Analyzing Social Media Data to Detect and Map Suspicious Individuals

Danayal Khan

I. Introduction

The use of Social Media in organizing and promoting protests was made popular during the Arab Spring revolution. Social Media is a powerful tool that allows users to communicate across borders easily; and to spread their propaganda effectively. As a result, it is easy to misuse the power of open communication.

Affiliates of terrorist organizations such as ISIS and the Taliban actively use Twitter to promote hate speech and spread their propaganda. Furthermore, they target and recruit susceptible individuals through the strong ideologies they spread online. It is essential for governments and social media companies to monitor and track these individuals in order to ensure the safety of their population and their sovereignty.

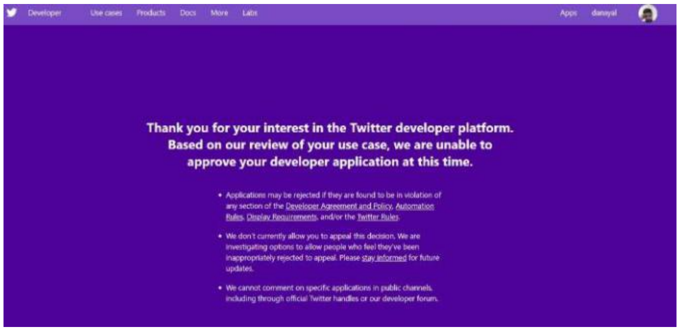
Before strict regulation of social media, affiliates of terrorist organizations would use their real names since there was little that was done to control the social media space. However, Twitter has recently enforced strict community guidelines and aims to enforce these guidelines effectively. As a result, a lot of accounts with pseudonyms and random characters have been created to spread hate speech. This makes it difficult to identify suspicious individuals.

II. Solution

I aim to use data mining and data analysis techniques to mine tweets with certain keywords and determine the users who most frequently use these keywords. The keywords to be used are common terms used by affiliates of terrorist organizations. Once a list of users is established, I aim to identify which users follow each other and make a social graph that connects these individuals. By mining for certain keywords and automatically connecting suspicious individuals, one can overcome the hindrance of finding suspicious individuals who use random characters as their username, furthermore, this technique can be applied to mine for any set of keywords. The social graph can then be used to determine who has the most influence in the social media space and governments or technology companies can take action accordingly.

III. Procedure

Twitter has restricted only approved developers to use their API. This is a good step in terms of security as it makes it more difficult for ill-intentioned developers, hackers, or journalists to collect tweets on a mass-scale. I tried requesting for their developer API but Ire rejected.

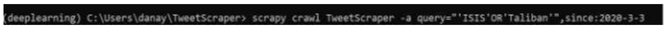


Thank you for your interest in the Twitter developer platform. Based on our review of your use case, we are unable to approve your developer application at this time.

- Applications may be rejected if they are found to be in violation of any section of the Developer Agreement and Policy, Automation Rules, Display Requirements, and/or the Twitter Rules.
- We don't currently allow you to appeal this decision. We are investigating options to allow people who feel they've been inappropriately rejected to appeal. Please stay informed for future updates.
- We cannot comment on specific applications in public channels, including through official Twitter handles or our developer forum.

Fig. 1. Rejection for Accessing Twitter Developer API

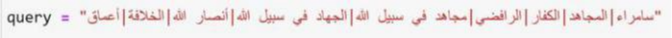
As a result, I resorted to a third-party API called TweetScraper to scrape user tweets using certain queries. The following is an example of a spider crawler that uses the query "ISIS" or "Taliban" that saves a list of tweets with those keywords in a folder.



```
(deeplearning) C:\Users\danay\TweetScraper> scrapy crawl TweetScraper -a query="'ISIS'OR'Taliban'",since:2020-3-3
```

Fig. 2. An example of a query used to scrape Tweets with keywords

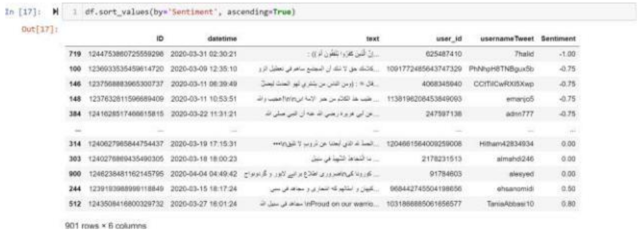
After some initial testing, I realized that English key- words do not return desirable results; the list of accounts mined from these keywords Ire not suspicious. After con- ducting research, I tried using Arabic keywords such as "دهاجملا" as my query and it returned significantly better results. The following is a query I have used for our initial results. The list of query was taken from [1].



```
query = "سامراء|المجاهد|الكفار|الرافضي|مجاهد في سبيل الله|الجهاد في سبيل الله|أنصار الله|الخلافة|أعماق"
```

Fig. 3. Query using Arabic keywords

By using a sentiment analyzer, the list of tweets returned Ire sorted according to the most negative tweets first in a Pandas dataframe.



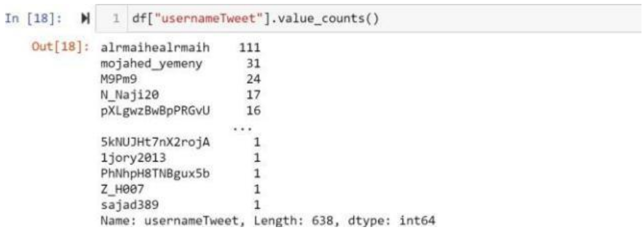
```
In [17]:  1 df.sort_values(by='Sentiment', ascending=True)
Out[17]:
```

	ID	datetime	text	user_id	usernameTweet	Sentiment
719	1244753860725559296	2020-03-31 02:30:21	[illegible]	625487410	7halid	-1.00
100	1236933535459614720	2020-03-09 12:35:10	[illegible]	1091772485643747329	PhNhpH8TNBgux5b	-0.75
146	1237568883965300737	2020-03-11 06:39:49	[illegible]	4068345940	CCfTilCwRX05Xwp	-0.75
148	1237632811596689409	2020-03-11 10:53:51	[illegible]	1138198208453849093	emanjo5	-0.75
384	1241628517466615815	2020-03-22 11:31:21	[illegible]	247597138	adm777	-0.75
...	...	...	...	...	...	...
314	1240627985844754437	2020-03-19 17:15:31	[illegible]	1204661564009259008	Hitham42834934	0.00
303	1240276869435490305	2020-03-18 18:00:23	[illegible]	2178231513	almahdi246	0.00
900	1246238481162145795	2020-04-04 04:49:42	[illegible]	91784603	alsayed	0.00
244	1239193988999118849	2020-03-15 18:17:24	[illegible]	968442745504198656	ehsanomidi	0.50
512	1243508416800329732	2020-03-27 16:01:24	[illegible] \nProud on our warrio...	1031868885061656577	TaniaAbbasi10	0.80

901 rows × 6 columns

Fig. 4. Tweets sorted according to their sentiment

The most frequent occurring users were determined for further analysis.

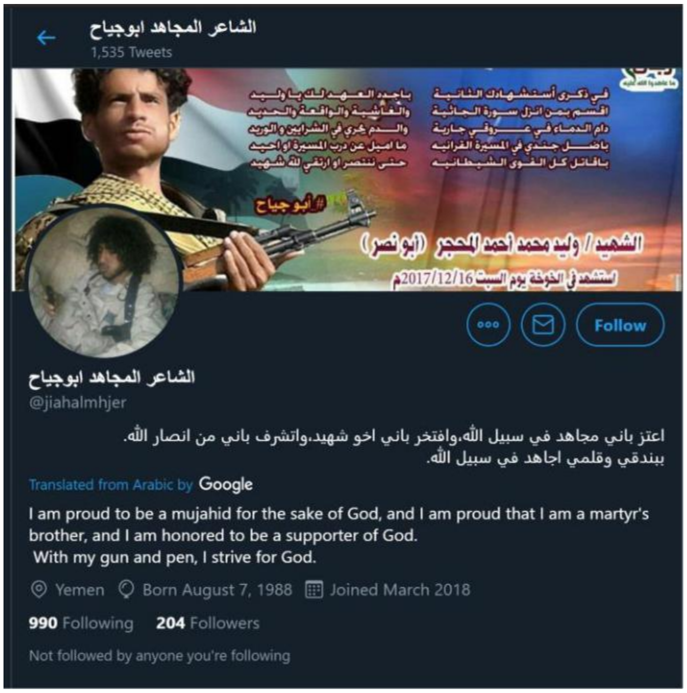


```
In [18]:    1  df["usernameTweet"].value_counts()

Out[18]:  alrmaihealrmaih    111
          mojahed_yemeny      31
          M9Pm9               24
          N_Naji20            17
          pXLgwzBwBpPRGvU     16
                             ...
          5kNUJHt7nX2rojA      1
          1jory2013            1
          PhNhpH8TNBgux5b      1
          Z_H007               1
          sajad389             1
          Name: usernameTweet, Length: 638, dtype: int64
```

Fig. 5. Most frequent occurring users

The following are the profiles of a couple of suspicious individuals determined by the analysis conducted.


#أبو جياح
الشهيد / وليد محمد أحمد المحجر (أبو نصر)
استشهد في الخوخة يوم السبت 2017/12/16م
•••
Follow

الشاعر المجاهد ابوجياح
1,535 Tweets

الشاعر المجاهد ابوجياح
@jiahalmhjer

اعتز باني مجاهد في سبيل الله،وافتخر باني اخو شهيد،واتشرف باني من انصار الله.
ببندقي وقلمي اجاهد في سبيل الله.

Translated from Arabic by Google

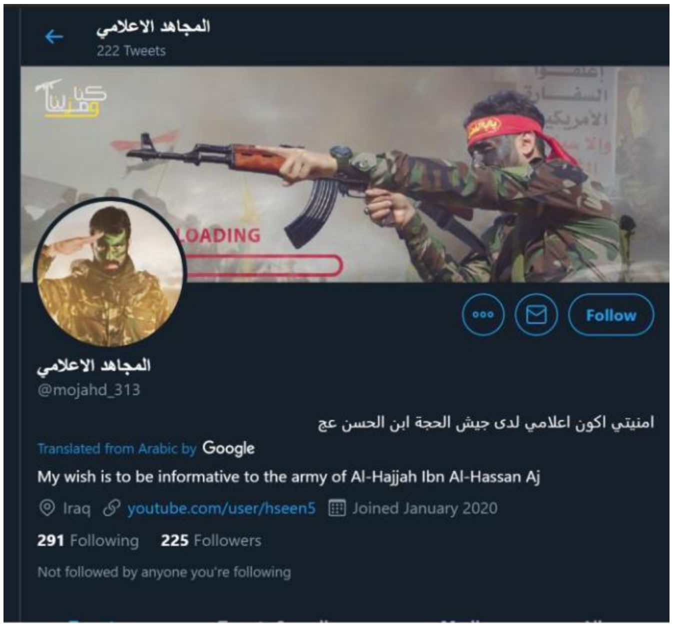
I am proud to be a mujahid for the sake of God, and I am proud that I am a martyr's brother, and I am honored to be a supporter of God.
With my gun and pen, I strive for God.

Yemen Born August 7, 1988 Joined March 2018

990 Following 204 Followers

Not followed by anyone you're following

Fig. 6. Twitter Profile of Suspect 1



المجاهد الاعلامي
222 Tweets

المجاهد الاعلامي
@mojahd_313

امنيتي اكون اعلامي لدى جيش الحجة ابن الحسن عج

Translated from Arabic by Google

My wish is to be informative to the army of Al-Hajjah Ibn Al-Hassan Aj

Iraq youtube.com/user/hseen5 Joined January 2020

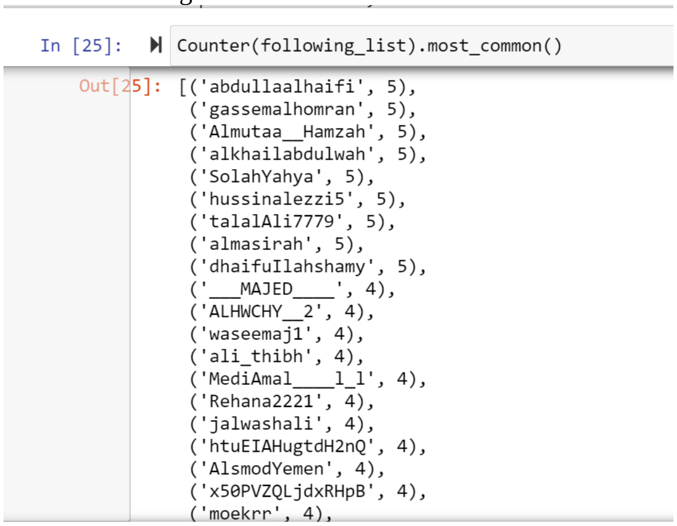
291 Following 225 Followers

Not followed by anyone you're following

Fig. 7. Twitter Profile of Suspect 2

IV. Relationships

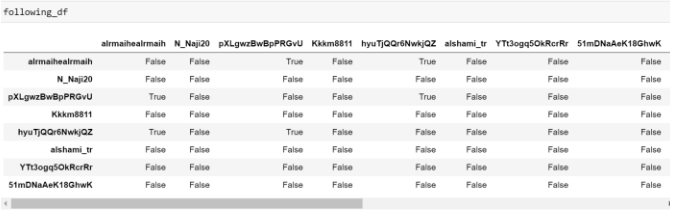
From the list of all the users retrieved, the followers of each user was analysed to see if any of the other users in the list follow them. The users with the most common followers amongst the list were determined and deemed to be the most influential amongst all users found.



```
In [25]:  Counter(following_list).most_common()

Out[25]:  [('abdullaalhaifi', 5),
           ('gassemalhomran', 5),
           ('Almutaa__Hamzah', 5),
           ('alkhailabdulwah', 5),
           ('SolahYahya', 5),
           ('hussinalezzi5', 5),
           ('talalAli7779', 5),
           ('almasirah', 5),
           ('dhaifuIlahshamy', 5),
           ('___MAJED____', 4),
           ('ALHWCHY__2', 4),
           ('waseemaj1', 4),
           ('ali_thibh', 4),
           ('MediAmal____l_l', 4),
           ('Rehana2221', 4),
           ('jalwashali', 4),
           ('htuEIAHugtdH2nQ', 4),
           ('AlsmodYemen', 4),
           ('x50PVZQLjdxRHpB', 4),
           ('moekrr', 4),
```

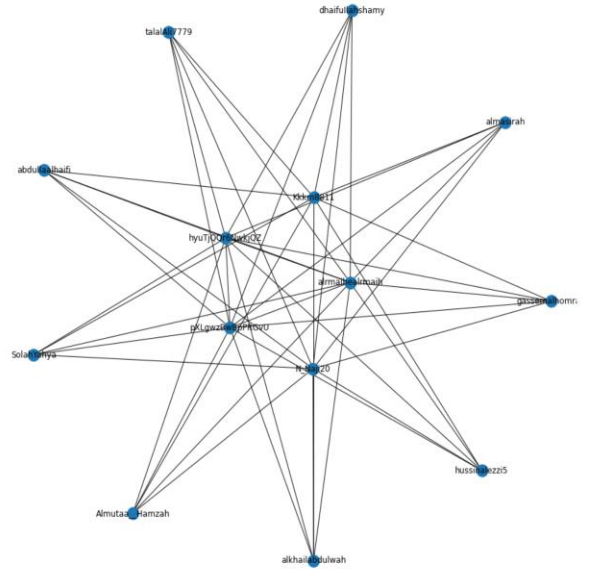
These influential users were separated in a new dictionary as keys and their follows were inserted as the value. Using these dictionaries, a matrix was created to visualize the follow/following relationship amongst the influential users and the followers. The users on the columns are followed by the users on the columns if the value is returned true.



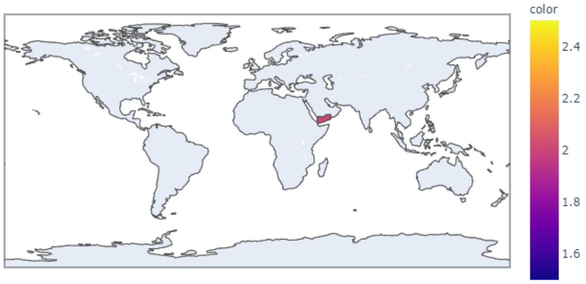
```
following_df
```

	alrmaihealrmaih	N_Naji20	pXLgwzBwBpPRGvU	Kkkm8811	hyuTjQQr6NwkjQZ	alshami_tr	YTt3ogq5OkRcrRr	51mDNaAeK18GhwK
alrmaihealrmaih	False	False	True	False	True	False	False	False
N_Naji20	False	False	False	False	False	False	False	False
pXLgwzBwBpPRGvU	True	False	False	False	True	False	False	False
Kkkm8811	False	False	False	False	False	False	False	False
hyuTjQQr6NwkjQZ	True	False	True	False	False	False	False	False
alshami_tr	False	False	False	False	False	False	False	False
YTt3ogq5OkRcrRr	False	False	False	False	False	False	False	False
51mDNaAeK18GhwK	False	False	False	False	False	False	False	False

From this analysis, it is easy to create a social network graph with end nodes being the users and the edges being a follow/following relationship. The users with the most connected edges were placed at the center of the graph since they are deemed to be the most influential suspicious indivuals.



Users on Twitter have an option to display their location. Since this is an opt-in feature, not a lot of users have their locations public by default. However, a heat graph depicting the most frequent location of these suspicious indivudals was created and shown in the following figure.



V. Relationships

The work conducted in this project highlights the impact that simple web scraping and data analysis have on finding insights. It also shows how easy and widespread it is to create an account with random alphanumeric combinations and promote hate speech without getting detected by Twitter. A social graph was produced that highlights the most influential suspicious individuals that should be a focus on an investigation by Twitter or relevant authorities.

For the full code, please visit: https://github.com/Danayal/Analyzing-Social-Media-Data-to-Detect-and-Map-Suspicious-Individuals/blob/master/security.ipynb

References

[1] R. Gupta and H. Brooks, *Using social media for global security*. Indianapolis, IN: John WIley & Sons, Inc, 2013.