Is Twitter a credible source of information?

5th December 2022

Presented by Sanchit Kumar

Contents

- Executive Summary
- Methodology
- Most Prolific Twitterers
  - General Population
  - Across Types of Organization
- Twitters Location Analysis
- Timeline of Tweets
- Message Duplication
  - General Population
  - Across Types of Organizations
- Conclusion and Actionable Recommendation

Executive Summary

Below are my observations from the analysis:

- *The most prolific twitterers of original tweets* by organization are seen among private institutions with financial incentives to tweet more*
- *Reputed institutions (news organizations, nonprofits, universities / schools, private) retweet significantly less relative to the volume of original tweets while individuals retweet significant more*
- *The most significant number of education-related tweets arise from the US, Nigeria, UK, and India primarily (a wide distribution geographically)*
- *There were a few peaks of tweets in August and November when analyzed by frequency in day intervals*
- *Using Jaccard Similarity on the entire population, we see that there is a notable amount of similarity / duplicates in tweets (~30%)*
- *Using Jaccard Similarity of 0.5, Reputed institutions have significantly less duplicates relative to the population (~7.5 - 8.5%)*

Ultimately, Twitter **cannot** be regarded as a primary source of information because of the following:

- *There are multiple private institutions and individuals who have financial interests in the messages they tweet / retweet while news organizations and verified users do not consistent of the large population of twitterers. However, if one were to look for original tweets by news organizations and other reputed organizations in particular, information may be relied on as they create original tweets.*
- *There is widespread geographical distribution of twitterers and consistent volumes across time, which does not reflect important trends in education*

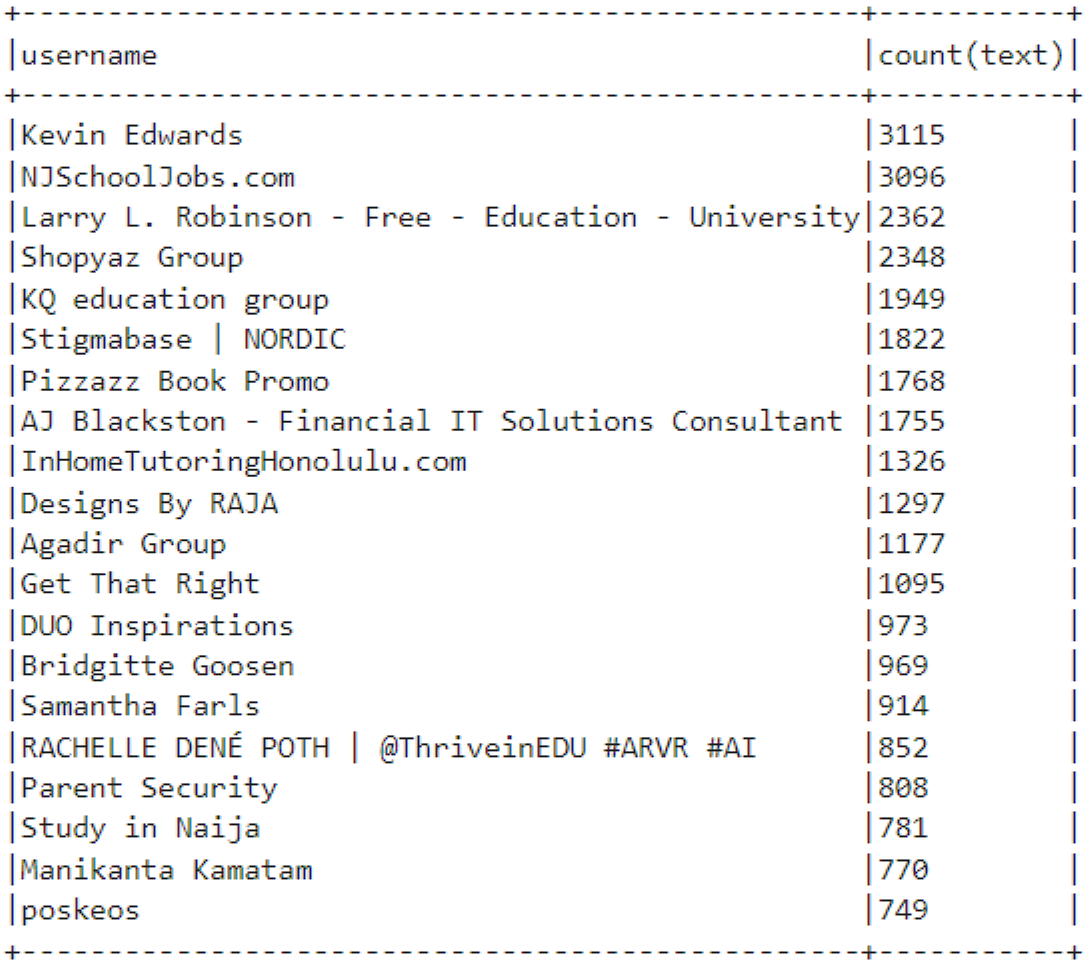
Methodology

1. Filter out observations for education
   - ➢ Out of 99992797 records, the analysis was run on 3838633 records by filtering out for education-related words such as education, k-12, grade inflation, learning, school, university, instruction, curriculum, etc.
2. Understand complex schema of data and select variables for analysis
   - ➢ I analyzed which variables contained significant null values, removing them from my analysis. For selection of variables analyzed, please see footnote.*
3. Parquet format was used for speed of analysis with big data (~38M records).
4. Most prolific twitterers were analyzed by tweets and retweets. Retweets was decided through retweeted column after analyzing multiple retweet variables and chose this as having the best data (if retweet column was equal to RT or not).
   1. Out of 672606 original twitterers, 23375 are selected with those over 5 tweets for most prolific analysis.
   2. Out of 1407996 retweeting twitterers, 27299 are selected with those over 10 tweets for most prolific analysis.
5. Tweets were also divided into types of organizations (news organizations, nonprofit, university/school, private institutions, and influencers) after analyzing the most prolific twitterers. Influencers were determined by those accounts that were verified and had above 50000 followers.
6. Twitters were then analyzed by location and tweets were analyzed across time.
7. Tweets were then analyzed for message duplication (across the entire dataset as well as types of organizations)

*We would like the columns created_at because we want to understand distribution of tweets over time. Entities do not matter to us since we are analyzing profiles of Twitterers. Favorited does not matter us since whether this Tweet has been liked by the authenticating user is not going to have significance on Twitterers profiles. Filter_level does not matter to us since the maximum value of the filter_level parameter which may be used and still stream this Tweet is not going to have significance on Twitterers profiles. ID is important as it references the user by a unique number. is_quote_status matters to understand originality/uniqueness. Lang will not be kept as we already filtered for 'en' in the previous file. Retweeted will be kept for retweet analysis. Source, timestamp_ms does not have valuable info. Text will be kept for analysis of words used but not tweet_text as they are similar. Large majority of truncated is false so is not useful. User will be kept for analysis of Twitterers - they key columns required from user are id, name, description, user verified and follower count.

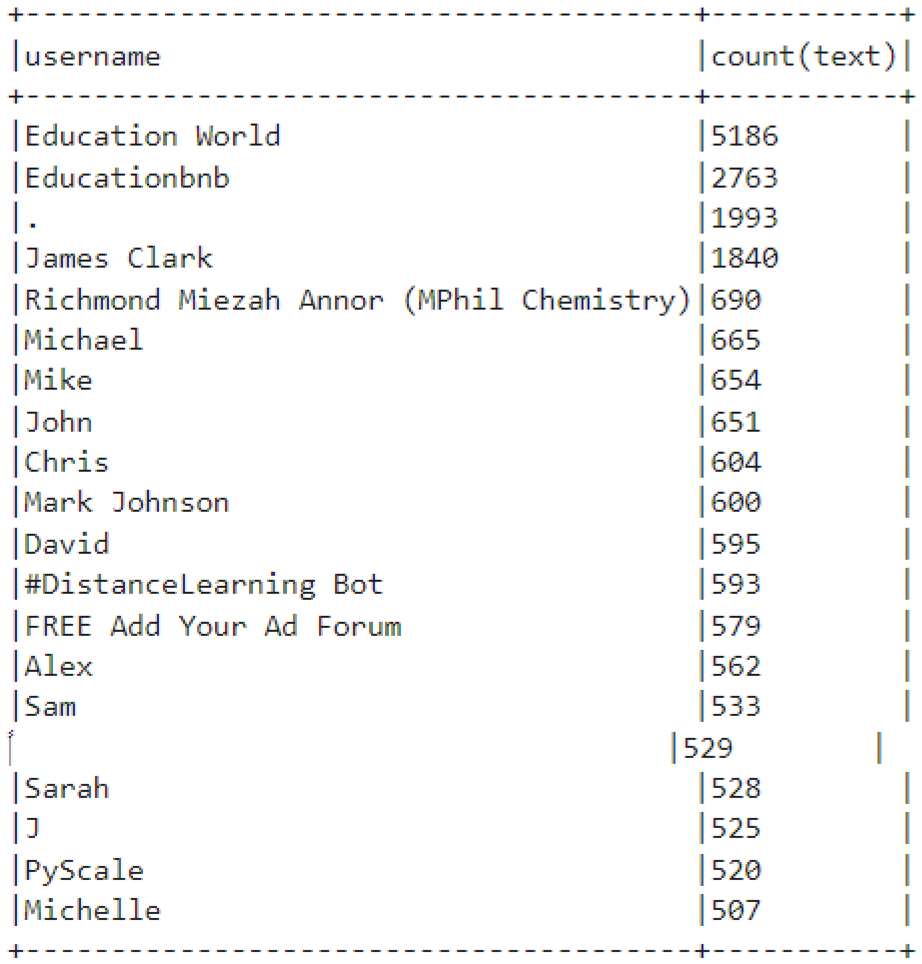
The most prolific twitterers of original tweets have more reputed organizations and influencers while retweeters have more individuals*

*Most prolific twitterers of original tweets**



```
+-------------------------------------------------+-----------+
|username                                         |count(text)|
+-------------------------------------------------+-----------+
|Kevin Edwards                                    |3115       |
|NJSchoolJobs.com                                 |3096       |
|Larry L. Robinson - Free - Education - University|2362       |
|Shopyaz Group                                    |2348       |
|KQ education group                               |1949       |
|Stigmabase | NORDIC                              |1822       |
|Pizzazz Book Promo                               |1768       |
|AJ Blackston - Financial IT Solutions Consultant |1755       |
|InHomeTutoringHonolulu.com                       |1326       |
|Designs By RAJA                                  |1297       |
|Agadir Group                                     |1177       |
|Get That Right                                   |1095       |
|DUO Inspirations                                 |973        |
|Bridgitte Goosen                                 |969        |
|Samantha Farls                                   |914        |
|RACHELLE DENÉ POTH | @ThriveinEDU #ARVR #AI      |852        |
|Parent Security                                  |808        |
|Study in Naija                                   |781        |
|Manikanta Kamatam                                |770        |
|poskeos                                          |749        |
+-------------------------------------------------+-----------+
```

*Most prolific twitterers of retweets**

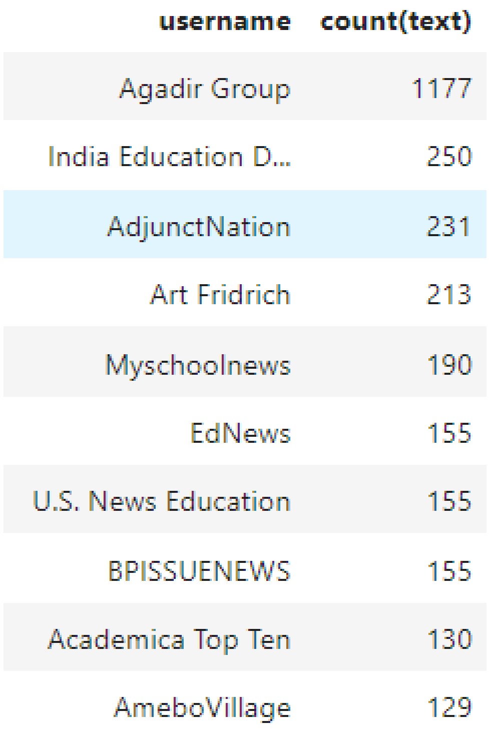


```
+---------------------------------------+-----------+
|username                               |count(text)|
+---------------------------------------+-----------+
|Education World                        |5186       |
|Educationbnb                           |2763       |
|.                                      |1993       |
|James Clark                            |1840       |
|Richmond Miezah Annor (MPhil Chemistry)|690        |
|Michael                                |665        |
|Mike                                   |654        |
|John                                   |651        |
|Chris                                  |604        |
|Mark Johnson                           |600        |
|David                                  |595        |
|#DistanceLearning Bot                  |593        |
|FREE Add Your Ad Forum                 |579        |
|Alex                                   |562        |
|Sam                                    |533        |
|                                     |529        |
|Sarah                                  |528        |
|J                                      |525        |
|PyScale                                |520        |
|Michelle                               |507        |
+---------------------------------------+-----------+
```

*Original tweets are considered those that are not retweets. Retweets and original tweets have been separated using retweeted column that denotes retweeted tweets as 'RT'.

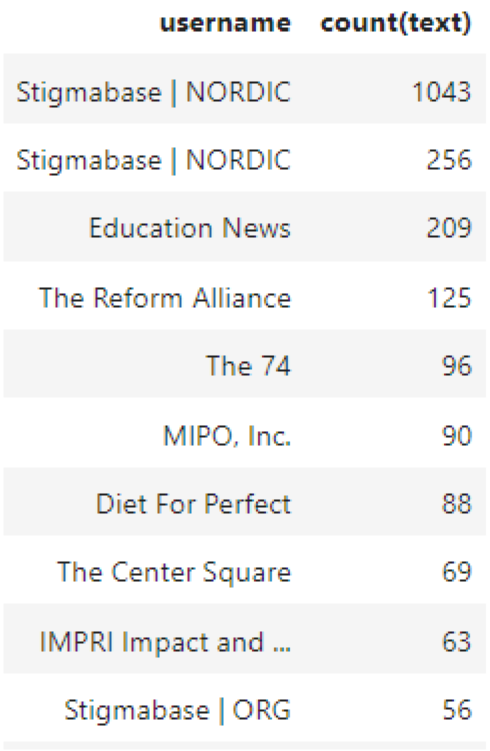
The most prolific twitterers of original tweets by organization are seen among private institutions with financial incentives to tweet more*

Top 10 News Organizations Twitterers



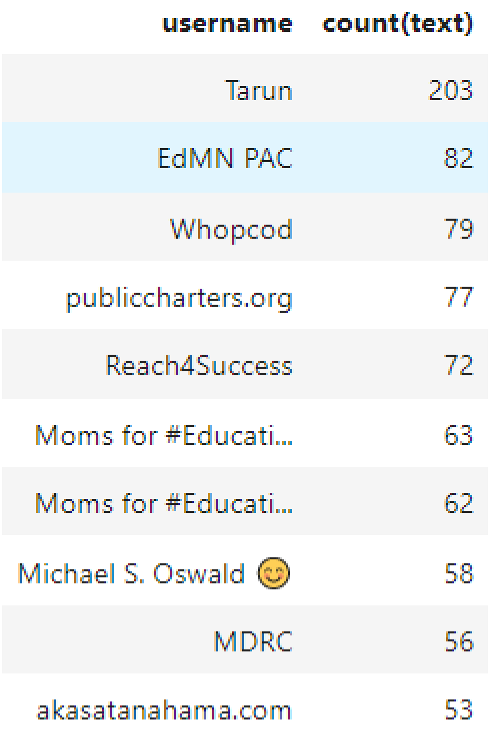
username	count(text)
Agadir Group	1177
India Education D...	250
AdjunctNation	231
Art Fridrich	213
Myschoolnews	190
EdNews	155
U.S. News Education	155
BPISSUENEWS	155
Academica Top Ten	130
AmeboVillage	129

Top 10 Nonprofit Twitterers



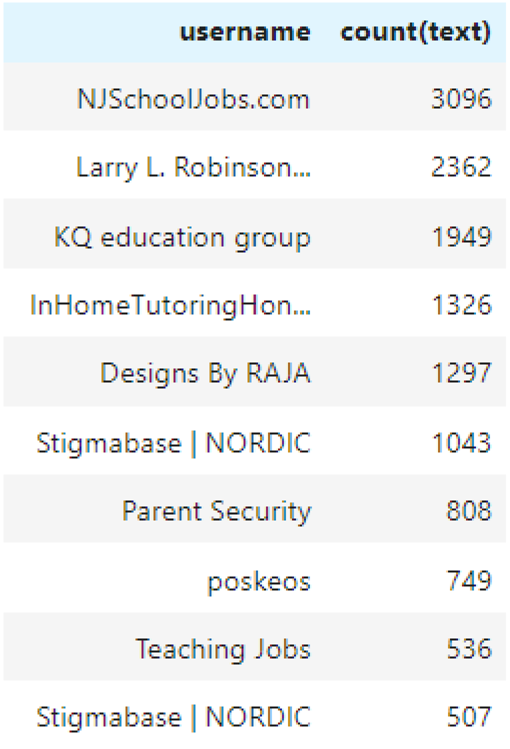
username	count(text)
Stigmabase | NORDIC	1043
Stigmabase | NORDIC	256
Education News	209
The Reform Alliance	125
The 74	96
MIPO, Inc.	90
Diet For Perfect	88
The Center Square	69
IMPRI Impact and ...	63
Stigmabase | ORG	56

Top 10 Uni/School Twitterers



username	count(text)
Tarun	203
EdMN PAC	82
Whopcod	79
publiccharters.org	77
Reach4Success	72
Moms for #Educati...	63
Moms for #Educati...	62
Michael S. Oswald 😊	58
MDRC	56
akasatanahama.com	53

Top 10 Private Institutions Twitterers

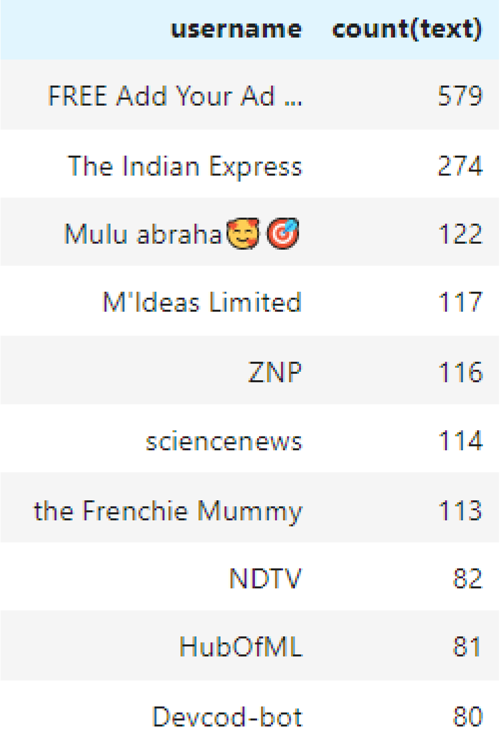


username	count(text)
NJSchoolJobs.com	3096
Larry L. Robinson...	2362
KQ education group	1949
InHomeTutoringHon...	1326
Designs By RAJA	1297
Stigmabase | NORDIC	1043
Parent Security	808
poskeos	749
Teaching Jobs	536
Stigmabase | NORDIC	507

*Original tweets are considered those that are not retweets. Retweets and original tweets have been separated using retweeted column that denotes retweeted tweets as ‘RT’.

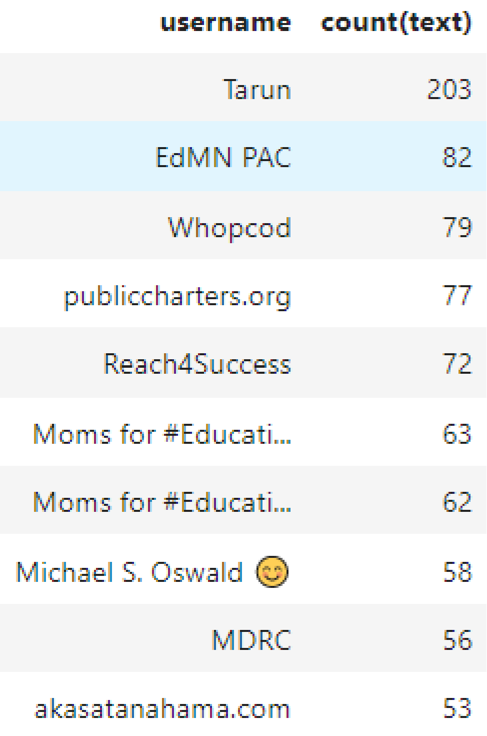
Different types of organizations have consistent volumes of retweets among the most prolific in each organization*

Top 10 News Organizations Twitterers



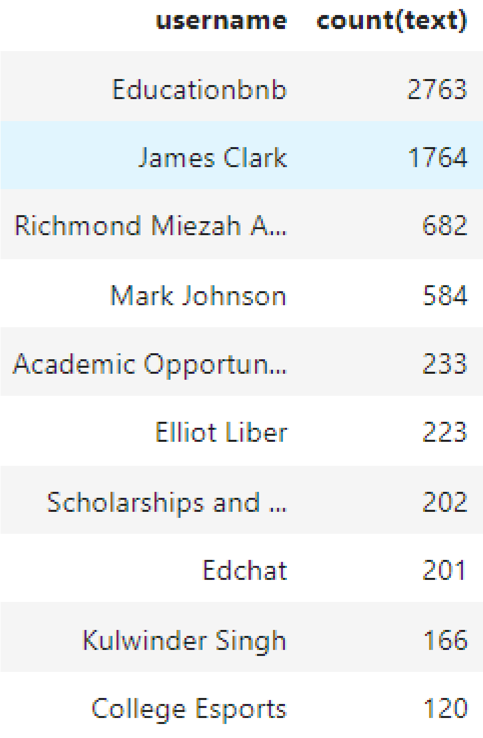
username	count(text)
FREE Add Your Ad ...	579
The Indian Express	274
Mulu abraha🥰🎯	122
M'Ideas Limited	117
ZNP	116
sciencenews	114
the Frenchie Mummy	113
NDTV	82
HubOfML	81
Devcod-bot	80

Top 10 Nonprofit Twitterers



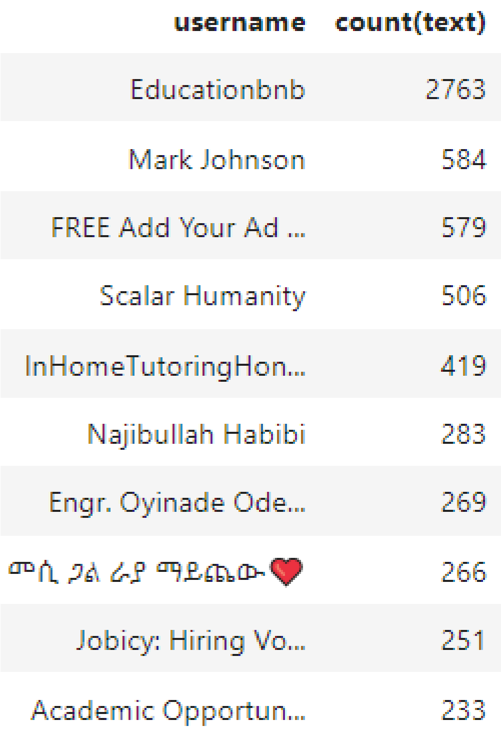
username	count(text)
Tarun	203
EdMN PAC	82
Whopcod	79
publiccharters.org	77
Reach4Success	72
Moms for #Educati...	63
Moms for #Educati...	62
Michael S. Oswald 😊	58
MDRC	56
akasatanahama.com	53

Top 10 Uni/School Twitterers



username	count(text)
Educationbnb	2763
James Clark	1764
Richmond Miezah A...	682
Mark Johnson	584
Academic Opportun...	233
Elliot Liber	223
Scholarships and ...	202
Edchat	201
Kulwinder Singh	166
College Esports	120

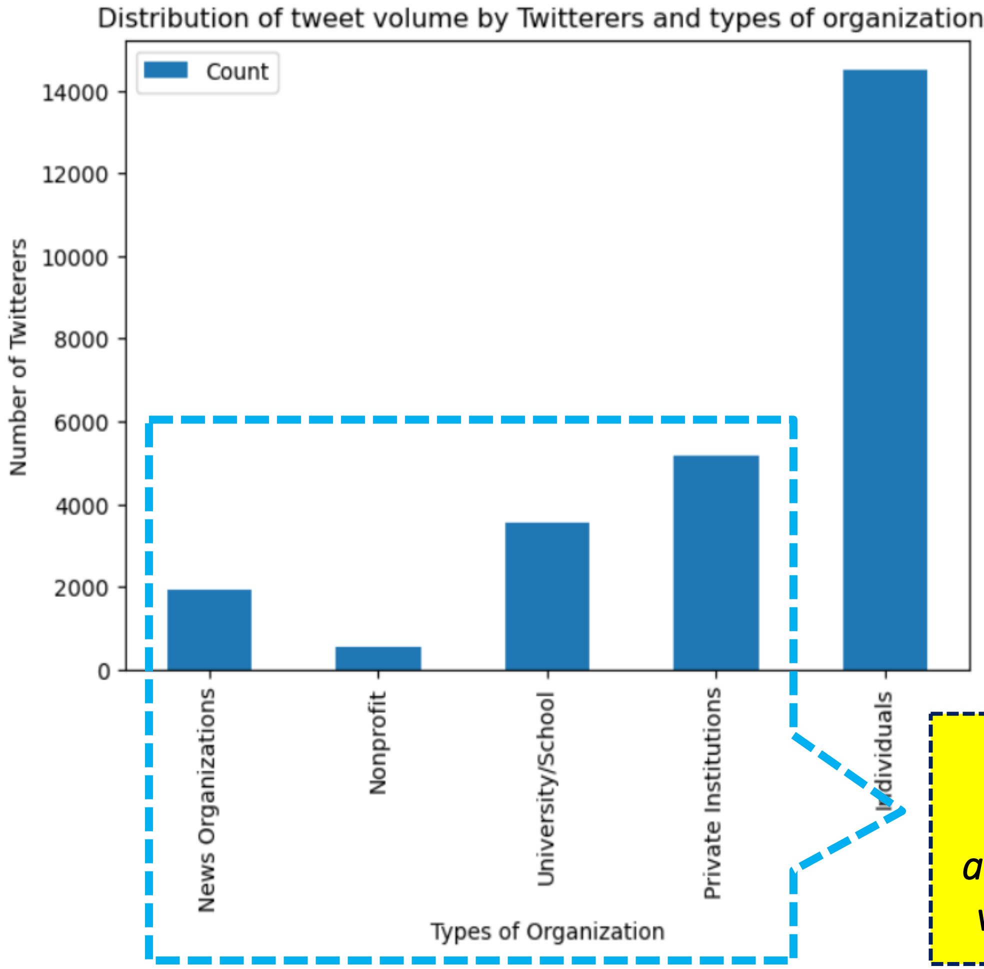
Top 10 Private Institutions Twitterers



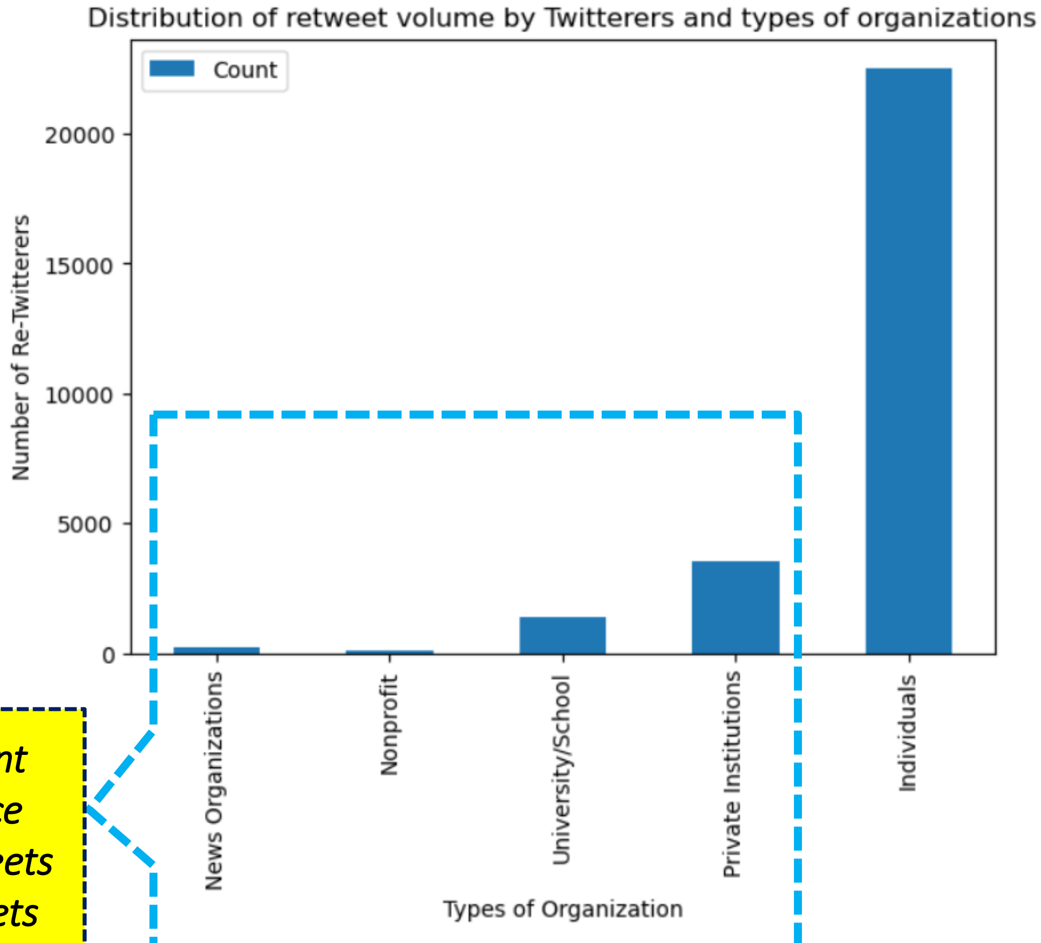
username	count(text)
Educationbnb	2763
Mark Johnson	584
FREE Add Your Ad ...	579
Scalar Humanity	506
InHomeTutoringHon...	419
Najibullah Habibi	283
Engr. Oyinade Ode...	269
መሲ ጓል ራያ ማይጨው❤️	266
Jobicy: Hiring Vo...	251
Academic Opportun...	233

*Original tweets are considered those that are not retweets. Retweets and original tweets have been separated using retweeted column that denotes retweeted tweets as 'RT'.

Classified organizations retweet significantly less relative to the volume of original tweets



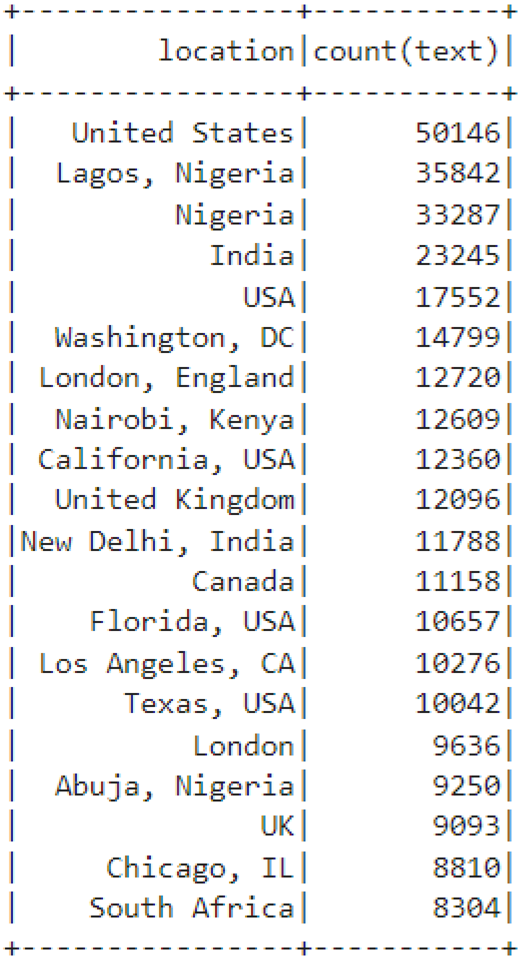
Significant difference across tweets vs. retweets



*Original tweets are considered those that are not retweets. Retweets and original tweets have been separated using retweeted column that denotes retweeted tweets as ‘RT’.

The most significant number of education-related tweets arise from the US, Nigeria, UK, and India primarily

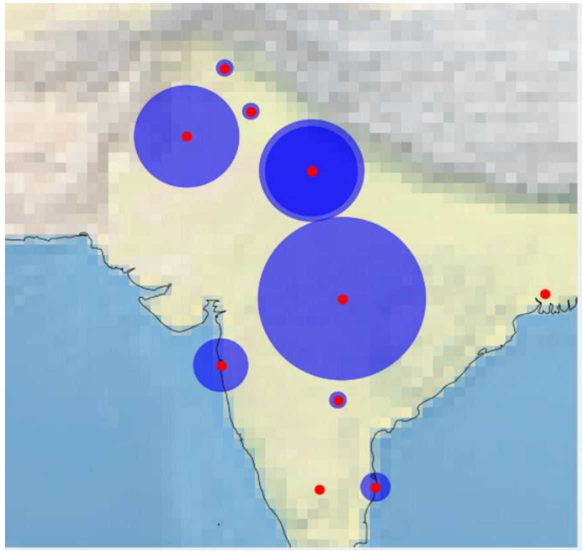
Locations with the highest tweets*



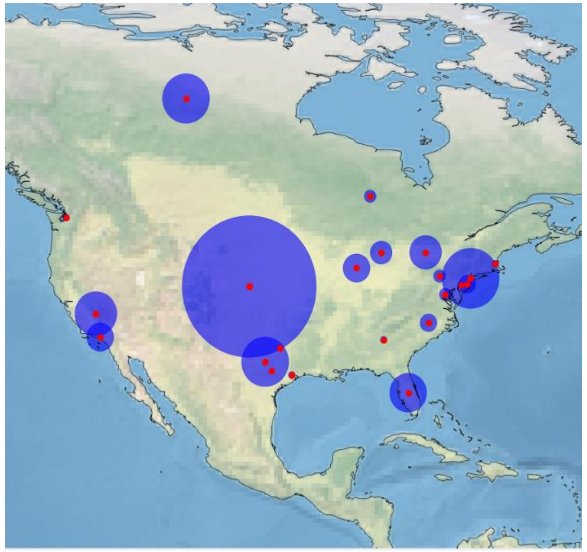
```
+----------------+-----------+
|        location|count(text)|
+----------------+-----------+
|   United States|      50146|
|  Lagos, Nigeria|      35842|
|         Nigeria|      33287|
|           India|      23245|
|             USA|      17552|
|  Washington, DC|      14799|
| London, England|      12720|
|  Nairobi, Kenya|      12609|
| California, USA|      12360|
|  United Kingdom|      12096|
|New Delhi, India|      11788|
|          Canada|      11158|
|    Florida, USA|      10657|
| Los Angeles, CA|      10276|
|      Texas, USA|      10042|
|          London|       9636|
|  Abuja, Nigeria|       9250|
|              UK|       9093|
|     Chicago, IL|       8810|
|    South Africa|       8304|
+----------------+-----------+
```

*For duplicates seen in the table, they were aggregated for the heatmaps.

India and Pakistan Heat Map

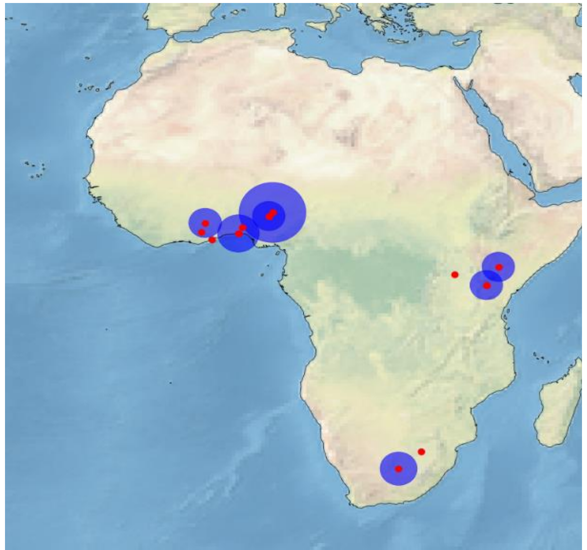


US and Canada Heat Map

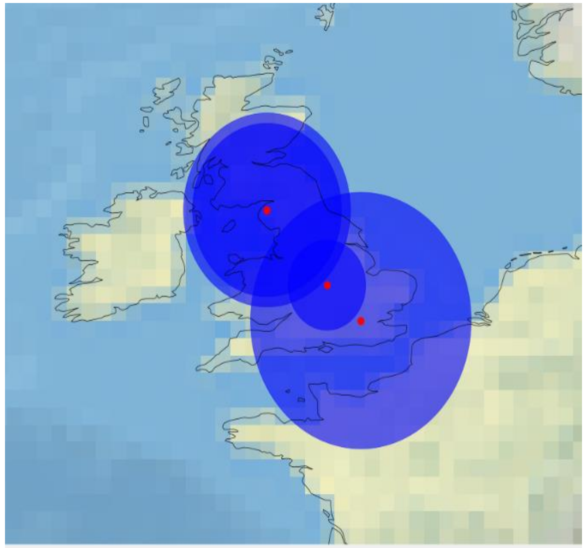


Tweet volumes are widely distributed across the US

Africa Heat Map



UK Heat Map



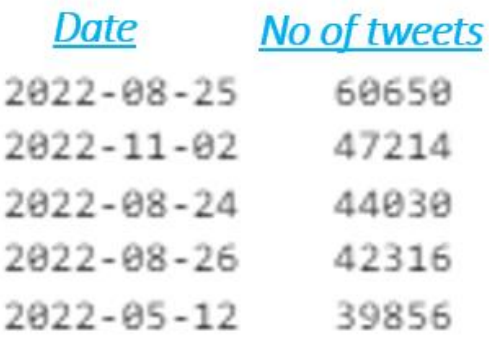
Prolific Twitterers are located across multiple continents

5th December 2022

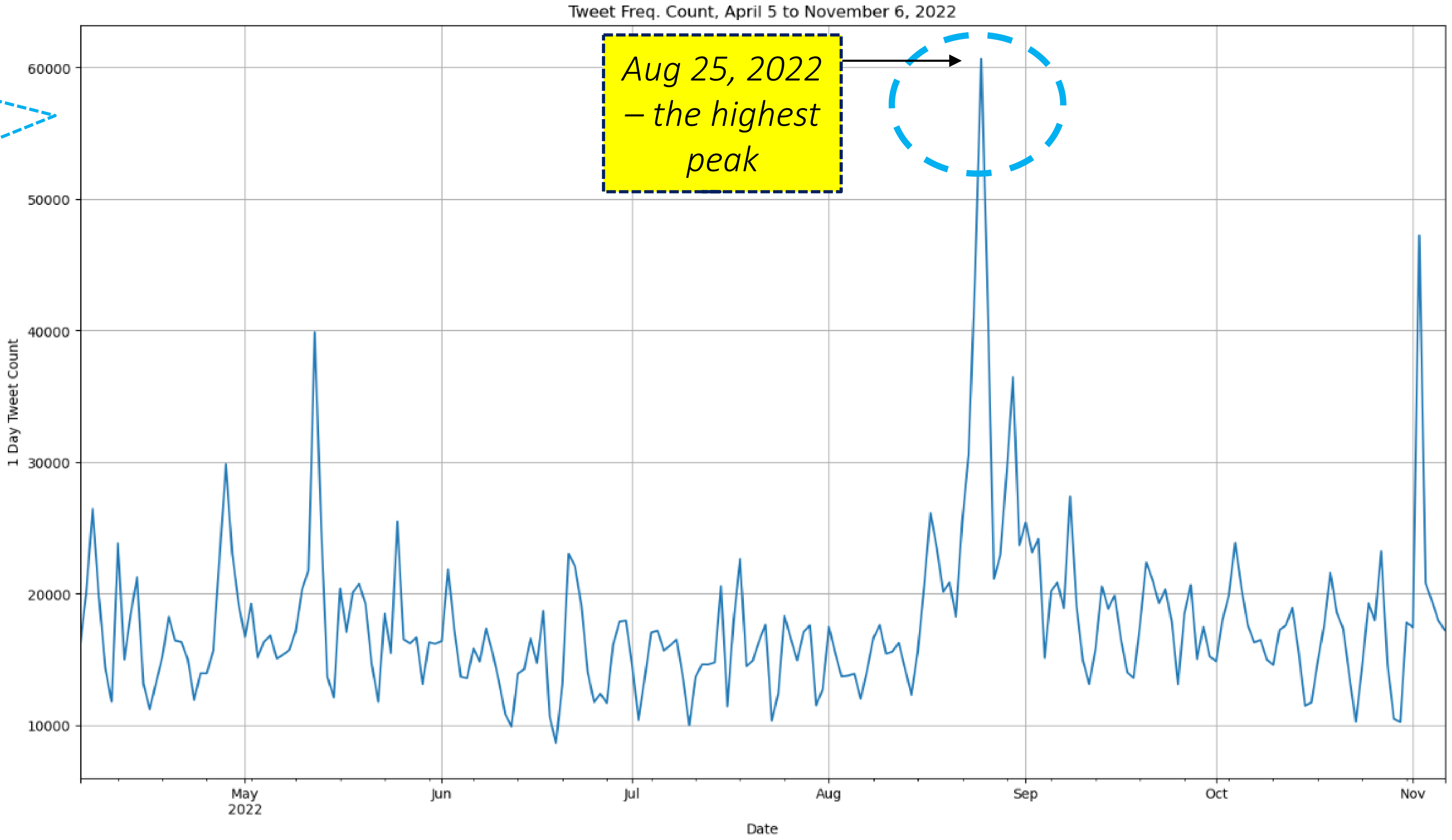
*There were a few peaks of tweets in August and November when analyzed by frequency in day intervals**

There do not seem to be any data collection gaps between April 5 and November 6

Days with the highest tweets



Date	No of tweets
2022-08-25	60650
2022-11-02	47214
2022-08-24	44030
2022-08-26	42316
2022-05-12	39856



Aug 25, 2022 – the highest peak

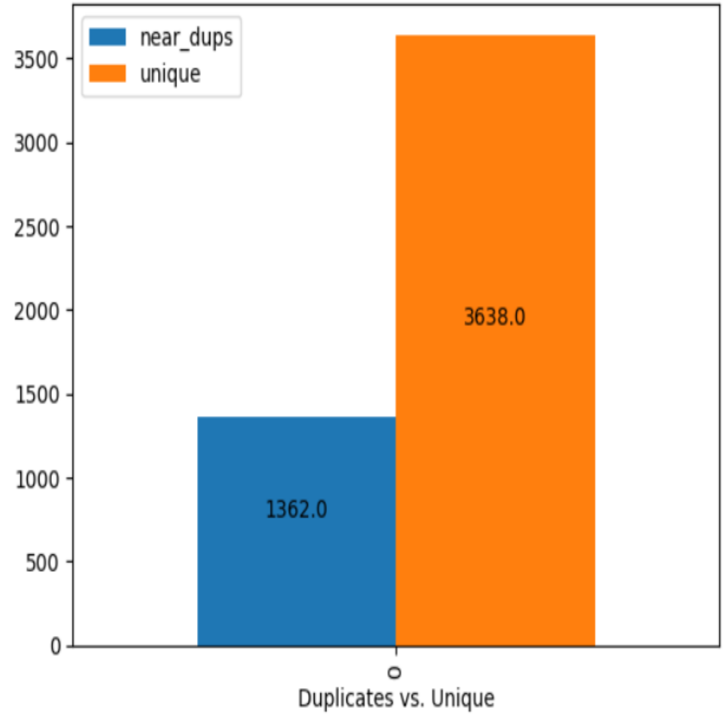
*My notebook has code with timeline of tweets in terms of 5 min frequency and 1 day frequency. Displaying 1 Day frequency was chosen here for interpretability. Faced Java memory errors when trying to analyze through filtering for description reasons for peaks / valleys.

Using Jaccard Similarity on the entire population, we see that there is a notable amount of similarity / duplicates in tweets*

Given limits of 5000, duplicate ratios of ~27%, 30%, 39% were seen for thresholds 0.3, 0.5, and 0.7, respectively.

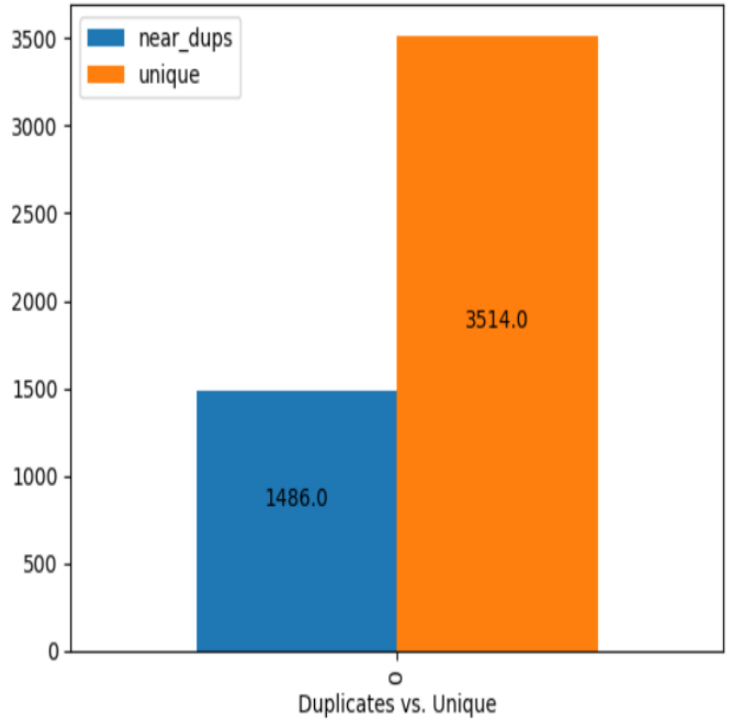
Tweets Duplication Analysis –

Jaccard Similarity (0.3)



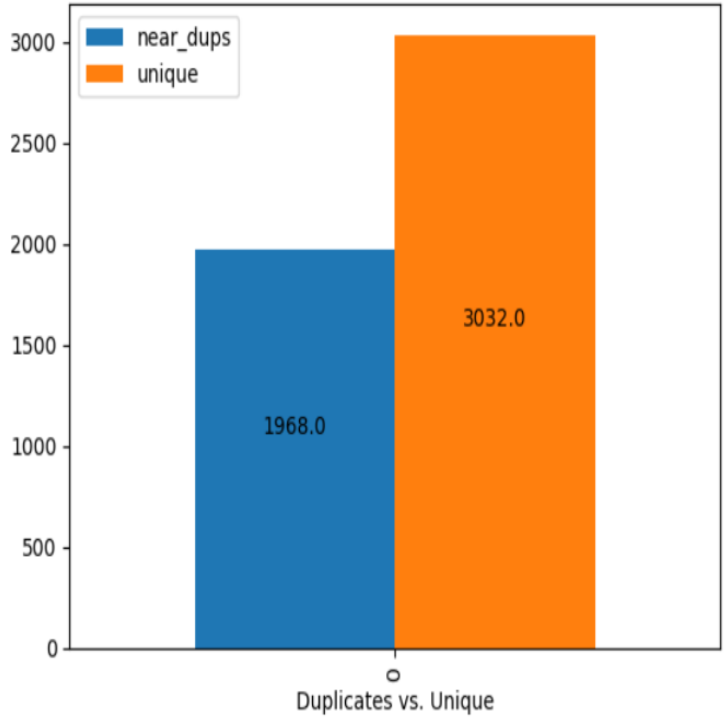
Tweets Duplication Analysis –

Jaccard Similarity (0.5)



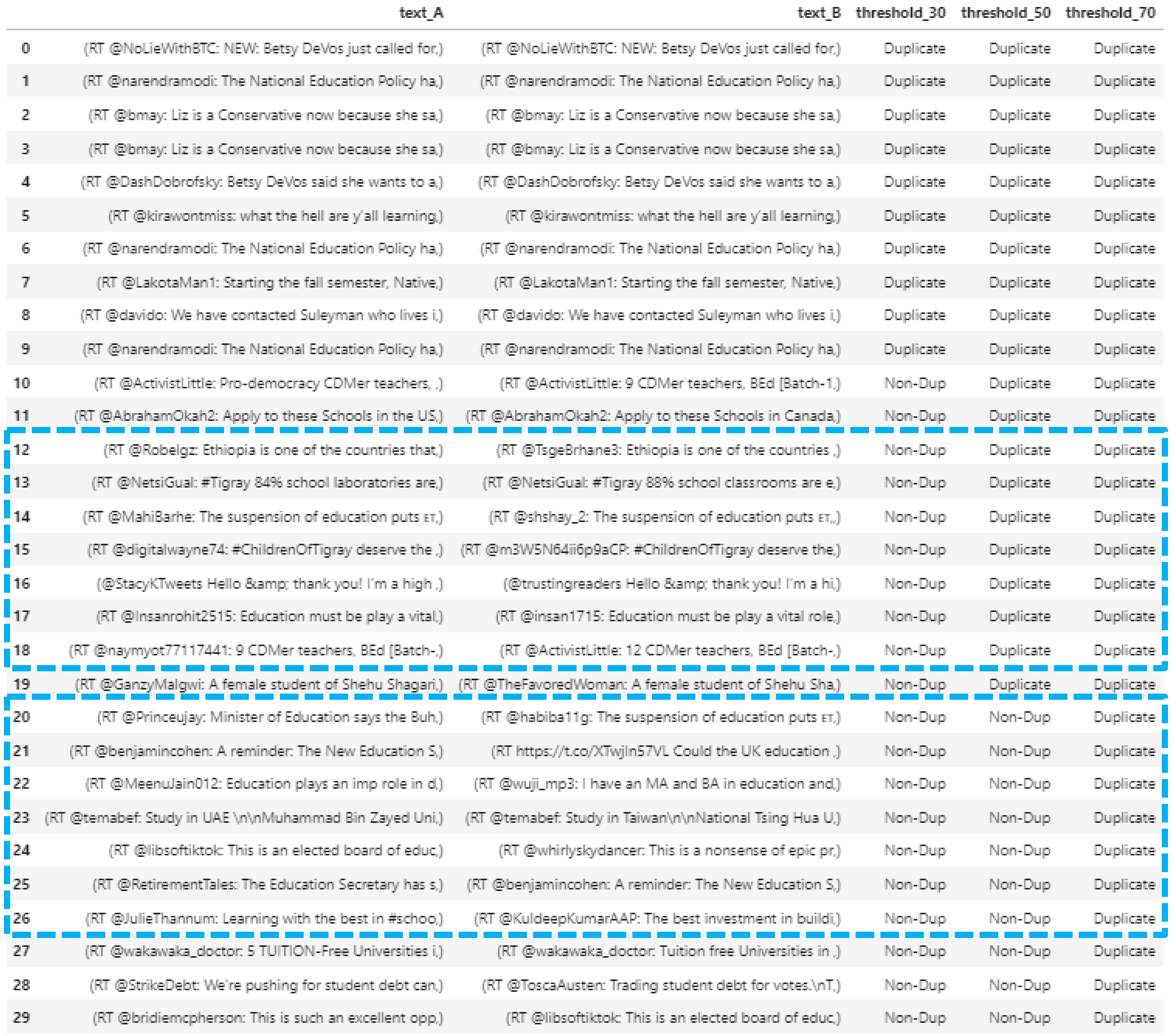
Tweets Duplication Analysis –

Jaccard Similarity (0.7)



*Subset of first 50 words from the tweets text were taken for similarity analysis.

*Although the subset is small, the optimal Jaccard distance of 0.5 appears to be the correct threshold.**



	text_A	text_B	threshold_30	threshold_50	threshold_70
0	(RT @NoLieWithBTC: NEW: Betsy DeVos just called for,)	(RT @NoLieWithBTC: NEW: Betsy DeVos just called for,)	Duplicate	Duplicate	Duplicate
1	(RT @narendramodi: The National Education Policy ha,)	(RT @narendramodi: The National Education Policy ha,)	Duplicate	Duplicate	Duplicate
2	(RT @bmay: Liz is a Conservative now because she sa,)	(RT @bmay: Liz is a Conservative now because she sa,)	Duplicate	Duplicate	Duplicate
3	(RT @bmay: Liz is a Conservative now because she sa,)	(RT @bmay: Liz is a Conservative now because she sa,)	Duplicate	Duplicate	Duplicate
4	(RT @DashDobrofsky: Betsy DeVos said she wants to a,)	(RT @DashDobrofsky: Betsy DeVos said she wants to a,)	Duplicate	Duplicate	Duplicate
5	(RT @kirawontmiss: what the hell are y'all learning,)	(RT @kirawontmiss: what the hell are y'all learning,)	Duplicate	Duplicate	Duplicate
6	(RT @narendramodi: The National Education Policy ha,)	(RT @narendramodi: The National Education Policy ha,)	Duplicate	Duplicate	Duplicate
7	(RT @LakotaMan1: Starting the fall semester, Native,)	(RT @LakotaMan1: Starting the fall semester, Native,)	Duplicate	Duplicate	Duplicate
8	(RT @davido: We have contacted Suleyman who lives i,)	(RT @davido: We have contacted Suleyman who lives i,)	Duplicate	Duplicate	Duplicate
9	(RT @narendramodi: The National Education Policy ha,)	(RT @narendramodi: The National Education Policy ha,)	Duplicate	Duplicate	Duplicate
10	(RT @ActivistLittle: Pro-democracy CDMer teachers, ,)	(RT @ActivistLittle: 9 CDMer teachers, BEd [Batch-1,)	Non-Dup	Duplicate	Duplicate
11	(RT @AbrahamOkah2: Apply to these Schools in the US,)	(RT @AbrahamOkah2: Apply to these Schools in Canada,)	Non-Dup	Duplicate	Duplicate
12	(RT @Robelgz: Ethiopia is one of the countries that,)	(RT @TsgeBrhane3: Ethiopia is one of the countries ,)	Non-Dup	Duplicate	Duplicate
13	(RT @NetsiGual: #Tigray 84% school laboratories are,)	(RT @NetsiGual: #Tigray 88% school classrooms are e,)	Non-Dup	Duplicate	Duplicate
14	(RT @MahiBarhe: The suspension of education puts ET,)	(RT @shshay_2: The suspension of education puts ET,,)	Non-Dup	Duplicate	Duplicate
15	(RT @digitalwayne74: #ChildrenOfTigray deserve the ,)	(RT @m3W5N64ii6p9aCP: #ChildrenOfTigray deserve the,)	Non-Dup	Duplicate	Duplicate
16	(@StacyKTweets Hello &amp; thank you! I'm a high ,)	(@trustingreaders Hello &amp; thank you! I'm a hi,)	Non-Dup	Duplicate	Duplicate
17	(RT @Insanrohit2515: Education must be play a vital,)	(RT @insan1715: Education must be play a vital role,)	Non-Dup	Duplicate	Duplicate
18	(RT @naymyot77117441: 9 CDMer teachers, BEd [Batch-,)	(RT @ActivistLittle: 12 CDMer teachers, BEd [Batch-,)	Non-Dup	Duplicate	Duplicate
19	(RT @GanzyMalgwi: A female student of Shehu Shagari,)	(RT @TheFavoredWoman: A female student of Shehu Sha,)	Non-Dup	Duplicate	Duplicate
20	(RT @Princeujay: Minister of Education says the Buh,)	(RT @habiba11g: The suspension of education puts ET,)	Non-Dup	Non-Dup	Duplicate
21	(RT @benjamincohen: A reminder: The New Education S,)	(RT https://t.co/XTwjln57VL Could the UK education ,)	Non-Dup	Non-Dup	Duplicate
22	(RT @MeenuJain012: Education plays an imp role in d,)	(RT @wuji_mp3: I have an MA and BA in education and,)	Non-Dup	Non-Dup	Duplicate
23	(RT @temabef: Study in UAE \n\nMuhammad Bin Zayed Uni,)	(RT @temabef: Study in Taiwan\n\nNational Tsing Hua U,)	Non-Dup	Non-Dup	Duplicate
24	(RT @libsoftiktok: This is an elected board of educ,)	(RT @whirlyskydancer: This is a nonsense of epic pr,)	Non-Dup	Non-Dup	Duplicate
25	(RT @RetirementTales: The Education Secretary has s,)	(RT @benjamincohen: A reminder: The New Education S,)	Non-Dup	Non-Dup	Duplicate
26	(RT @JulieThannum: Learning with the best in #schoo,)	(RT @KuldeepKumarAAP: The best investment in buildi,)	Non-Dup	Non-Dup	Duplicate
27	(RT @wakawaka_doctor: 5 TUITION-Free Universities i,)	(RT @wakawaka_doctor: Tuition free Universities in ,)	Non-Dup	Non-Dup	Duplicate
28	(RT @StrikeDebt: We're pushing for student debt can,)	(RT @ToscaAusten: Trading student debt for votes.\nT,)	Non-Dup	Non-Dup	Duplicate
29	(RT @bridiemcpherson: This is such an excellent opp,)	(RT @libsoftiktok: This is an elected board of educ,)	Non-Dup	Non-Dup	Duplicate

In this outlined subset, Jaccard distance of 0.5 and 0.7 appears to correctly classify into non-duplicates relative to 0.3.

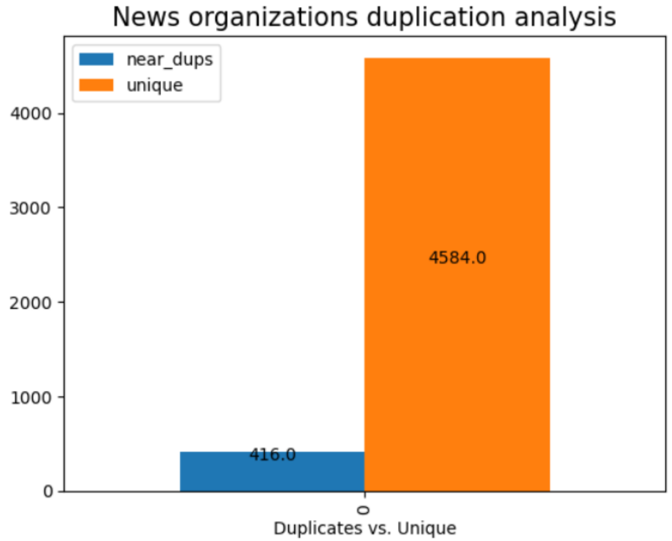
In this outlined subset, Jaccard distance of 0.5 appears to correctly classify into non-duplicates relative to 0.7.

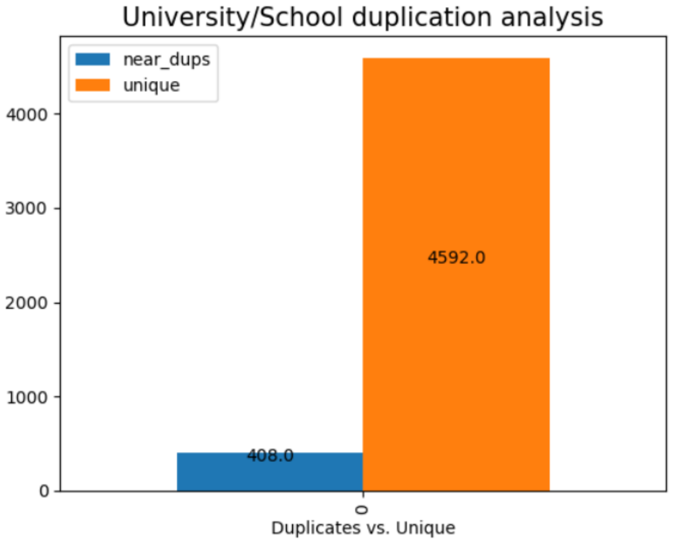
*Subset of first 50 words were taken for similarity analysis.

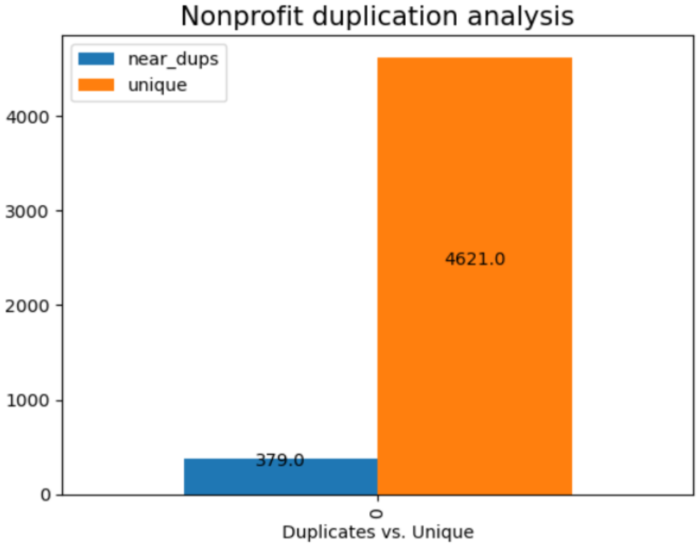
5th December 2022

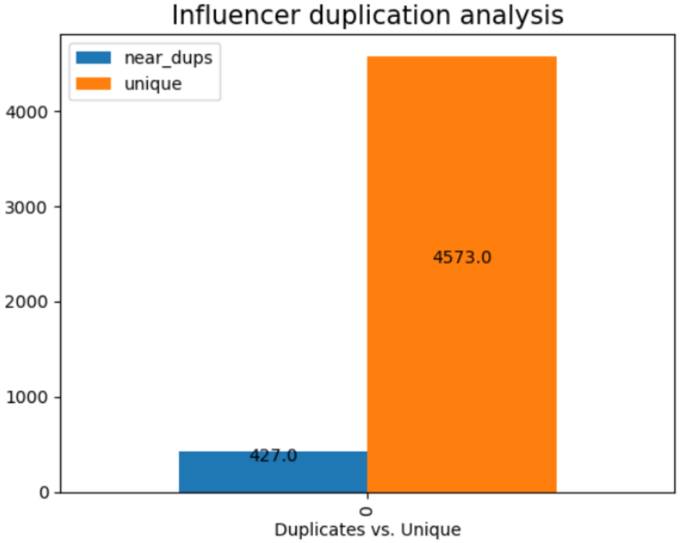
Using Jaccard Similarity of 0.5, classified organizations have significantly less duplicates relative to the population*

Given limits of 5000, duplicate ratios were very similar across types of organizations (news organizations, university/school, nonprofit, and influencers) and are in the range of 7.5% to 8.5%. Given that the general population of tweets that had a 30% duplication ratio at Jaccard similarity of 0.5, it is important to note that highly reputed organizations are less likely to copy and paste the same text while other individual twitterers are more likely to do that.


News organizations duplication analysis


University/School duplication analysis


Nonprofit duplication analysis


Influencer duplication analysis

*Subset of first 50 words from the tweets text were taken for similarity analysis.

Conclusion

- *Twitter is not a reliable source of information because of variety of organizations and individuals in the ecosystem – difficult for users to differentiate between reliable and unreliable*
- *The most prolific twitterers with original tweets are private institutions and individuals who have financial interests in the messages they tweet / retweet while reliable news organizations and verified users do not consistent of the large population of twitterers.*
- *There is widespread geographical distribution and consistent volumes of tweets across time*
- *Tweet similarity should be analyzed further to see if important information is being shared*

Actionable Recommendations

- *Twitter could create newsfeeds, where verified users and reputed organizations like news organizations can tweet – creating a new ecosystem for gathering reliable information*
- *Twitter could identify authors they believe to have financial incentives to market products / etc by issuing warnings to users while also highlighting reliable users through a filtering or verifiable mechanism*
- *Twitter could also highlight through text similarity of a tweet has been copy and pasted numerous times to suggest it is important information*