Twitter Analysis

- BDP Final Project

Neal Xu

Agenda

- Executive Summary
- Methodology
- Source Data Overview & EDA
- Feature Engineering
- Data Cleaning-Up
- Topic Selection
- The Analysis (Author, Location, Timeline, and Uniqueness)
- Summary Conclusion & Recommendation for Future Work

Executive Summary

Problem:

Whether Twitter can be considered a credible source of information, reflects the emergence of important trends or topics in education, specifically: “Biden’s college student debt relief”.

Solution:

Analysis of approximately 100 million Tweets (~500GB) using Google Cloud Platform.

Value:

This project helps us to understand if we should rely on social media such as Twitter for major decision-making that requires the gaining of the latest news.

Next Steps:

How credible are the original tweets that could be taken for topic knowledge gaining from non-authority entities?”

Methodology

1

Platform

Google Could Platform

2

Language & File Type

PySpark

Json

3

Data Frame

PySpark DataFrame

Spark RDD

Pandas

4

Main Functions & Packages

.select()
.filter()
.withColoumn()
.groupBy()
.agg()
rlike()
contains()

5

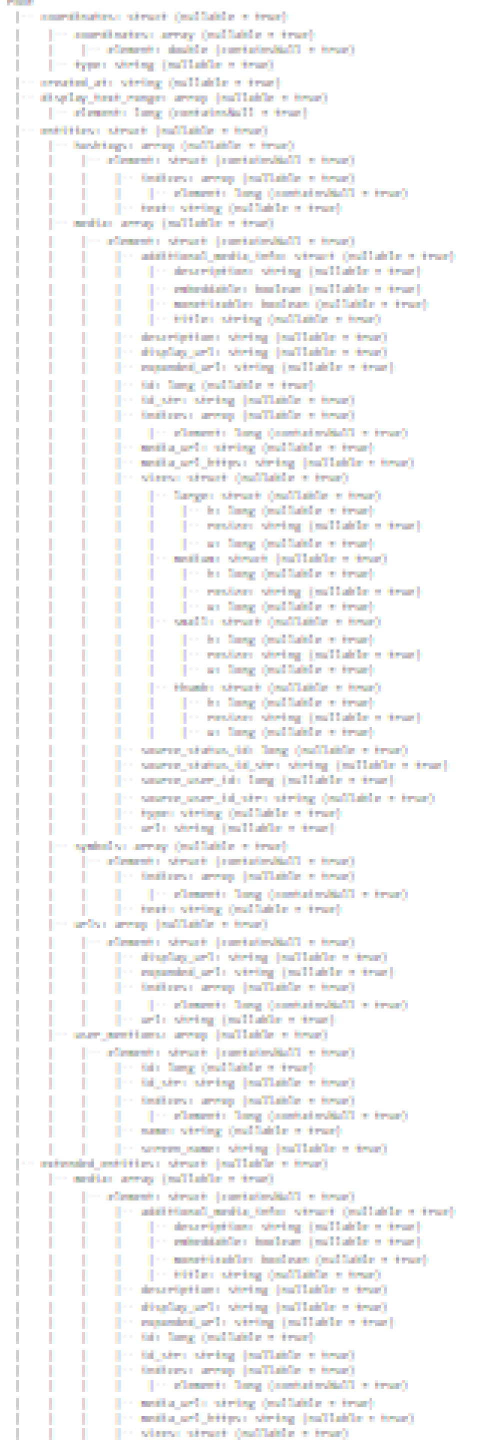
Method

Pyspark.ml.feature

MinHashLSH
With
Jaccard Similarity = 0.5

Source Data Overview & EDA

Too much irrelevant columns from: .printSchema()



```
[8]: twitter.count()
[8]: 99992797
```

Original Data Count:
99992797
(almost 100 million)

[158]:

	year	count
0	2022	81496566

All from year:
2022

```
[ ]: filtered.filter(col("coordinates").isNotNull()).count()
[161]: 97554
```

Available coordinates (Bad location variable):
97554/99992797 = 1%

Bad raw data variable:
retweet_count (all "null" values)

```
[10]: # Bad retweet count data:
filtered.groupby('retweet_count').count().limit(20).toPandas()
```

[10]:

	retweet_count	count
0	0	81496566

Text language:
English only

```
[8]: filtered.groupby('lang').count().limit(20).toPandas()
```

[8]:

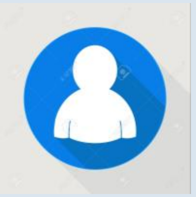
	lang	count
0	en	81496566

- Six new features were generated from the original data set and will be used for further analysis.

Feature Engineering

“user_name”

From: user['name']



“user_description”

From: user['description']

“user_screen_name”

From: user['screen_name']

“user_location”

From: user['location']

“place_location”

From: place['full_name']

“place_country”

From: place['country']

“retweet_count”

From: retweeted_status['retweet_count']

“retweeted”

From: retweeted_status['retweeted']

DATA Cleaning-Up

Filtering out tweets that are irrelevant to education

```
# Filtering for text that is related to education topic based on education-related key words:
twitter = twitter.withColumn("lowered_text", lower(col("text")))
filtered = twitter.filter(twitter.lowered_text.contains('school')\
                         |twitter.lowered_text.contains('learn')\
                         |twitter.lowered_text.contains('knowledge')\
                         |twitter.lowered_text.contains('college')\
                         |twitter.lowered_text.contains('kids')\
                         |twitter.lowered_text.contains('university')\
                         |twitter.lowered_text.contains('professor')\
                         |twitter.lowered_text.contains('children')\
                         |twitter.lowered_text.contains('child')\
                         |twitter.lowered_text.contains('higher')\
                         |twitter.lowered_text.contains('secondary')\
                         |twitter.lowered_text.contains('primary')\
                         |twitter.lowered_text.contains('public')\
                         |twitter.lowered_text.contains('education')\
                         |twitter.lowered_text.contains('elementary')\
                         |twitter.lowered_text.contains('class')\
                         |twitter.lowered_text.contains('student')\
                         |twitter.lowered_text.contains('course')\
                         |twitter.lowered_text.contains('degree')\
                         |twitter.lowered_text.contains('department')\
                         |twitter.lowered_text.contains('private'))
```

Selecting helpful variables

```
# Refining useful data from the dropped data frame to get only useful information:
# (Additional useful information could be retrieved from json subset data)

cleaned = dropped.select([dropped.created_at,
                          dropped.id_str,
                          dropped.user['name'].alias('user_name'),
                          dropped.user['description'].alias('user_description'),
                          dropped.user['screen_name'].alias('user_screen_name'),
                          dropped.user['location'].alias('user_location'),
                          dropped.place['full_name'].alias('place_location'),
                          dropped.place['country'].alias('place_country'),
                          dropped.quoted_status_id_str,
                          dropped.retweeted_status['retweet_count'].alias('retweet_count'),
                          dropped.retweeted_status['retweeted'].alias('retweeted'),
                          dropped.lowered_text,
                          dropped.retweeted_from,
                          dropped.timestamp_ms])
```

Filter out 18 million irrelevant data rows

```
cleaned.count()
```

```
81496566
```

14 cleaned variables are kept

```
cleaned.printSchema()
```

```
root
 |-- created_at: string (nullable = true)
 |-- id_str: string (nullable = true)
 |-- user_name: string (nullable = true)
 |-- user_description: string (nullable = true)
 |-- user_screen_name: string (nullable = true)
 |-- user_location: string (nullable = true)
 |-- place_location: string (nullable = true)
 |-- place_country: string (nullable = true)
 |-- quoted_status_id_str: string (nullable = true)
 |-- retweet_count: long (nullable = true)
 |-- retweeted: boolean (nullable = true)
 |-- lowered_text: string (nullable = true)
 |-- retweeted_from: string (nullable = true)
 |-- timestamp_ms: string (nullable = true)
```

Author Identification Analysis

Top five most prolific Twitterers (Original Contents):

user_screen_name	count
education_24x7	10317
educationbnb	6235
techysaavy	4450
WorkAcademic	4194
jc_james_clark	4088

Top five most prolific Twitterers (Retweets):

user_screen_name	max(retweet_count)
kalvin_stevens	516954
8d1jay	516951
malikgoodwin58	516928
savvh12	516795
Dinasor22	516772

Twitterers by the five entities (total)

	entities	count
0	Others_social_media_influencers	75664035
1	schools	1923196
2	government_entities	1263961
3	universities	941791
4	news_outlet	1516533
5	nonprofit_organizations	187050

The most prolific retweet count is significantly more than the original counts and "education_27" is the most active original content creator that is 40% more than the second-ranked user. In terms of the top retweet count, the top users have a similar count of around 500,000 retweets.

Most twitters are "other social media influencers", the rest of them are equally distributed not including non-profit organizations and universities for around 1.2 to 1.9 million. Universities and non-profit organizations are only around 1/12 of the schools, governments, and news outlets.

Topic Selection

The education topic selected is:

“Biden’s college student debt relief”

The attached code below is used to label tweet text based on if the text is related to the selected topic.

(Around 4% of the tweets are related)

```
# Selecting Topic: Student Loan Debt Relief
```

```
Debt_Relief = '|'.join(["debt(s)?", "relief", "loan", "forgive(ness)?","biden","president","tuition","reimbursement","credit","consolidate","income","salary","tax","low-income","application","Sup
Debt_Relief
#Standardized_testing
```

```
'debt(s)?|relief|loan|forgive(ness)?|biden|president|tuition|reimbursement|credit|consolidate|income|salary|tax|low-income|application|Supreme |court|block(ing)?'
```

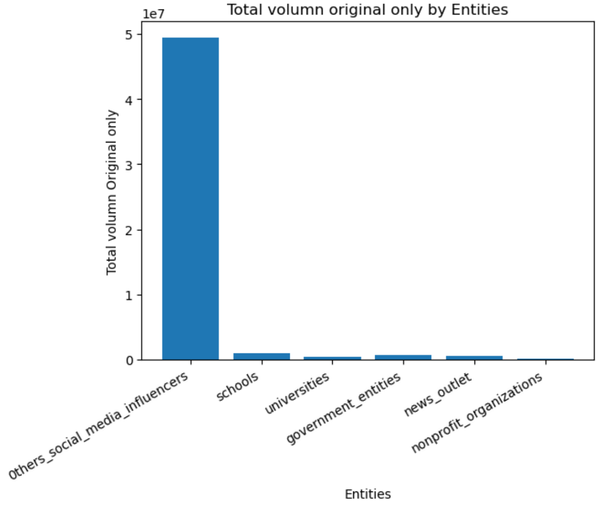
```
cleaned = cleaned.withColumn('Debt_Relief', when(cleaned.lowered_text.rlike(Debt_Relief),'Related').\
                                            otherwise('Not Related'))
```

```
# Related tweets / Not Related ratio:
cleaned.select('Debt_Relief').where(cleaned.Debt_Relief == "Related").count()/\
cleaned.select('Debt_Relief').where(cleaned.Debt_Relief == "Not Related").count()
```

```
0.04580198741291551
```

Distribution of tweet / retweet volume by Organizations

	entities	count
0	Others_social_media_influencers	49441905
1	schools	926111
2	universities	471616
3	government_entities	762710
4	news_outlet	561242
5	nonprofit_organizations	113562

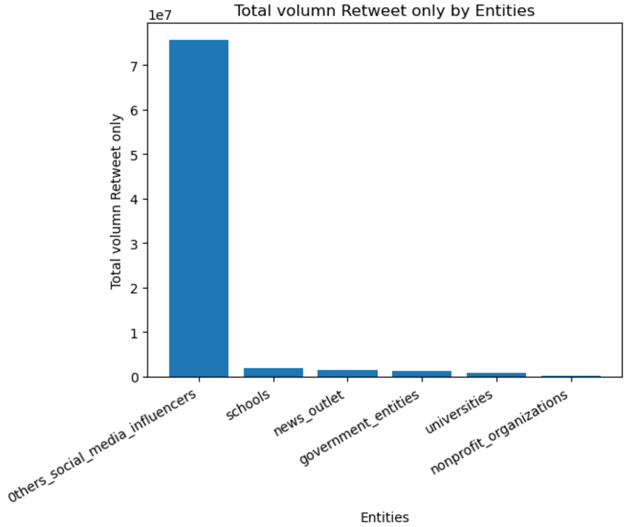


*On the left side is the distribution of **original** tweets by different entities.*

It is consistent with the previous analysis that the "other social media influencers" has the most significant original tweet counts since most accounts are grouped into this sector.

Here, schools and governments ranked the top two counts again and non-profit organizations still being the least active group.

	entities	count
0	Others_social_media_influencers	75664035
1	schools	1923196
2	news_outlet	1516533
3	government_entities	1263961
4	universities	941791
5	nonprofit_organizations	187050



*On the left side is the distribution of **retweets** by different entities.*

The distribution has a similar pattern to the original tweets by different entities. However, the difference is that the retweet count is almost twice the original content for each of the entities.

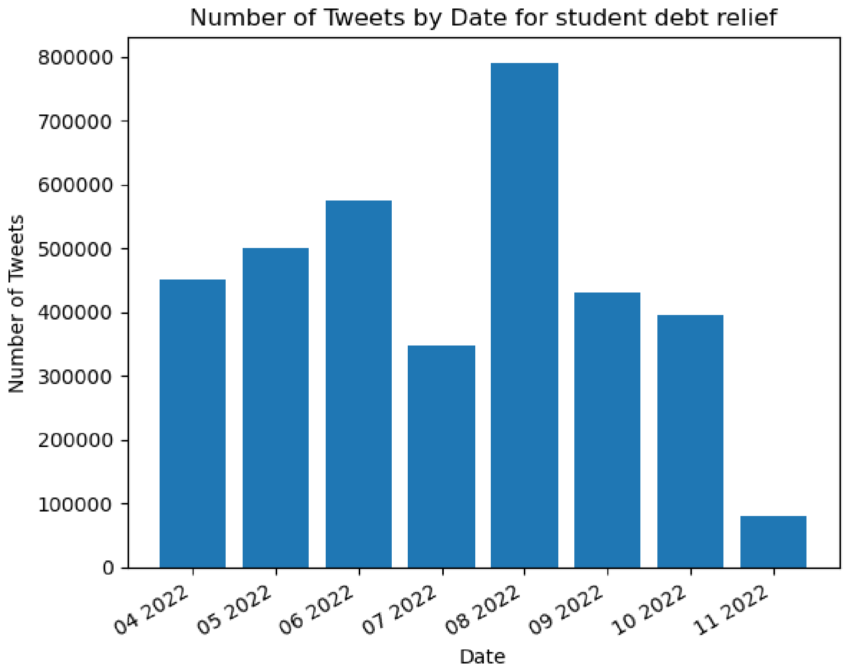
Here, schools and news outlets ranked the top two counts. The retweet for news outlets is almost three times its original content counts. While the non-profit organization still being the least active group.

Time Analysis & Location

-The topic chosen for timeline and location analysis is “Student Loan Debt Relief”

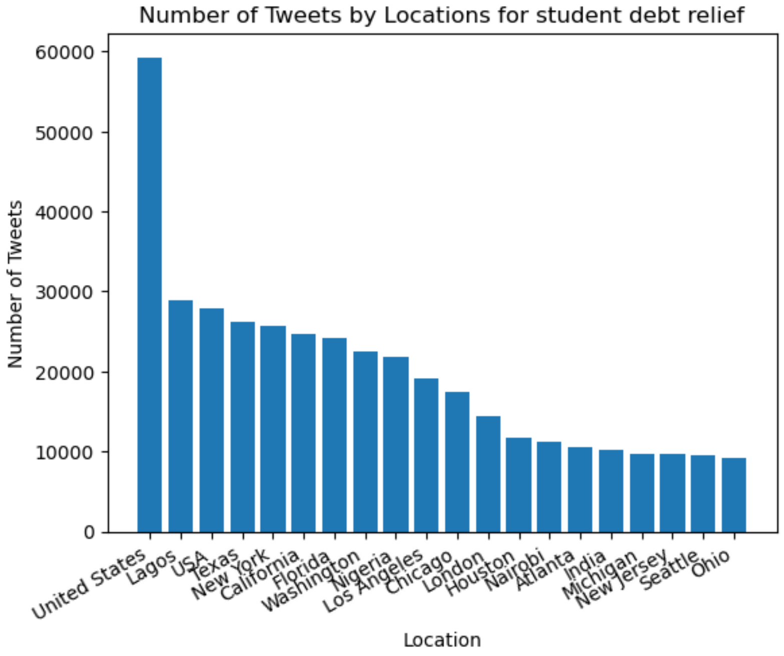
The timeline below shows that the tweet amount related to the student loan debt relief surged during August 2022, the same time the news about Biden’s student debt relief came out.

(There are large gaps between months for the amount. A significant peak appeared during August 2022)



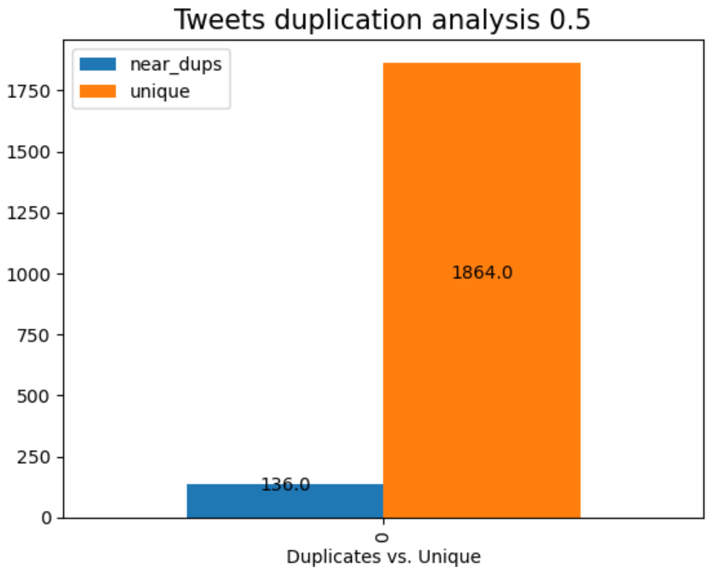
The visualized location analysis below shows that most of the tweets and twitters are located in the United States and in the cities where higher education institutes are located. It is consistent with the debt relief topics since it is a relief for American college students.

(The only location variable that could be used is from the location under “users”, however, the problem of mess still exists due to the locations users entered. Below is the best representation of the location.)



Uniqueness Analysis

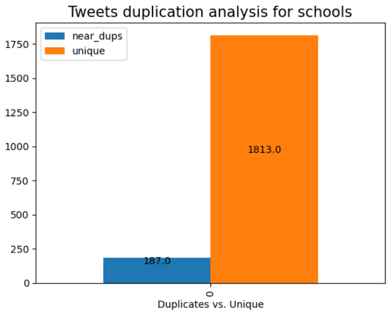
Most Tweets are unique (2000 random samples):

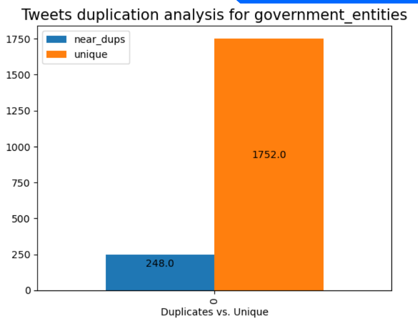


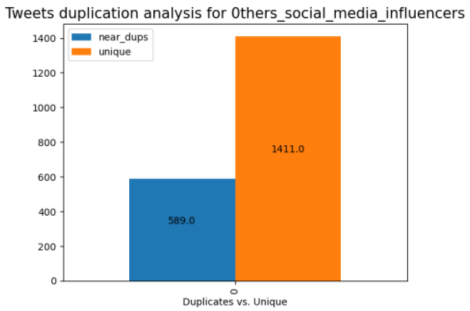
The threshold is set as 50 for maximum accuracy (Jaccard distance = 0.5) based on the Duplication text analysis.

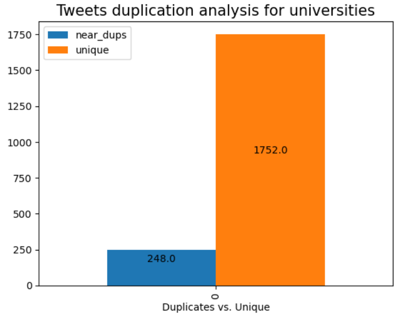
The tweet uniqueness analysis by different organizations is visualized on both the lift and right sides and they have a similar pattern to the overall analysis that most tweets are unique tweets.

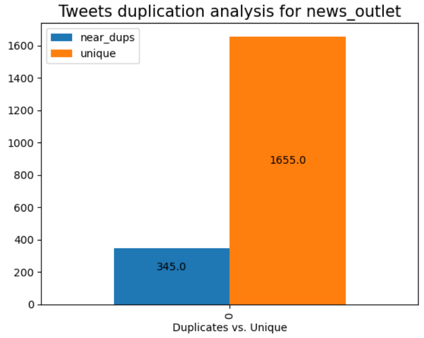
Most organizations including schools, governments, non-profit organizations, and news agencies mainly have unique tweets while other social media influencers have a higher rate of retweet count compared to the rest of the entities.

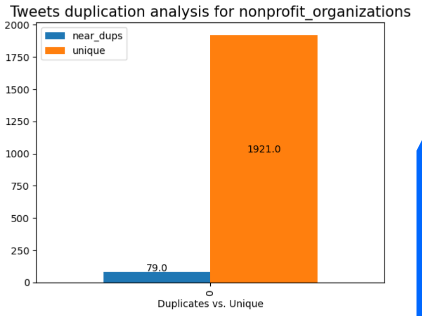












Summary Conclusion & Recommendation for Future Work

Conclusion:

Twitter could be considered a source of information that can reflect the emergence of important trends or topics in education. It also could be useful for understanding the public's opinion on a certain topic or trend in education.

However, based on the current analysis, it should not be considered a creditable source to obtain knowledge for the topic in general until further tweet analysis. Since most tweets are original content created by social media influencers other than authority agencies such as governments, schools, news, and non-profit organizations.

Future Work:

In addition to the current analysis, the analysis of the credibility of original tweets created by social media influencers could be the aim for the next step:

"How credible are the original tweets could be taken for topic knowledge gaining from non-authority entities?"