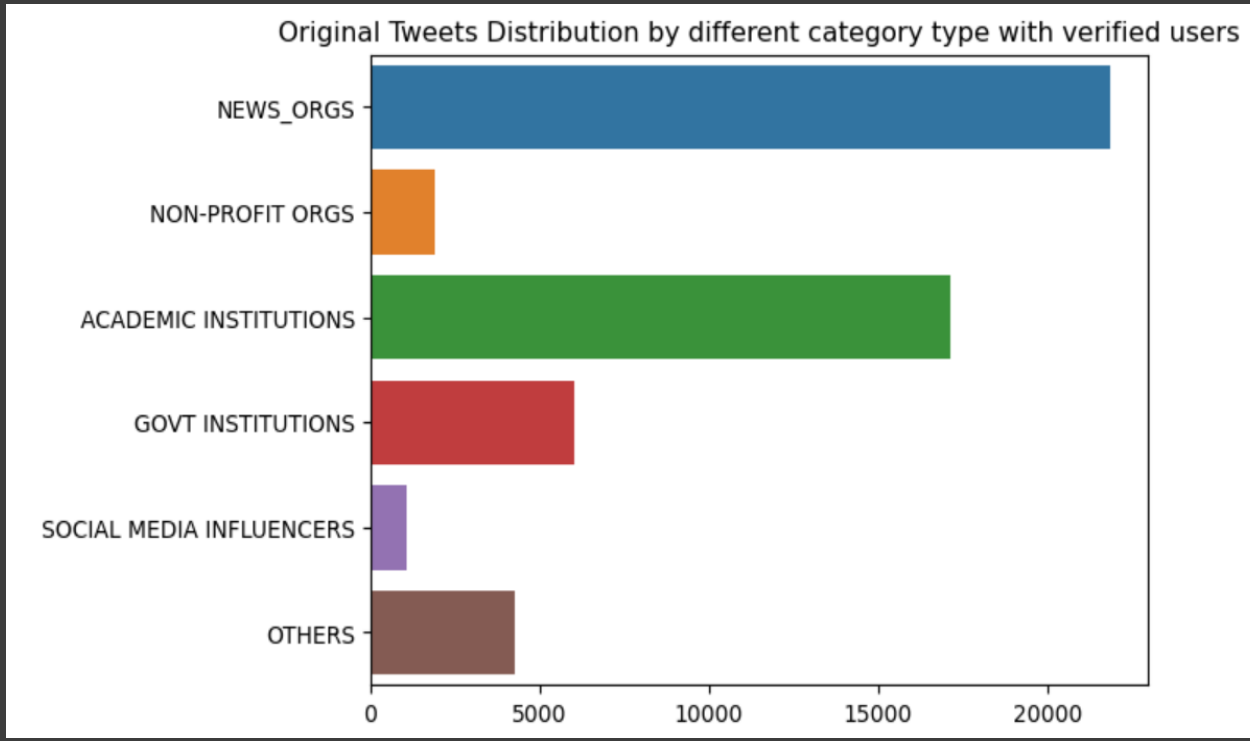


Fig. 1.6
Original Tweets volume distribution by different categories

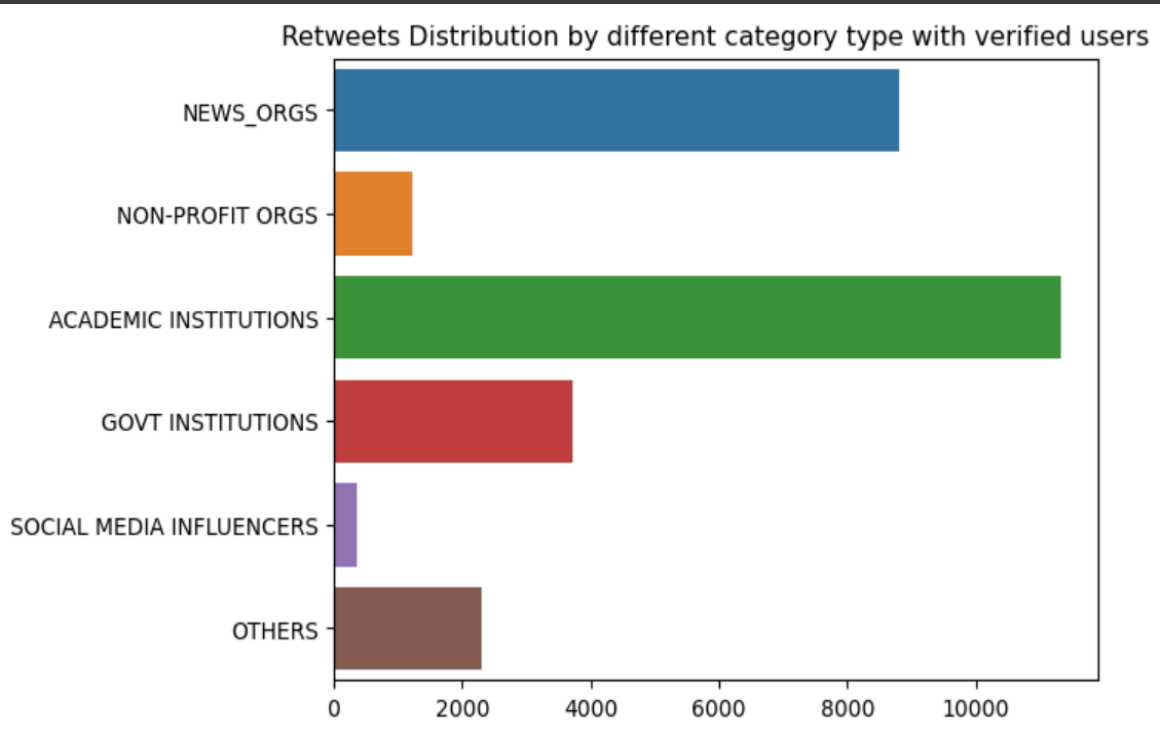


Fig. 1.7
Retweets volume distribution by different categories

- Original Tweets: In Fig. 1.6, NEWS organizations share the most authentic and original content, followed by academic institutions.
- Retweets: In Fig. 1.7, it is interesting to see that academic institutions messages get retweeted the most which is an indication of emerging trends in education.

05 Author Identification – by total tweets

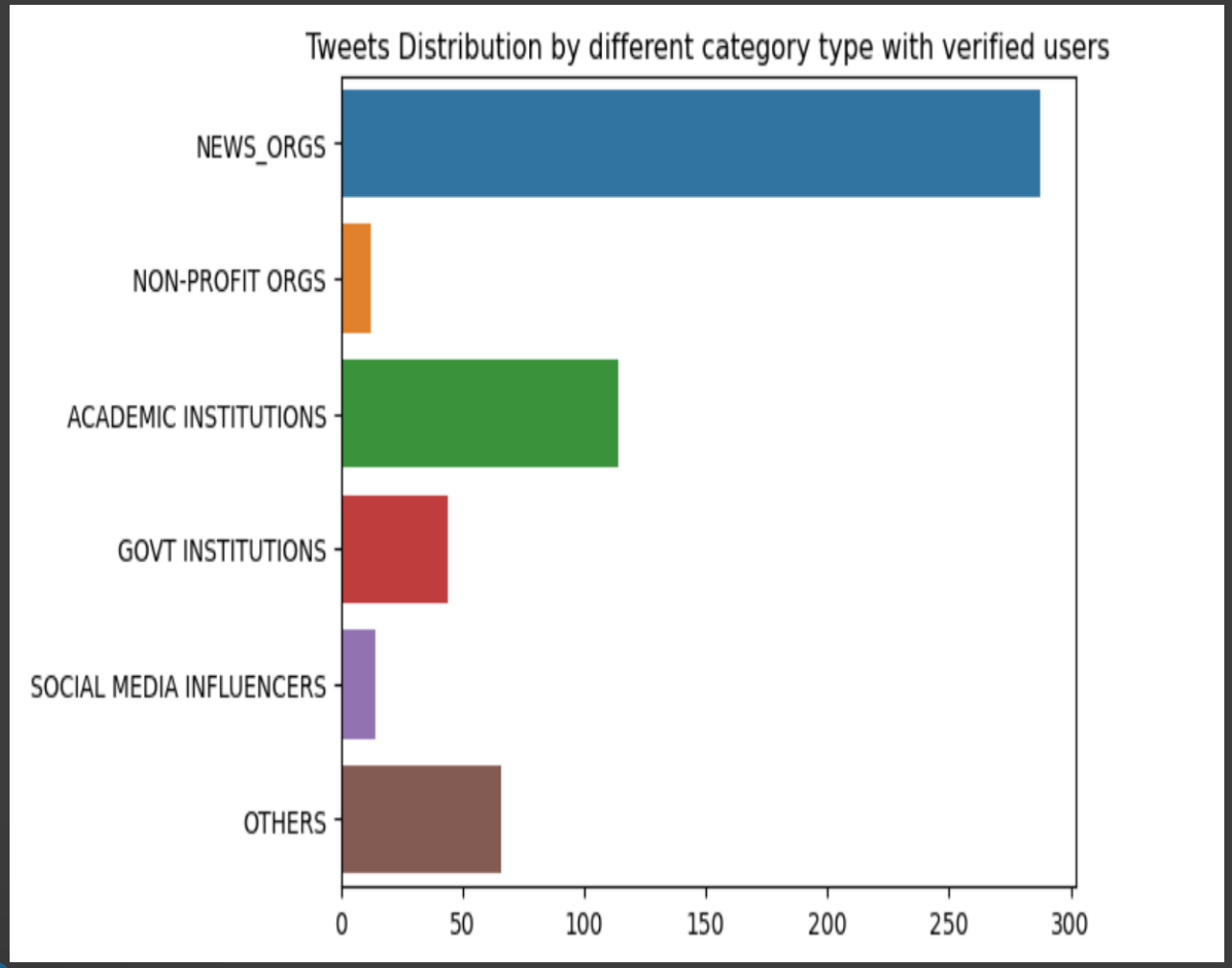
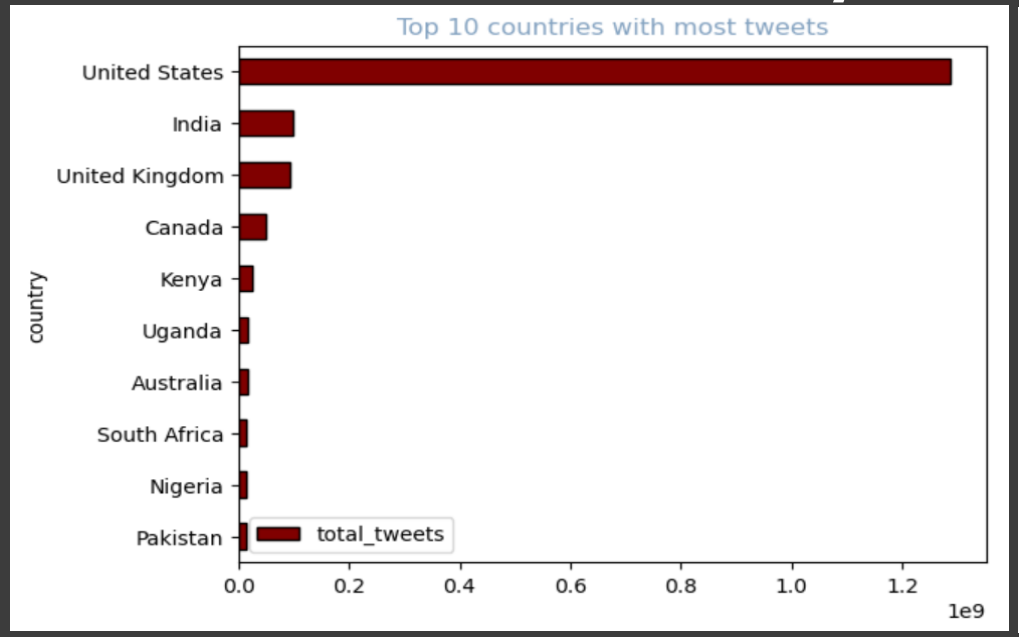


Fig. 1.8
Total Tweets volume distribution by different categories

- For most prolific twitterers, it can be seen in Fig. 1.8 that NEWS Organizations has the most tweet messages (both original and retweet included).
- Academic institutions being the second top most category.

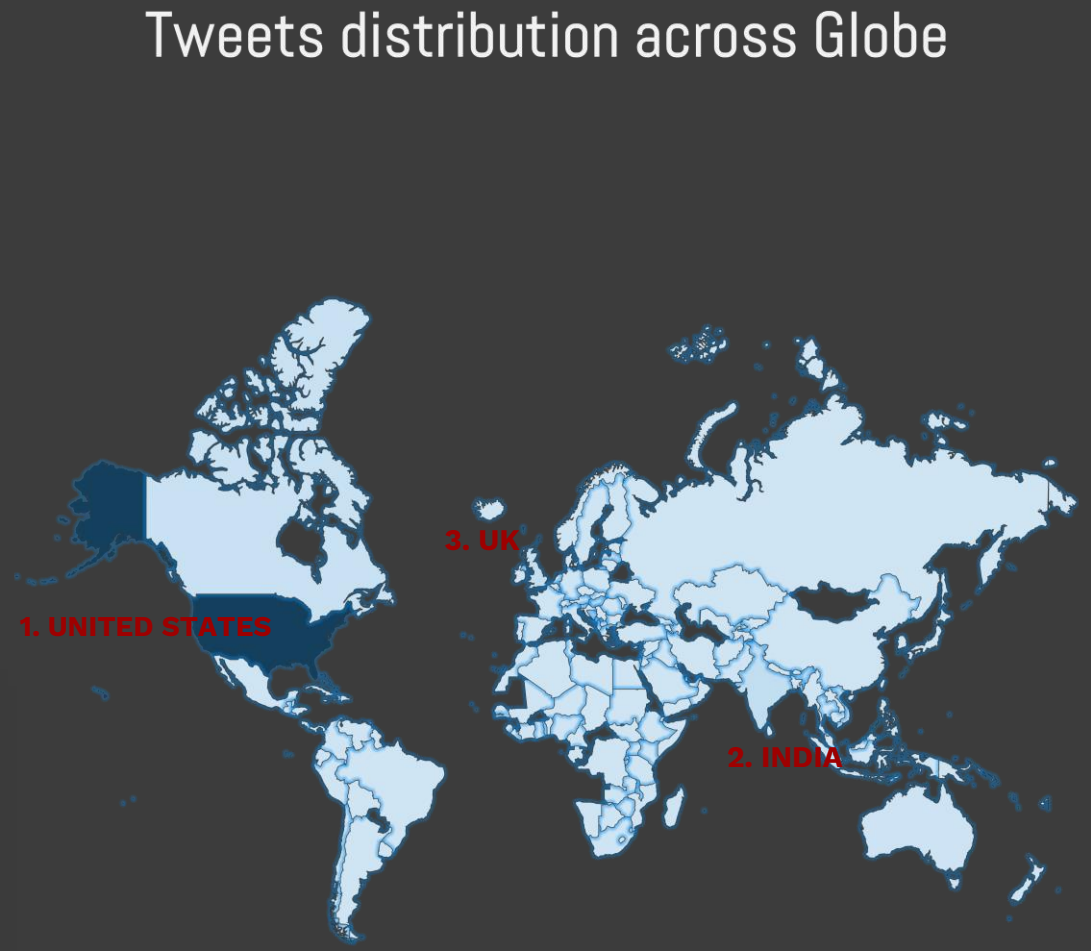
```
Total tweets by news organizations:  287.605321MM
Total tweets by non-profit organizations:  12.481969MM
Total tweets by academic institutions:  114.076251MM
Total tweets by govt institutions:  43.987955MM
Total tweets by social media influencers:  14.30525MM
Total tweets by others:  65.586208MM
```

06 Location Analysis



country	total_tweets
United States	1287886041
India	98093550
United Kingdom	91733253
Canada	49040137
Kenya	23803518
Uganda	16278705
Australia	15927858
South Africa	13818584
Nigeria	13383993
Pakistan	13230285

Fig. 1.9 Top-10 countries with most tweets



- In Fig. 1.9, United States is the top country with most tweets (both original and retweets), followed by India and United Kingdom.
- In United States, Farmington Hills, MI (Michigan) is the top User Location.

```
0    Farmington Hills, MI
1           United States
2                Novi, MI
3                     USA
4                   India
Name: user_location, dtype: object
```

07 Timeline Analysis

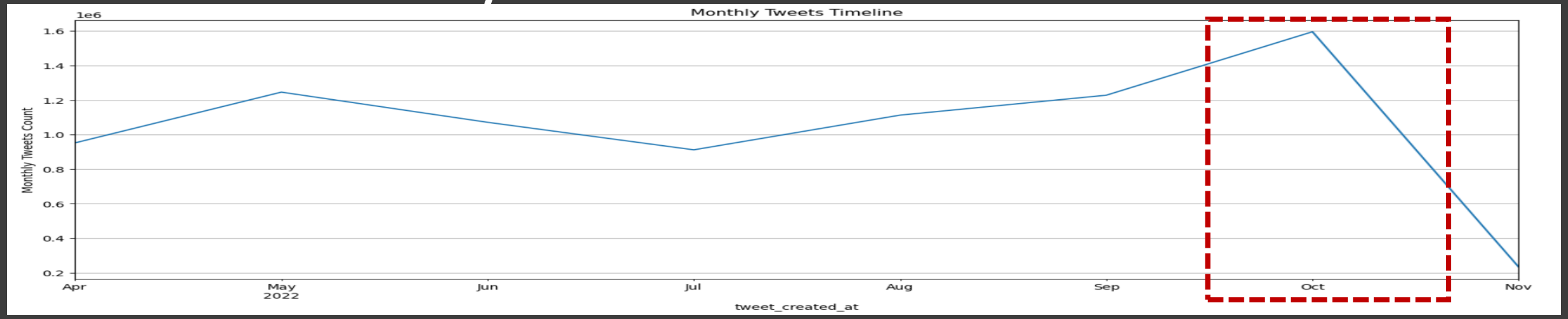


Fig. 1.10 Monthly Tweets Timeline Trend

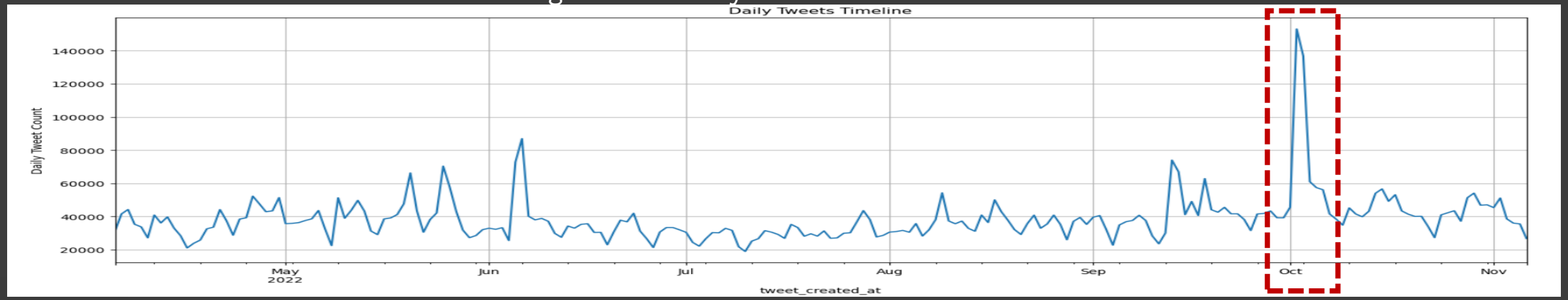


Fig. 1.11 Daily Tweets Timeline Trend

- In Fig. 1.11, tweets increased a little in the month of May. Further, it can be seen that in October, there is a drastic increase in the number of tweets created.
- For more details in terms of daily frequency, refer Fig. 1.11. In this figure, it can been seen for October, tweets were heavily created just in the first few days of the beginning of the month, which conveys a trend started and dissolved in a few days.

08 Message Uniqueness Analysis

- Performed MinHash Algorithm on a subset of 10K tweets.
- Implemented uniqueness analysis with different jaccard distances.
- It was found that jaccard distance = 0.3 was giving the optimal results.

- Out of 10K, ~24.5K tweets were duplicates (25% of sample).
- This outcome makes sense as these 25% tweets can have the retweet messages. Thus, making them duplicate messages.

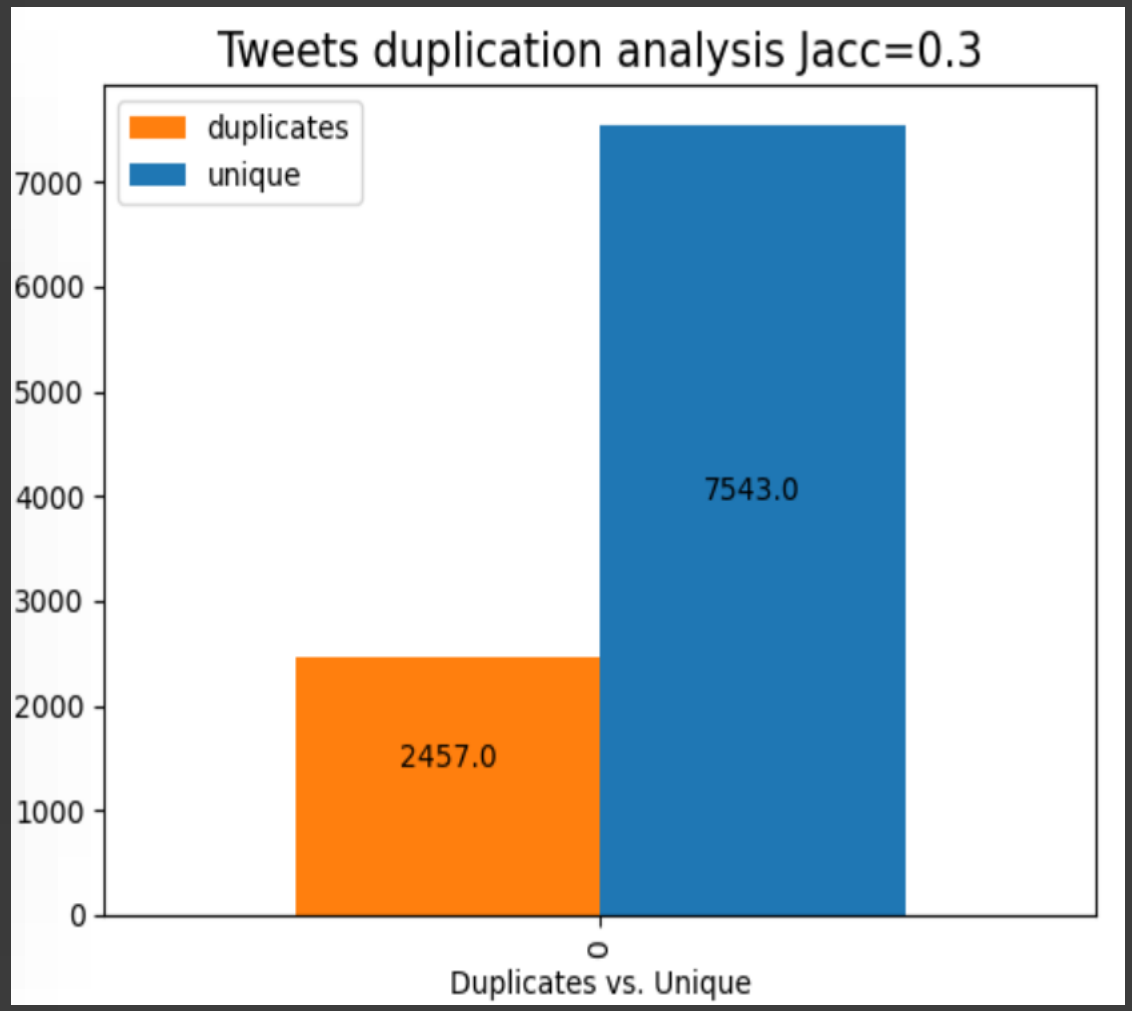


Fig. 1.12 Tweet Text Uniqueness Analysis

09 Conclusions & Recommendations

Conclusions

- News Organizations post the most authentic and original content. Thus, twitter as a platform can be trusted as a credible source of information.
- Educational Institutions got the most retweets in last months which indicates an academic trend in the tweets.
- United States is the country where most tweets are created.
- In the beginning of October, tweets volume spiked up and got dissolved after a few days.
- Lastly, 25% tweets were duplicates in a sample of 10K, indicative of the presence of retweets messages in the sample.

Recommendations

- Users/Organizations with authentic, verified and original messages should be stacked up in the system so that there messages doesn't dissolve with random users, making Twitter more reliable source of information.
- To prevent spread of mis-information, additional indicator should be provided stating this message is not verified yet.

Thank You