Analyzing Tweets about the Islamic State

Aaron Pulver
MS Geography
University of Utah, UT USA
aaron.pulver@utah.edu

I. Introduction and Overview

Micro-blogging platform Twitter has been a major distributor of information and opinions over the past five years. Founded in 2006 in San Francisco, Twitter has over 280 million active monthly users [1]. The premise of Twitter is 140 character text statements known as tweets. Each day, Twitter receives over 500 million tweets [1]. These tweets can be about anything from personal information, to music, to current events. There have been many studies which analyze Twitter data to better understand social networks and to observe patterns in the spread of information [2] [3] [4]. In this paper I look to understand how and where information and news about the Islamic State (ISIS) are spread through Twitter.

ISIS was chosen as the topic of interest due to its current prevalence within the news around the world. ISIS is an unrecognized state that is attempting to establish a worldwide Islamic caliphate. ISIS has occupied or currently occupied areas in four countries: Iraq, Syria, Libya, and Nigeria [5]. ISIS, a branch off of Al Qaeda, has become known for their brutal tactics in an effort to establish a caliphate.

Tweets were strategically mined from twitter using the Twitter Streaming API. The tweets were then filtered so that only those which contained variations of the word "ISIS" and had verified location data were remaining. The tweets were then mapped to show where in the world people were tweeting about ISIS. A density map was created to show the most common areas for tweets about ISIS. Other interesting statistics about languages, devices, and timing were calculated about the tweets. The main purpose of this project was to see if major news stories about ISIS were correlated with the overall twitter traffic about ISIS. It was hypothesized that major stories such as mass murders, beheadings, or successful air-strikes would result in a significant increase in twitter traffic.

II. Literature Review

Data mining has become a very popular field in the social sciences. There are hundreds of papers which do social media or social network analysis. The most important aspect of this project is obtaining geo-located tweets. Twitter provides a streaming API which allows anyone with proper authentication to download tweets in near real-time. Filters can be applied to the data to restrict tweets to those which contain certain phrases. Location filters can also be applied to remove tweets which are outside a specified bounding box [6]. According to a study by Takhteyev et al., around seventy percent of tweets have some sort of publically available location associated with them whether that is coordinates, vague user location descriptions, or specific user locations [7]. Although, some location information is provided, only 1-5% of tweets contain geographic coordinates.

It has been shown that tweets can accurately predict breaking news stories faster than many common news aggregators such as Google News [8]. Jackoway et al. showed how collecting and aggregating tweets based on similar keywords and location can be used to predict future major stories. Information on Twitter spreads very quickly through retweets. Retweets are tweets, not authored by the current user, which are shared to the current user's followers. It has been shown that any retweeted tweet is likely to be seen by an average of 1,000 other users [2]. This allows information to quickly spread from one person to another who may create their own tweet about the subject. Finally, a study comparing the spread of information between Digg, an online news aggregator where users submit links for others to view, and Twitter concluded that although news articles initially spread more quickly on Digg, news articles spread at a much steadier rate on Twitter and reached more people due to the architecture of Twitter [4]. These studies show that monitoring twitter traffic is a reliable method for detecting major world events and news stories.

III. Data Collection Architecture

A system architecture was developed to obtain tweets and store them as shown in Figure 1. A python script was written which used the tweepy library to connect to the Twitter Streaming API. The streaming api continuously returned tweets which met specific criteria. These tweets were then stored in a MySQL database. The database stored the user, tweet, hashtags, urls, and other metadata which was useful for analysis. This part of the system ran on a Raspberry Pi microcomputer so that data would be continuously collected and stored.

The other major part of the system was done after all data were collected. The data were extracted from the MySQL database as a CSV file which could then be processed in Microsoft Excel and imported into ArcMap as XY data for spatial analysis.

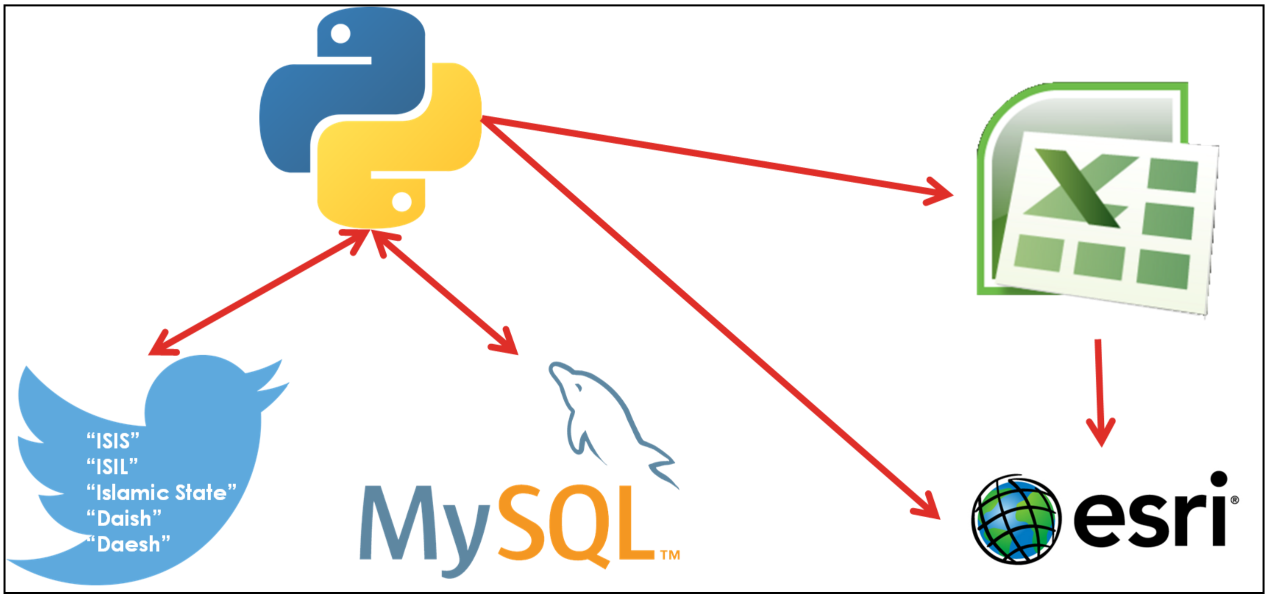

Figure 1. System Architecture

IV. Methodology

Tweets were collected using a filtered twitter stream. Since I only wanted tweets related to ISIS, only tweets containing "ISIS","ISIL", "Islamic State", "Daesh", or "Daish." All of these phrases refer to the Islamic State. The information contained within each tweet as well as the user who tweeted it were stored in the database. Collection was done for a period of 17 days from March 21th to April 6th 2015. After the collection period was over, the tweets were extracted from the database using SQL queries to join related tables so a single record contained the relevant information for analysis. Only tweets which had verified geographic coordinates were extracted and saved to a CSV file.

The CSV file was then opened Excel and any duplicates were removed. The tweets were then grouped into four different categories: language of the tweet, device or application used to submit the tweet, the date of the tweet, and the time of the tweet. I was able to generate several histograms which showed the trends of the collected tweets. Then the grouped and labelled tweets were imported in ArcMap as XY data. Two kernel density maps were derived; one using a 50km cell size, and one using a 1km cell size. These maps show where the most tweets occurred. A choropleth map showing the majority language, of the top 6 overall languages per country was also created.

V. Results

Over 2 million tweets were collected over the course of the collection period. Of these, only 11,156 tweets had geographic coordinates. The first step was to group and label the tweets. There were a total of 36 different languages used. The most popular languages were English, Indonesian, Spanish, Italian, Portuguese, and French (Figure 2). This was not surprising other than Indonesian being the second most popular. A map showing the most popular language per country is shown in Figure 9.

A kernel density map (Figure 8) was generated from the raw tweet location dataset (Figure 7). The primary areas of twitter activity are in the New York metro area, Southern California, Texas, London, and Jakarta. These are highly populated areas which have easy access to the internet.

There were 63 different applications used to submit tweets (Figure 3). The most popular were the official Twitter apps for Android and iOS. This was not surprising as these are the two most used mobile operating systems.

Figure 4 shows the number of tweets per hour of the day. All times are shown in Eastern Standard Time (EST). There is a clear trend toward more tweets in the morning and then later at night.

Figure 5 shows the number of tweets per day. There is one day with very few tweets: March 23rd. This is due to the power going out and stopping data collection. The average number of tweets per day, excluding March 23rd, was 692. The median number of tweets was 617. There are three peaks: March 26th, March 27th, and April 1st. I investigated the March 26th and 27th peaks to see what caused this increase of tweets to 1197 and 1181 respectively.

VI. Discussion

I read many of the tweets which occurred on either March 26th or 27th and noticed what looked to be a trend. A satirical article was published online which claimed that One Direction star, Zayn Malik, was leaving to join ISIS [9]. I selected tweets which had “Zayn”, “Malik”, “One Direction”, or “1d” in them. I found that on both days there were 229 tweets containing at least one of these phrases. This appeared to explain a large portion of the increased Twitter traffic. I then plotted the total number of tweets per day versus the number of tweets about Zayn and performed a linear regression (Figure 6). The residual error was 0.47 which means that news articles definitely affected the amount of twitter traffic about ISIS. The news articles about Zayn do not explain the complete increase in twitter traffic. This may be due to using certain key phrases which inherently may have filtered out tweets which were still about Zayn.

I also briefly looked into the spike on April 1st. There were several articles and tweets about Iraqi forces taking back the city of Tikrit from ISIS.

VII. Conclusion

After conducting this brief analysis of tweets regarding the Islamic State, I can safely conclude that it is possible to correlate twitter traffic with specific events. Much more analysis and data mining has to be done to automatically extract tweets which are about similar events. This project confirms what many other studies have shown.

References

[1] Twitter, "About Twitter, Inc.," Twitter, [Online]. Available: https://about.twitter.com/company. [Accessed 9 March 2015].

[2] H. Kwak, C. Lee, H. Park and S. Moon, "What is Twitter, a Social Network or a News Media?," in *19th international conference on World wide web*, 2010.

[3] D. Moncanu, A. Baronchelli, N. Perra, B. Goncalves, Q. Zhang and A. Vespignani, "The Twitter of Babel: Mapping World Languages through Microblogging Platforms," *PLoS ONE,* vol. 8, no. 4, 2013.

[4] K. Lerman and R. Ghosh, "Information Contagion: An Empirical Study of the Spread of News on Digg and Twiter Social Networks," in *Fourth International AAAI Conference on Weblogs and Social Media*, Washington D.C., 2010.

[5] N. Thompson, R. Greene and S.-G. Mankarious, "ISIS: Everything you need to know about the rise of the militant group," *CNN,* 10 February 2015.

[6] Y. Takhteyev, A. Gruzd and B. Wellman, "Geography of Twitter Networks," *Social Networks,* vol. 34, pp. 73-81, 2012.

[7] K. H. Leetaru, S. Wang, G. Cao, A. Padmanabhan and E. Shook, "Mapping the global Twitter heartbeat: The geography of Twitter," *First Monday,* vol. 18, no. 5, 2013.

[8] A. Jackoway, H. Samet and J. Sankaranarayanan, "Identification of Live News Events using Twitter," in *3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, 2011.

[9] Huzzlers, "Zayn Malik Leaves One Direction to Join ISIS," Huzzlers, 26 March 2015. [Online]. Available: http://huzlers.com/zayn-malik-leaves-one-direction-to-join-isis-2/. [Accessed 22 April 2015].

Appendix

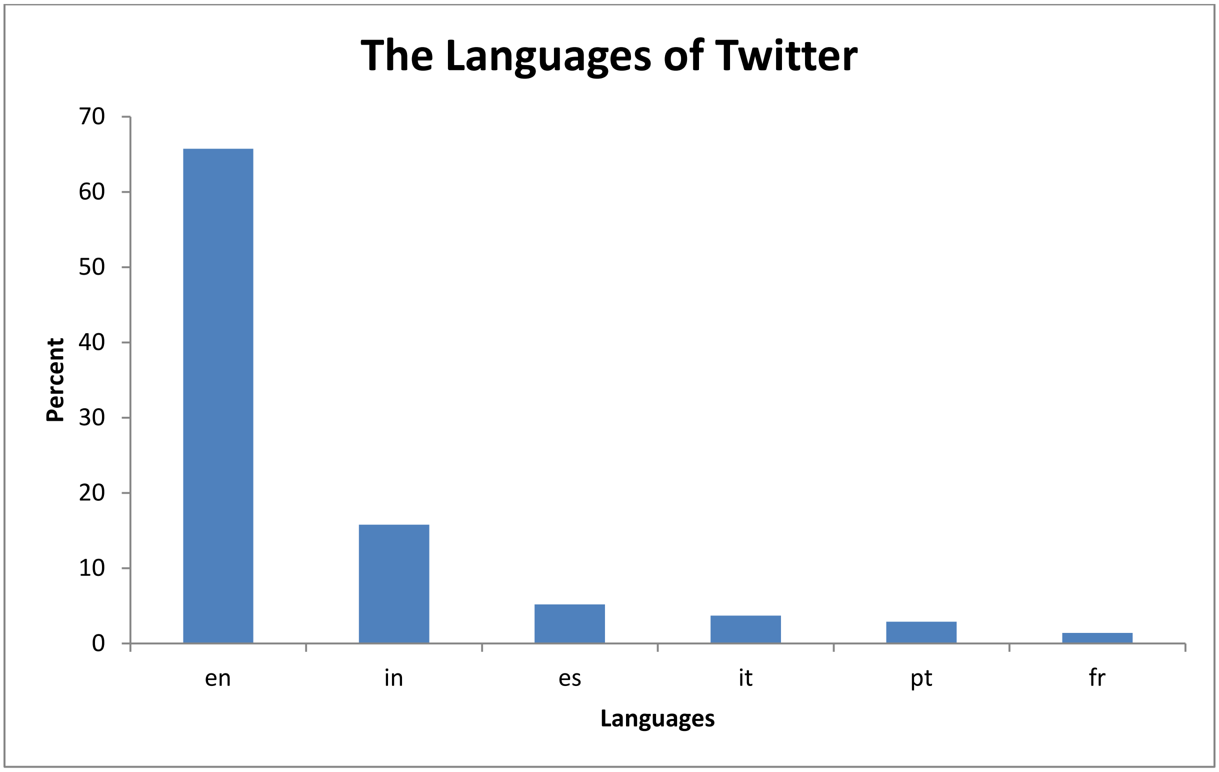

Figure 2. The languages of tweets.

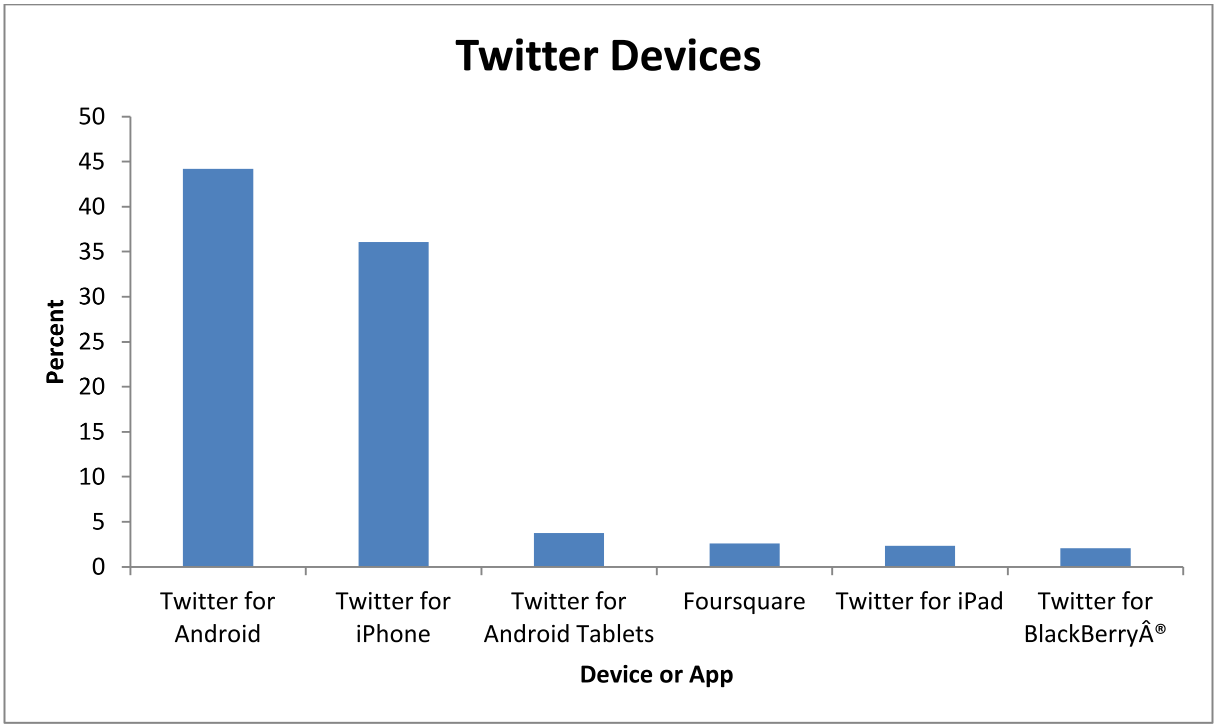

Figure 3. The devices used to tweet.

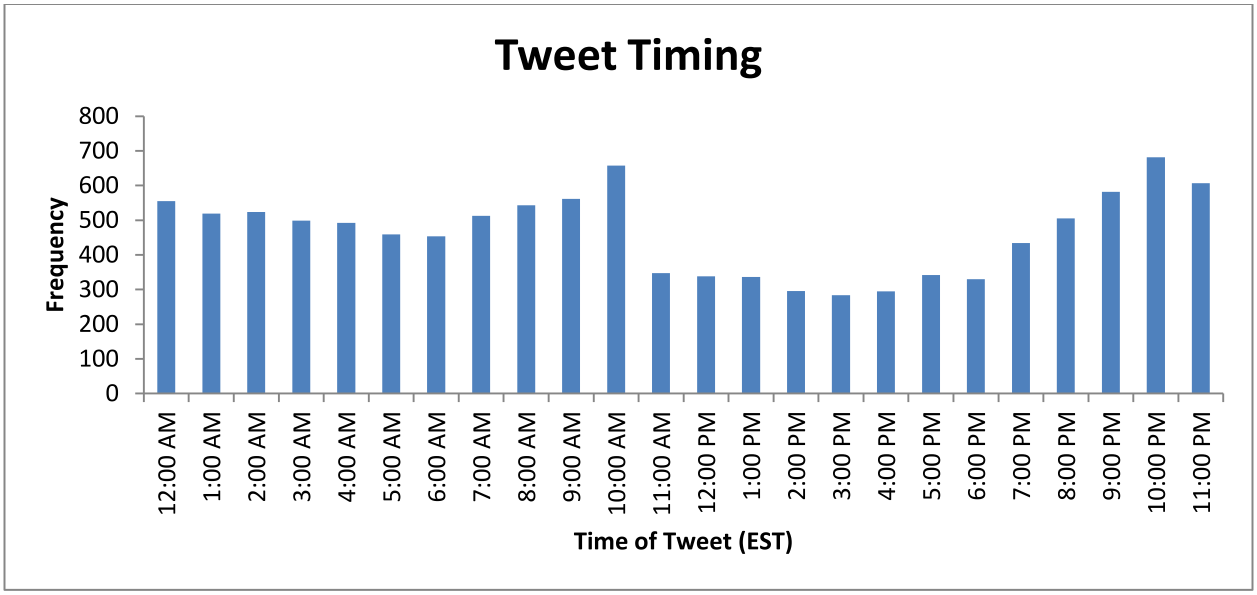

Figure 4. The times of tweets

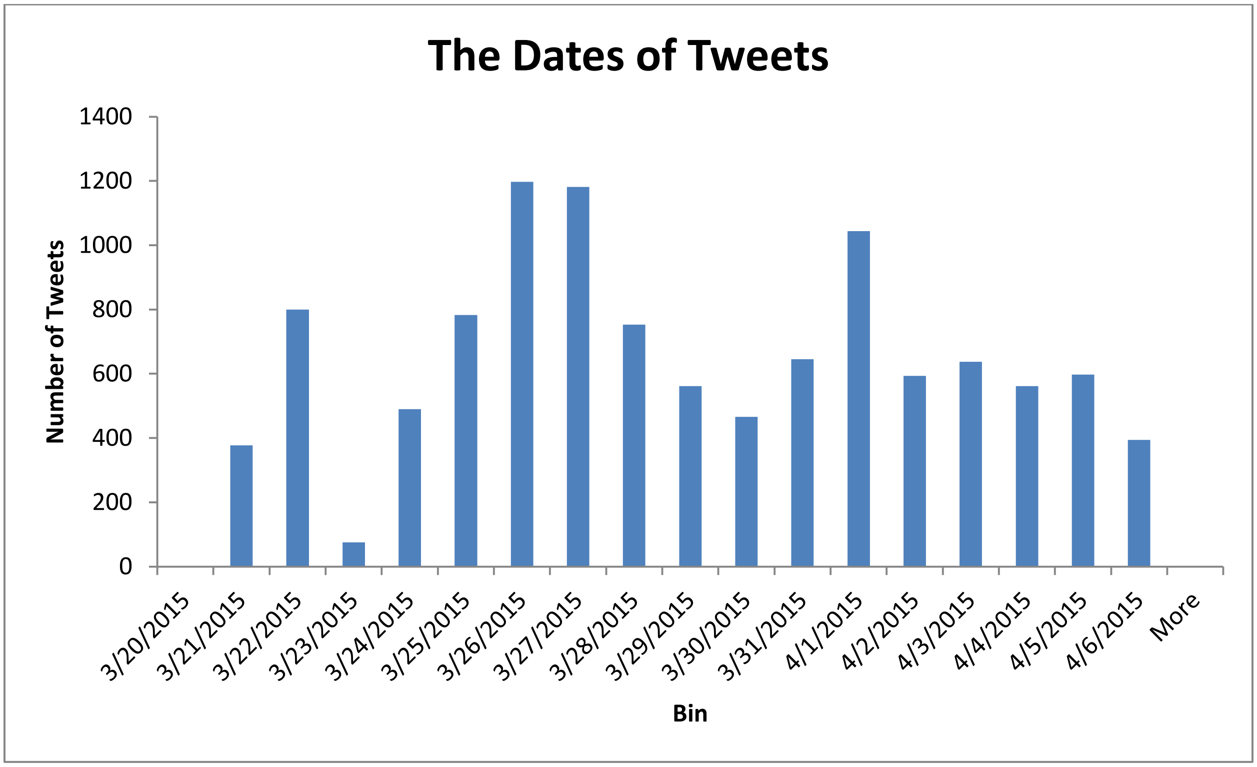

Figure 5. Dates of Tweets.

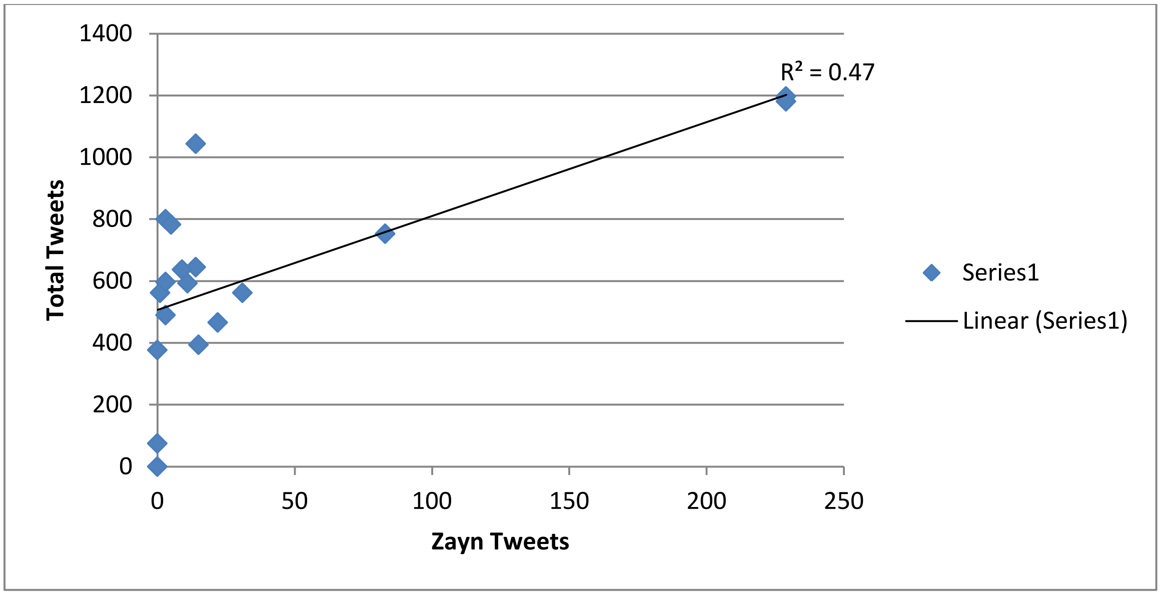

Figure 6. Regression Analysis

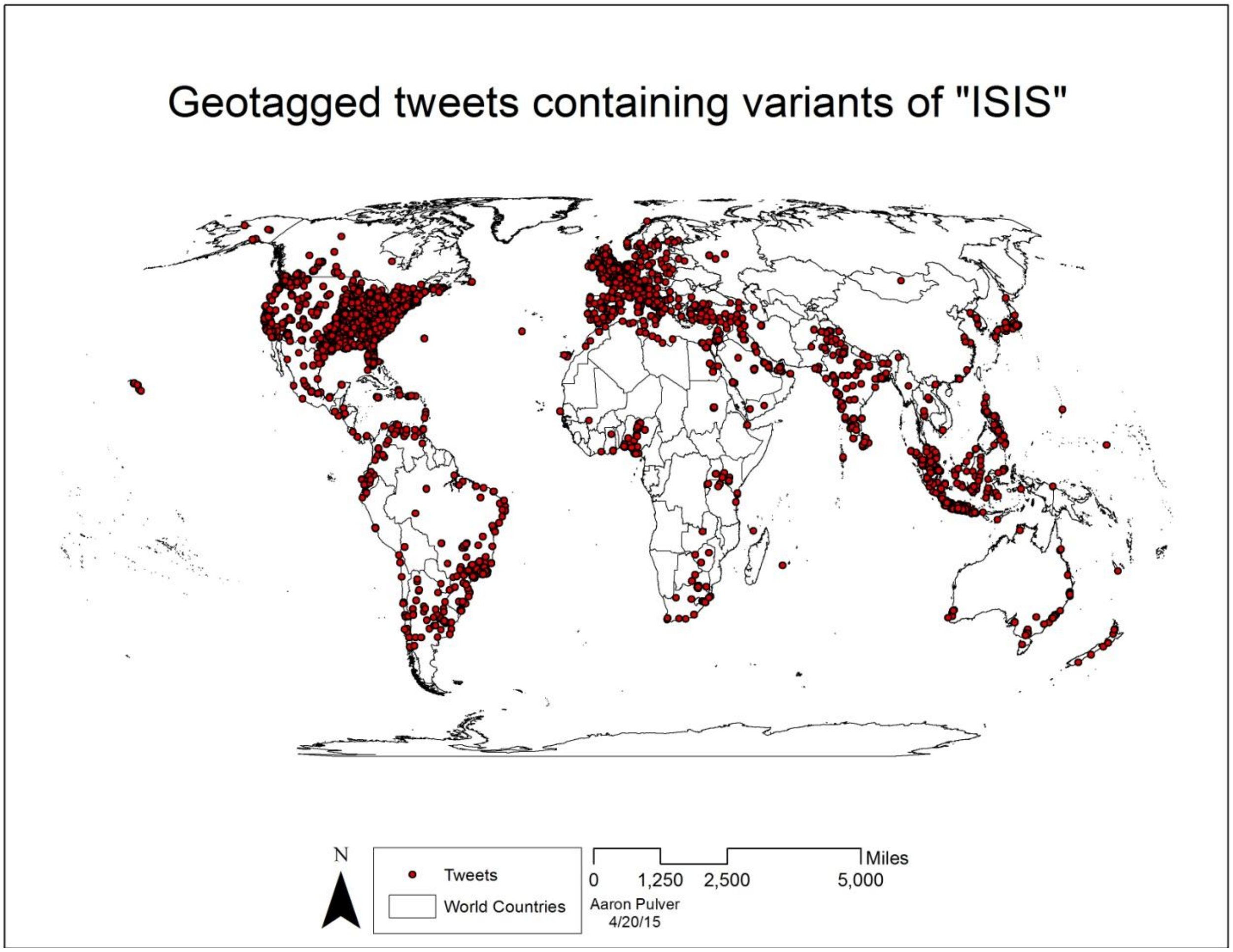

Figure 7. Geo-tagged Tweets

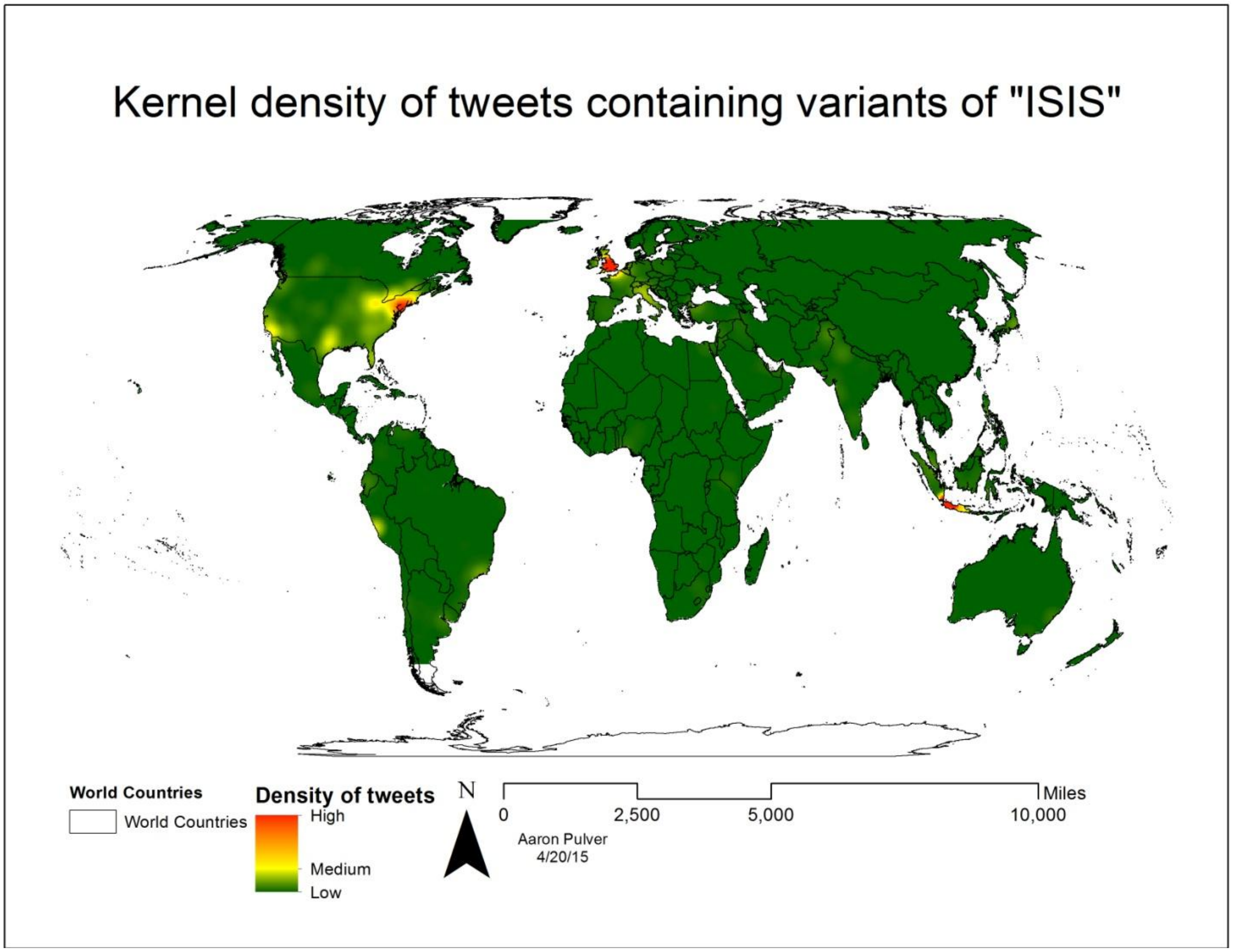

Figure 8. Kernel density map of tweets.

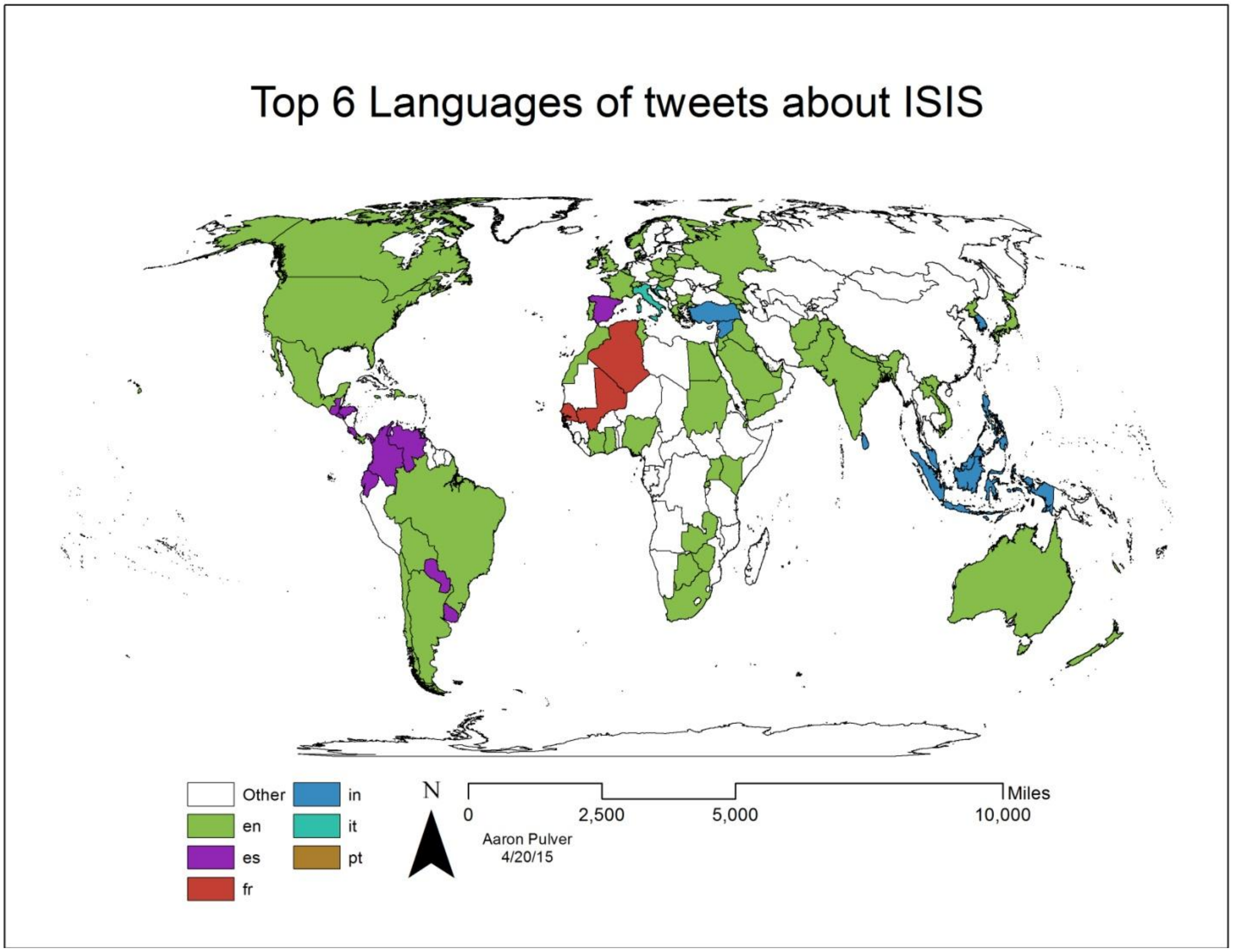

Figure 9. Map of Languages.