Twitter Posts and Twitterers Analysis

Big Data Platforms
Final Project - Education

March 2023 , Chicago IL

Submitted by Kshitij Mittal

Tweet

Search Twitter

Executive Summary

@welcome

Twitter is a micro-blogging platform that enables users to share short-form text-based messages, known as tweets, with their followers. Launched in 2006, it has since become one of the most popular social media platforms with over 450 million active users (2023). Twitter API facilitates real-time access to its streaming data, and makes it a valuable source for data mining and big data analysis.

In our current purview, we are focusing on identifying if Twitter can be considered a credible source of information to gauge important trends and topics in education.

Key Findings

- An extensive social network platform, twitter exhibits both pros and cons for quality data analysis
- While it does provide large quantities of data, several critical data quality checks need to be placed to make it analyzable
- Being high dimensional and deeply nested, twitter data does give highly sparse datasets which increases computational loads
- For the field of education, it provided a diverse set of twitterers. However, a minor chunk of them were verified and yielded credible and unique insights
- Data showed insightful geography and time series trends which correlated with recent educational developments
- While retweets as a metric helps in gauging user influence, it adds a lot of data and had to be filtered out for many analyses

Tweet

Methodology

@methodology

Following platforms and tech stack were used for this analysis:

Google Cloud Platforms	Cloud storage platform
PySpark	Large Scale Distributed data processing
Python (Pandas, Numpy)	Final data processing (not distributed) and storing notebooks
Matplotlib Seaborn	Data Visualizations
Locality Sensitive Hashing	Text similarity analysis for original tweets (across different user groups)
Parquet Format	Storing intermediate data sets

- All computational tasks were performed using PySpark and Pandas Dataframes on a GCP Dataproc Hub Cluster (with auto-scaling enabled)
- All intermediate processed data was stored on GCP in parquet format
- Raw tweets were filtered extensively during the EDA process. Identifying only relevant columns; eliminating sparsely populated columns greatly helped reduce the computational loads
- Tweets were divided into original tweets, retweets, replies and quotes using twitter reach variables ('is retweet', 'reply original user', 'quote original user')
- For author identification and message uniqueness analysis, verified users were prioritized using 'is verified' column
- User location ('user.location') was analyzed for the geographic analysis for original tweets. Tweet geographies (coordinates) were majorly null and discarded during the analyses
- 'Created at' column was parsed into a date time function to tease out timeline trends for original tweets
- Locality sensitive hashing and Jaccard distance were implemented to determine uniqueness of posts (by verified users)

Tweet

Tweet clean-up and filtering

@filter

To streamline our data analysis, tweets pertaining to only education were filtered and prioritized. The clean-up process continued through multiple stages of analysis, as more tweets were analyzed in the process.

99,994,342

Tweets in the original dataset

39,848,346

After **filtering** for educational keywords

36,344,266

Tweets after **removing** sports tweets

27,102,994

Tweets after **removing**
- Bots accounts
- Employment portals (for teachers)
- More sports words

Removed post EDA

Original Dataset

Raw Data

- Included variety of topics not relevant to education (like sports, governance, crime)
- Tweet data was reduced to lower case and stripped off of any special characters

Show more

Key Words used*

Education

primary school schools education k12 high school teacher higher secondary higher education, senior secondary, sophomore math mathematics science physics chemistry biology Humanities history philosophy....

~60 different words

Show more

Key Words used*

Sports

“baseball", "volleyball", "association football", "varsity", "high school”, “softball”, “playoff", "varsity sports", "university sports", "vs.” ,"college football", "college basketball", "basketball", "track and field", "swimming"....

~42 different words

Show more

**Keywords were obtained through web searches and manual tweet checks, and websites like relatedwords.org*

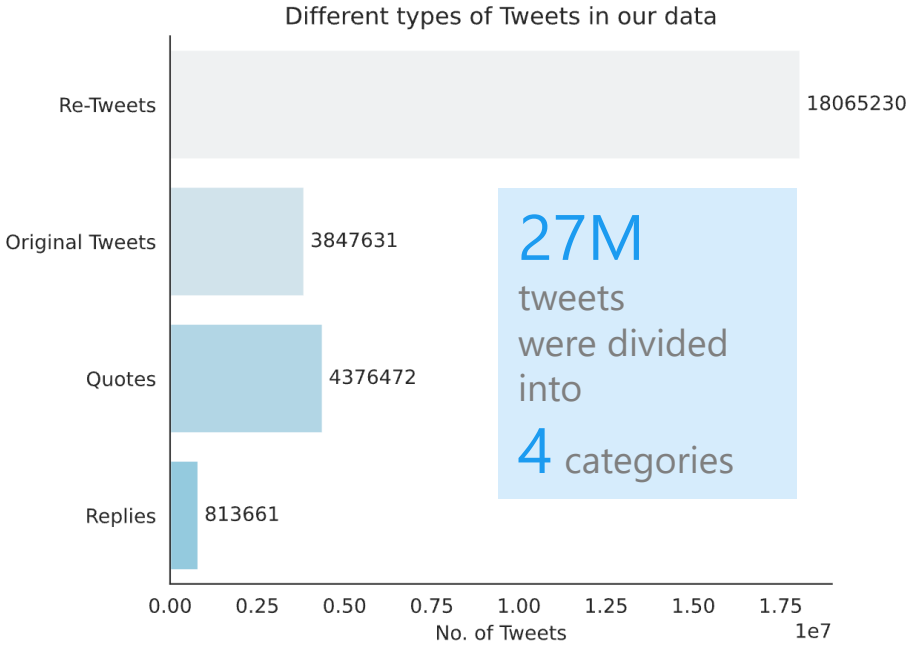
Home
Executive Summary
Methodology & Overview
Clean-Up & Filtering
Exploratory Analysis
Author Identification
Location Analysis
Timeline Analysis
Message Uniqueness
Conclusion & Insights
Tweet

Exploratory Data Analysis

@eda

Search Twitter

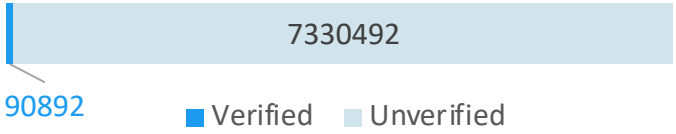
Original Twitter data includes 40 deeply nested variables. Many of these variables showed high percentages of null values and were rendered less optimal for further analysis. This greatly reduced the file load and decreased the computational resources.



27M tweets were divided into **4** categories

7.4M

Distinct twitterers (~1.2% verified)



Apr'22 to Feb'23

Timeframe of our data collection

Variables kept in the final dataset (with fill %)

id	100.00%
created_at	100.00%
text	100.00%
tweet_text	100.00%
tweet_country	0.86%
tweet_country_code	0.86%
tweet_place_full_name	0.86%
tweet_place_type	0.86%
is_retweeted	100.00%
reply_count	100.00%
quote_count	100.00%
retweet_count	100.00%
favorite_count	100.00%
rp_original_id	16.25%
rp_user_id	16.72%

rp_original_user	16.72%
qu_original_id	8.99%
rt_original_id	66.64%
rt_original_user	66.64%
account_id	100.00%
account_name	100.00%
account_description	100.00%
account_location	69.68%
total_followers	100.00%
total_friends	100.00%
total_listed	100.00%
total_favourites	100.00%
total_tweets	100.00%
account_created_at	100.00%
account_profile_picture	100.00%

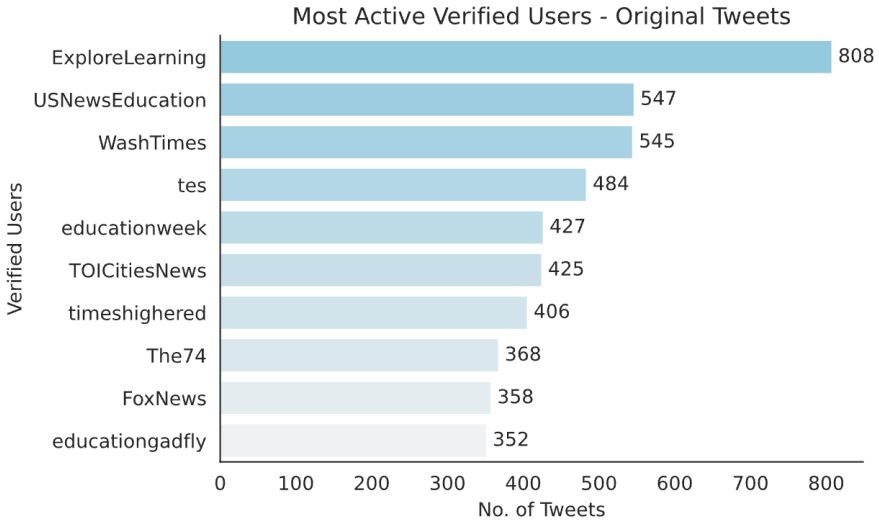
Tweet Information | Tweet Reach | Twitterer Information

Tweet

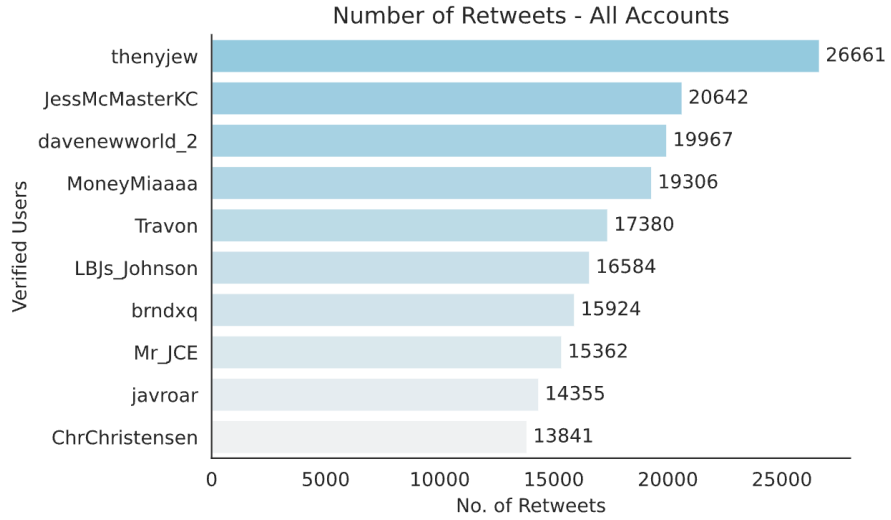
Author Identification

@author

On analyzing ~90K verified users, we found news and education related accounts to be the most influential:



However, unverified accounts and posts were being retweeted more (sometimes even for irrelevant posts) – Verification does not always translate to influence



While twitter does have credible accounts for education information, there is still a huge room to grow while maintaining engagement and adding more information. Unverified accounts tend to take away major attention from credible sources

**Time frame: April 2022-Feb 2023*

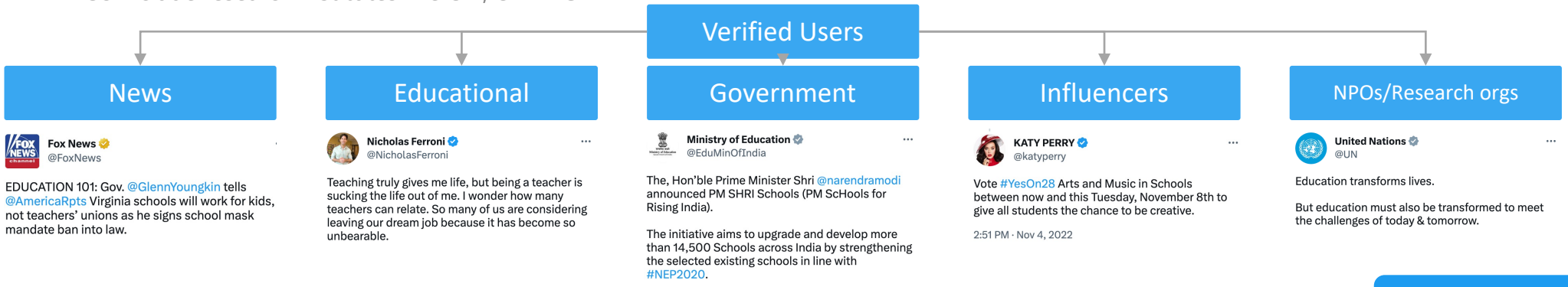
Tweet

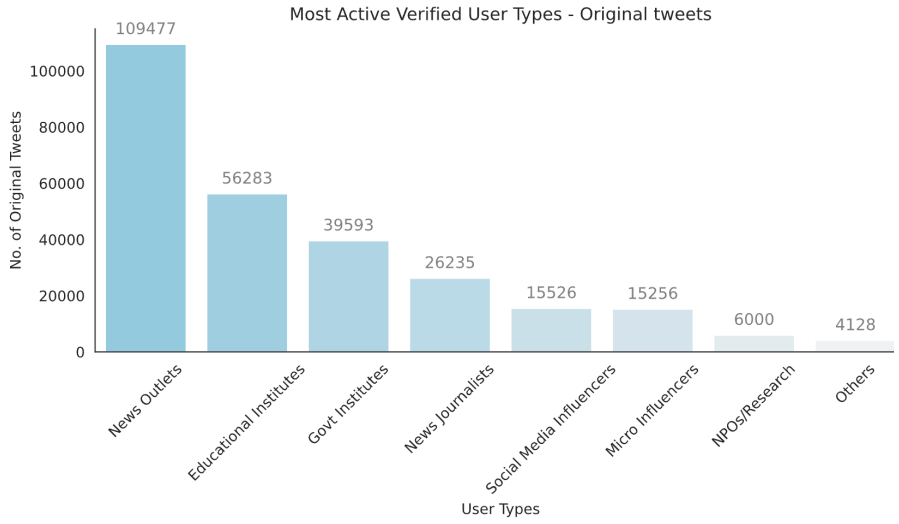
Author Identification

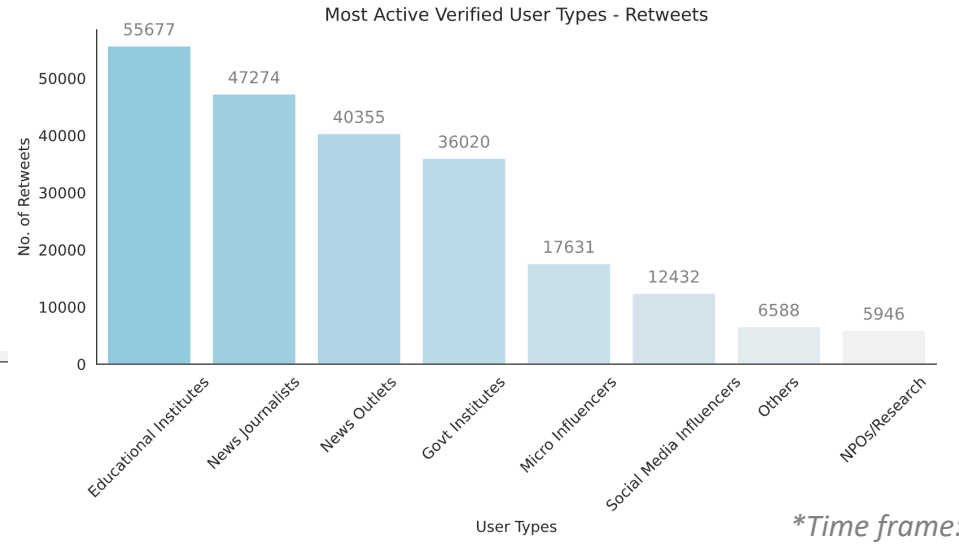
@author

Drilling down on verified users, five broader organizations/user types were extracted by analyzing account descriptions and names

- Schools and Universities were showing high similarity in content upon manual inspection, and were merged into Educational Institutions
- News category was given two layers – News Outlets and News Journalists
- Influencers were given two layers – Micro Influencers (>5000 followers) and Major Social Media Influencers (>50000)
- NPOs include research institutes like UN, UNHRC







News Outlets remain the most active sources of education information on twitter, where as educational institutes are more broadly retweeted

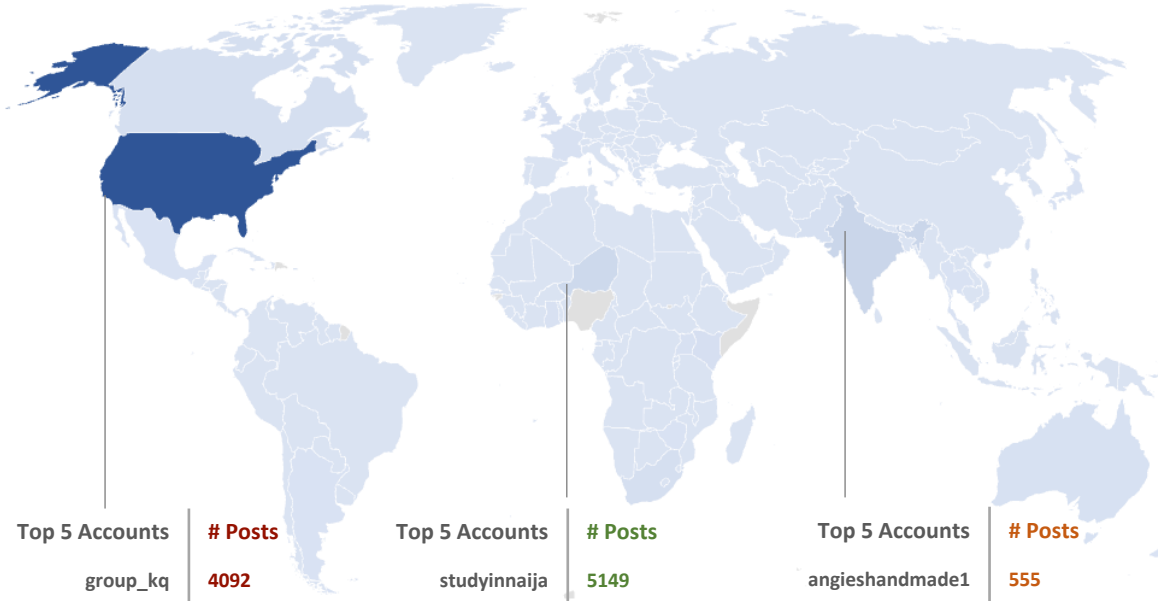
**Time frame: April 2022-Feb 2023*

Tweet

Location Analysis

@geo

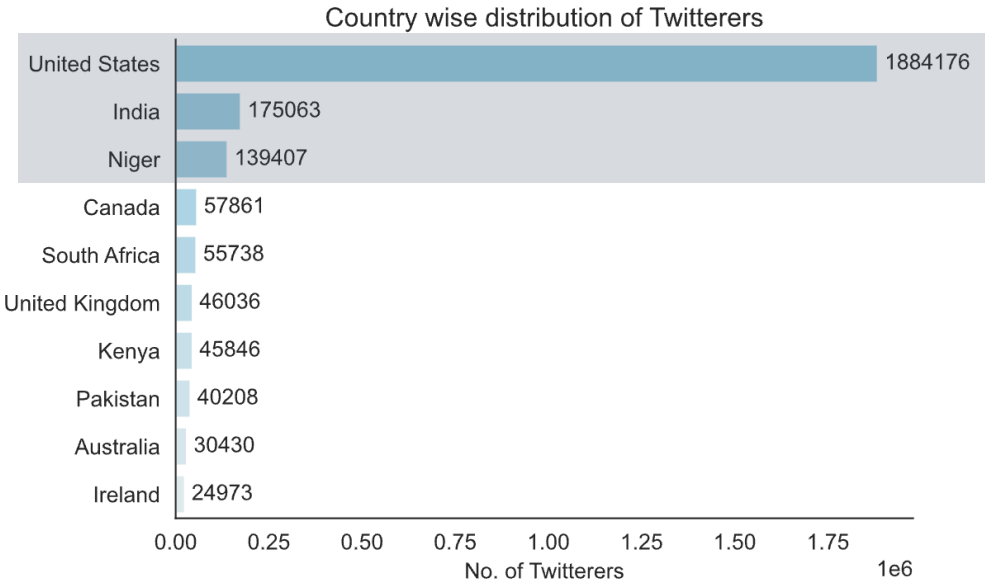
Account location data from 8 million tweets was cleaned and analyzed to yield processable locations. United States locations were easier to process due to well-defined location standards (state codes and city abbreviations)



Top 5 Accounts	# Posts
group_kq	4092
DennisStemmle	2676
dealsoftakis	2475
Rdene915	1554
abook_and_bev	1384

Top 5 Accounts	# Posts
studyinnaija	5149
GraceSmartsBlog	575
StrawberryNG	403
myschoolnewstv	331
realminablog	331

Top 5 Accounts	# Posts
angieshandmade1	555
malpani	411
PophaleSamarth	378
school_finds	367
sirimahanthesh	365



User location column in twitter is manually-filled, and does not confide to international geo tags. Due to this, many user locations could not be processed without advanced NLP

For some accounts, location also changes with time.

United States still remains the most active twitter market. However, developing countries like India and Nigeria have leapfrogged advanced economies like Canada and UK in twitter usage for education

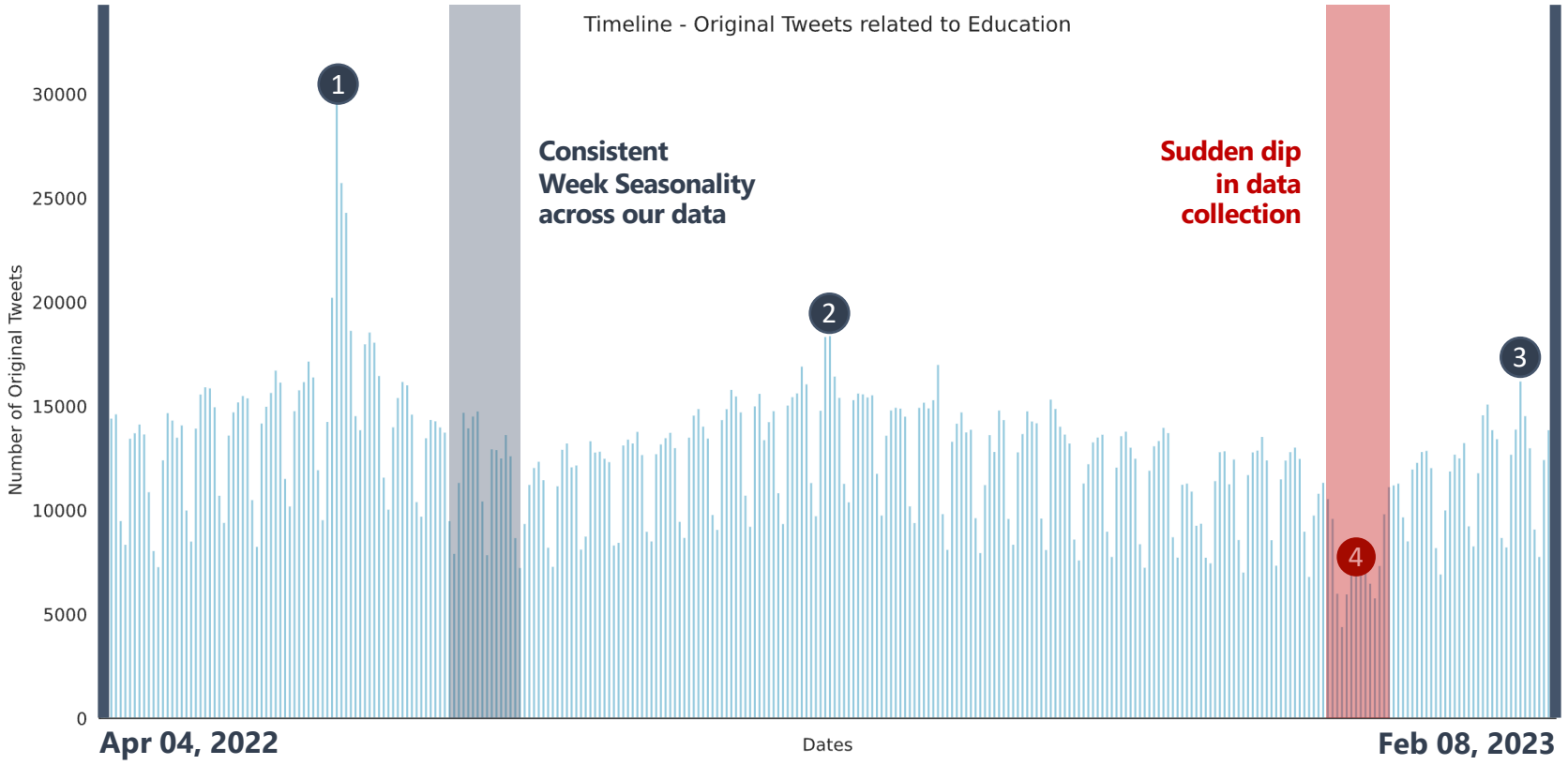
**Time frame: April 2022-Feb 2023*

Tweet

Timeline Analysis

@time

In our timeframe, we see three major peaks and 1 major valley when accounted for original tweets related to education. We do not see any major data collection gaps, but in the latter half of December 2022 twitter suffered from some outages (visible by sudden dip in tweet counts)



1
May 5, 2022
Texas Elementary School Shooting

2
Aug 25 – Sep 06, 2022
President Biden announces student debt relief

3
Feb 1, 2023
Governor DeSantis proposes elimination of diversity studies in Florida

4
Dec 25, 2022*
Twitter outage (during Elon Musk takeover)

Education stays a hot topic throughout the year, and can be driven by different headliners as per specific timespans. We also see a seasonal effect where tweet counts considerably decrease over the weekends

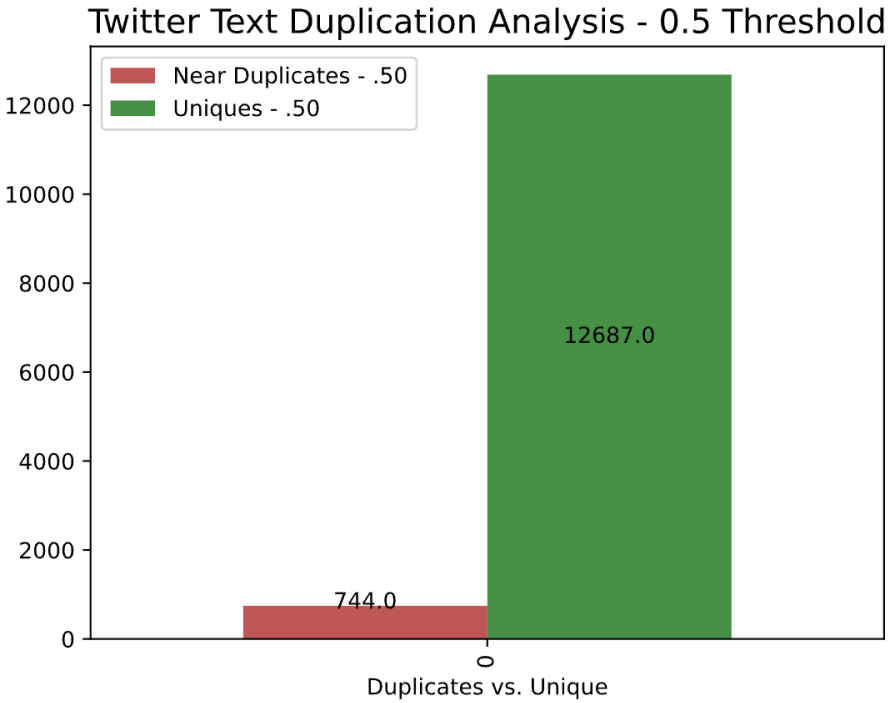
*https://www.nytimes.com/2023/02/28/technology/twitter-outages-elon-musk.html

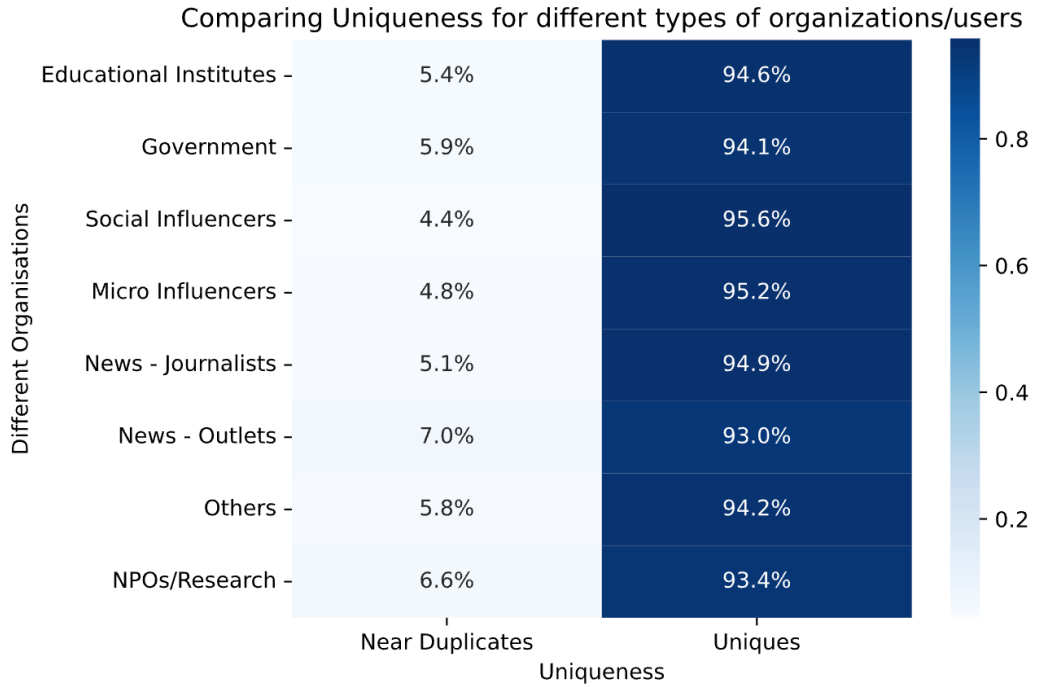
Home
Executive Summary
Methodology & Overview
Clean-Up & Filtering
Exploratory Analysis
Author Identification
Location Analysis
Timeline Analysis
Message Uniqueness
Conclusion & Insights
Tweet

Message Uniqueness Analysis*

@unique

- Emphasized tweets posted by verified twitterers for credible sources of information
- Original tweets were selected because re-tweets by default will have a high duplicity
- Because of verified posts, the duplicity between tweets was low. A Jaccard distance of 0.5 was employed to find near duplicates using count-vectorizer and locality sensitive hashing
- Within different user/organization categories, the duplication ratio was still hovering around 5-7%





Posts from verified users show high levels of uniqueness, and should be prioritized while searching for new information points

*Time frame: April 2022-Feb 2023

*Conducted on a random sample of data (n=13,664 tweets)

Conclusion and Recommendations

@insights

Search Twitter

Twitter is an effective tool for gauging important trends in education, but it comes with its own set of caveats

While the data included ~7M unique IDs, only 1% of them were verified users. Many accounts were bots as well, due to which a large chunk of data was not usable.

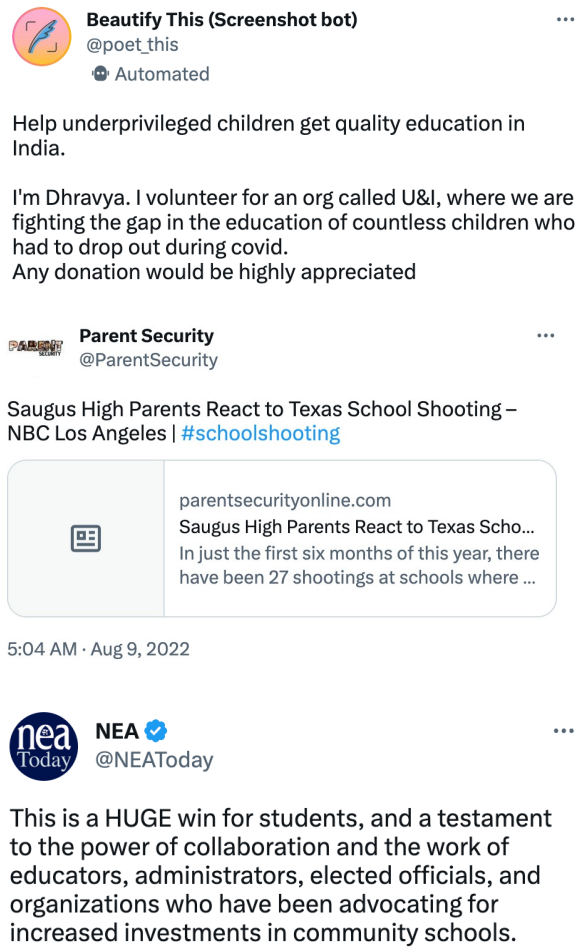
Credible source of education information is the primary criterion for this analysis. While twitter can provide horizontal breadth for this analysis, it should be supported by other verified data sources (eg: NYTimes) for vertical depth

Many tweets were not specific to education topics, but social media updates. (Example: scoreboard updates, job openings, school shootings).

School shootings, even though not a directly education related topic, provide an opportunity to broaden our analysis.
Advanced NLP techniques NLP should be employed to identify these through our data filters

United States, due to its sheer internet population size, remains one of twitter's most active userbases.
This biased some analysis to US centric topics (like student debt)

Twitter must standardize its user location tagging. Currently it is manually entered, and adds a lot of noise to location analysis. With a standardized input, more filters can be used to narrow on other geographies for a global point of view
Gathering individual Tweets coordinate should also improve



Conclusion and Recommendations

@insights

News is often categorized as a single user type.
However there is a considerable difference between news outlets and news/independent journalists.

Journalists often provide more in-depth, critical and unbiased analyses into education topic on twitter.
News channels are regulated and can often have commercial interests, which might not give the best picture.

Rajdeep Sardesai ✓
@sardesairajdeep

ASER status of education report (rural): good news: school enrolment is up despite pandemic; not so good news: reading and maths skills have declined for junior school children. A country's future rests on quality of education/healthcare and not Hindu Vs Muslim! #GetRealIndia.

Government accounts are a good source for education updates, but they can provide a biased view as per the incumbent regime

NPOs and Research organisations keep a better track at latest issues in the field. However their current influence is extremely limited, and should be emphasized upon for high quality topics.
(Eg: this UNICEF post had only 15 retweets)

UNICEF Education ✓
@UNICEFEducation

It is not enough for children with disabilities to simply attend school.

Education systems need to be inclusive to ensure all students learn and grow side by side, to the benefit of all.

15 Retweets **3** Quote Tweets **48** Likes

There are very original tweets in the data, when compared with re-tweets, quotes and replies

While original tweets provide new education topics, their fewer occurrences limited the analysis to a small subset. Retweets provided a measure for twitterer influence, and can be analysed for nuances using more computational resources

Thank You!