Project 7: Studying ISIS Twitter influence with social network analysis from the pro-ISIS fanboy tweet data

*Note: Sub-titles are not captured in Xplore and should not be used

1st Niklas Saari
dept. Computer Science and Engineering
University of Oulu
Oulu, Finland
niklas.saari@student.oulu.fi

***Abstract*—This document describes project work for the course Social Network Analysis in Spring 2021. Radical and extreme groups have taken social networks as part of their toolchain to spread propaganda messages, get attention, look support for their actions and recruit more members. One of the most brutal of these groups, ISIS, which is also designated as terrorist organization, has been forerunner on using these social platforms successfully. To possibly prevent future terror events, it would be important to study these social networks on how they are used by these radical groups. ISIS Twitter dataset of around 17 000 tweets was selected as dataset to identify important characters from the network, to construct communities and to see how the most influencers behave, who they are, and how they connect to other people. Further communities were constructed to see how networks behave internally.**

***Index Terms*—Twitter, Social Network Analysis, Terrorism, Networks, ISIS**

I. Group information

This project has been done alone. The original project from the given project list was 7 with the title of ”Analysis of ISIS Twitter dataset”.

II. Introduction

Social networks have become part of the most people's everyday life. These networks, such as Facebook, Twitter or Reddit are base for many kinds of groups and people for communicating with each other. They are being used widely for expressing opinions or advertising services and for many other kinds of things. This has naturally risen special interest in radical and terrorism organisations because of the provided possibility for influencing different kind of people with a large scale. While the brutality of terrorism has become even more severe over recent time based on the data gathered by Global Terrorism Database and seen in Appendixes in the figure 1, this raises specific interest in terms of identifying and preventing possible future incidents.

Terrorism is identified to be especially brutal from Islamic State of Iraq and Levant (ISIL), which is also known ISIS or Daesh. [1] Their usage of the social media is also known to be ”probably more sophisticated than [that of] most US companies”, [2] and has been one of their main campaigning tactics in Syria and Iraq. Twitter has been their primal social network [3] so far. It is known that ISIS has been previously organising for example specific hashtag campaigns to get their topics trending and gain more visibility [3].

Because the social networks have begun to be in key role to motivate support for their actions, raise funds and even for recruiting foreign fighters with huge success [4] allowing organisation to reach worldwide recognition and impact, it is important to study how these networks are formed and what we could learn from them and how we could act based on this information.

Social Network Analysis (SNA) is a process for studying social life by social structures which is constructed by relations and patterns formed by these relations. This is usually done with networks and applying the graph theory. [5], [6]

In this article we will focus particularly on the Twitter platform and for the Social Network Analysis of how ISIS has been using this specific platform to spread their propaganda and organizing recruitment. Twitter data of over 17 000 tweets of pro-ISIS fanboys has been used as dataset for this study.

We will try to identify major characters from the provided data, for example which characters have the most influence and what kind of networks they are constructing. This is evaluated based on different values specific for Twitter platform, such as usage of mentions, retweets or hashtags; how are different Twitter users using them.

We further try to estimate the sentiment of the tweets in terms of negative, neutral and positive and estimate the most frequent hashtags and their possible context and characteristics. Different kind of graphs will be constructed to develop our understanding of underlying network. Social network is constructed by using hashtags, and their relations to other tweets based on their appearances.

This document is structured as following: in the section III the main problem has been described. In the section IV the exact details of the used dataset has been described. In the

section V general methodology has been described and further continued in the section VI in more precise matter. The results are presented in the section VII and paper has been finally concluded in the section VIII.

III. Problem Description

The main problem is to study and tell how we can find specific communities from underlying dataset and identify interesting numerical values from the constructed network. Can we detect specific patterns or behaviours related to specific Twitter users, is there connection between them and how powerful their influence actually is. We further try to find a way to tell about what kind of messages they are distributing in overall. Dataset was not collected by itself, instead it was given in the project assignment.

IV. Dataset description

Twitter dataset of tweets collected from pro-ISIS fanboys of all over the world has been used as a base for this study. This dataset was provided with the project assigment. Based on the same project assigment, the origin of the dataset is unknown, as it is stated to be published in dark web website. However, after doing some research, it seems that this dataset is probably collected by Fifth Tribe digital agency, and published originally on the *Kaggle.* [7] Data is under Creative Commons 0 (CC0) license and can be used without restrictions to the fullest extent allowed by law. Data was created originally with the intention of "to develop effective counter-messaging measures against violent extremists at home and abroad." [7]

Tweets are located during the period of 1st of June 2015 and 13rd of May 2016, which contains the November 2015 Paris attack as interesting point of event regarding the context of this study. Tweets are had been written with multiple languages, but in general they are in English. Their content is varying a lot; they could be text with varying context, external links to other places, images and videos or rewteets.

Dataset contains total of 17410 different tweets by 112 different users, and was originally given in the newer Excel format (**.xlsx**). Following data columns can be found from the raw data:

- name
- username
- location
- number of followers
- number of statuses
- time (month/day/year 24-hour clock)
- tweet (multilingual)

Location is user supplied data and can be therefore anything.

A. Pre-processing of the data

As the initial dataset was given in Microsoft Excel Open XML (.xlsx) format and it required some conversion to be more suitable for processing with programmatically with selected programming language (Python in this case) and in general for easier handling and compatibility. Dataset was converted to basic **.csv** file format by using Python **pandas** [8] library with **openpyxl** [9] engine. Successful conversion was verified later by checking that there were no null data shells for columns which are considered as "important" and the amount of rows matches with original and converted data. "Important" means in this context that every tweet should have at least username and tweet content to be meaningful.

The data has been on some cases further pre-processed as following to extract some specific information and features. This information is stored programmatically on the run-time-memory by creating specific Python class object to represent single line from the dataset data, which also contains extracted additional data. Extracting methods have been discussed in more detail on the section VI.

1) Mentions: Mentions of different users have been extracted from the every tweet based on the '@' symbol in tweet data. Twitter usernames are case-insensitive and therefore as and additional step, they are stored in lowercase to improve accuracy of the data and also to reflect real world behaviour when linking to other tweets. This is implemented by using specific regex patterns.

2) Retweets: Retweets are identified from the data based on 'RT' as first word in the tweet.

3) Hashtags: Hashtags have been extracted from the every tweet based on the '#' symbol in tweet data. Twitter hashtags are case-insensitive and therefore as and additional step, they are stored in lowercase to improve accuracy of the data and also to reflect real world behaviour when linking to other tweets. This is implemented by using specific regex patterns.

4) Sentiment analysis: Sentiment analysis is applied for every tweet in the dataset to describe the potential category in terms of *negative, neutral and positive.* Python package named as VADER Sentiment, which was originally presented in the article "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text" [10] has been used as tool for scoring the data for these categories. This is discussed in more details on the section VI.

B. Data verification after some pre-processing methods

As there were many methods implemented for extracting information from the dataset, some testcases were applied and random data was selected to be sure, that extraction is working as expected. This was also applied for data conversion. Testing was implemented by using Python package **pytest** [11], and based on the limited test cases data is extracted as intended.

V. General methodology

VI. Detailed methodology

Power law

A. Degree Centrality

Degree centrality is the sum of in-degree and out-degree. It is representing the amount of edges entering and leaving the nodes respectively. The most important nodes have the most direct connections with others under degree centrality. The value can be computed as,

$$C_D(v_i) = \sum\nolimits_j A_{ij} \tag{1}$$

B. Betweeness Centrality

Betweeness centrality is another way to measure importance of nodes. It is describing the amount of the shortest path passing the node. Important nodes have high betweeness centrality, information is flowing through them, and they are connecting multiple nodes into the network.

$$C_B(v_i) = \sum_{v_s \neq v_i \neq v_t \in V, s<\mathrm{t}} \frac{\sigma_{st}(v_i)}{\sigma_{st}} \tag{2}$$

VII. Results and Discussion

VIII. Conclusion and Perspectives

A. Authors and Affiliations

B. Identify the Headings

C. Figures and Tables

Acknowledgment

References

[1] Wikipedia contributors, "Islamic state of iraq and the levant — Wikipedia, the free encyclopedia," https://en.wikipedia.org/w/index.php?title=Islamic_State_of_Iraq_and_the_Levant&oldid=1024074085, 2021, [Online; accessed 19-May-2021].

[2] R. Khalaf and S. Jones. (2014) Selling terror: how isis details its brutality. [Online]. Available: https://www.ft.com/content/69e70954-f639-11e3-a038-00144feabdc0

[3] J. Berger. (2014) How isis games twitter. [Online]. Available: https://www.theatlantic.com/international/archive/2014/06/isis-iraq-twitter-social-media-strategy/372856/

[4] J. Stern and J. Berger. (2015) Isis and the foreign-fighter phenomenon. [Online]. Available: http://www.theatlantic.com/international/archive/2015/03/isis-and-the-foreign-fighter-problem/387166/

[5] A. Marin and B. Wellman, "Social network analysis: An introduction," *The SAGE handbook of social network analysis*, vol. 11, p. 25, 2011.

[6] E. Otte and R. Rousseau, "Social network analysis: a powerful strategy, also for the information sciences," *Journal of Information Science*, vol. 28, no. 6, pp. 441–453, 2002. [Online]. Available: https://doi.org/10.1177/016555150202800601

[7] F. Tribe. (2019) How isis uses twitter. [Online]. Available: https://www.kaggle.com/fifthtribe/how-isis-uses-twitter

[8] W. McKinney *et al.*, "Data structures for statistical computing in python," in *Proceedings of the 9th Python in Science Conference*, vol. 445. Austin, TX, 2010, pp. 51–56.

[9] E. Gazoni and C. Clark. (2021) openpyxl - a python library to read/write excel 2010 xlsx/xlsm files. [accessed 19-May-2021]. [Online]. Available: https://openpyxl.readthedocs.io/en/stable/

[10] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8, no. 1, May 2014. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/14550

[11] H. Krekel, B. Oliveira, R. Pfannschmidt, F. Bruynooghe, B. Laugher, and F. Bruhin, "pytest 6.2.3," 2004. [Online]. Available: https://github.com/pytest-dev/pytest

Appendix

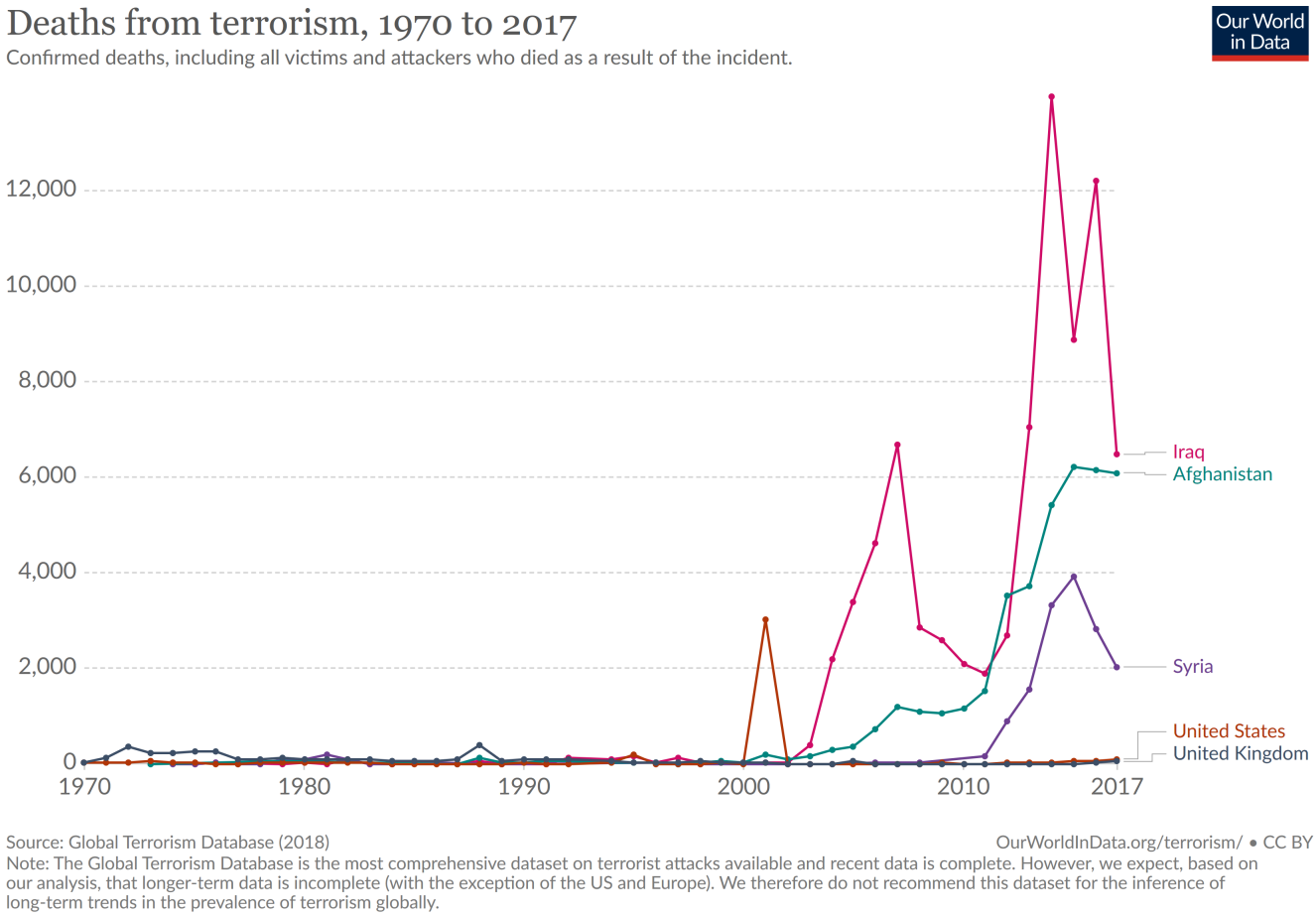

Fig. 1. Deaths from terrorism, 1970–2017.