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.

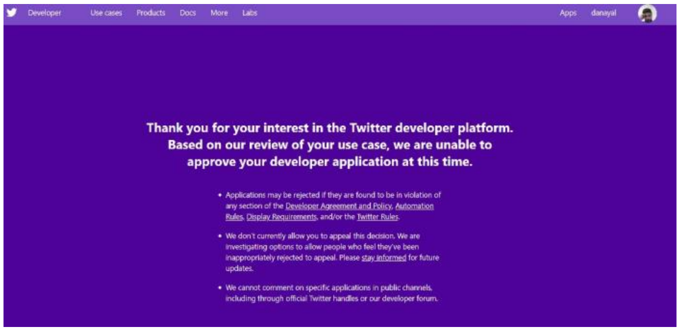
Before strict regulation of social media, affiliates of terrorist organizations would use their real names since there was little that was done to control the social media space. However, Twitter has recently enforced strict community guidelines and aims to enforce these guidelines effectively. As a result, a lotof 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 individual 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.

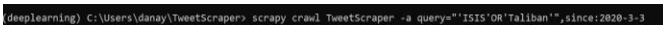


Thank you for your interest in the Twitter developer platform. Based on our review of your use case, we are unable to approve your developer application at this time.

- 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

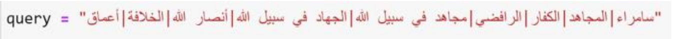
As a result, I resorted to a third-party API called TweetScraper to scrape user tweets using certain queries. The following is an example of a spider crawler that uses the query "ISIS" or "Taliban" that saves a list of tweets with those keywords in a folder.



```
(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

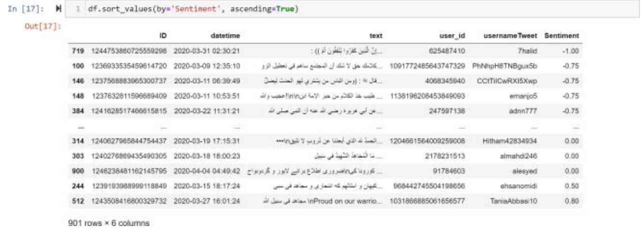
After some initial testing, I realized that English key- words do not return desirable results; the list of accounts mined from these keywords Ire not suspicious. After con- ducting research, I tried using Arabic keywords such as "المجاهد" as our query and it returned significantly better results. The 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.

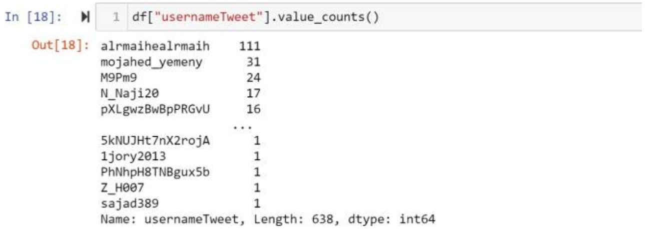


```
In [17]:  df.sort_values(by='Sentiment', ascending=True)
Out[17]:
                     ID             datetime                                          text              user_id     usernameTweet  Sentiment
 719  1244753860725559298  2020-03-31 02:30:21                 ...إنّ الذين كفروا ينفقون أم )) :            625487410            7halid      -1.00
 100  1236933535459614720  2020-03-09 12:35:10  ...كلامك حق لا شك أن المجتمع ساهم في تعطيل الزو  1091772485643747329  PhNhpH8TNBgux5b      -0.75
 146  1237568883965300737  2020-03-11 06:39:49   ...قال = : (ومن الناس من يشتري لهو الحديث ليضل           4068345940  CCrTiICwRXI5Xwp      -0.75
 148  1237632811596689409  2020-03-11 10:53:51  ... طيب هذا الكلام من خير الامة ابن\n\n!عجيب والله  1138196208453849093           emanjo5      -0.75
 384  1241628517466615815  2020-03-22 11:31:21     ...عن أبي هريرة رضي الله عنه أن النبي صلى الله            247597138           adnn777      -0.75
 ...                  ...                  ...                                           ...                  ...               ...        ...
 314  1240627965844754437  2020-03-19 17:15:31    ...الحمد لله الذي أبعدنا عن ترويب لا تليق\n•••  1204661564009259008    Hitham42834934       0.00
 303  1240276869435490305  2020-03-18 18:00:23                       ... ما المجاهد الشهيد في سبيل           2178231513         almahdi246       0.00
 900  1246238481162145795  2020-04-04 04:49:42  ... کورونا کی\nضروری اطلاع برائے لاہور و گردونواح             91784603            alesyed       0.00
 244  1239193988999118849  2020-03-15 18:17:24        ...کیهان و امثالهم که انتحاری و مجاهد فی سبی   968442745504198656          ehsanomidi       0.50
 512  1243508416800329732  2020-03-27 16:01:24     ...مجاهد في سبيل الله \nProud on our warrio  1031866885061656577       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.

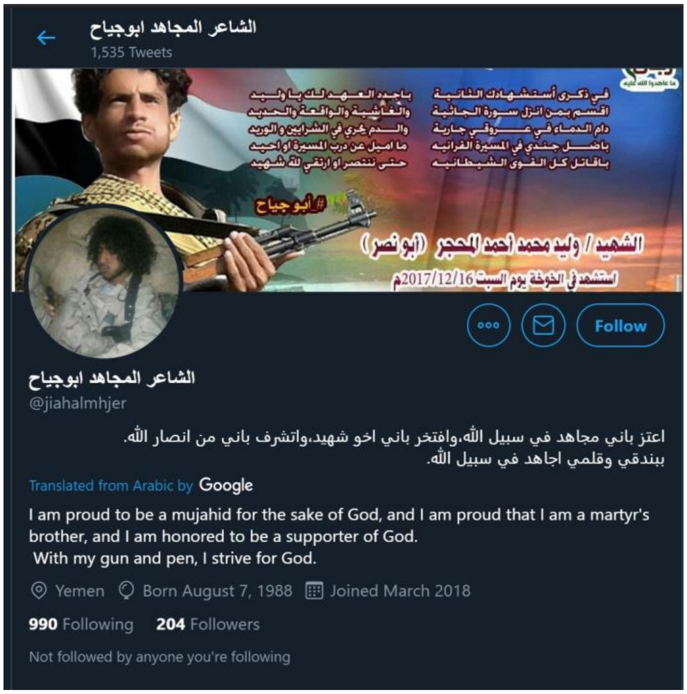


```
In [18]:  df["usernameTweet"].value_counts()

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.



الشاعر المجاهد ابوجياح
1,535 Tweets

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

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

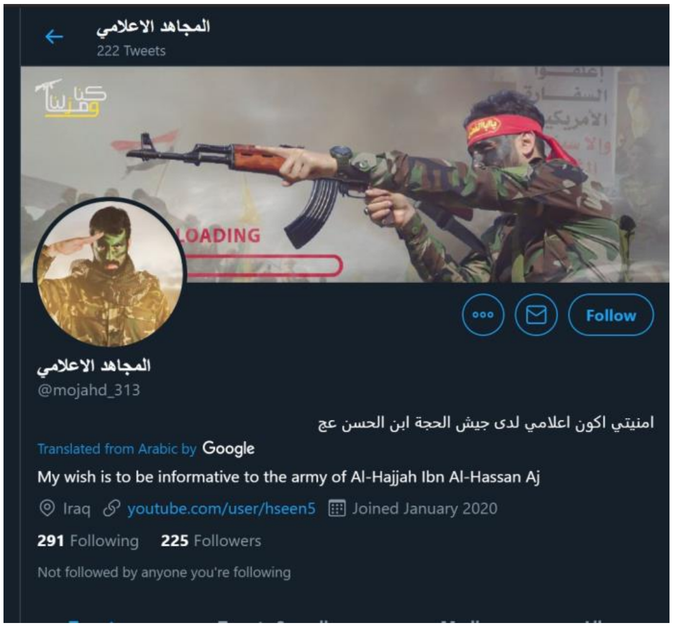
I am proud to be a mujahid for the sake of God, and I am proud that I am a martyr's brother, and I am honored to be a supporter of God.
With my gun and pen, I strive for God.

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

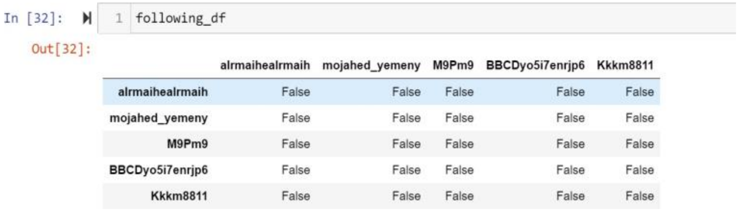
291 Following 225 Followers

Not followed by anyone you're following

Fig. 7. Twitter Profile of Suspect 2

IV. Work To Be Conducted

During our initial analysis, I analyzed five accounts but did not find any follow/following relationship amongst them as is described in the figure below:



```
In [32]:  following_df

Out[32]:
```

	alrmaihealrmaih	mojahed_yemeny	M9Pm9	BBCDyo5i7enrjp6	Kkkm8811
alrmaihealrmaih	False	False	False	False	False
mojahed_yemeny	False	False	False	False	False
M9Pm9	False	False	False	False	False
BBCDyo5i7enrjp6	False	False	False	False	False
Kkkm8811	False	False	False	False	False

Fig. 8. Follow/Following Relationship

This was expected as so far, I have not collected a large amount of data. My goal is to collect sufficient data to successfully analyze relationships amongst suspects and derive a social graph that depicts the most influential suspects.

References

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