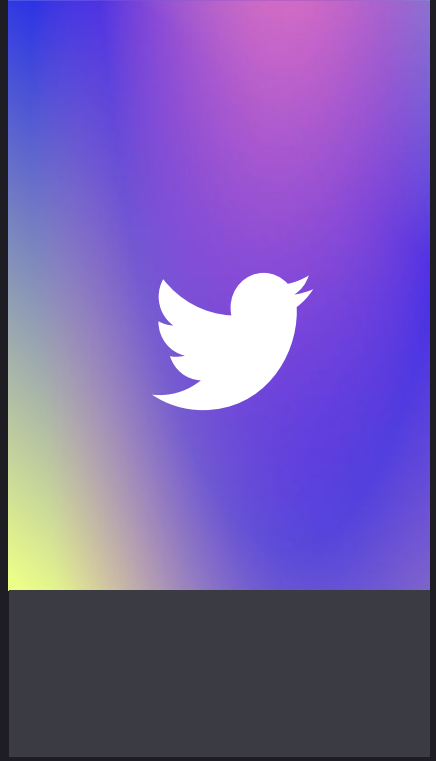
BDP Final Project

Understanding the reliability of Twitter as a source of information on Higher Education costs



Monica Zhang
March 5, 2023

TABLE OF CONTENTS

After examining millions of tweets on the cost of Higher Education, we demonstrate Twitter is not a credible source of information

The cost of Higher Education in the US has become a major topic of discussion in recent years. High tuition fees and the student debt crisis have sparked heated conversations among policymakers, educators, and the general public.

$1.757 trillion

The amount of student loan debt in the United States as of 2023[1]

43.4 M

The number of Americans with outstanding student loans[1]

Unsurprisingly, Twitter has been a hub for discussion on the topic. With millions of users expressing their opinions and sharing information, we aim to investigate the validity of information being shared by Tweeters.

After analysis on tweet authors and their location, timeline of tweets, and message uniqueness we find that Twitter is a platform with emphasis on real-time, short-form, and repetitive content that can lead to incomplete or biased perspectives. For these reasons, Twitter cannot be considered a reliable source of information.

[1]Hanson, Melanie. "Student Loan Debt Statistics" EducationData.org, February 10, 2023, https://educationdata.org/student-loan-debt-statistics

Multiple methodologies were used to unpack 500 GB of data

Source Data Overview

- 100M tweets stored in JSON files
- 500 GB of data
- 22 columns containing information on Tweet, User, Place, and Entities objects
- Data schema is deeply nested, and each object contains multiple nested fields or properties
- Tweets have been created from Apr 2022 to Feb 2023

Methodologies

- The analysis was performed using a Dataproc cluster in Google Cloud Platform for distributed and efficient computation
- PySpark, Pandas, Matplotlib, Folium, AWOC, and other packages were used to gain insights from the dataset.
- The analysis was first performed on a small sample of the data, and then it was scaled through modular code
- Multiple independent notebooks were used

Analysis Type	Approach
Author identification	Text mining and keyword filtering based on Tweeters profile name, description and url.
Location	Mapped Tweeters location field to countries in the world and US states and visualized geographical distribution.
Timeline	Analysis on created_at field. Queries using SQL and plots through Matlibplot.
Message uniqueness	Locality-sensitive hashing (LSH). Used CountVectorizer, Tokenizer, StopWordsRemover to analyze original tweets and plotted results through Matlibplot.

An extensive list of keywords was used to filter the tweets

The word cloud showing the words and hashtags used.

Tweets filtering

1. The analysis is focused on tweets related to the costs of attending universities. This includes topics such as *student debt, financial aid, graduate student pay, loan forgiveness, college affordability, university profit, etc.*

2. In order to find tweets related to the topic, over 370 keywords were chosen.
   a. The keywords were selected by doing research on these topics and picking words from news articles, as well as researching hashtags on Twitter. ChatGPT was also used to increase the number of keywords.

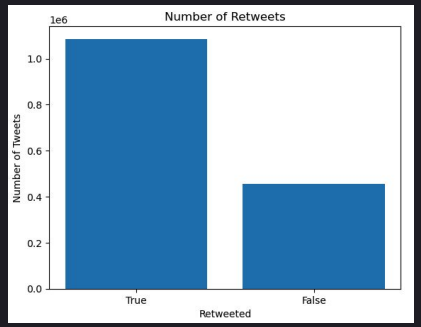
3. After filtering, **~1.58M** tweet were used for further analysis.

Tweets clean up

Only tweets in English language were used, and all rows with empty tweet text and user details were removed.

Exploratory Data Analysis allowed us to uncover important variables

- During EDA, we discovered many useful attributes related to the Tweets
  - *created_at:* date the tweet was created
    tweet_text: text of the tweet
- However, to fully utilize the information we needed to extract attributes from nested fields.
- Many fields had similar names and potentially similar information, but only specific ones carried relevant information for the analysis.



EDA Finding: 70.4% tweets are retweets

Below are examples of fields that were extracted and details on their usage:

USER	user.name, user.screen_name, user.description, user.verified, user.location	Useful for user identification and location analysis
PLACE	place.country, place.name, place.place_type	Sounded promising for location analysis, but mostly contained null data
RETWEETED_STATUS	retweeted_status.retweet_count, retweeted_status.user.screen_name, retweed_status.user.verified,	Used to calculate retweet volume. Many other fields contained retweet counts but this one is the one providing relevant information.

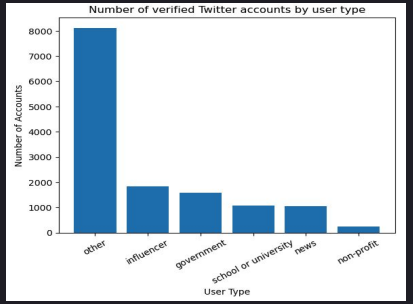
A user_type column was created to identify the type of authors

user_type	Definition
Government	Verified accounts with keywords such as *senator, congressman, department of, etc.*
Universities or schools	Verified accounts containing keywords such as *school, university, K-12, etc.*
Non-profit	Verified accounts containing keywords such as *501c3, charity, foundation, etc.*
News	Verified accounts containing keywords such as *times, journal, post, chronicle, etc.*
Influencers	Any account with more than 100k followers
Other	All other accounts

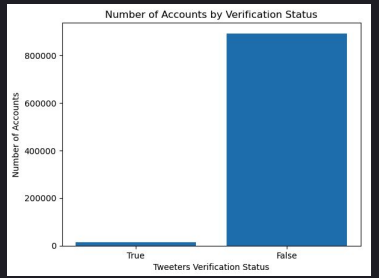
We created a new column for user_type. This was added to categorize the type of users into different categories: *government entities, universities or schools, non-profit organizations, news outlets, social media influencers, and others.*

Definition considerations

- We had to rely on keyword filtering on user.description, user.name, and user.url fields to define the type of user.
- Any user can create an account on Twitter and claim to be part of the government, a school, or a non-profit organization.
- To avoid false positives, we decided to only look at verified accounts.
  - This filters out accounts such as *screw the government* or *swag university* that would otherwise be categorized as government and university accounts.
  - Since only 1.5% of accounts are verified, most Tweeters will fall into the Other category.



Verified accounts breakdown by user_type



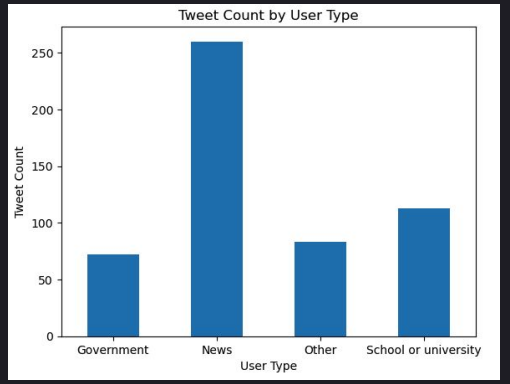
Number of accounts by verification status

The most prolific verified Twitterers by tweet volume are News accounts

Tweet count	Account Name and Account Screen Name	User type
113	ExploreLearning @ExploreLearning	School or university
72	Federal Student Aid @FAFSA	Government
57	U.S. News Education @USNewsEducation	News
46	Higher Ed Dive @HigherEdDive	News
45	Fastweb Scholarships @PayingForSchool	Other
45	The Washington Times @WashTimes	News
40	Times Higher Education @timeshighered	News
38	Zack Friedman @zackafriedman	Other
37	The Hill @thehill	News
35	Forbes @Forbes	News

When looking at the top verified accounts by tweet volume, **News** accounts are the most prolific. This indicates that News accounts share a lot of information through tweets, along with School or University and Government accounts.

The top 3 accounts are related to schools, federal student aid, and education. It is not surprising to see they post a lot of tweets related to the cost of higher education.



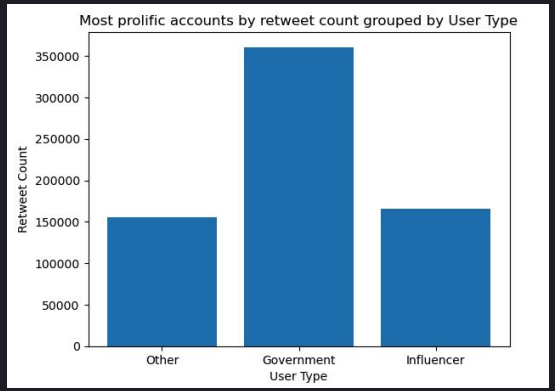
Tweet Count by User Type
(Top 10 accounts)

Note: when including non-verified accounts, the most prolific ones belonged to the **Other** user type category. This is expected given our category definitions, and it was mostly spam accounts who tweeted the same message multiple times.

Government accounts are the verified accounts with the most retweets

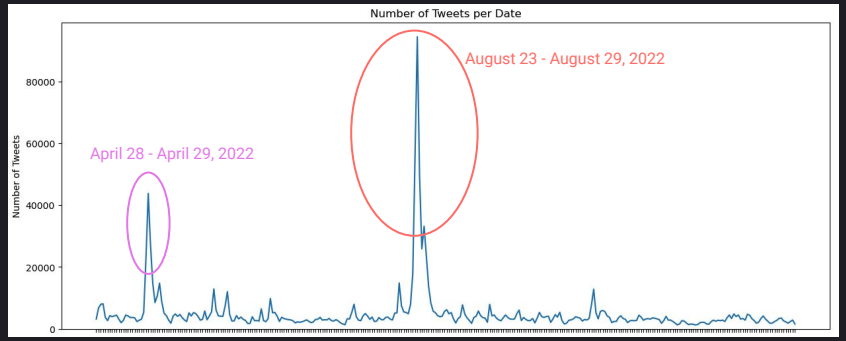
Retweet count	Name	User type
85003	Jon Schwarz @schwarz	Other
80179	President Obama @POTUS44	Government
78430	Elizabeth Warren @ewarren	Government
70668	Lacy M. Johnson @lacymjohnson	Other
69280	Tom Cotton @TomCottonAR	Government
67309	President Biden @POTUS	Government
65499	Alexandria Ocasio-Cortez @AOC	Government
56458	Steve Hofstetter @SteveHofstetter	Influencer
55254	Charlie Kirk @charliekirk11	Influencer
53818	David Hogg @davidhogg111	Influencer

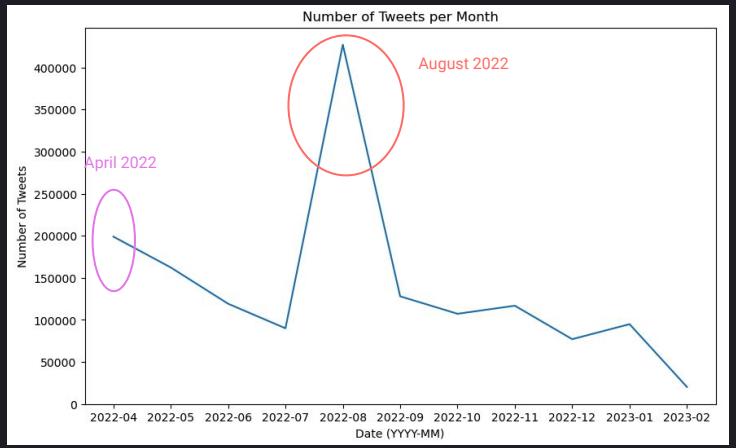
When looking at the most prolific verified accounts by retweet volume, **Government Entities** accounts are the most prolific.
While News and School and university accounts post a large amount of Tweets, their information is not shared and retweeted in large amounts.



- Seeing top government entities in the list highlights the highly political nature of our topic.
- Student loans forgiveness has been strongly supported by the Democratic Party in the US, and their Twitter accounts on this list shows the extent of their support.
- However, seeing 5/10 Democratic government accounts in the list shows that information may be one-sided and biased

A large amount of tweets was created in April and August 2022





In the time period between April 2022 to February 2023 when the tweets were created, we noticed a relatively low number of tweets per month with peaks in April 2022 and August 2022.

The exact dates with the highest number of tweets created correspond to significant developments in the student loan forgiveness journey in the US.

- April 28 - April 29, 2022: ~94K tweets created
  - On 4/28 Biden announces he is ‘ a couple of weeks’ away from a decision on student loan forgiveness
- August 23 - August 29, 2022: ~275K tweets created
  - On 8/23 the DOJ issues an opinion stating the president has the authority to forgive student loan debt
  - On 8/24 President Biden directs the Department of Education to forgive student loans

We also notice certain periods where the number of tweets created is low or close to zero. This raises concerns about the depth and sustainability of discussion on the platform, as it seems that conversations fizzle out quickly, waiting for the next trending topic to emerge.

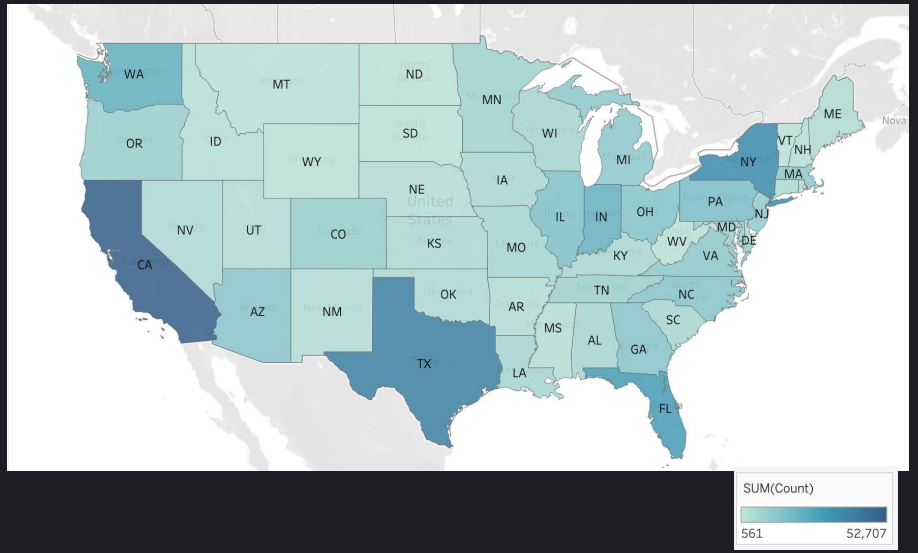
Investopedia. (2022). Student Debt Timeline: Key Events and Policy Changes. Retrieved from https://www.investopedia.com/student-debt-timeline-7112128#:~:text=Aug.,31%2C%202022.

The highest amount of tweets come from Tweeters located in California, Texas, Florida, and New York

Analysis by US State

The highest amount of tweets on the topic of Higher Education cost come from users located in California, Texas, Florida, and New York.

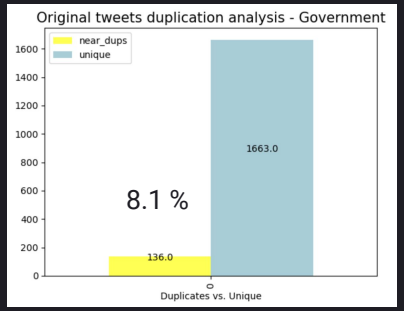
This can be explained by the fact that the 4 states are the most populous state in the US with a large number of college students. Additionally, these states have a strong political climate, so residents may be more politically engaged and active on social media.

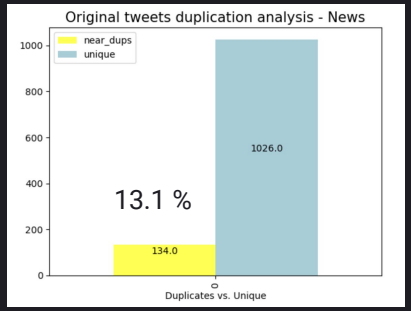


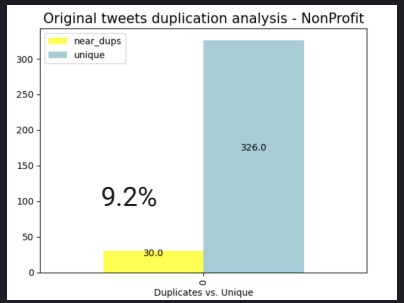
Analysis by Country

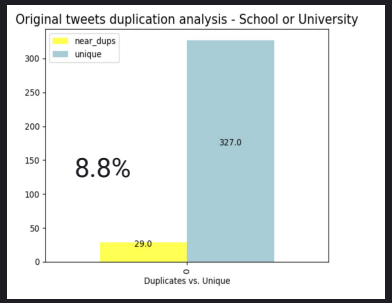
Since our topic is highly related to the American population, we focused our analysis on US states. However, it may be interesting to know that top tweeters were located in the United States, Nigeria, India, Canada, and United Kingdom.

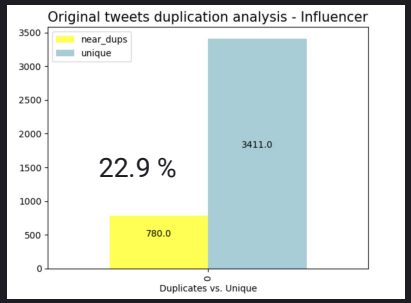
Influencers and News outlets had the highest rate of duplicative tweets, which may lead to the spread of misinformation

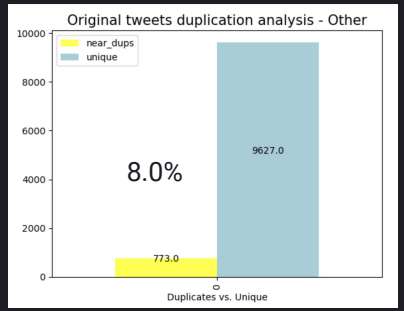












We ran a duplication analysis on original tweets (not considering retweets and quotes) to understand how unique each tweet is for every category of user.

Influencers and **News** outlets had the highest rate of duplicate tweets, meaning their tweets are often very similar, if not duplicates, with other tweets.

This makes sense as News outlet share similar news, and influencer may want to just repeat or re-use information they read about from other accounts. However, if the repeated information is not accurate, this leads to the **rapid spread of unreliable information**.

It is worth noting that differences in the amount of tweets available for analysis in each category can impact the conclusions drawn from the analysis.

To be trusted as a credible source of information, Twitter should increase the number of verified accounts, limit spam, and share unbiased information

Conclusion

Based on the analysis, it appears that Twitter may not be a reliable source of information. Here are the 3 main reasons:

1. The top government profiles with the most retweets are exclusively from the Democratic Party which suggests a potential political **bias**.
2. Popular topics such as student loan forgiveness are often discussed only when major events occur, but the conversations tend to fizzle out quickly. This highlights the highly viral nature of the platform and raises concerns on discussion sustainability.
3. A percentage of messages on Twitter lack uniqueness, particularly among News and Influencer accounts, which suggests that users may be encountering repetitive and potentially **unreliable** information.

Actionable recommendations

To become a more reliable source of information, here are some recommendations for Twitter:

1. **Address political bias** on the platform by breaking the social media bubble. This can be tackled by revising the algorithm used, so that users are exposed to a wider range of perspectives.
2. **Encourage meaningful conversations** by increasing the number of characters allowed in a tweet. Users would be able to convey more nuanced and complex thoughts.
3. **Promote verified and trusted sources**. Only 1.5% of accounts are verified, leading to the spread of information from unreliable sources.
4. Improving the fact-checking system and **penalize accounts that share misinformation**