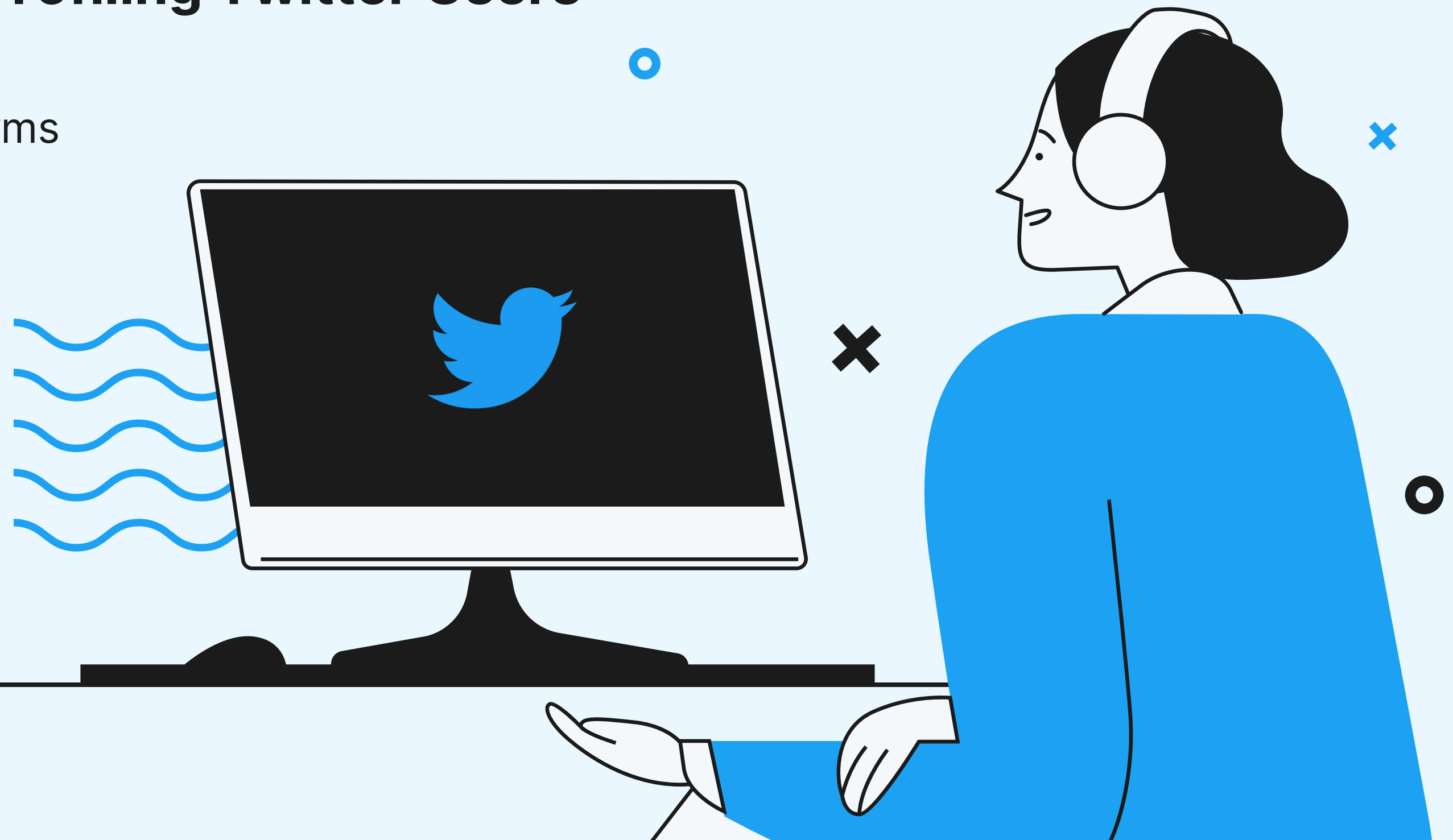


Educational Trends on Twitter

Big Data Analysis on Profiling Twitter Users

MSCA 31013 Big Data Platforms
Final Project



SHIJIA (SILVIA) HUANG

Agenda



Executive Summary

High-level summary of the background and project objectives

Methodology and Source Data Overview

Techniques utilized and overview of the data source

Data Preparation and Exploratory Data Analysis

Data preprocessing, cleaning, filtering, and exploration

Author Identification

Profiling the Twitter users and the types of organizations they belong

Location Analysis

Geographical distribution of Twitterers and relationship with hot topic emergence

Timeline Analysis

Timelines of tweets and data collection gaps investigation

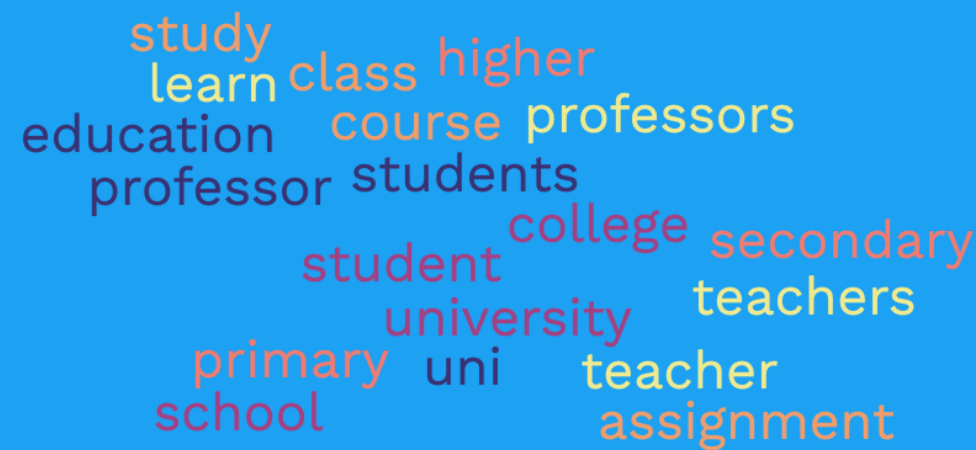

Message Uniqueness Analysis

Tweet similarity analysis for all Twitter users and by types of organizations

Conclusions and Recommendations

Analysis summary and making actionable recommendations

Executive Summary



study
learn
class
higher
education
course
professors
professor
students
student
college
secondary
university
teachers
primary
uni
teacher
school
assignment

Background and Context

- As one of the world's most popular social networking sites, Twitter allows users (i.e., Twitterers) to broadcast short posts known as **tweets**.
- As of 2022, Twitter has **450 million** monthly active users (MAU) worldwide with over **500 million** tweets sent per day.
- Twitter has become one of the most relevant **discussion platforms** for news and popular events happening worldwide, and people on Twitter are avid news consumers to **stay informed**.

Problems and Questions

- Given the diverse backgrounds of Twitter users, it is **questionable** whether Twitter can be considered a **credible source of information** that accurately reflects the emergence of important **trends or topics**.
- Who are those people that are tweeting about things like **education**? Does their activity relate to the emergence of a new hot topic in education? Do their tweets offer **insightful** information on these topics?

Resolution and Next Steps

- To address those questions, I will profile these Twitter users by performing analysis on a large amount of educational-related Tweets data based on **user organization, location, timeline, message uniqueness**, etc.



Methodology and Source Data Overview



Processing Big Data on Google Cloud Platform (GCP)

Large Volume of Source Data with Complex Structure

- Around 100 million raw tweets (~ 500GB) are stored in deeply nested individual JSON files on GCP.
- Only a small subset of fields were properly populated that contain valuable information.
- Tweets were collected on topics like education, schools, universities, learning, etc.
- Not all of them were directly related to the topic of interest (i.e., primary, secondary, or higher education).

Methodology to Profile Twitterers

- **Author Identification:** Identify the most prolific Twitterers based on the original tweet and retweet volume and find what organization they belong to based on users' self description.
- **Location Analysis:** Find where these Twitterers are and whether it is related to the emergence of new issues where they are located.
- **Timeline Analysis:** Identify when tweets were posted to discover patterns such as peaks, valleys, and gaps in data collection.
- **Message Uniqueness Analysis:** Identify the tweets' uniqueness based on Jaccard similarity and Minhash for all tweets and by types of organization.



Data Preparation and Exploratory Data Analysis



The following steps were performed before conducting the analysis

Data Preprocessing

- Read multiple JSON files on GCP into a single Spark DataFrame and parse each field appropriately for further processing.
- Make sure the schema was maintained as nested columns by checking a small sample.

Data Cleaning

- Print the data frame for the root and also each nested column to see the actual values.
- Discard irrelevant columns such as URLs and entities that are unrelated to the question of interest to decrease the data complexity.

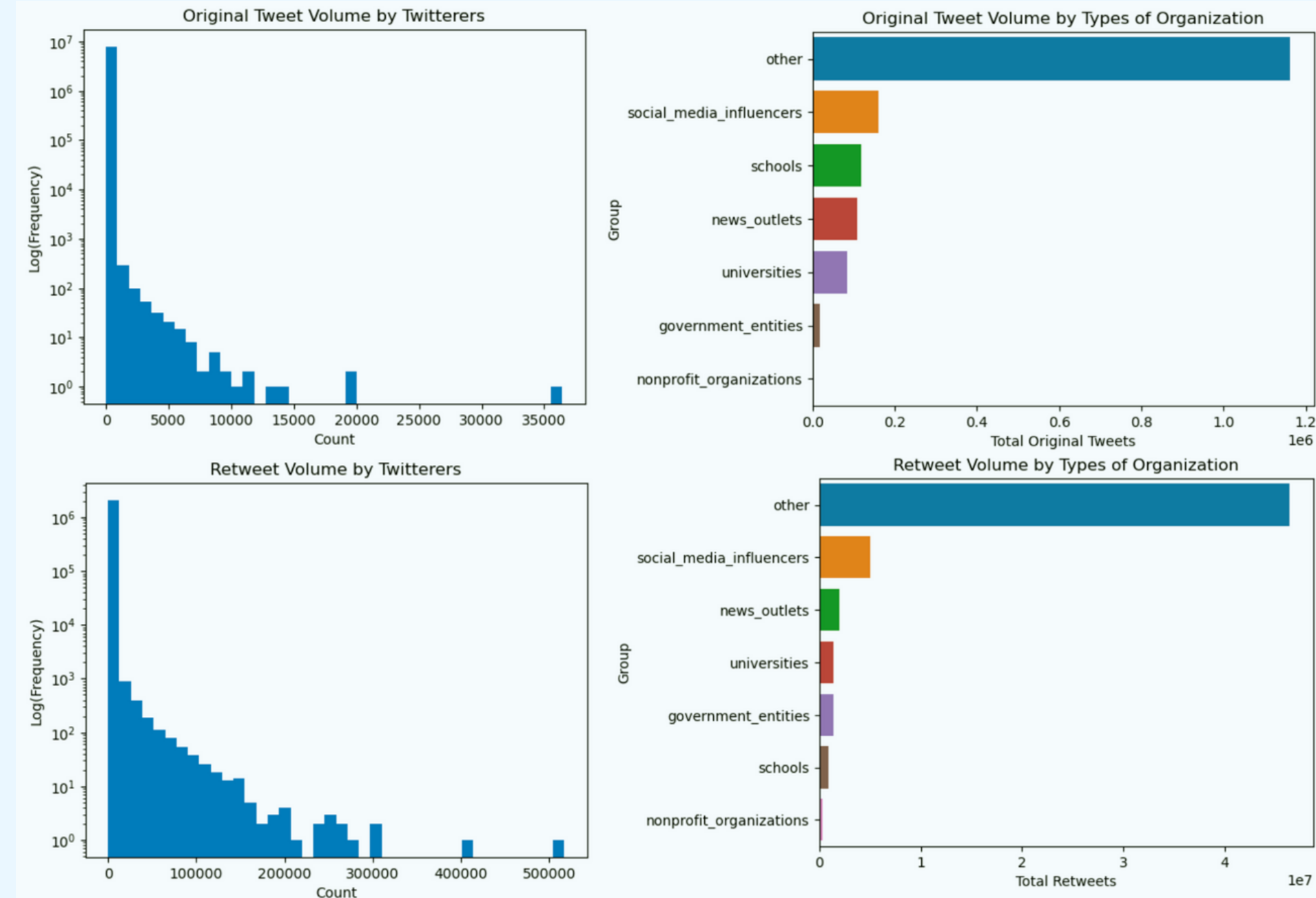
Data Filtering

- Clean the tweet texts and remove mentions, hashtags, stopwords, etc. then tokenization.
- Filter tweets related to primary, secondary, or higher education based on token words and keywords.
- Save the filtered tweets as Parquet files for next steps.

Exploratory Data Analysis

- Simple plot on relevant variables and noticed many of them were badly populated (most nulls):
 - favorites count
 - followers count
 - friends count
 - statuses count

Author Identification

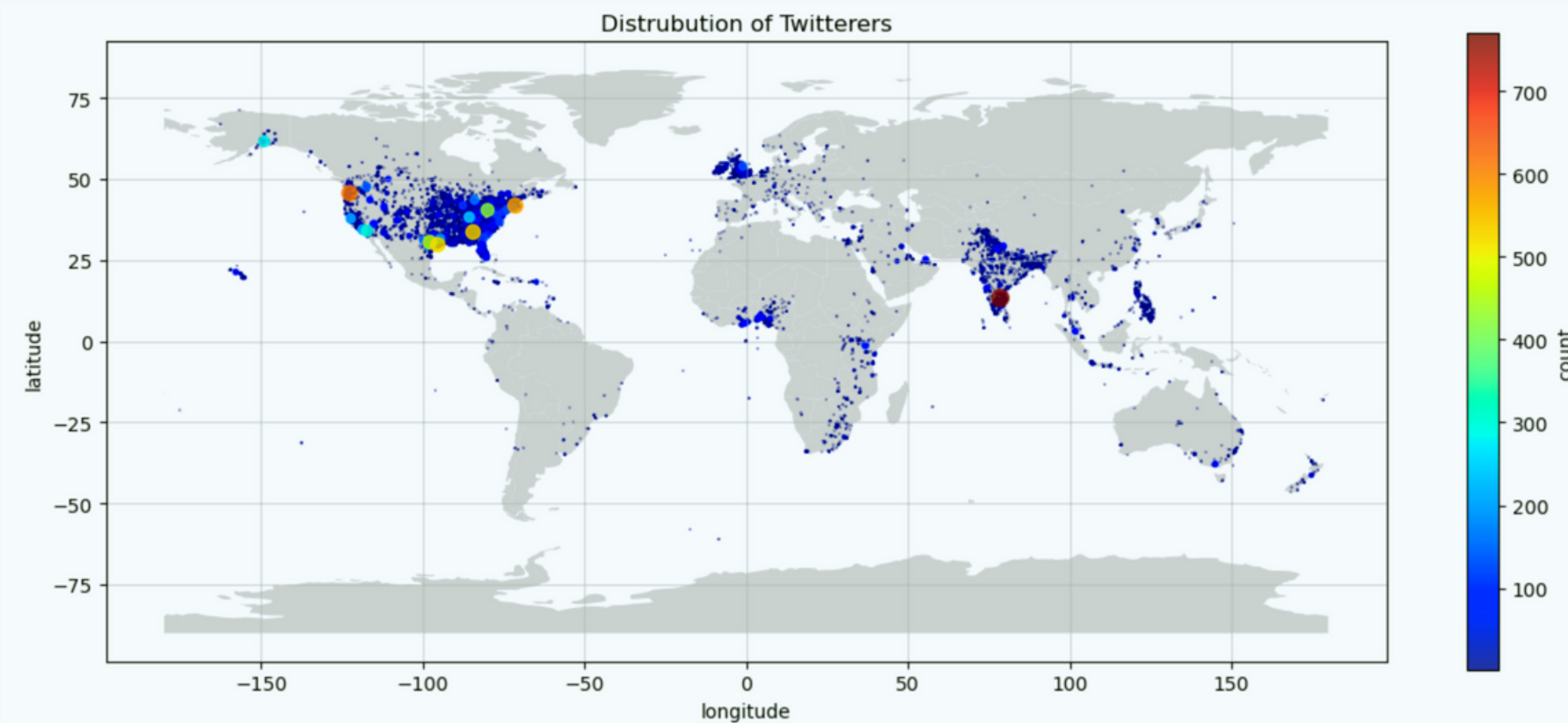


- The original tweets and retweets volume were highly skewed across all Twitterers. A small number of Twitterers contributed to the majority of Twitter activity.
- Social media influencers were the largest identified contributor to tweets and retweets.
- Schools and universities tend to tweet more original content than retweets.
- Governments and nonprofits tweets less than schools and universities on education in general.

Location Analysis

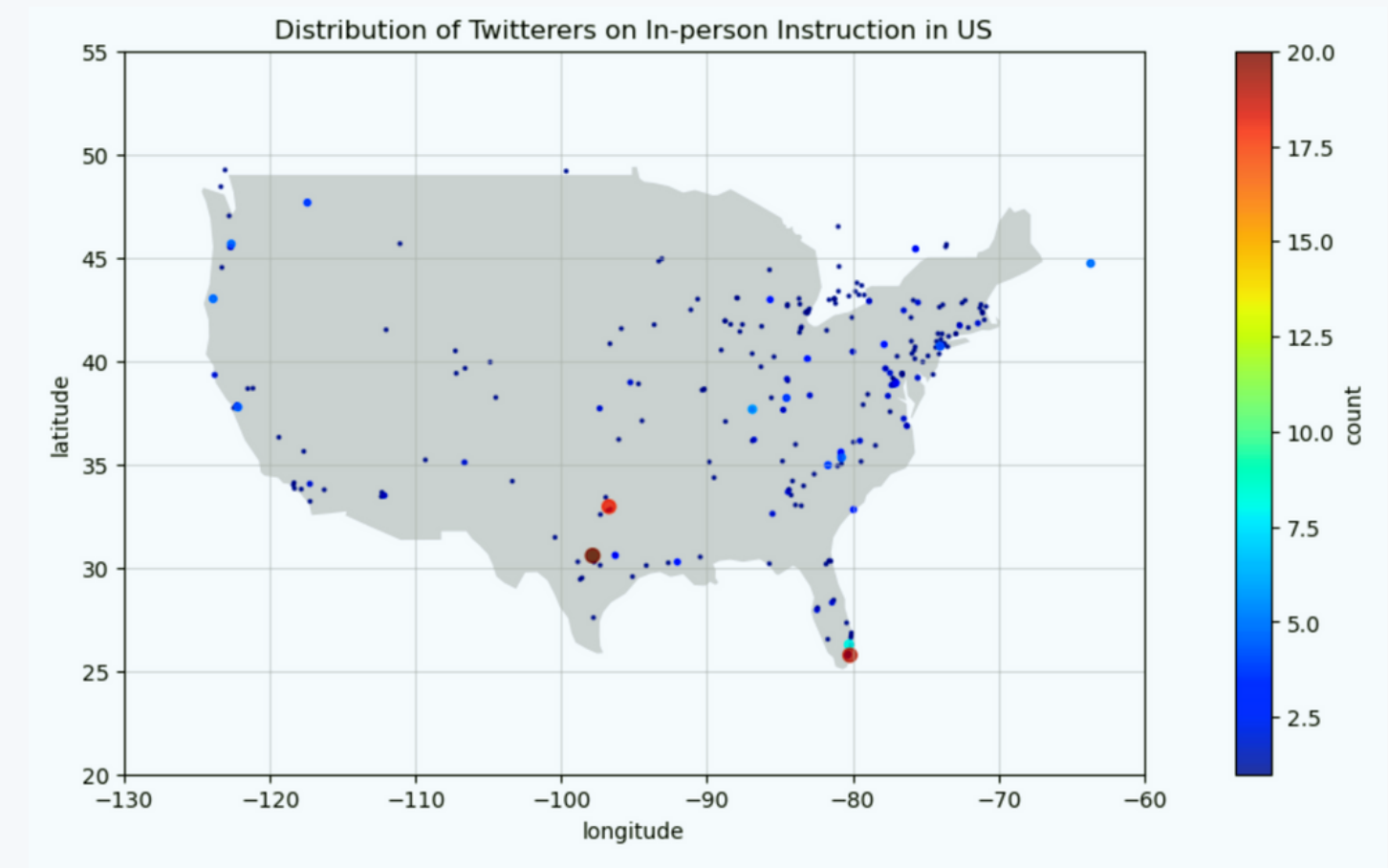
Twitterers who are talking about '*In-person Instruction*' are mostly in the United States

By filtering out Twitterers who are discussing topics like '**In-person Instruction**', I noticed that most of them are located in the US. Taking a closer look, many of them were from Florida, Texas, and North East where schools were specifically ordered open for the 2021-2022 school year (source: EducationWeek).



The Majority of Twitterers are Located in the United States, India, and Europe

Specifically, most Twitterers are located on the East and West Coasts of the United States. There are also big user groups in South India and West Europe where there are big cities.



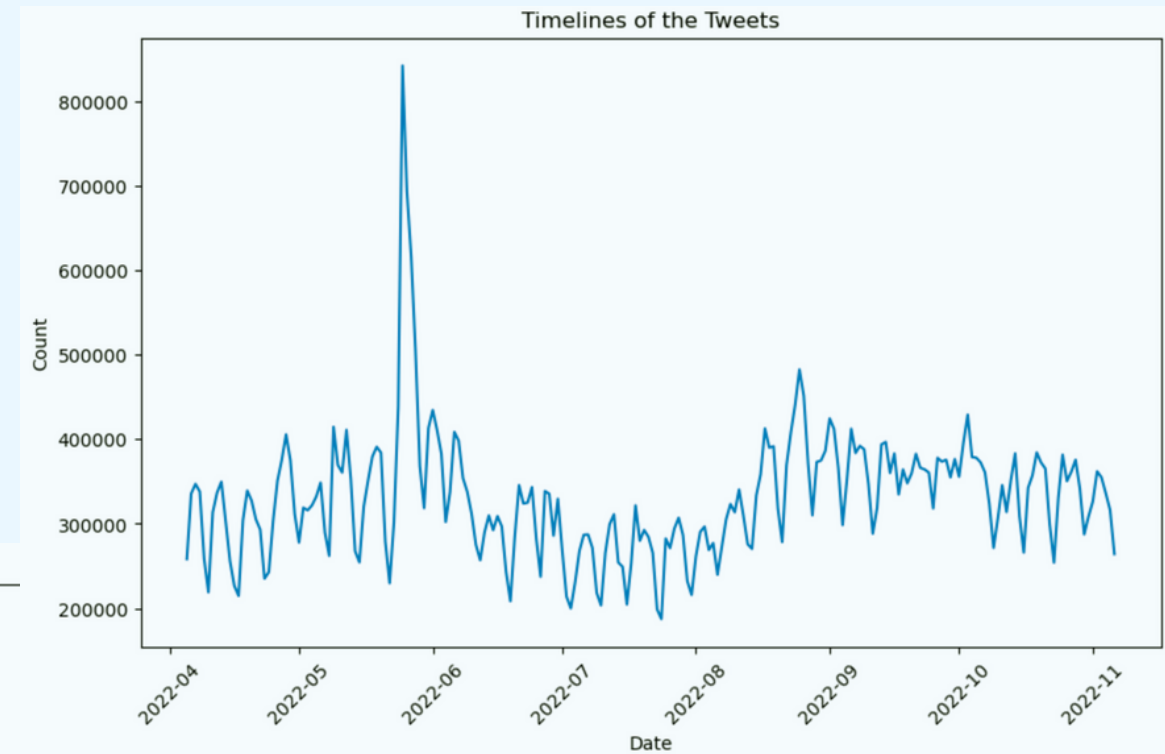
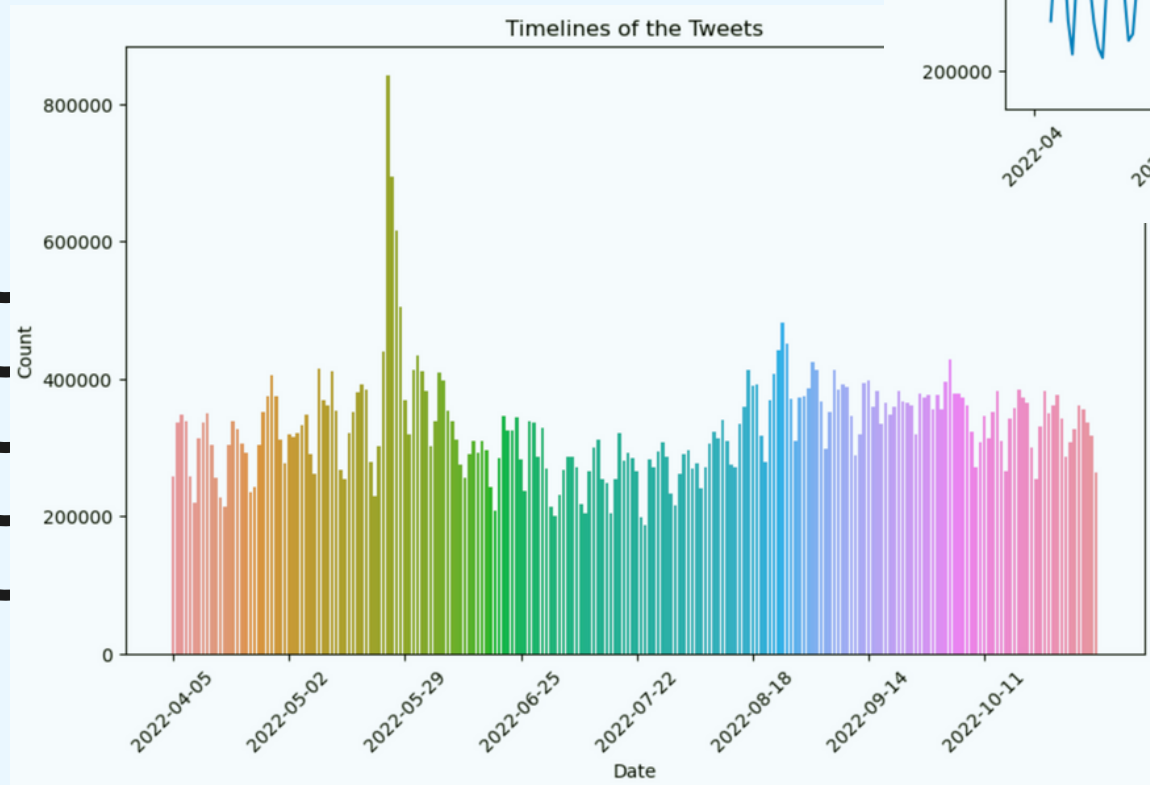


Timeline Analysis



Timeline of the Tweets

The Tweet data were collected on posting time from April to November 2022.



Peaks and Valleys on Tweets Volume

Midway through May 2022, during Teacher Appreciation Week and the UNESCO World Higher Education Conference, there were significant tweet surges.

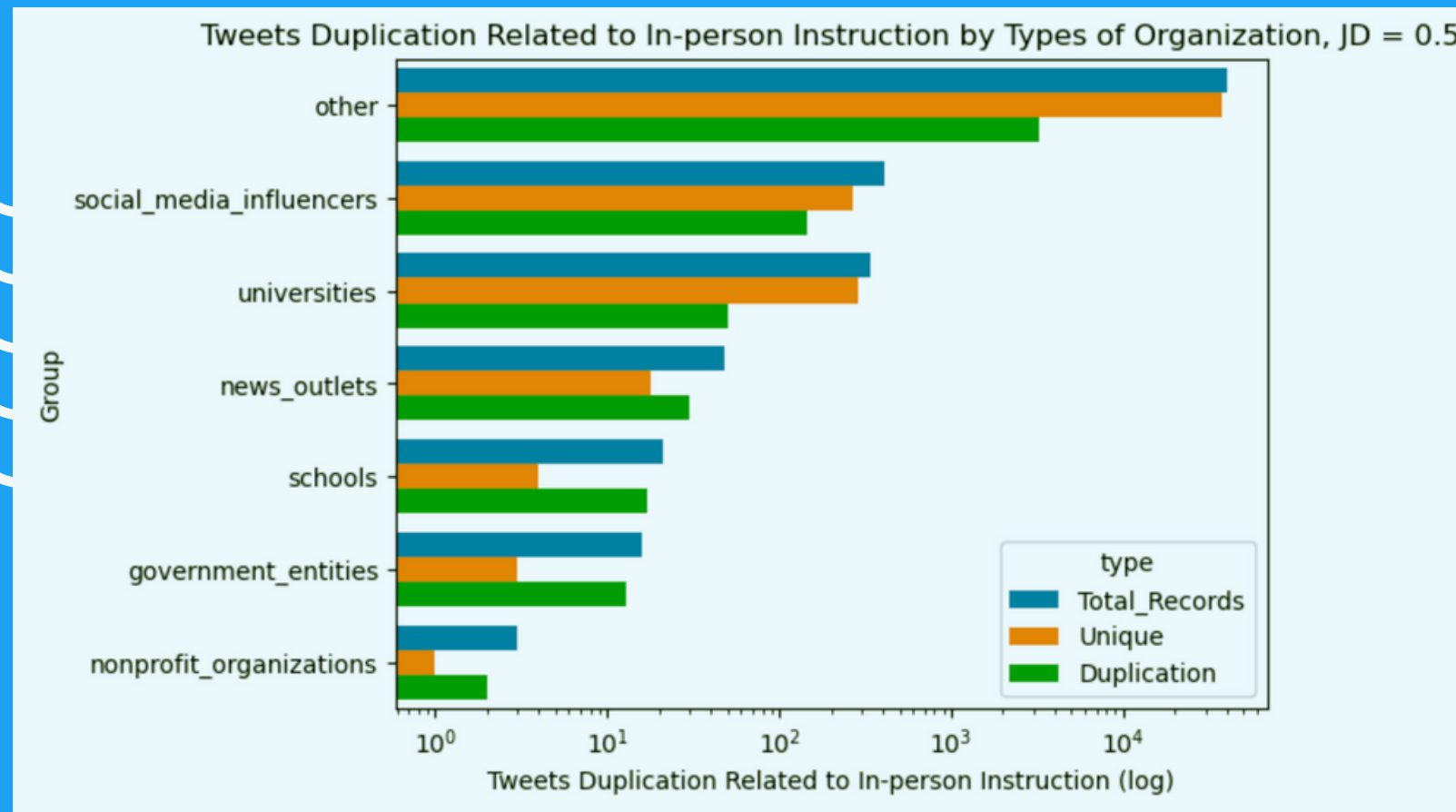
When summer vacation was in session, from June to August 2022, a brief dip followed the peak.

Data Collection Gaps

There were no significant gaps in the data collection process, albeit there were several days with no tweets, which could be the consequence of data cleaning and filtering.



Message Uniqueness Analysis



Message Uniqueness for All Tweets

- In general, most of the tweets are unique instead of copy-pasting the same text.
- With a Jaccard Distance of 0.5, 92% of tweets were unique and 8% were duplicates.

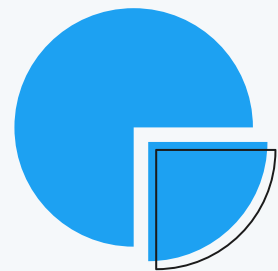
Methodologies

- The tweets are all for the topic '**In-person Instruction**'.
- Compare message uniqueness for all tweets using MinHash based on low, medium, and high Jaccard Distance.
- Decided to choose **Jaccard Distance = 0.5** since it best distinguished unique and duplicate tweets.
- Then classify the types of organization for each twitterer and then compare message uniqueness **within each group**.

Message Uniqueness for Each Twitterer Group

- Social media influencers and universities tend to tweet more unique messages than copying text from Twitterers in the same group.
- While news outlets, schools, governments, and nonprofits tend to copy tweets from users in the same group.

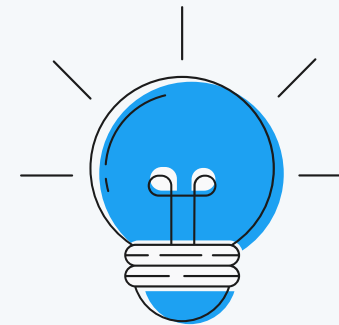
Conclusion & Recommendation



Higher tweet volumes can partially reflect the emergence of important trends or topics in education since information on Twitter is biased toward opinions from users in developed countries.

Recommendation

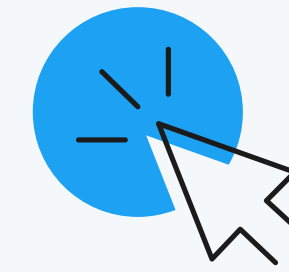
When gathering information about global issues regarding education, we should also look at other sources along with Twitter to get an objective understanding of the problems.



Tweet and retweet volumes can be greatly influenced by big educational events or academic calendars (i.e., summer break), and social media influencers can greatly drive the discussion on education.

Recommendation

Be careful when there is a heated discussion on Twitter since it can be initiated by social media influencers. A deeper investigation of the discussion topic should be conducted.

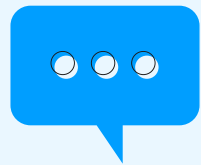


Based on user segmentation and message uniqueness, Twitter can be a valuable source of information on public opinion on education as opposed to official propaganda and advertisement.

Recommendation

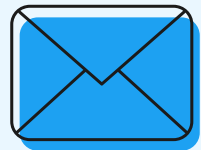
Tweets can be used to understand general people's opinions using techniques such as sentiment analysis after filtering out tweets related to the topic of interest.

Thank You!



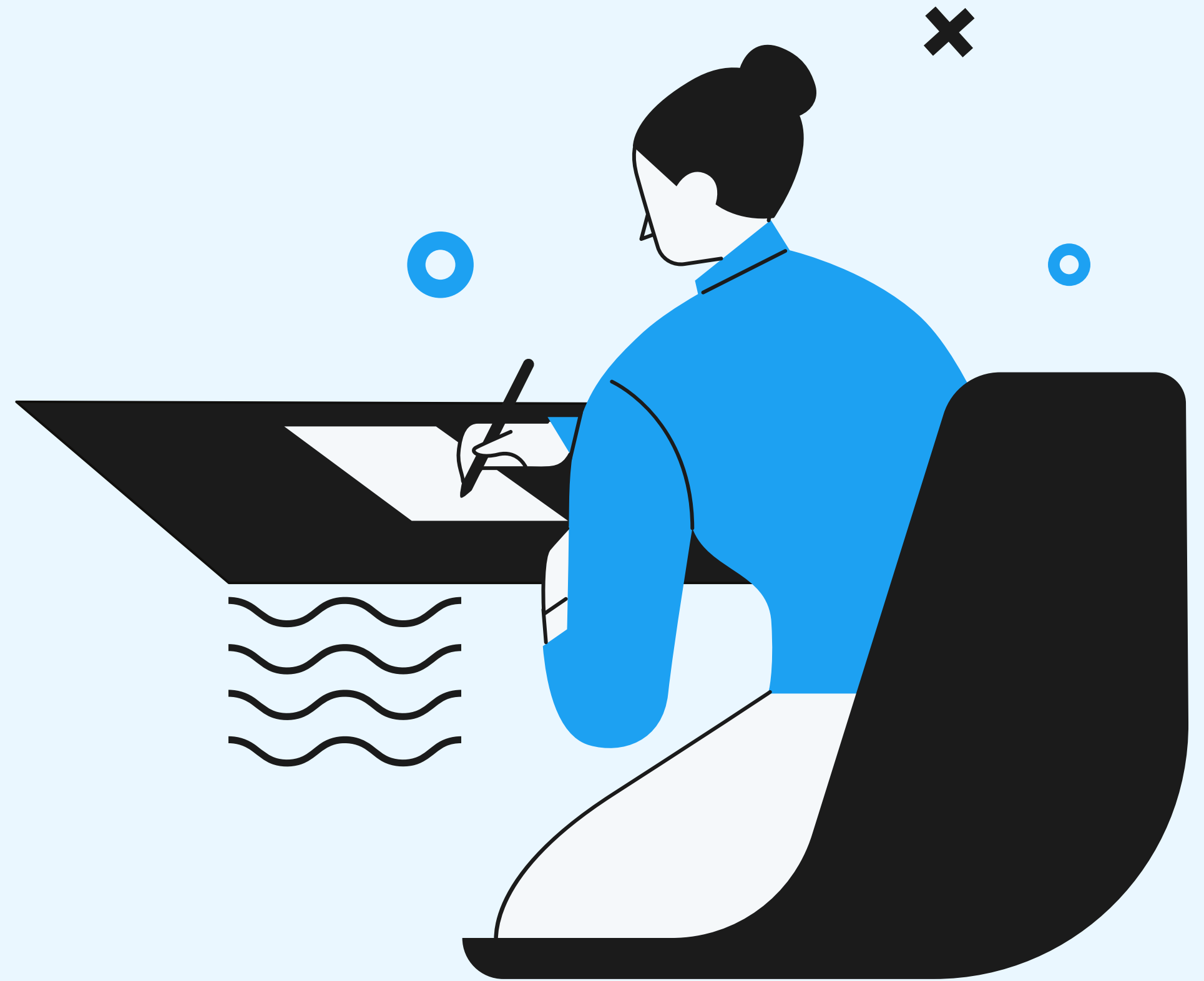
Info

Shijia (Silvia) Huang, University of Chicago



Email

shijia@uchicago.edu



References

1. <https://www.demandsage.com/twitter-statistics/>
2. <https://www.edweek.org/leadership/map-where-are-schools-closed/2020/07>

