Capstone Report 1:

Social Media Sequential influence around incidents of police use of force against unarmed Black victims

Andrea Navarrete, Omar Pardo, Jose Lopez, Shadi Fadaee, Shimeng Feng

Professor Colin Wayne Leach & Courtney Cogburn

1. Problem Overview

One of the key interests of researchers in data science and behavioral sciences is understanding the temporal sequence of activities across multiple media sources. Grasping this relationship would allow them to trace the influence of a specific type of activity across these platforms. In this project, we intend to understand Twitter activity around the time of events related to police use of force against unarmed black victims. In other words, our goal is to understand the influence that Twitter activity around issues of violence against black victims has on the sentiment around them, as well as the changes that follow regarding these events. Thus, the overall objective of this capstone project is to analyze sentiment, social networks, topic modeling and image categorization around issues surrounding these specific events.

2. Data

Datasets and Details

The data was provided by professor Leach, who in turn got it from the Beyond the Hashtags project (Freelon et al. (2016)). The original set consists of all public tweets posted between June 1st, 2014, and May 31st, 2015, which contain at least one of 45 keywords. Such keywords are mainly related to police murders of black people which contributed to the rise of the #BlackLivesMatter movement.

The original dataset contains more than 40M tweets, but for the present analysis, we will only consider the 8.5M tweets that were written in August 2014. This month was relevant for many reasons:

- Eric Garner was murdered 13 days before (July 17th) after being put in a chokehold by Daniel Pantaleo (NYPD officer).
- Michael Brown was fatally shot by a white police officer (Darren Wilson) on August 9th, leading to several protests in his hometown, Ferguson, Missouri.
- In addition to Michael Brown, five black folks died because of incidents with the police.

Hydratation - Parsing Data

The data hydration process involved retrieving the twitter data when given a tweet ID. Using the hydrator desktop application we were able to obtain 8.5M tweets in JSON format distributed across 31 files. In these files, each row corresponds to a tweet in JSON format, so we needed to parse the rows into columns to be able to work with a relational database. Some of the functions we needed to apply were:

- Retrieve the data as list: hashtags, url, user_id_mentions, user_screen_name_mentions, symbols, latitude, longitude
- Generate new columns for user and entities information

After parsing the files, we uploaded a table containing 78 columns for all 8.5 Million tweets to BigQuery, for cloud access.

Data Cleaning

Once the data was stored in the cloud, the next task was to clean it so that it was useful for our analyses. To achieve that, we first considered the problem of having data that was not related to our problem. This was possible because the data was collected by matching keywords, but there was no human supervision. It is almost impossible to review tweet by tweet, but there are some general actions we implemented.

Firstly, we analyzed the keywords used to collect the data. We grouped them into categories and concepts, to handle them in a better way. For instance, "*victor white*" and "*#victorwhite*" are two different keywords, but they are both about the same concept: the murder of Victor White. Also, this murder happened in March 2014, - before the period we are studying. Consequently, this concept can be grouped with others into a larger category that is *Murders of Black People (before August)*.

There is one category called *Murders of Black People (after August)* which groups 9 concepts and 18 keywords. We wouldn't expect to find any tweets brought by these keywords in our dataset since those killings happened later. However, some of these victims shared their names with other people who were popular before. For instance, Walter Scott was a black victim shot on April 4, 2015, in North Charleston, South Carolina. Walter Scott was also a Scottish historical novelist who died in 1832.

Therefore, we kept only the tweets that contain at least one of the other 27 keywords, and for the concepts that happened during August, only the tweets that were written after the incident. Henceforth, we will call these tweets, **valid tweets**. The dataset also contains repeated tweets, since every retweet counts as a different tweet. To analyze content we will consider all of them as one, and we will call them **unique tweets**.

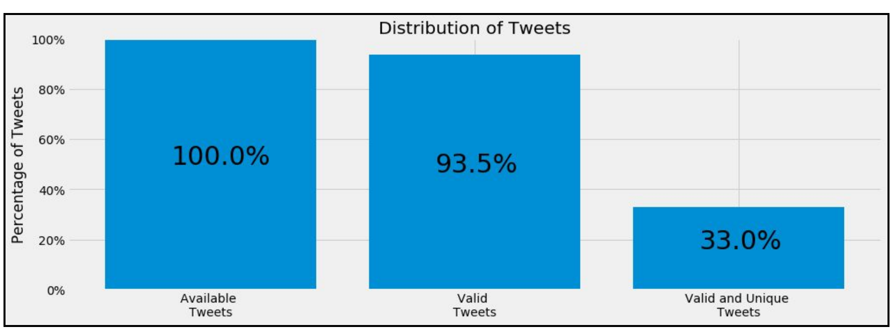

Figure 1. Distribution of tweets adhering to validity and uniqueness rules

3. Related Research

1. Twitter Data on Social Problems

Since social media has become the main channel of communication and have significant influence on social events, it is reasonable to discover societal behavior and response based on social media activities, Twitter, specifically in our case. Through analyzing different hashtags and understanding users' interactions, the research by Jelani Ince, et al. found that Twitter users were most likely to express approval of the movement and assert their solidarity In regards to the the hashtag black lives matter.

https://www.tandfonline.com/doi/pdf/10.1080/01419870.2017.1334931?needAccess=true

2. Topic modeling

Given the faculty's interest in discovering the underlying topics in the corpus, the team deep dived on the Latent Dirichlet Allocation algorithm. This statistical model allows the explanation of elements through unobserved clusters that represent their similarity. Under this approach, each tweet will be representing a document, and the team will find the underlying topics that can group the whole dataset.

3. Sentiment analysis

The following are some corpus / lexicon that we have identified and need to refine:

LIWC (Linguistic Inquiry and Word Count): Is a text analysis program that calculates the degree to which various categories of words are used in a text. The application relies on an internal default dictionary that defines which words should be counted in the text files.The sources for the dictionary words include social media such as Twitter and Facebook. The default LIWC2015 Dictionary is composed of almost 6,400 words, word stems, and select emoticons. For each text file, approximately 90 output variables are given. This data record includes outputs such as the word count and - of specific interest to our project - 41 word categories around psychological constructs (e.g., affect, cognition, biological processes, drives). The psychological processes section of the categories include positive and negative emotion, anxiety, anger and sadness. For some categories, the Stanford Natural Language Toolkit (NLTK;Toutanova, Klein, Manning, & Singer, 2003) was used to find the common words.

NRC Word-Emotion Association Lexicon: The lexicon includes two sentiments and 8 emotions such as anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. The lexicon can be applied to both English and over 100 other languages. Lexicon can be downloaded directly for non-commercial usage.

Valence Aware Dictionary and Sentiment Reasoner (Vader): Sentiment is a rule-based sentiment model for social media/microblog-like context (Hutto, C.J. & Gilbert, E.E. (2014)). The lexicon contains both the polarity and the intensity of sentiments and has benchmarked towards some most popular algorithms.

4. Exploratory Data Analysis

Categories of hashtags & keywords

For this section, we will consider 5 categories of keywords:

- **#Ferguson**: considers only *#Ferguson*.
- **#MichaelBrown**: links the keywords related to the name of Michael Brown.
- **Other Murders (Aug)**: groups 5 additional murders of black people occurred in August 2014
- **#EricGarner**: joins the keywords related to Eric Garner since he was killed just a few days before August.
- **#BLM**: *Black Lives Matter* and *#BlackLivesMatter*
- **Other Murders (before Aug):** gathers tweets mentioning the killings of 4 black people, which happened before August 2014

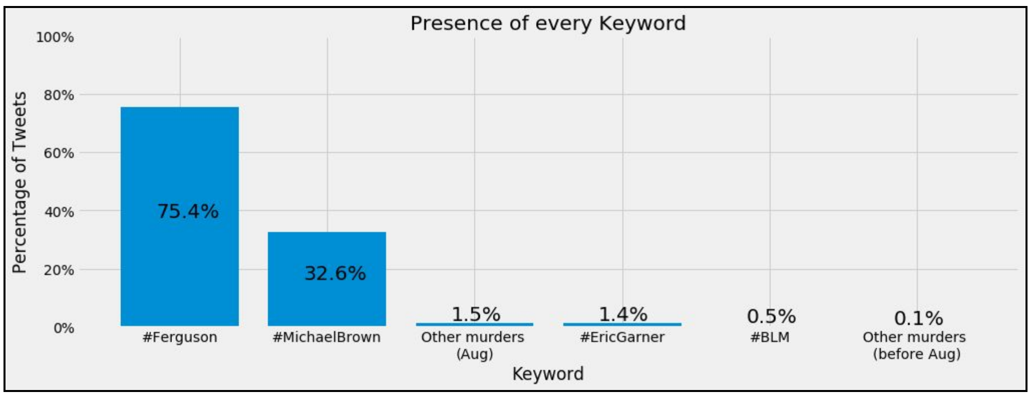

Figure 2. Percentage of tweets containing keywords from each category

We can notice that 3 out of 4 tweets in our dataset mention #Ferguson, and 1 out of 3, #MichaelBrown. However, the other categories have much less presence in our dataset.

We can also see that percentages sum up more than 100%. The reason is that a single tweet can contain multiple keywords. Consequently, with the intention of assigning a tweet to only one group, we defined the next partition. This will be important for the topic modeling we will do in the next phase of the project.

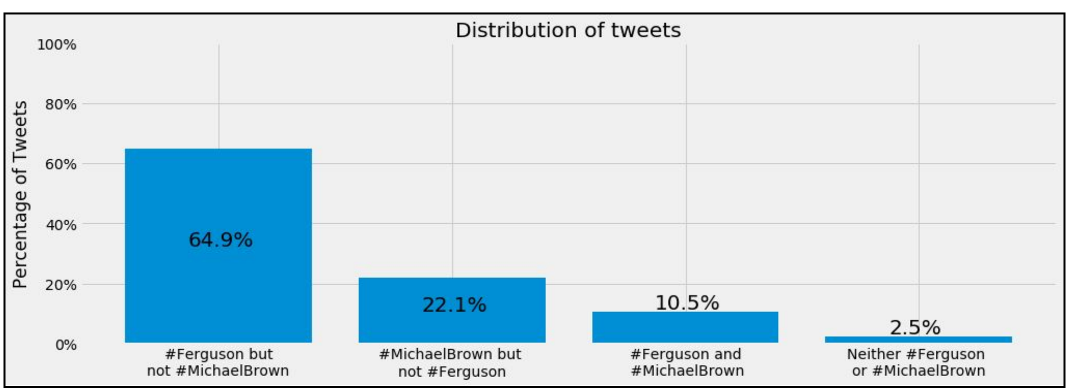

Figure 3. Distribution of tweets under unique buckets

For the next analysis, we will introduce the time-series for the most popular categories.

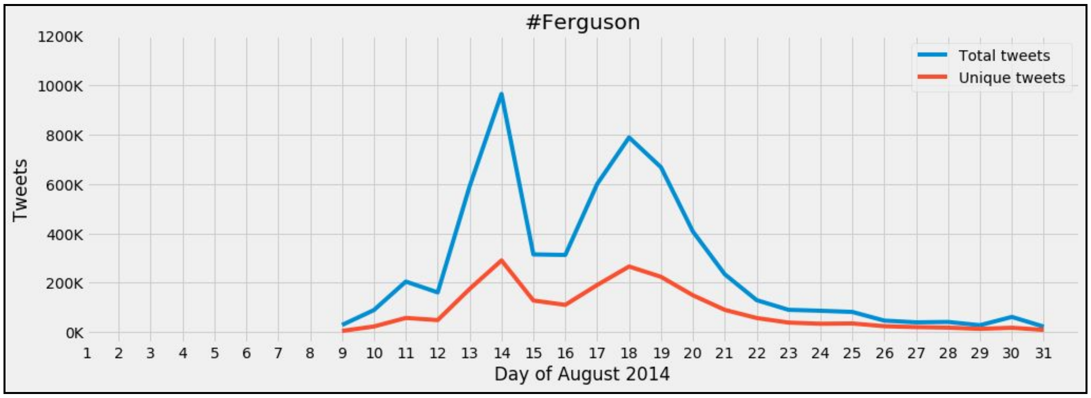

Figure 4. Number of daily tweets with #Ferguson keywords, in August, 2014

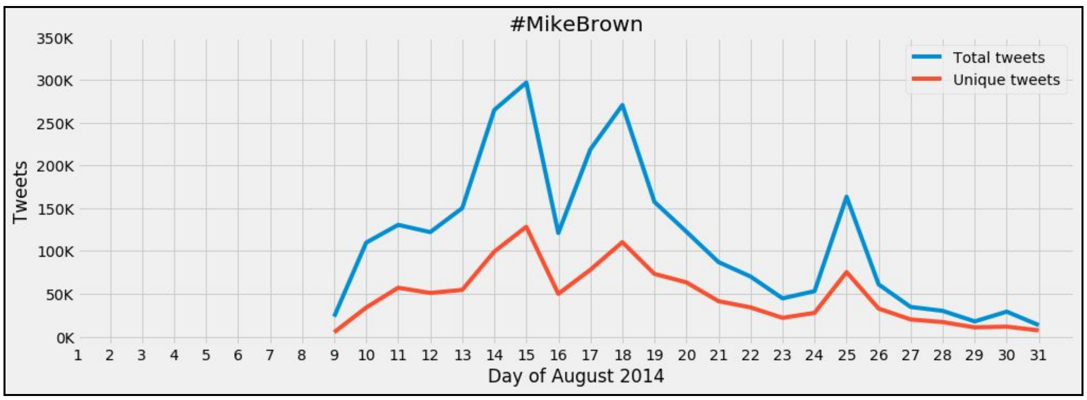

Figure 5. Number of daily tweets with #MikeBrown keywords, in August, 2014

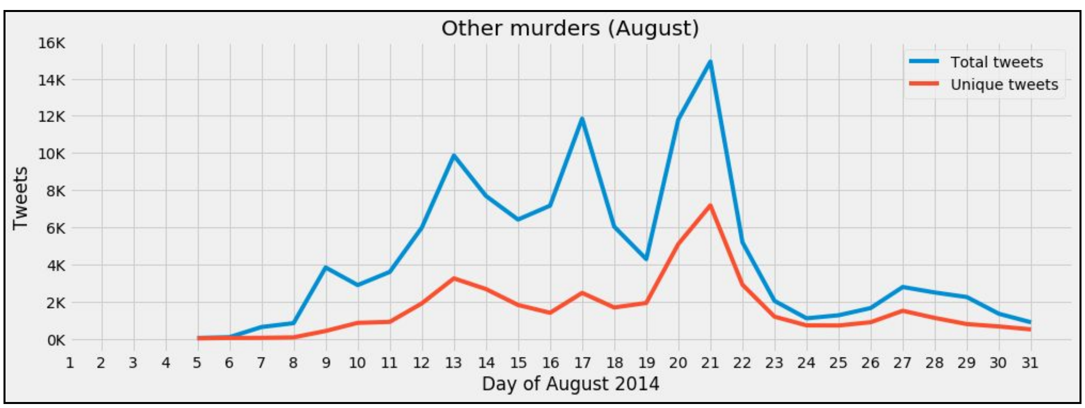

Figure 6. Number of daily tweets with keywords related to other murders,in August, 2014

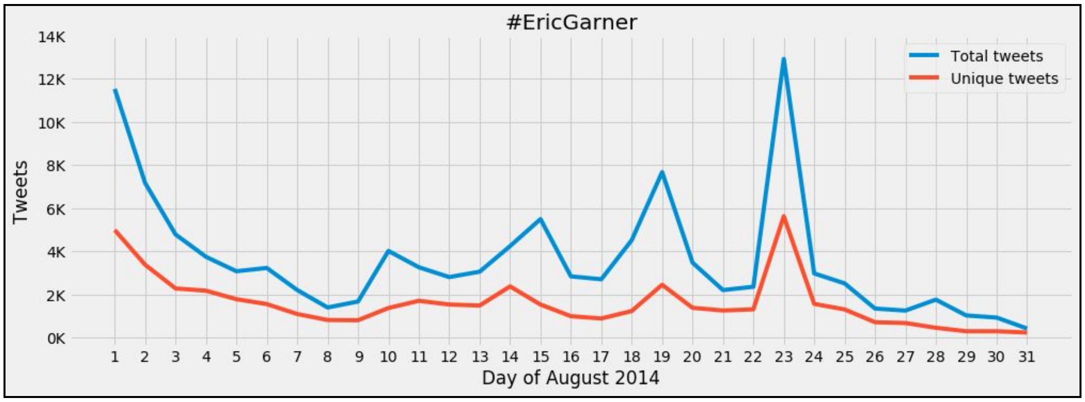

Figure 7. Number of tweets with #EricGarner keywords, by day of August, 2014

We can notice general things, like the irruption of #Ferguson five days after the murder of Michael Brown. It is interesting to notice that the days following his murder #Ferguson was as popular as #MichaelBrown. However, the incidents and protests that followed his death, made the hashtag of his hometown much more popular.

Talking about the series of Eric Garner we can see that it was slowly declining during the early days of that month. Nonetheless, once Michael Brown's killing happened, the number of tweets increased again, having a high pick on the 23rd. The reason is that more than 2,500 people marched that day protesting his murder.

The series related to the other murders has different picks, each of them related to one of the five killings. It is interesting to note that there seems to exist a cumulative effect, with the number of tweets reaching a new maximum every time a new death occurs.

User behavior

In this section we wanted to understand the behavior of the users and try to see how it changed overtime. We first plotted the total number of tweets across time, total number of unique users and unique verified users. From here, we are able to see that close to the spikes on number of tweets the separation between unique users and number of tweets increases, which means that the same users are the ones increasing the activity of tweets they are making, so the same users are tweeting more around those dates. Nevertheless, we can see that most of the users that increase the activity are not verified users, given that the behaviour of verified users does not change over time. This type of analysis is important as it enables us to see who is tweeting and when.

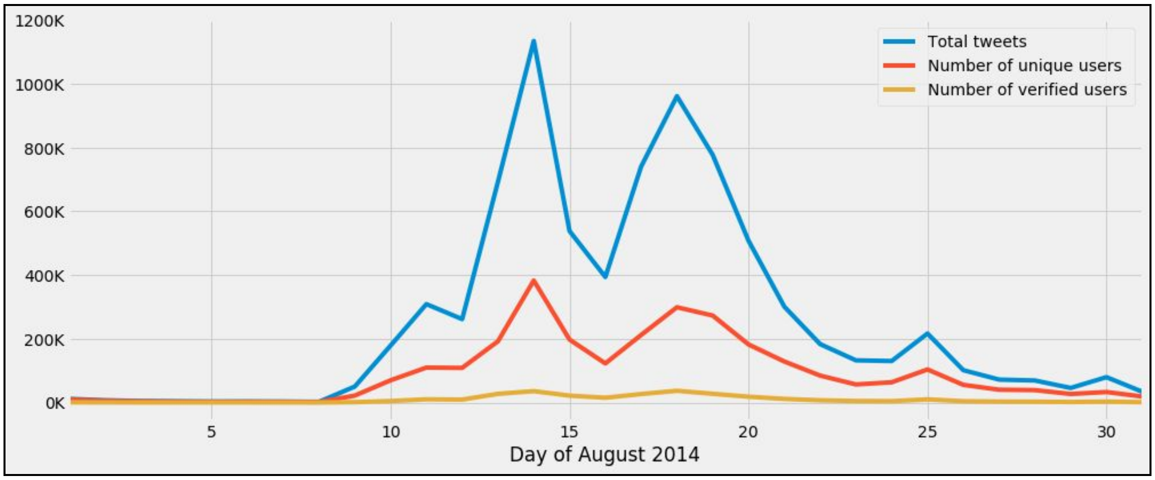

Figure 8. Activity of users and tweets during August, 2014

At the same time, given that we are able to see if a user is mentioned in a tweet from another user we can start constructing a Network Graph where each node is a user and an edge (a,b) will exist if node *a* mentions node *b* on the full text. For this, we remove the RT texts, to just create edges from users that write the full text of the tweet instead of just retweeting. In this particular case, we got a network of:

- Number of Nodes: 1,239,929
- Number of Edges: 130,908
- Average degree: 0.1056

If we want to know which users were mentioned the most, we can use the in degree metric for each node (user), which calculates how many users mentioned it. The Histogram below shows that there are users who are being mentioned more than 500 times (not including RT). If we explore who is the user that is being mentioned the most, we find that it is a YouTube account (verified), with 3,397 mentions from distinct users.

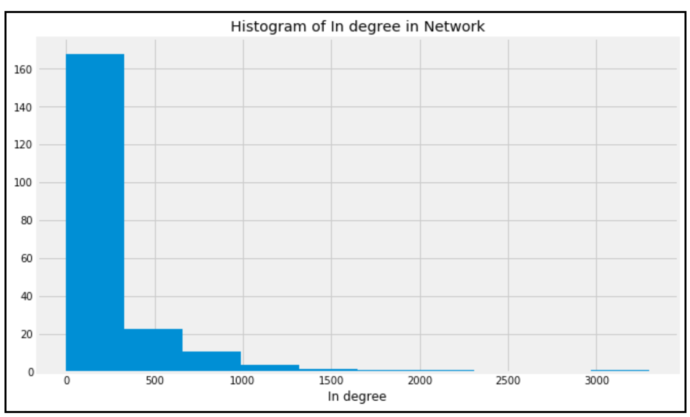

Figure 9. In-degree for users in network

With this type of structure we can start exploring most influential users and analyze possible clusters around users in the next parts of the analysis. Nevertheless, this gives us a first approach on how to deal with interactions among users.

Encapsulated Retweet behavior changes

As indicated by earlier research by [Microsoft](),11% of the tweets are encapsulated retweets. In the case of our dataset, we have that 67% of the tweets are retweets. Thus, it is important to understand retweet behavior.

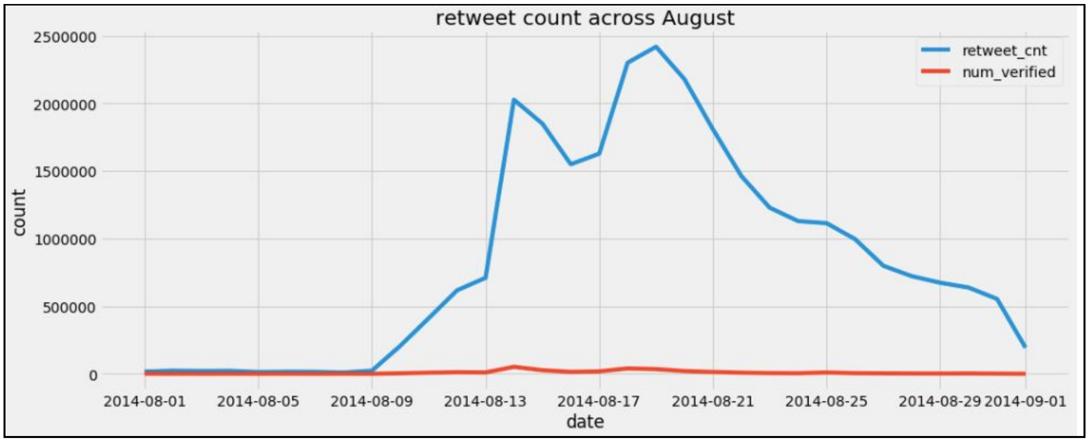

Figure 10. Number of retweets on dataset by day for August, 2014.

The above figure shows that retweets have a significant increase starting on August 8th and have two surges in sync with the events. However, we can see that the retweets show a longer and higher lagging time after the events.

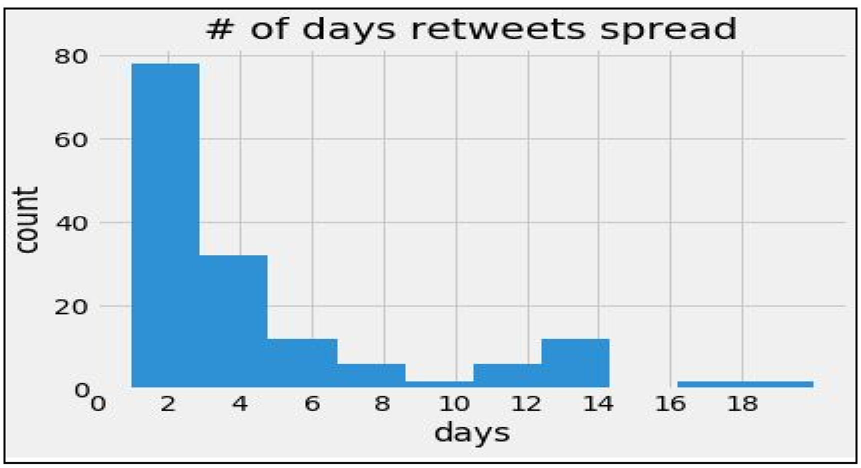

Figure 11. Number of days that a Tweet was Retweeted

Another interesting discovery is the retweet lifecycle - the number of days a tweet is retweeted. There is a polarized behavior where tweets either survive multiple days among users or get quite popular within one day and die down. Figure 12 shows a sample of tweets that lasted for multiple days, one by American Journalist Chris Hayes and one by public speaker Jackie Summers.

Figure 12. Sample of Tweets that were retweeted for multiple days

This leads to a deep dive into the retweet behavior of influencers (users with multiple followers). We have discovered that an average user has about 10,000 followers in this case. Thus, we have separated users as influencer - those with more than 10,000 followers - and then we have adjusted the total retweets by dividing the total number of influencers. Figure 13 shows the user behavior of retweets has three surges during the time period between August 9th and 21st. Influencers posted an average of 4.5 times during the time of the incident.

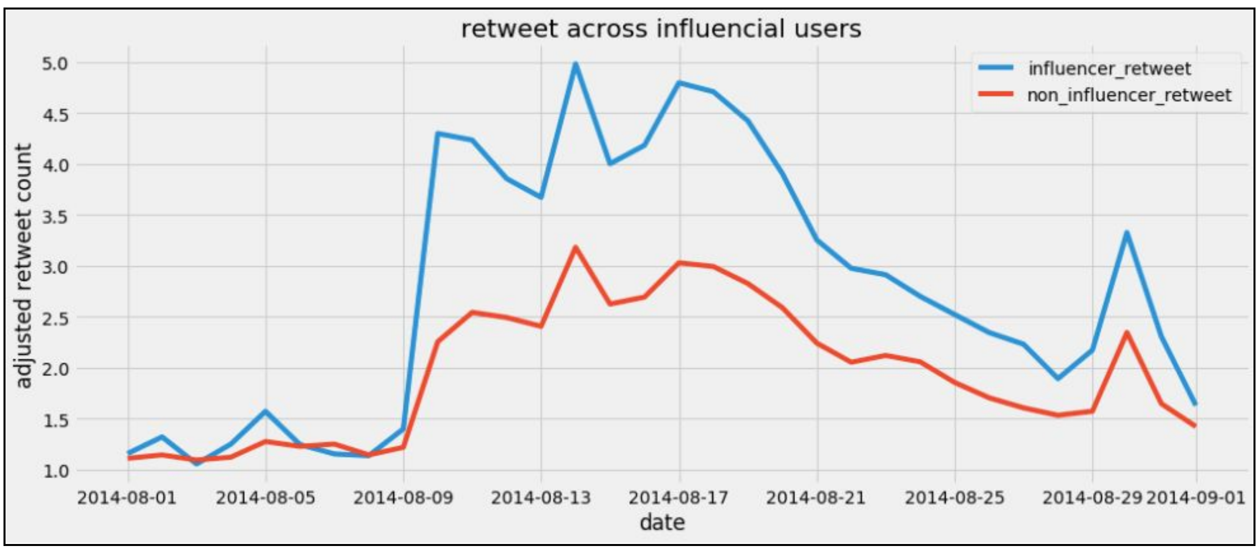

Figure 13. Behavior of retweets for influencers and non-influencers

4. Goals and Next steps

From the beginning of the project, professors Leach and Cogburn laid out their expectations on what would be interesting to analyze in the data, from a sociological perspective. These subjects were summarized on the following 4 work streams:

1. Topic modeling

One of the objectives of this project is to perform a topic modeling analysis that contributes to the understanding of the underlying themes discussed through Twitter during the period under study. An interesting aspect of this is how much bias in the topics is included due to the construction of the database from keywords. The team is building a pipeline to analyze the dataset using the Mallet Model. Some preliminary results indicate that coherence is maximized when modeling the corpus under 10 topics, and some of the clearest themes have a relation with media and police. Further efforts will be made during the next phase of the project.

2. Sentiment analysis

The goal of sentiment analysis is to 1) understand emotions within the tweets 2) explore how emotions change throughout different periods and incidents 3) understand how audiences interpret the tweets and react to them. The complexity of understanding emotion within tweets is to explore beyond the basic (positive, negative, and neutral) sentiment classes. We are aiming to expand the classes of emotion to ensure a more comprehensive coverage of the social influence of tweets. In addition to the multidimensional lexicon, the team is in the process of communicating with professors to determine a domain-specific lexicon. In addition, we are aiming to include the output from topic modeling as a feedback loop to refine and modify the aforementioned lexicon to cover a comprehensive output. Once a completed lexicon is created, we are planning to map a time series to explore the changes.

3. Network analysis

An additional form of analysis that the team started exploring is social network. Given the amount of interactions that take place between Twitter users surrounding our topics of interest, we modeled users and mentions in tweets as a network in our initial attempt at network analysis. For the next phase of the project, we want to explore how these interactions change over time following the events that are covered in the keywords.

4. Image Analysis

One additional topic to discover is the relation between specific tweets and attached images. As indicated by the paper '*Integrating Text and Image: Determining Multimodal Document Intent in Instagram Posts*', we can potentially segment different tweets based on both text and image into intention, semiotic, and contextual relationships. This additional information extracted from images allows us to gain more context on the potential influence on a broader audience.

Contribution

Each team member has contributed during every step of the process, from reading and discussing the background information to agreeing on a way to store the information where it is accessible to make the analyses and exploration that will guide the capstone. Specific efforts include Andrea and Shimeng working on preparing the dataset as a single file, Omar and Jose working on the topic modeling efforts and Shadi working on drafting the sentiment analysis. Everyone contributed to the exploratory analysis, on choosing the most interesting insights, and ultimately to the construction of this report.