**Title:** How ISIS used Twitter

**Team members** : Geethika Marru ([gjm5597@psu.edu](mailto:gjm5597@psu.edu))

Lakshmi Kausalya Kavuri ([lkk5341@psu.edu](mailto:lkk5341@psu.edu))

Sai Surya Sudha Soundarya Mantha ([spm6677@psu.edu](mailto:spm6677@psu.edu)

**Abstract**:

With the increased use of social media by all communities, Twitter plays prominent role in interaction and expression of ideas.  ISIS is one of the most barbaric and technology savvy armed forces that is using media to project political and religious institutions. ISIS has a new pool of potential supporters in social media. This project focuses on analysis of tweets by ISIS fanboys in 2016 following the Brussels attacks.

**Introduction**:

Text mining or text analytics is technology that uses unstructured text data from documents and databases, normalizes to structured data suitable for analysis. Examining large document collection to discover new information/ patters/ emotions/ keywords help answer research questions. Text analytics systems today with advanced algorithms analyze large text-based data in an unbiased manner. Data from over 17,000 tweets from 100+ pro-ISIS fanboys is scraped from all over the world post the November 2015 Paris Attacks.  The data is collection of large number of Arabic tweets referring to ISIS. The project uses this data to check the networks, most frequent words and the sentiment of each tweet.

**Characteristics of dataset:**

Source: <https://www.kaggle.com>

Size: 17410 rows, 8 columns

Data Description: Columns present -

1. Name
2. Username
3. Description
4. Location
5. Number of followers at the time the tweet was downloaded
6. Number of statuses by the user when the tweet was downloaded
7. Date and timestamp of the tweet
8. The tweet itself

The ‘tweets’ column is the main column to perform text analytics.

Visualization: The two numeric columns number of followers and number of statuses. We can find correlation between the two variables:

Graphical user interface

Description automatically generated with medium confidence

**Data pre-processing steps:**

Data pre-processing is an important step in data mining process to transform raw data into an understandable format.

The pre-processing steps performed on the text data are:

* Convert to lower case
* Remove punctuation
* Remove apostrophe, single characters, punctuation
* Convert numbers
* Remove stop words
* Remove punctuation

**Analysis Performed using the data:**

1. ***Time series Analysis***: Time series analysis is analyzing data points collected over an interval of time. The particular data set considered has tweets collected/ scraped over 2016 over a span of 5 months.

The number of tweets per day across the time span:

Chart, line chart, histogram

Description automatically generated

Identifying the trend using rolling average to check if there is any possible trend:

Chart, line chart

Description automatically generated

Checking for the 2 attacks made by ISIS during the first half of the year and looking at the trend of the tweets:

Chart, line chart

Description automatically generated

***Comments:*** We find an increasing trend in the number of tweets made over the considered time period. We can also find an extreme hike in the number of tweets that are being made during the immediate days of the attack.

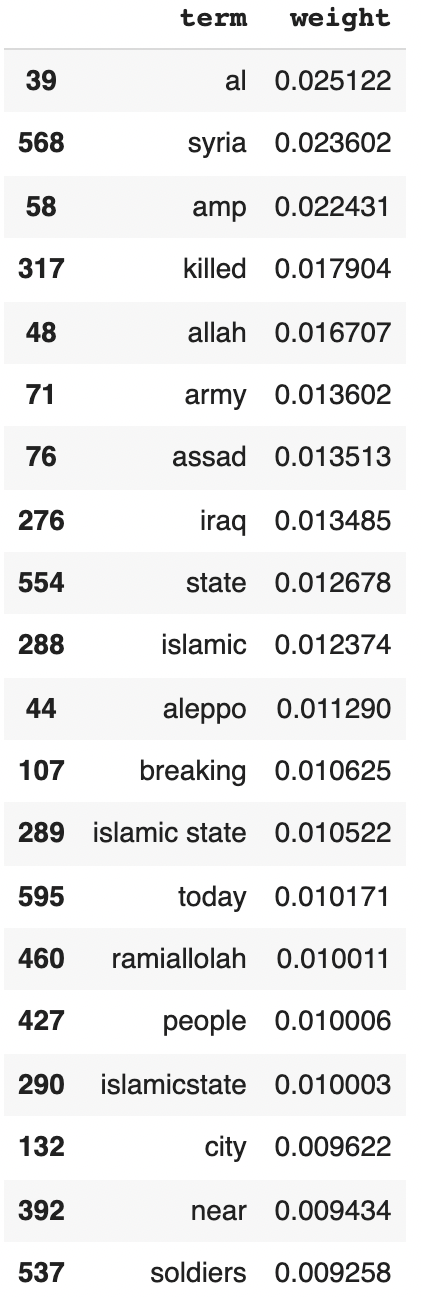
1. ***Users with tweets having the maximum followers:***

Chart, bar chart

Description automatically generated

The average value of followers = 3975. This value might be biased due to the extreme values.

1. ***Using TF-IDF to check the top 20 weighted words from the tweets column:* Term Frequency-Inverse Document Frequency** is a statistical measure that evaluates how relevant a word is to a document in a collection of documents.



1. ***Word Cloud:*** Word clouds are valuable way to communicate important underlying raw text-based data

A picture containing text

Description automatically generated

1. ***Network Analysis:*** Network refers to a structure representing a group of objects/people and relationships between them. Network consists of nodes and edges where nodes represent objects we are going to analyze and edges represent the relationships between those objects. In Twitter, edges can be following/follower relationships.

***Comments:***

* Most mentioned user is RamiAlLolah with 565 times mentioned by the other users.
* Most active user is Uncle\_SamCoco with 1578 times mentioning other users.
* Top 10 frequent pairs of taggers and mentions:

***Table

Description automatically generated***

* Network graph of ISIS tweets:

Diagram

Description automatically generated

1. ***Tweets Classification:*** This part of the analysis focuses on classifying the tweets as a positive tweet/ negative tweet or neutral using set of negative and positive words.

***Comments:*** The rule based approach gives the following results:

* # Positive tweets = 1200
* # Negative tweets = 6858
* # tweets with both positive and negative words = 6179
* # tweets with no negative and positive words = 3173

Eg:

SHEIKH FATIH AL JAWLANI 'FOR THE PEOPLE OF INTEGRITY, SACRIFICE IS  EASY'

**Results:**

From all the analysis made using the data, following are the results:

* Who are the major players in the ISIS twitter network?

Major Players:

* + RamiAlLolah
  + Warreporter
  + Nidalgazaui
  + Mohi\_ayubhi
* What are the keywords that are commonly used by ISIS fanboys

Keywords used:

* + Syria
  + Amp
  + Killed
  + Islamic
* Is there any trend we observe in the tweets pattern

There is no specific trend observed in the number of tweets being tweeted each day

* What is the sentiment we observe in the tweets
  + Highest percentage of tweets have negative words
  + We can also observe tweets with aggressive, violent and religious

**Limitations:**

* Take care of all kind of special characters while working with data
* Classification is more complex and needs advanced NLP to properly classify
* Need to have more data to use clustering and categorization