Proceedings of the 2020 Annual General Donald R. Keith Memorial Capstone Conference
West Point, New York, USA
April 30, 2020
A Regional Conference of the Society for Industrial and Systems Engineering

Evaluating the Utility of Analyst-Level Access to Authenticated APIs

Sara Klena, Xochitl Fernandez, Thomas Cruz, Noah Stewart, Gonzalo Borrego Acosta, Noah Hanau, and Ian Kloo

Department of Systems Engineering
United States Military Academy, West Point, NY

Corresponding Author: sara.klena@westpoint.edu

Author Note: The authors listed above are seniors at the United States Military Academy at West Point and conducted this project as part of a graduation requirement for their senior capstone course. They worked under their advisor, Mr. Ian Kloo, in combination with advisors from Intelligence Command, United States Army, who provided insight into the current status of data science in their organization and field. The authors would like to thank INSCOM personnel, the Department of Systems Engineering, and Mr. Kloo for their support and assistance.

Abstract: The United States Army Intelligence and Security Command (INSCOM) is responsible for operational intelligence and security, conducting multi-discipline operations internationally (U.S. Army Intelligence & Security Command). In the digital age, INSCOM gathers significant data via commercial applications, social media sites and internet searches. INSCOM analysts are working to implement data science techniques in their work, but current government policies restrict their ability to efficiently access information on the open internet. Our cadet team served as a proxy for an intelligence analyst team without policy restrictions, using various data science techniques to conduct analysis and develop applications. Through this project, we developed an example methodology that demonstrates the significant analytic value of having access to authenticated APIs.

Keywords: Web Scraping, R, Shiny, App Development, API, INSCOM, OSINT, Intelligence

1. Background

This project supported the United States Army Intelligence and Security Command (INSCOM) in exploring the use of Open Source Intelligence (OSINT) in daily work. INSCOM's mission statement is:
"INSCOM executes mission command of operational intelligence and security forces; conducts and synchronizes worldwide multi-discipline and all-source intelligence and security operations; delivers linguist support and intelligence-related advanced skills training, acquisition support, logistics, communications, and other specialized capabilities in support of Army, Joint, and Coalition commands and the U.S. Intelligence Community" (U.S. Army Intelligence & Security Command).

INSCOM is a large organization that conducts intelligence (including OSINT) and security operations for the U.S. Army. Open source is defined as "any person or group that provides information without the expectation of privacy—the information, the relationship, or both is not protected against public disclosure… Open sources refer to publicly available information medium and are not limited to physical persons" (ATP 2-22.9, 1-1). Through OSINT, analysts are not collecting specific information on the general public but instead are simply looking at available, open source data on news sites and social media. These findings provide analysts with insight pertaining to current events, national concerns, geospatial information and other INSCOM concerns. OSINT has become an even more important form of intelligence as the scope of publicly available information continues to increase in the technological age. This work analyzes the importance of OSINT from a baseline perspective, with minimal access to the personal work of INSCOM personnel due to privacy concerns. We work within the baseline definition of open source data and the basic capabilities of an academic individual conducting OSINT work without any government involvement.

Some OSINT analysts are trained in web scraping, which allows them to gather large sets of this data to analyze its scale and sentiment. However, INSCOM analysts are largely unable to access social media platforms due to the potential of infringing the public's privacy. Most social media sites require users to register their affiliation before accessing bulk information through an Application Programming Interface (API). APIs are "a set of definitions and protocols for building and integrating application software" (Red Hat). APIs allow an individual to communicate with news and social media interfaces to simplify app development (Red Hat). Social media platforms like Facebook and Twitter will not allow government officials to access their data due to implemented policies. Furthermore, current INSCOM policies limit any data collection from social media sites that goes beyond traditional web scraping (i.e., automated parsing of a publicly accessible web page). Even

traditional scraping requires significant individualized training and labor to yield useful results (Wilkinson, 2019). The inability to use data from authenticated APIs and the current learning curve for traditional web scraping create a gap in OSINT analysts' ability to accomplish their mission.

2. Global Social Media

Because INSCOM's areas of interest are global, we explored the state of social media in the United States, Western Europe, the Philippines, Russia, China and Hong Kong. Through our research we found that Russia and China have the most restrictions when it comes to their state of social media. These restrictions come from limitations set by their governments. Foreign governments strictly censor their social media platforms, limiting the utility of scrapping and using data found on social media (Bamman et al., 2012).

Within each country, there are different social media platforms that are more popular amongst the population that personnel within the United States do not have access to. For example, Sina Weibo, analogous to Twitter, is a popular Chinese platform (Nanjing Marketing Group, 2014). An individual seeking access to these platforms must register with a Chinese cell-phone number and an identification number, similar to an American SSN. Additionally, gaining authenticated API access would require proof of citizenship. Therefore, because social media platforms in heavily restricted countries are difficult to gain access to, it is difficult to scrape their popular social media platforms.

Due to the difficulties of gaining access to foreign social media platforms, we concluded that Twitter is the most useful platform for our research. Though Twitter is more popular in the United States and Western Europe, it is still used and accessible throughout the world. Twitter's international nature, easy-to-access authenticated API and large user base provide useful data and insight to analysts. Twitter serves as a comparable proxy for other authenticated APIs in the social media space as most sites share similar data structures.

3. Methodology

With a focus on Twitter, we developed a three-step plan for data analysis using the company's authenticated API. Our objective was to conduct real-time analysis to give an overview of a current known issue (i.e. an election, a natural disaster, etc.). Figure 1 below shows the steps in the process as well as the outputs from each step. Using this framework, INSCOM analysts could supplement current capabilities without the need for commercial tools or data. These enhanced capabilities, as we demonstrated with our own results (described below), require a simple and easily learned understanding of coding and the ability to access the authenticated API.

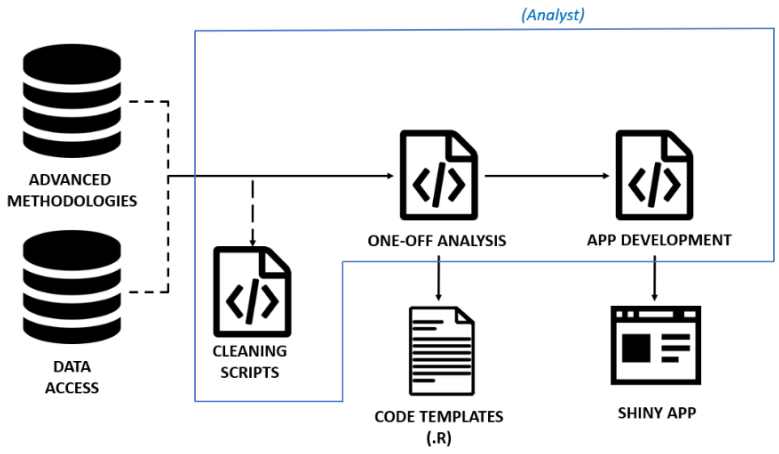

Figure 1. Methodology Flow Chart

3.1 Advanced Methodologies & Data Access

The first step of our methodology is data collection. This step involves gaining access to the data source (Twitter's API) and implementing a process to organize the data as we scrape it using R scripts written by our team. The first task in cleaning the data is filtering it by location, language and topic. Although OSINT work can involve translation capabilities, the breadth of our research focuses on English as a simplifying assumption. We view the data by the latitudinal and longitudinal coordinates of each tweet and change the topic of interest as needed. Additionally, we remove URLs, hashtags, and stop-words (common words that hold little information) from each tweet. The *Tidytext* R package has stop-word dictionaries that contain words like "the," "and" and "or," and access to these dictionaries allows us to easily remove these words from a dataset (Silge & Robinson, 2019).

3.2 One-Off Analysis

In order to perform our one-off analysis, we need a large number of tweets to determine the overall sentiment towards an event and frequency of certain words being used to describe it. For each analysis we pull 3,200 tweets using keywords and hashtags relating to each area of interest. Using a large number of tweets allows us to find a general trend amongst the tweets and eliminate any outlying trends. With properly formatted and processed data, we evaluate several analytic approaches to determine what would be feasible. Under the guidance of INSCOM, we focus on sentiment analysis and word frequency analysis which were both straightforward to implement in R.

3.2.1 Sentiment Analysis

We leverage the *Sentimentr* package in R to take a dataset of tweets and determine if the overall sentiment of the tweet is positive or negative. The S*entimentr* package uses a dictionary of words, each marked with a positive or negative sentiment value corresponding to the magnitude of the word's sentiment level (Fuchs, 2020). *Sentimentr* then sums these values to determine the tweets' overall sentiment. Using this technique, analysts can determine the aggregate sentiment of tweets regarding an event, individual, or idea (Fuchs, 2020).

One issue with sentiment analysis is that it cannot detect sarcasm in tweets. As a result, sarcastically positive tweets might be interpreted as demonstrating positive sentiment, or vice versa. In Figure 2, we show a histogram depicting the results of one of our sentiment analyses. The x-axis represents the sentiment of a dataset, ranging from -1 (negative) to 1 (positive). Tweets with a positive sentiment have positive values and fall to the right of zero, while tweets with negative sentiment fall to the left. The y-axis indicates the frequency of tweets with a specific sentiment level. Ultimately, the example topic of "Coronavirus" in Figure 2 shows an overall negative sentiment associated with the data set.

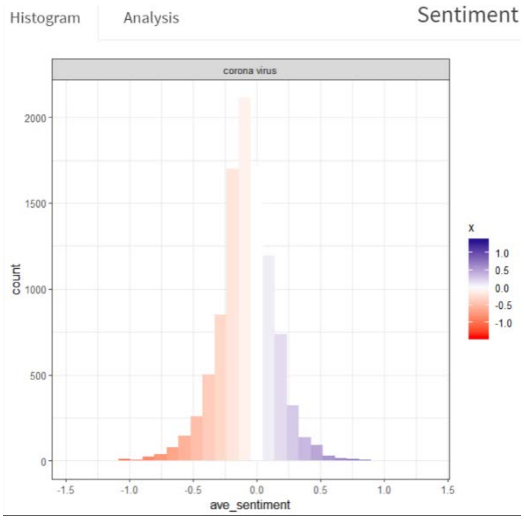

Figure 2. Sentiment Analysis

3.2.2 Frequency Analysis with Word Clouds

We use frequency analysis to determine how often certain words show up in a dataset of tweets. While text frequency is the simplest form of text processing, it can give an accurate general sense for the overall content of collections of text. As such, analysts can identify the most common words to provide a general depiction of a conversation.

We use word clouds as our main visualization technique for text frequency as they are easy to interpret. While word clouds are useful, we caution that there are inherent flaws with this technique: analysts should remove non-English words, stop words (described in 3.1 above), and slang to prevent skewing of frequency data. Figure 3 shows an example of a word cloud for a search of the phrase "Coronavirus." The phrase "Coronavirus" is trending across the world, so this word cloud will inevitably include words of many languages. In order to resolve this, we limit the search to English words only, allowing us to easily analyze the word cloud without potentially confounding words.

Figure 3. Word Cloud

3.3 App Development

After developing scripts for data access, cleaning data, and text analysis, we present an application to perform the entire analytic process. The goal is to create a user interface that allows non-coders to use our methodology using R and the Shiny Dashboard package, which compiles all objectives into an easy to use interface (*Shiny Dashboard,* 2014). For an OSINT analyst these scripts and apps provide large amounts of data that can be quickly analyzed to gain valuable information on specific regions around the world and social trends amongst the people from that region. This information highlights areas where OSINT analysts should further their research to understand the source of these trends. At a bare minimum these scripts and app provide a large amount of data that can analysts can quick access. These methodologies can prove to be more effective than traditional data collection like reading news articles by reducing the amount of time it takes to gain valuable information.

4. Results and Discussion

When compared to a traditional research effort, our results are atypical. We did not set out to solve a specific problem with our work, but instead sought to demonstrate the potential output one could expect from a minimally trained but technologically enabled team of analysts. Analysts can use the large collection of data to draw conclusions on the public, levels of social unrest, using frequency analysis and sentiment analysis. To this end, we will present our results and discussion from the perspective of the analyst team.

Our work with authenticated APIs allowed us to become familiar enough to quickly replicate the One-Off Analysis and App Development stages of our methodology and create tangible and transferrable products. This allows intelligence analysts to promptly create a product that acclimates to a dynamic environment allowing for a rapid assessment of a current specific environment. Overall, we concluded that access to authenticated APIs would enable previously untrained analysts to develop analytic methodologies and provide significant value. As untrained analysts, we were able to use R as a platform for basic coding. We found that working with R required little formal education as many previous examples and scripts are available. Even with rudimentary R skills, we were able to create an app that gathered data and displayed it on an easy to read interface. Our methodology demonstrates that minimally trained analysts could develop useful tools using authenticated APIs that provide significant value.

4.1 Unauthenticated (Traditional) Web Scraping

Before developing a concise methodology, we conducted several technical sprints to gain familiarity with various web scraping methods and analyzed the quality of information gathered. We initially focused on traditional web scraping techniques that are currently available to INSCOM. While these methods demonstrated some promise, we ultimately found them difficult to implement with our level of expertise. We looked at multiple media outlets to determine which would provide the most utility.

Since public news websites do not have authenticated APIs, we scraped them with a traditional approach to determine their utility. These sites should be among the simplest to work with due to their consistent HTML formats. Additionally, news websites typically involve objective articles without the slang and sarcasm found on Twitter, making them amenable to analysis. In practice, however, scraping news sites proved quite difficult. Many sites leverage modern web design frameworks by loading the least amount of HTML possible for each page view to still convey information (this helps with page load speed and other design considerations). Web scraping under these circumstances is possible, but much more involved than scraping traditional HTML websites and is too difficult to reasonably expect a minimally trained analyst team to master.

4.2 Authenticated APIs

Authenticated APIs provide several advantages over unauthenticated web scraping, primarily because large amounts of information are much easier to access with an API. On a platform like Twitter, a continuous stream of high-volume, clean information was easy for our team to attain without any previous training in this area.

With authenticated APIs, information generally exists in more accessible and uniform states. Consequently, analysts can collect and organize this data efficiently. We found that for these reasons, authenticated APIs allowed us to apply our processes to a more diverse problem set without constant modification. Additionally, in comparison to traditional web scraping, authenticated API access allowed us easy access to historical data. While this may not exist on every site at every privilege level, it is a potentially important piece of functionality as it allows analysts to do retrospective analysis. We also found that authenticated APIs typically provide access to large quantities of information. For example, Twitter allows individuals to scrape a maximum of 18,000 tweets per 15 minutes (Kearney, 2019). This was significantly more data than we were able to collect using traditional web scraping methods.

In sum, through authenticated APIs, we were able to efficiently collect and organize information in a way that provides utility in analysis. As described in our One-Off Analysis section, our analysis required high volumes of information over a period. Twitter's API allowed us to easily collect data at a scale that would have been difficult with other methods.

5. Limitations & Future Work

Army policy limits INSCOM's ability to collect bulk data through APIs. We offer two recommendations. First, the Army should recognize INSCOM's ability to leverage API access and relax its policies, allowing the government to work with social media sites in a limited capacity. As mentioned previously, this will also involve discussion with the social media sites to allow for limited government access since current policies prevent government personnel from accessing APIs. Second, INSCOM should find alternative methods to obtain data collected through APIs. One possibility is to attach non-INSCOM data engineers to INSCOM analysts. These data engineers would be able to work with Army and INSCOM regulation to develop a plan for specific data collection that avoids the gathering of sensitive information on US persons. Even with unrestricted access to authenticated APIs, one of the biggest challenges analysts will face is accurately interpreting data in a foreign language. Not only do analysts need to discern denotation, but also connotation and figurative language all through the lens of a different culture. In many data science projects in the natural language processing space, language is a problem because coders typically are not linguists. In INSCOM's case, however, our research suggests the data science capability could be implemented by the analysts with cultural expertise. Building these data science skills into a traditional analyst team, which just consists of a small group of 2-5 OSINT analysts and an analyst supervisor, in a non-disruptive way should be explored in future research efforts.

6. Conclusion

Our findings and methodology demonstrate the potential benefit of allowing OSINT analysts to access authenticated APIs; however, we also recognize the potential risks analyst-level access would create. Predominant risks involve the invasion of privacy of US persons and the violation of current Army directives that prohibit government access to authenticated APIs for work. INSCOM (and potentially follow-on researchers) should investigate ways to provide this important capability to

analysts in a way that appropriately mitigates these risks. Even without specialized training, giving analysts access to this wealth of valuable information via policy change and the addition of specifically tailored capabilities via data engineering would improve the analytic ceiling of the organization. Any future data science investment (i.e. additional personnel or training) would amplify these gains and provide valuable insight to the warfighter.

7. References

ABS-CBN News. (January 31, 2019). "Filipinos Still World's Top Social Media User - Study." *ABS.*

"Background: Shiny and HTML." (2014) *Shiny Dashboard*.

Bamman, D., O'Connor, B., & Smith, N. (2012). Censorship and deletion practices in Chinese social media. *First Monday*, *17*(3).

E-commerce Nation. (January 2, 2019). "Odnoklassniki, the Most Popular Social Network among Russians over the Age of 25." *E-Commerce Nation*.

Fuchs, M. (2020). "Doing Your First Sentiment Analysis in R with Sentimentr." *Towards Data Science*.

Headquarters, Department of the Army. (July 2012). "ATP 2-22.9: Open Source Intelligence."

"Internet censorship in Hong Kong". (August 2019). *Wikipedia*, Accessed October 6, 2019.

Kearney, M. (2019). rtweet: Collecting and analyzing Twitter data. *Journal of Open Source Software*, *4*(42), 1829.

Lam, L. (2014). Facebook is Hong Kong's top digital platform in survey commissioned by company. *South China Morning Post*, *22*.

Matsnev, O. (April 11, 2019). "Kremlin Moves Toward Control of Internet, Raising Censorship Fears." *The New York Times.*

Nanjing Marketing Group, (May 16, 2014). "Complete Guide to Chinese Social Media."

"Philippines." (February 11, 2019). *Philippines Country Report.*

"Russia." (February 11, 2019). *Russia Country Report.*

Sharma, R., Wigginton, B., Meurk, C., Ford, P., & Gartner, C. E. (2017). Motivations and limitations associated with vaping among people with mental illness: A qualitative analysis of reddit discussions. *International journal of environmental research and public health*, *14*(1), 7.

Silge, J. and Robinson, D. (November 24, 2019). "Text Mining with R." *The tidy text format.*

"Social Media Stats Philippines." (2019) *StatCounter Global Stats*. Accessed October 1, 2019.

"Social Media Stats Europe." (2020) *StatCounter Global Stats*. Accessed March 3, 2020.

Rainford, C. (February 27, 2018). "Trust in Traditional Media Increase Across Europe." *EBU European Broadcasting Union.*

Rinker, T. (2019) "Package *sentimentr*", *CRAN*.

"United States Army Intelligence and Security Command Mission Statement." (2020) *United States Army Intelligence and Security Command*. Accessed February 25, 2020.

"We are Social." (2019) *Digital 2019 in Hong Kong*. Accessed December 17, 2019.

Wilkinson, R. Personal communication (December 12, 2019).