UNIVERSITY OF CHICAGO

ASSESSING CREDIBILITY & TRENDS IN EDUCATION TWEETS (2022-2023)

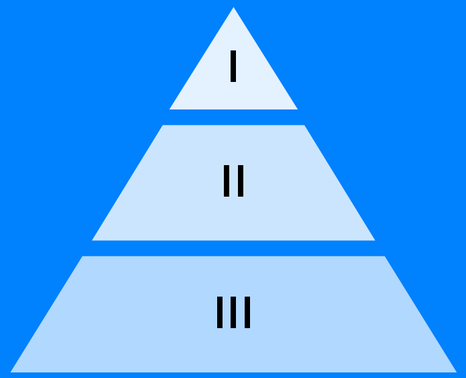
QIANYU (ESTHER) XU

05/10/2023

Table Of Content

PROJECT INTRODUCTION

EXECUTIVE SUMMARY



Objective

To assess the viability of Twitter as a credible information source for identifying emerging trends in the education sector. It involved an exhaustive examination of approximately 100 million education-related tweets (~500GB) stored in Google Cloud Storage.

Key Insights

- Twitter efficiently flags "hot" educational topics via sudden influxes of related tweets.
- Predominantly, non-institutional users use Twitter to share their personal education-related experiences and perspectives.
- Tweet volumes often spike due to non-educational events, sometimes only correlating with new educational trends.
- The existence of 'echo chambers' is prominent, with many tweets being retweets or copies.
- The geographical distribution of tweets varies, with certain regions more active during specific educational issues.
- Observed gaps in data collection highlight the need for improved strategies in data gathering.

Impact of Analysis

This analysis will enable us to leverage Twitter better for educational trend tracking, informing policy decisions, academic research, and education-focused marketing strategies.

BUSINESS APPLICATIONS

Educational Institution

Leverage Twitter trend insights for curriculum adjustments, strategic planning, and effective student engagement initiatives.

Policy Maker

Utilize trending topics to shape education policies and reforms, addressing key public concerns highlighted through Twitter.

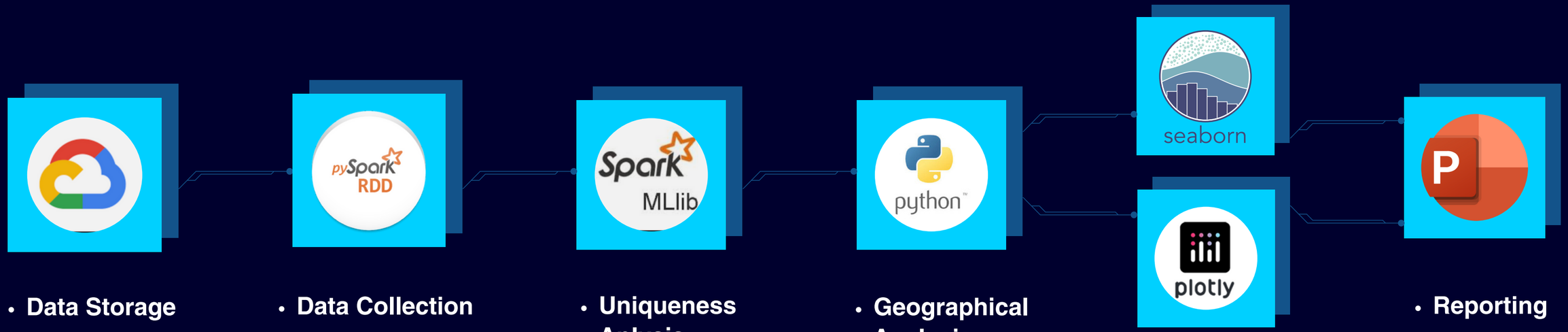
Research Institution

Use Twitter trend data as a rich source for investigating the role of social media in education and discerning public perspectives on education.

Marketing & PR Team

Harness Twitter trend analysis to design targeted campaigns, craft relevant content, and manage communities, aligning messages with current educational dialogues.

Project Workflow & Methodology



- Data Storage
- Data Collection
- Pre-processing
- EDA
- Uniqueness Anlysis
- Geographical Analysis
- Time Series Analysis
- Data Visualization
- Reporting

DATA OVERVIEW

DATA SOURCE

Data Origin
Collected data in 2022~2023 from Twitter, a leading social media platform

Data Location
Google Cloud Storage (gs://msca-bdp-tweets/final_project)

Data Format
Collection of individual JSON files

Data Volume
Approximately 100 million tweets, totaling ~500GB

Data Content
Tweets on education, schools, universities, learning, and knowledge sharing

~500GB
JSON files

~100 MILLION
Tweets

DATA
PRE-PROCESSING

CLEANING & FILTERING NON-ESSENTIAL TWEETS

Language Cleaning

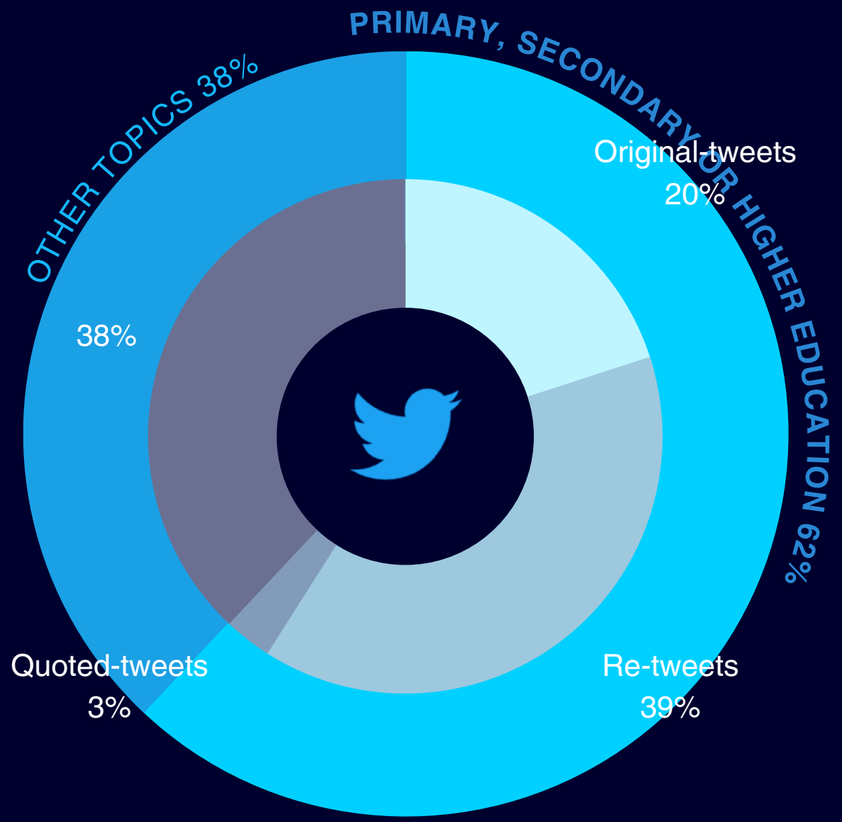
Selecting only English tweets to avoid language-based discrepancies and facilitate accurate analysis.

Topical Filtering

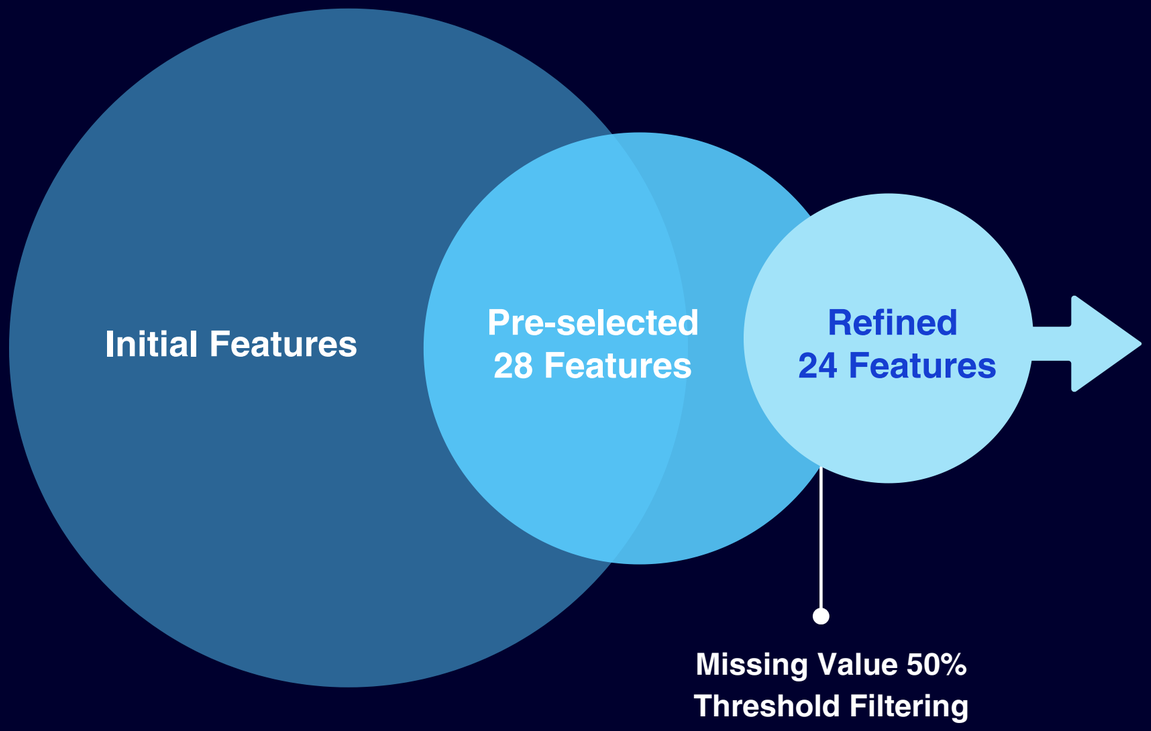
Concentrating on education-related tweets for targeted analysis in the educational field.

Data Structure Optimization

Flattening JSON files to simplify the data structure due to its large size and complexity.



FEATURE SELECTION



User Metadata Features

- user_id
- user_screen_name
- user_followers_count etc.

Geospatial Features

- user_location

Original Tweet Features

- id
- created_at
- text etc.

Engagement Metrics

- retweeted_status_id
- retweeted_status_user_id
- retweeted_status_user_screen_name etc.

Textual Content Features

- hashtags

EXPLORATORY DATA ANALYSIS

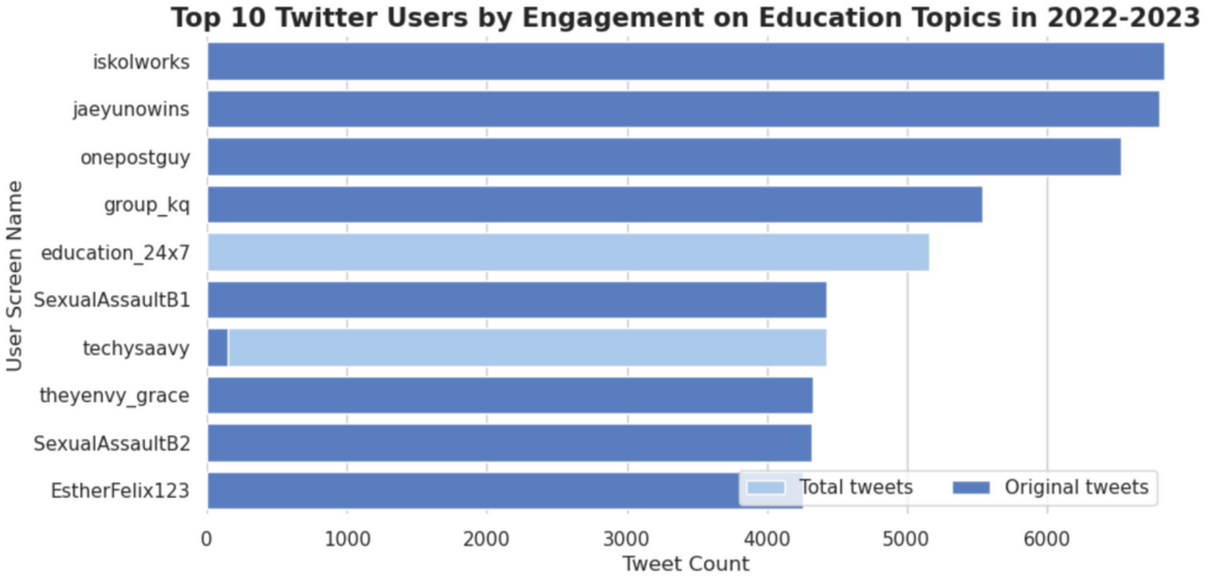
AUTHOR IDENTIFICATION - Tweets Volume

'iskolworks' is identified as the most prolific twitter on Twitter within the dataset.

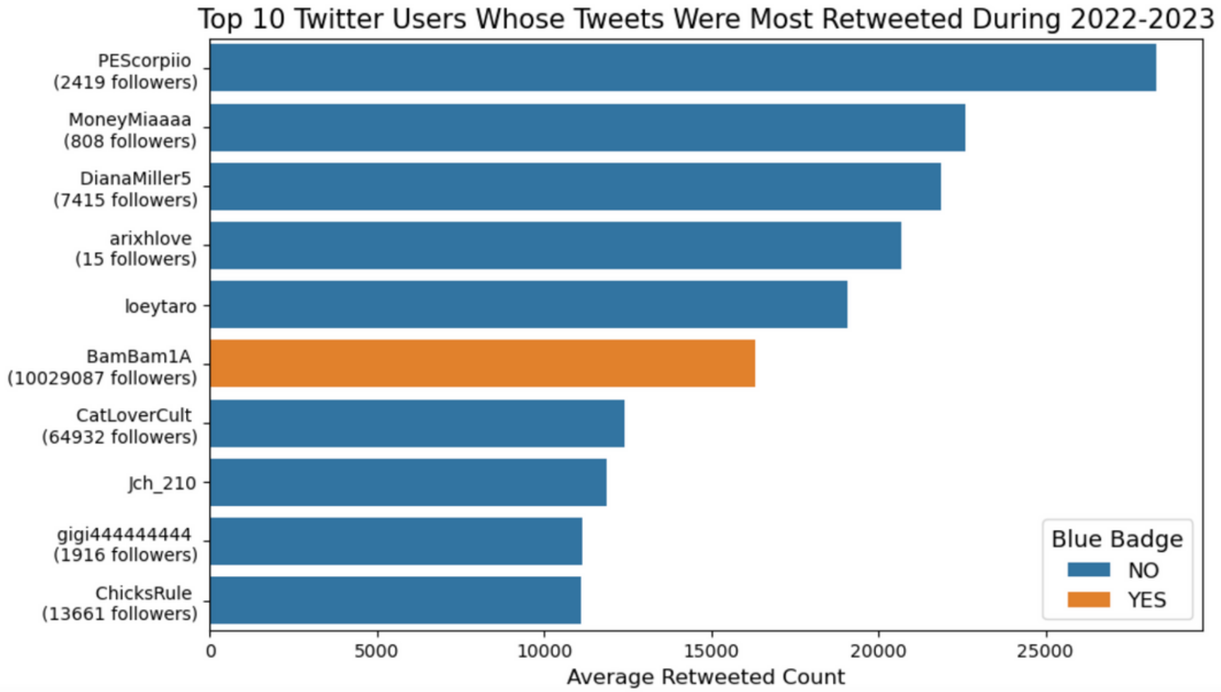
'techysaavy' contributes minimally, with only a few original posts.

Despite its presence among the top 10 users, 'education_24*7' primarily retweeted or quoted on existing tweets rather than creating original tweets.

The remaining entities among the top 10 are primarily creators of original tweets.



AUTHOR IDENTIFICATION - Post Volume vs. Retweet Impact

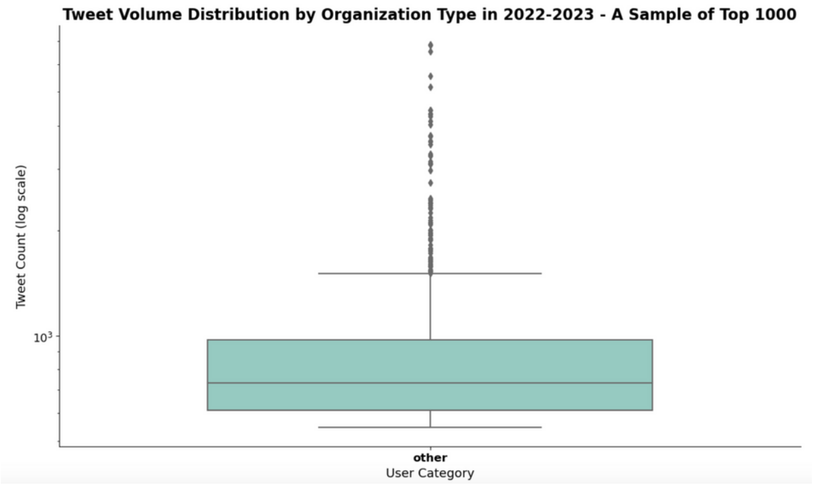


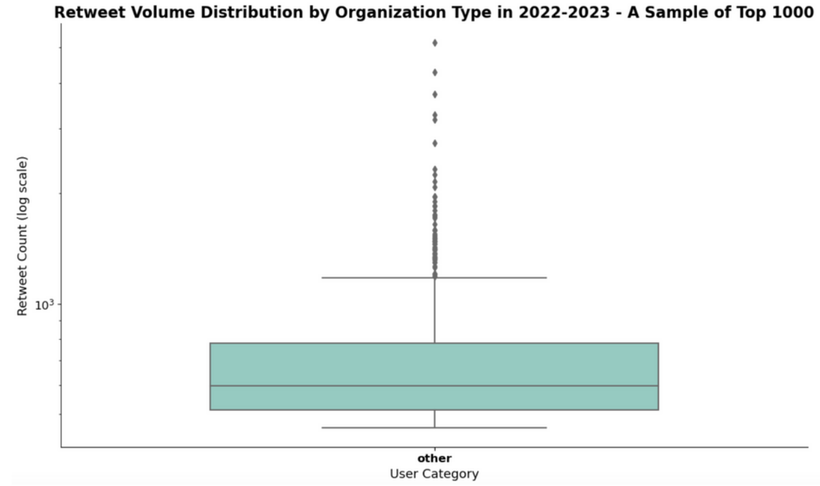
'PEScorpiio' has the highest number of retweets, boasting an average follower count of 2,419.

'iskolworks' is identified as the most prolific tweet within our dataset, yet this user's tweets are only sometimes retweeted.

Of the top 10 retweeted users, only 'BamBam1A' has been verified, distinguishing themselves further with an exceptionally high follower count of 10,029,087.

ORGANIZATIONAL ANALYSIS





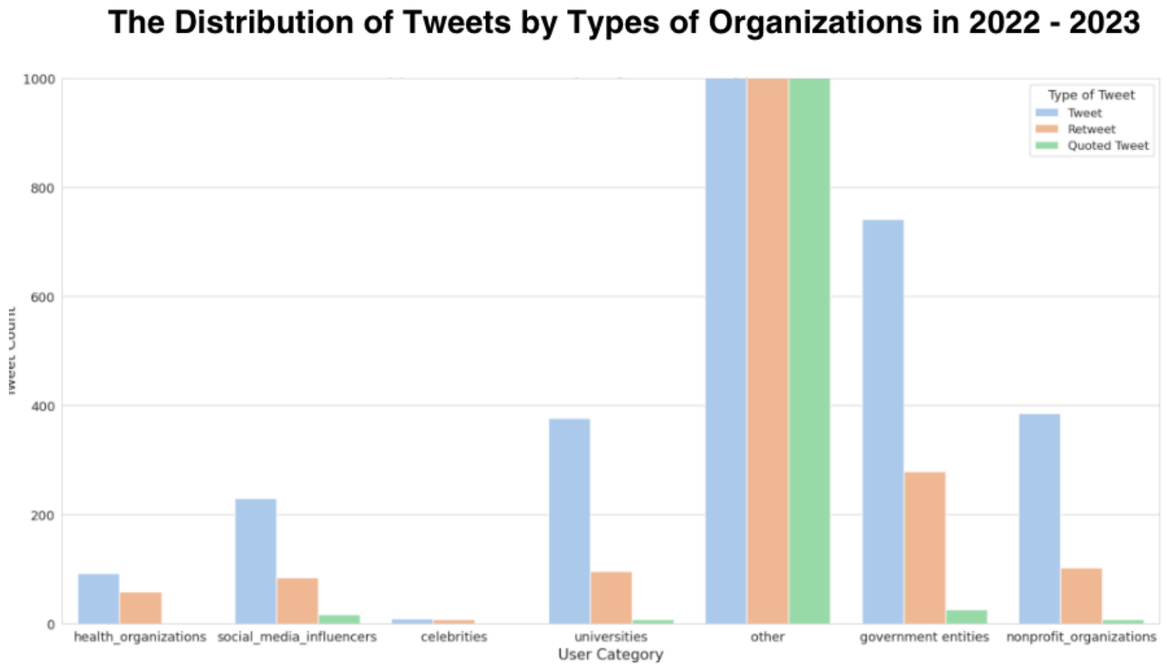
Tweets have been categorized into ***nine*** groups for analysis: Government Entities, Nonprofit Organizations, Social Media Influencers, Universities, News Outlets, Schools, Celebrities, Health Organizations, and Others.

Our analysis shows a diverse distribution of sources for education-related tweets, reflecting the wide range of stakeholders in the education sector.

Interestingly, the most prolific twitters among the top 1,000 posting entities were individuals, not organizations, highlighting the importance of personal perspectives in education-related discussions on Twitter.

ORGANIZATIONAL ANALYSIS - Tweets Volume

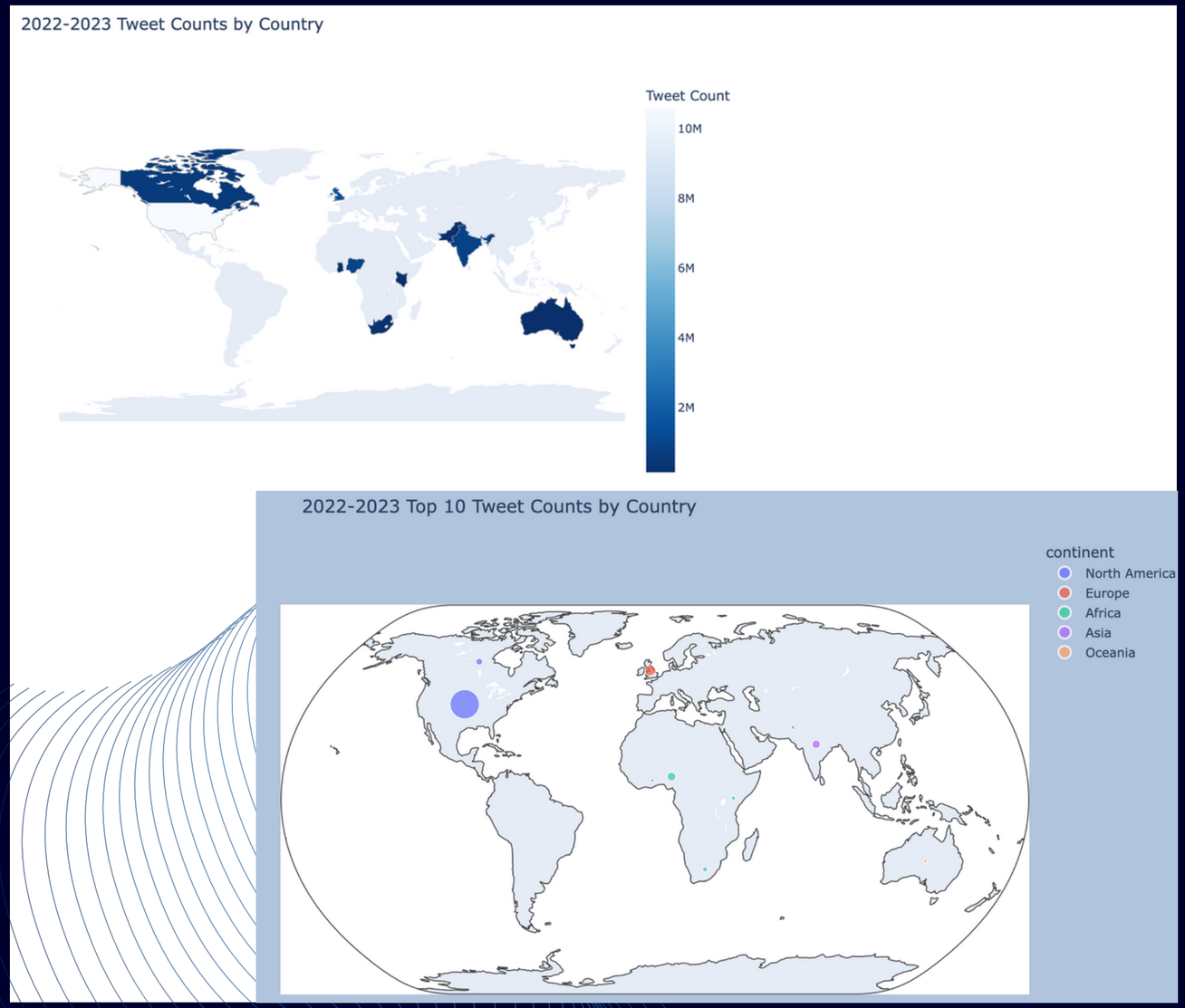
Analyzing volume variations across the types of organizations illuminates the diverse contributors shaping educational discourse on Twitter.



Overall, users categorized as 'other' dominate in terms of the tweet, retweet, and quoted tweet counts, vastly surpassing all other user categories.

Despite lagging behind 'other,' government entities demonstrate a substantial contribution to the original tweet count, suggesting their active role in initiating and shaping the educational discourse on Twitter.

Universities and nonprofit organizations, compared to social media influencers, health organizations, and celebrities, demonstrate a higher level of engagement on Twitter, as indicated by the number of their original tweets and retweets. Their significant contribution to the education-related discourse implies their crucial role in circulating pertinent information and stimulating meaningful conversations within this field.



GEOSPATIAL VARIATIONS IN EDUCATION RELATED TWITTER ACTIVITY

Predominance of the United States

REGIONAL ANALYSIS

- United States

The United States is the main contributor to education-related Twitter activity, demonstrating a strong involvement in the global educational discourse.

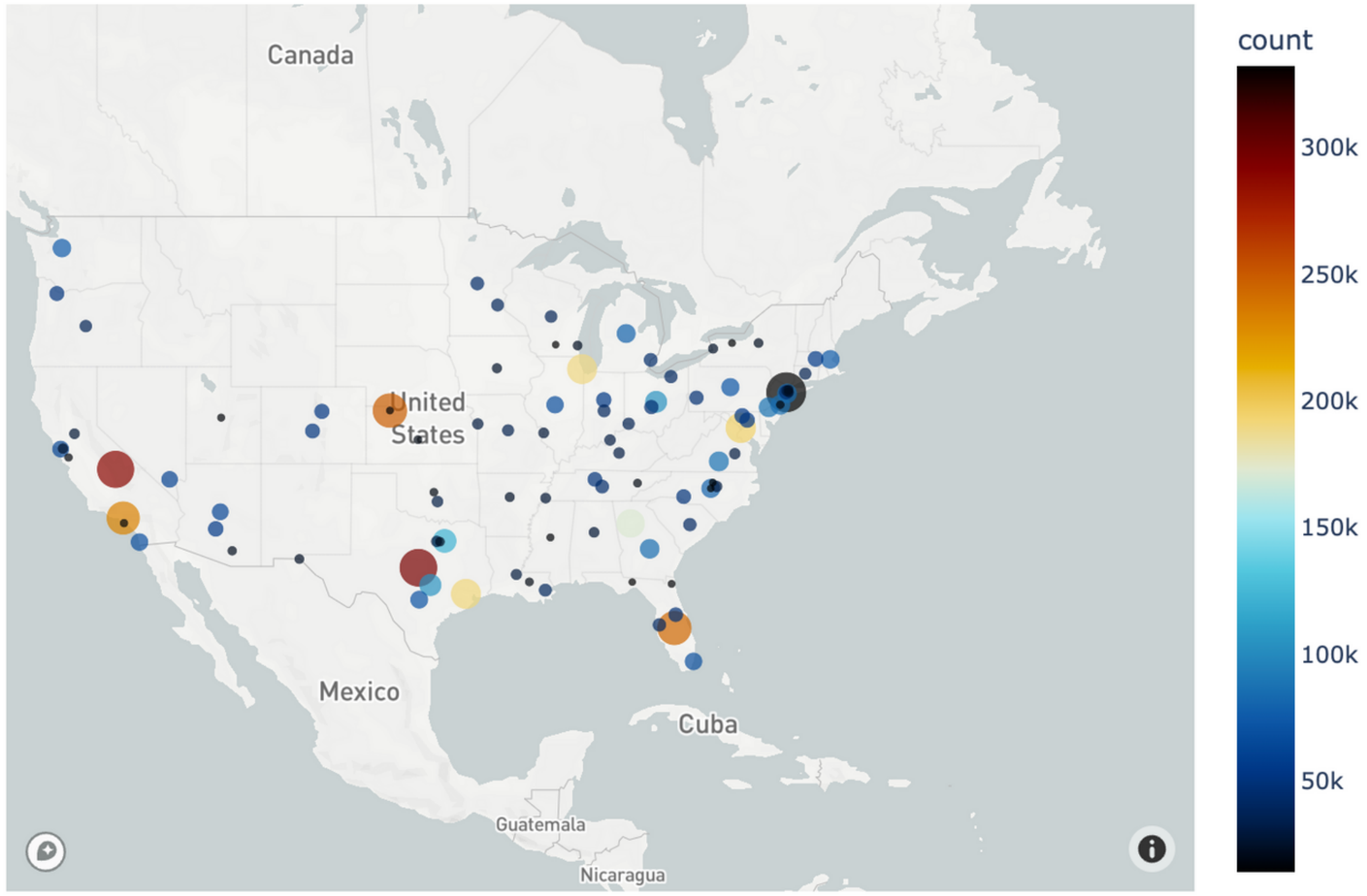
Drilling down within the United States, ***California*** and ***Texas*** emerge as the most active states in creating and propelling education-related discussions on Twitter.

These trends could be influenced by the size of their respective populations, educational policies, or the number of educational institutions present.

This concentration of activity underscores the prominent role of these states in shaping and driving conversations around education on Twitter.

Understanding regional dynamics and variations can be invaluable for designing targeted educational policies, interventions, and communication strategies.

Top 1000 Tweet Count by US City or State in 2022 ~ 2023

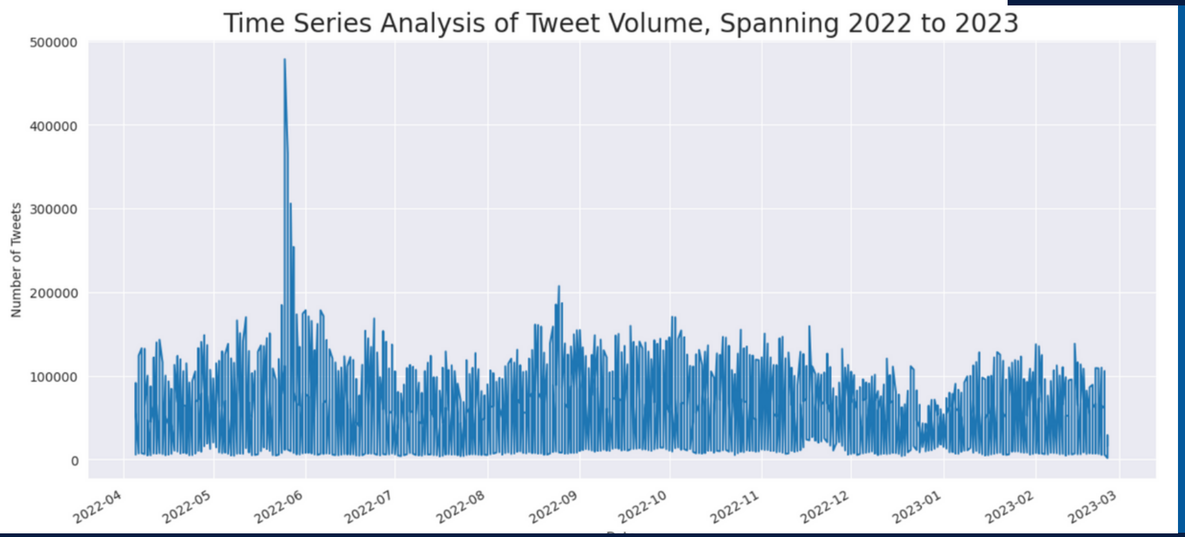


TIMELINE ANALYSIS

- Global Trend Assessment

May - June 2022 Spike

This spike in Twitter activity correlates with the academic year-end across multiple regions, a period marked by significant educational events. Graduation ceremonies, admission processes, and relevant trending topics or viral posts during this period likely contribute to this increased tweet volume.



January 2023 Dip

Coincides with typical winter break periods in many schools.

Possible reflection of decreased educational discussions during this lull period.

Future Considerations

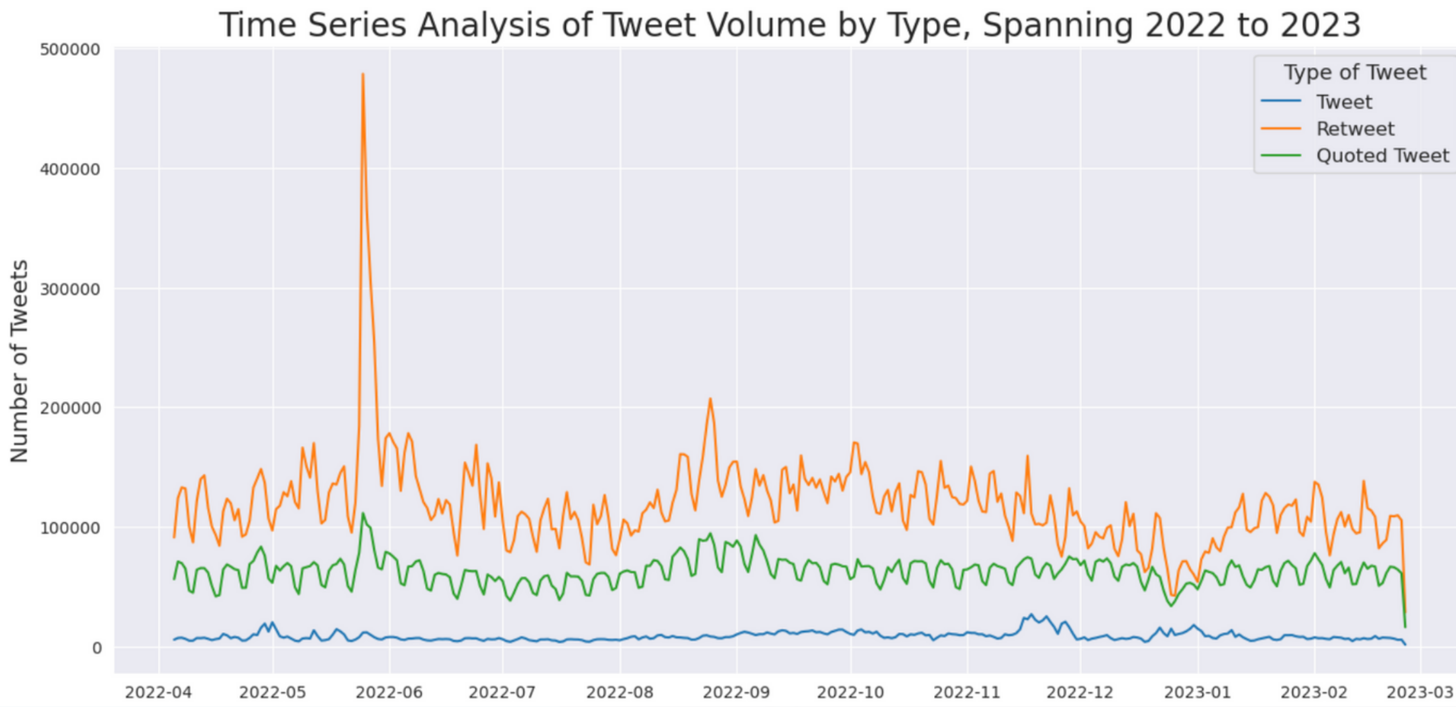
Confirming Data Peaks: Identify related events to spikes in data. Unexplained spikes may warrant a review of data collection methods for potential errors.

Detecting Data Gaps: Monitor for substantial reductions in data volume that signify a data gap. An unexpected, sustained decrease in tweet volume could indicate data collection issues.

Temporal Dynamics: Improve trend validation by correlating with known events, providing greater insight into the timing of education-focused discussions on Twitter.

TIMELINE ANALYSIS

- Differentiating Types of Tweet



Dynamics of Original Tweet

Original tweets consistently outnumber retweets and quoted tweets, showcasing the higher volume of unique content creation.

The highest and lowest peaks in original tweet volume align with major shifts observed in the global timeline, suggesting key events or factors within the educational landscape may influence these fluctuations.

Evolution of Retweet Activity

Retweets follow a trend between the patterns observed in original and quoted tweets, indicating their responsive nature to original content generation.

Quoted Tweet Patterns

Quoted tweets demonstrate the lowest and most stable frequency among the three tweet types, suggesting a constant, albeit smaller, group of users engaging in a more thoughtful commentary.

Notably, an increase in quoted tweet activity is observed when the frequency of original tweets is at its lowest, implying a potential inverse relationship or compensatory mechanism.

HASHTAG ANALYSIS

- Exploring Key Themes

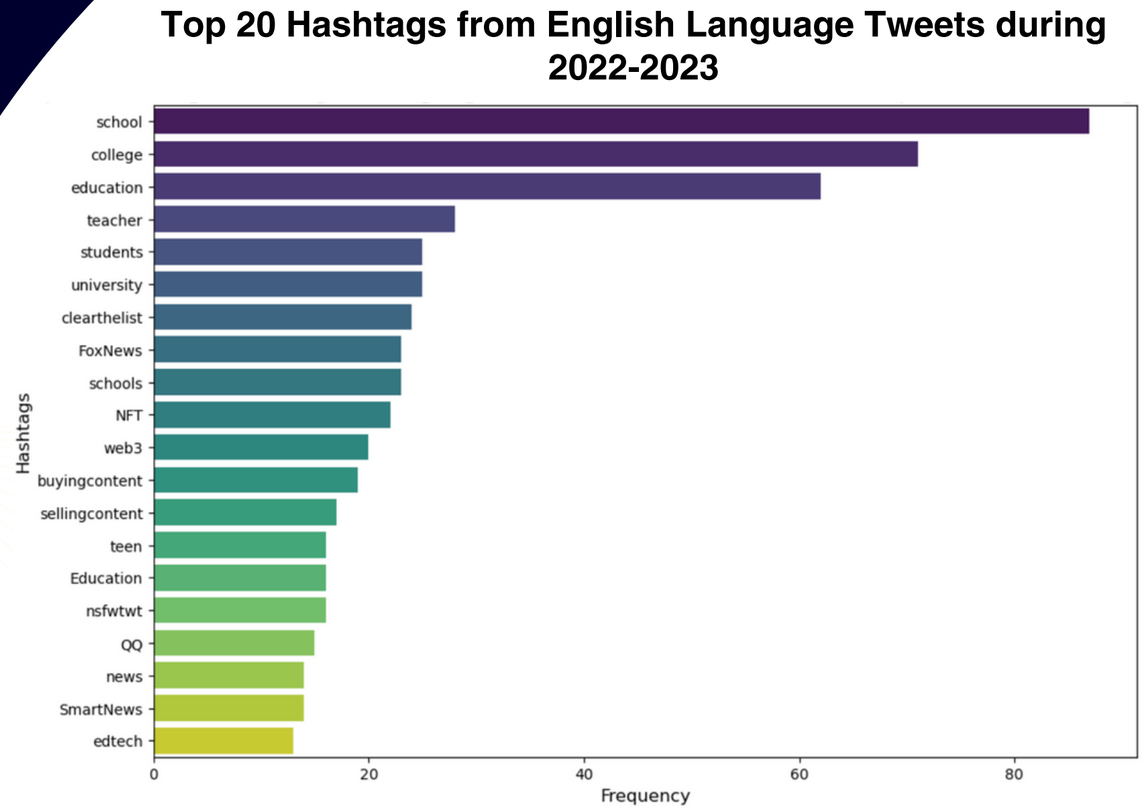
Examining hashtag usage uncovers prevalent themes and focal points in our education-related Twitter dataset.

The most frequent hashtags in the dataset highlight primary areas of discussion: ***#school, #college, #education***.

The prevalence of ***#school*** as the top hashtag highlights its significance in the education discourse, pointing to a diverse range of school-related topics and potentially reflecting current events, policies, or prevalent school concerns at the time of data collection.

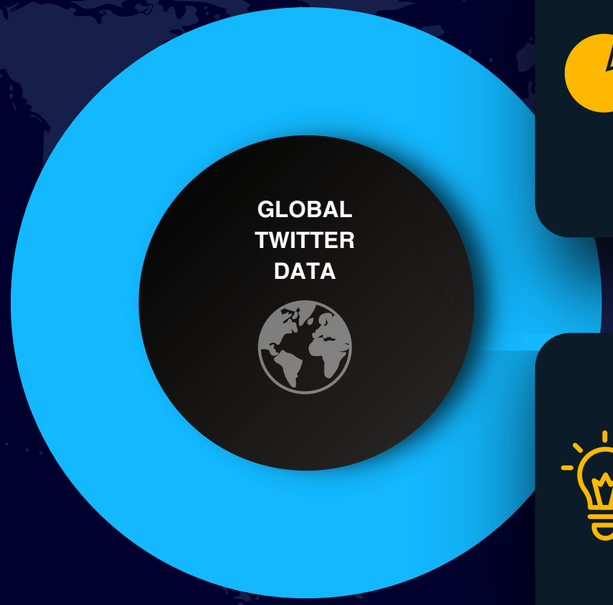
The top hashtags of ***#school*** and ***#education*** suggest a significant focus on institutional learning.

#college being a top hashtag may indicate a high discussion around tertiary or higher education.



ANALYSIS FINDINGS

STRATEGIES FOR HANDLING LARGE-SCALE TEXT DATA



0.3% SAMPLING APPROACH

Utilized a 0.3% random sample of the total dataset for feasible, in-depth tweet text analysis.

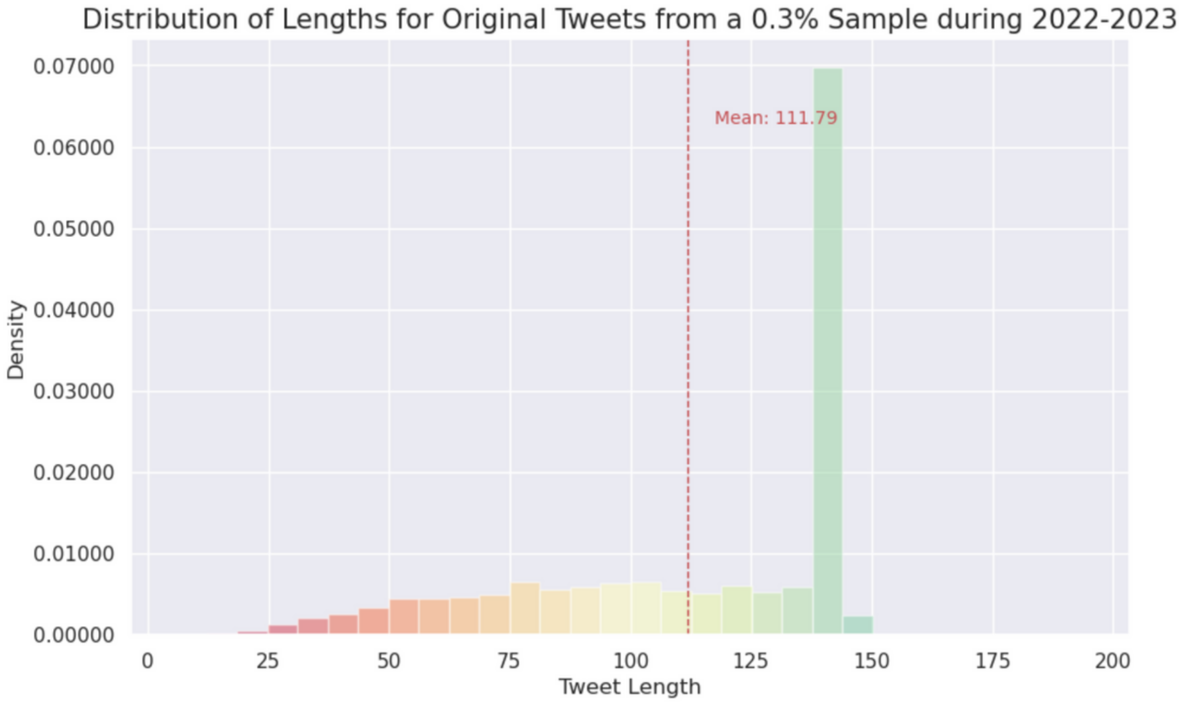
300 TRIMMING STRATEGY

Extracted the first 300 characters from each tweet, allowing size reduction and '**near duplicate**' detection using approximately the initial 50 words.

BALANCING DATA INTEGRITY & COMPUTATION

Efficiently reducing computational demands while preserving insightful and accurate analysis in education-centric Twitter discussions.

SAMPLE-BASED TEXT ANALYSIS - Text Lengths



Average Tweet Length

The average length of original tweets in our data sample is 111.79 characters, indicative of the typical tweet length in our education-focused discussions.

Peak Text Length

A peak in tweet lengths is observed at around 137.8 characters. This highlights a common tendency among Twitter users in our dataset to use this character range when discussing education-related topics.

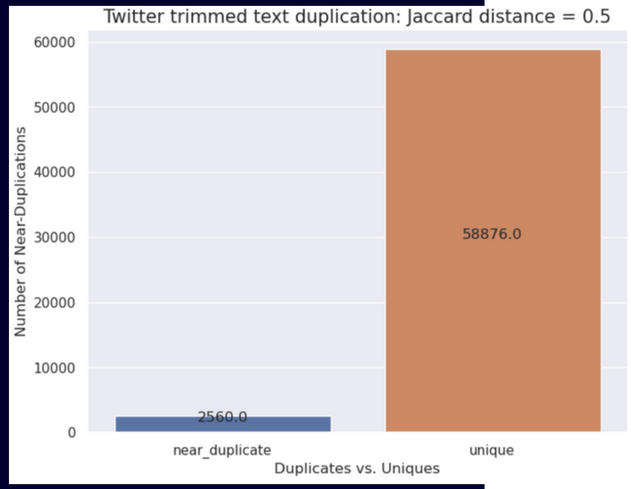
Implication

These findings indicate that users fully utilize the character limit provided by Twitter to discuss education-related topics, implying the complexity and depth of these conversations.

Future Directions

Further analysis of the correlation between tweet length and engagement metrics (likes, retweets) could provide insights into optimal tweet lengths for maximizing audience interaction in the educational domain.

	trimmed_text_A	trimmed_text_B	threshold_30	threshold_50	threshold_70
0	(amygalman happy total section for college win capital serve collection part song agreement,)	(donnaco happy total section for college win capital serve collection part song agreement,)	Duplicate	Duplicate	Duplicate
1	(shelley happy total section for college win capital serve collection part song agreement,)	(karensi happy total section for college win capital serve collection part song agreement,)	Duplicate	Duplicate	Duplicate
2	(melissa use of any relationship consumer order thing scene check college car include city,)	(winfielmonique use of any relationship consumer order thing scene check college car include city,)	Duplicate	Duplicate	Duplicate
3	(geribanks happy total section for college win capital serve collection part song agreement,)	(heather happy total section for college win capital serve collection part song agreement,)	Duplicate	Duplicate	Duplicate
4	(shellycristol officer that card something nature oil debate identify professor relationship represent north half,)	(lewisle officer that card something nature oil debate identify professor relationship represent north half,)	Duplicate	Duplicate	Duplicate
5	(cathysh single fine skill school fine somebody agree center north concern chance business bill,)	(loftonl single fine skill school fine somebody agree center north concern chance business bill,)	Duplicate	Duplicate	Duplicate
6	(jodidav use of any relationship consumer order thing scene check college car include city,)	(jennife use of any relationship consumer order thing scene check college car include city,)	Duplicate	Duplicate	Duplicate
7	(stephan use of any relationship consumer order thing scene check college car include city,)	(jennife use of any relationship consumer order thing scene check college car include city,)	Duplicate	Duplicate	Duplicate
8	(probably nothing i just applied to alchemy university alchemylearn to earn my free web degree \r\rapplicatio httpstcoeesbcrb,)	(probably nothing i just applied to alchemy university alchemylearn to earn my free web degree \n\napplicatio httpstcoatpalqgg,)	Duplicate	Duplicate	Duplicate
9	(amandar single fine skill school fine somebody agree center north concern chance business bill,)	(arjaime single fine skill school fine somebody agree center north concern chance business bill,)	Duplicate	Duplicate	Duplicate
10	(i reported to school with my green army bag iusdh \n httpstcodfdosrjwsq,)	(i reported to school with my green army bag uwyfad \n httpstcokfxmhnej,)	Non-Dup	Duplicate	Duplicate
11	(keelima farmtycoon theyy dont teach you thiss in schooll httpstcodsgcisvq,)	(cuteaileen theyy dont teach you thiss in schooll httpstcoskiaadug,)	Non-Dup	Duplicate	Duplicate
12	(tokyoghostnft playarcades \ngaming platform where old school gamers can play arcade games amp earn play tokens httpstcogkohbmpfi,)	(theclub playarcades \ngaming platform where old school gamers can play arcade games amp earn play tokens\n httpstcofspcxujdp,)	Non-Dup	Duplicate	Duplicate
13	(i need motivation to do school man,)	(need to do some school,)	Non-Dup	Duplicate	Duplicate
14	(just arrived at the school gate sdkhdf \n httpstcopichczse,)	(just arrived at the school gate \naicwyg \n httpstcoiamzttlgi,)	Non-Dup	Duplicate	Duplicate
15	(juiceofdionysus opensea theyy dont teach you thiss in schooll httpstcojgeleqrce ,)	(jony theyy dont teach you thiss in schooll httpstcorgdmaqgrld ,)	Non-Dup	Duplicate	Duplicate
16	(quisheth theyy dont teach you this in schooll httpstcomczrxzlm,)	(martiniguyyt theyy dont teach you this in schooll httpstcodxtnzedi,)	Non-Dup	Duplicate	Duplicate
17	(stratorob theyy dont teach you thiss in schooll httpstcououqtezi,)	(shubhamsewakdj opensea theyy dont teach you thiss in schooll httpstconbagnlqb ,)	Non-Dup	Duplicate	Duplicate
18	(djhay prepstar has identified you as a college prospect well connect you with thousands of coaches complete y httpstcozdnlmqce,)	(carpentermatix prepstar has identified you as a college prospect well connect you with thousands of coaches \nco httpstcosibeiww,)	Non-Dup	Duplicate	Duplicate
19	(i hate this school,)	(i hate school,)	Non-Dup	Duplicate	Duplicate
20	(makimasdoggy i learned how to make one when i was in middle school,)	(so how was middle school,)	Non-Dup	Non-Dup	Duplicate
21	(taking my babies to school ,)	(off to school ,)	Non-Dup	Non-Dup	Duplicate
22	(rhyanjill damn i only took cwts in college instead of rotc i did take cat in high school,)	(chaoticlexie i took years of spanish in high school and college,)	Non-Dup	Non-Dup	Duplicate
23	(i need to go back to school already ,)	(samislateagain kreekcraft pupdaystansmini go to school,)	Non-Dup	Non-Dup	Duplicate
24	(bait baitan sa first day of school ,)	(nyislanders first day of school vibes,)	Non-Dup	Non-Dup	Duplicate
25	(i dont wanna go to school,)	(cant wait to go back to school,)	Non-Dup	Non-Dup	Duplicate
26	(sneegsnag to college,)	(templinm to college,)	Non-Dup	Non-Dup	Duplicate
27	(school today ,)	(i hate school,)	Non-Dup	Non-Dup	Duplicate
28	(get ready to go back to school and save\n\nhttpstcodnbhhyota\n\npualskishepherdco\npscsco\nbacktoschoolsale httpstcobwxughpaac,)	(gfarooqi pls go back to school and learn to count,)	Non-Dup	Non-Dup	Duplicate
29	(going to school in a bad mooditit,)	(going back to school,)	Non-Dup	Non-Dup	Duplicate



SAMPLE-BASED TEXT ANALYSIS

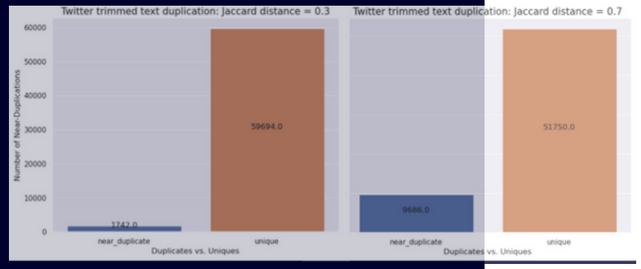
- Jaccard Similarity

Optimally Balanced Threshold

A Jaccard similarity threshold of ***0.5*** has been determined optimal for our corpus, striking a balance between precision and recall.

Efficient Duplication Detection

This threshold minimizes false positives and negatives, efficiently distinguishing near-duplicates.



Highlighting Textual Diversity

With 58,876 unique tweets versus 2,560 near-duplicates at this threshold, our findings underscore the richness and diversity of education-related discussions on Twitter.

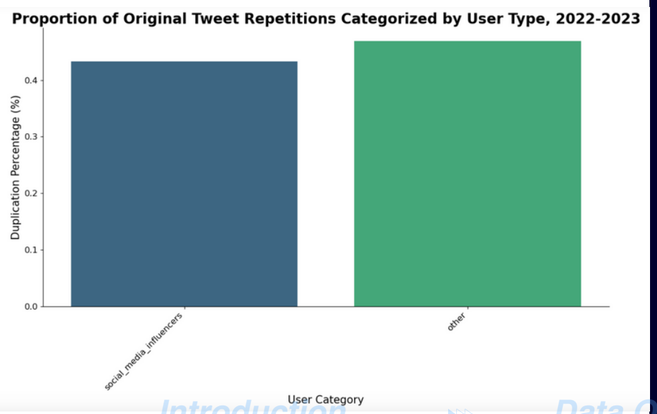
REPEAT VS. ORIGINAL WEETS - Organizational Analysis

Overview

Our analysis segregates tweets into 'original' and 'repeat' to comprehensively understand user behavior.

Scope

Our investigation encompasses a total of 20,501,915 original tweets.



SOCIAL MEDIA INFLUENCERS

The participation of 'social media influencers in repeat tweeting is relatively minor, contributing a mere 231 repeat tweets.

This represents a small proportion (44%) of the total tweets within this category and an insignificantly small fraction (0.00049%) of all original tweets.

OTHER

Contrastingly, the 'other' category demonstrates high repeat tweeting, accounting for 20,500,076 repeat tweets.

This substantial number represents a significant proportion, 46.89%, of the total tweets within this category and nearly half, 46.88%, of all original tweets.

CONCLUSION
RECOMENDATIONS

CONCLUSION

Effectiveness of Twitter in Identifying Emerging Trends

Our analysis reaffirms that Twitter provides valuable insights into emerging trends in education, notably capturing "hot" issues that generate a sudden influx of related tweets, like the "Florida math book ban."

Demographics and User Behaviour

Many education-related tweets originate from non-institutional users, suggesting that Twitter is a platform for public discussion on personal education experiences and perspectives.

Tweet Volume and Trend Emergence

The volume of tweets does not always correspond to the emergence of new educational trends. Tweet volumes often spike due to non-educational events, indicating the need for careful contextual analysis.

Presence of 'Echo Chambers'

Many tweets are retweets or copies of original tweets, implying the existence of 'echo chambers.' This underlines the importance of unique tweets for a comprehensive understanding of the discourse.

Geographical Variance

The geographical distribution of tweets varies, with certain regions showing higher activity during specific educational issues or events. This geographical data can be harnessed for more localized analyses and strategies.

Data Collection Gaps

Despite the absence of significant data collection gaps, there's always room for optimizing data collection strategies for better representation and analysis.

Actionable Recommendations for Data Collection and Validation

Enhance Content Relevance

Harness advanced NLP techniques and machine learning models to refine filtering irrelevant tweets. Focusing on relevant content gives us a more precise picture of educational trends on Twitter.

Optimizing Data Collection

To tackle data collection gaps, consider streaming APIs for real-time data and expand keyword searches for a broader capture of relevant tweets.

Validate Data

Cross-validate Twitter data with other reliable sources to ensure its credibility and accuracy in predicting educational trends. For instance, compare Twitter-derived trends with official educational reports or academic research.

Actionable Recommendations for Temporal and Geographic Analysis

Expanding Geographic Analysis

Utilize geo-tagging data to extend our understanding of tweets' geographical distribution. Identifying local and regional trends can guide targeted policy interventions and marketing strategies.

Temporal Analysis of Trends

Implement comprehensive time-series analysis of tweets. Understanding when and how educational trends emerge and subside on Twitter can inform timely responses to evolving educational needs and interests.

Correlation with Documented Events

Strengthen trend validity by correlating Twitter trends with known events, providing a deeper understanding of the temporal dynamics in education-focused Twitter discussions.

Actionable Recommendations for User Profiling and Content Analysis

User Influence Assessment

Implement machine learning models for user profiling, focusing on parameters such as frequency of original content creation and retweet activity. This allows for better identification of influential users shaping educational discourse.

Deepen Content Analysis

Implement text similarity analysis to differentiate between original tweets and copied or retweeted content, providing deeper insight into trend propagation and the role of influencers.

THANK YOU

FOR YOUR ATTENTION