GDELT News Analysis

Andrew Burkard, Kanav Chandra, Connor Laird, and Jiahao Xu

May 11, 2017

Abstract

The Global Data on Events, Location and Tone (GDELT) data set tracks media coverage of events across the world. We benchmarked several methods of analyzing this data and found a solution that uses free tools (BigQuery and Tableau) that was sufficiently capable of developing insight from the data. We developed three case studies of major international events to assess the actionability and descriptive abilities of the GDELT data set. These included the 2016 U.S. election, the 2015 attacks in Paris, and the recent relationship between the U.S. and North Korea. We gained empirical insight and context for these events and also uncovered some data quality issues with GDELT.

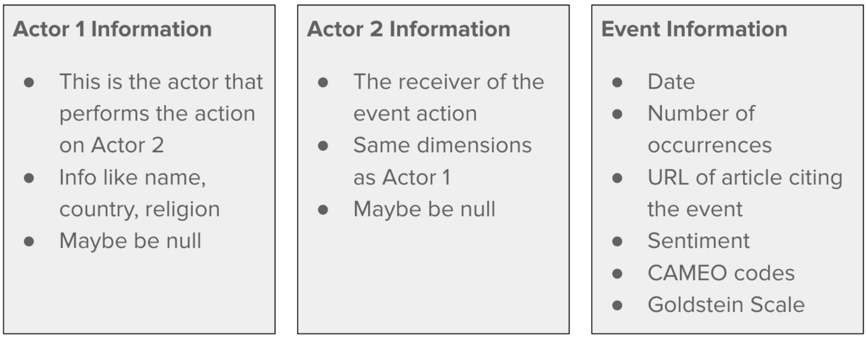
1 Introduction

Our group decided we wanted to focus on news media for our analysis given the recent surge of interest in journalism and high news throughput election cycle. One of our group members suggested that we investigate using the GDELT (The Global Data on Events, Location and Tone) project as a source of data, as he had some previous experience using this data for a project at work. GDELT can be found at this address - http://www.gdeltproject.org/. GDELT monitors all forms of media and tracks information about actors and sentiment. It was created by a Georgetown School of Foreign Service fellow Kalev Leetaru and is free and open source with support from Google Jigsaw. Jigsaw is a project that uses technological solutions to solve problems in society. We set out to do three things over the course of the project:

1. Use big data to examine how the sentiment of news coverage changes after major events
2. Develop an understanding of the GDELT dataset, and benchmark ways to access and analyze the data
3. Use case studies of specific major events to assess the actionability and descriptive abilities of the GDELT data set.

Prior to settling on the GDELT data set we investigated other possible sources for gathering data on news media subject matter, like the New York Times API, but concluded that we could find no better alternative that would provide the scale and robustness of GDELT. Not only is it a truly massive data set, but complex analyses involving natural language processing are built into the data.

The GDELT project is comprised of many data sets, but we focused on the events database. GDELT also maintains graph structured data among other sets. The events table encompasses approximately 90 GB as of May 2017, and is continuously updated. It has 61 columns (of types string, int, and float) and the schema, as seen in the figure below, focuses on two actors performing an event. Actor 1 is the primary focus of an event, while Actor 2 is the recipient of Actor 1's action.



Actor 1 Information

- This is the actor that performs the action on Actor 2
- Info like name, country, religion
- Maybe be null

Actor 2 Information

- The receiver of the event action
- Same dimensions as Actor 1
- Maybe be null

Event Information

- Date
- Number of occurrences
- URL of article citing the event
- Sentiment
- CAMEO codes
- Goldstein Scale

Figure 1: Schema of GDELT events database.

2 Literature Review

At the onset of our project we sought to verify that the GDELT dataset was a reliable and useful source. We turned to the academic and scientific community and found extensive use of the data and confirmation that we would not be hampered by issues with the data set over the course of our project. Some journals simply reviewed the data set as a resource, without delving into any specific analysis or use case. *Choice* recommended use of the data set by graduate students and professionals, saying that despite some shortcomings GDELT is an "unrivaled source of data for analysis by experts interested in furthering their understanding of the connections between events" [Iul14].

Not only is it reliable, but it is a preferred data set amongst its competitors. An analysis of GDELT in comparison to EventRegistry by Haewoon Kwak and Jisun An of the Qatar Computing Research Institute noted that GDELT has a scale and completeness that is unmatched by EventRegistry. [KA16]

We found some analysis that was similar to the ideas we came up with in our initial brainstorming sessions. Also pursuing a case study approach with the GDELT data, a group of researchers from Texas State University and Peking University looked at the image of China painted by international media organizations. [YLW16] While GDELT was their primary data set, they also incorporated social media data to get even more support to their findings. Specifically they examined how the 15 countries with the strongest socioeconomic connections to China portrayed it in media.

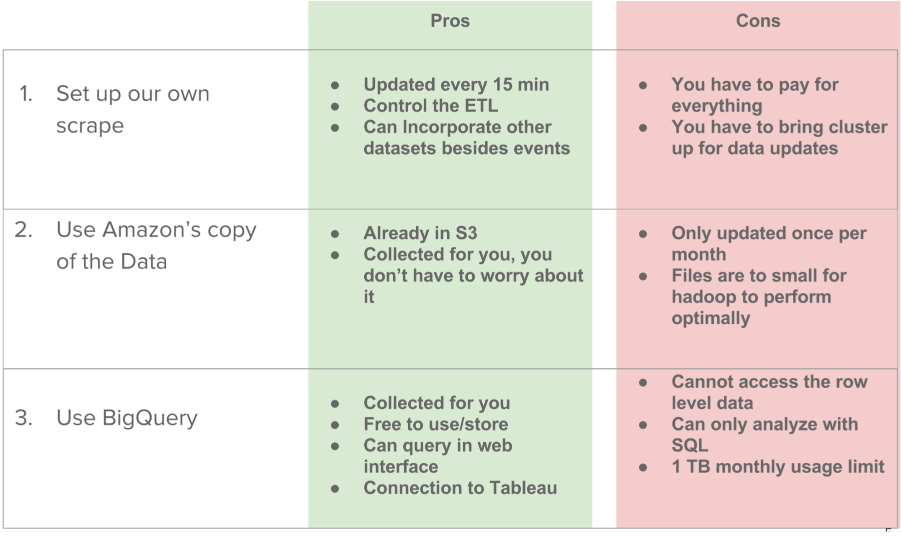
They not only looked at the temporal aspects, but also truly examined the spacial relationships between countries. This paper was written in partnership with the geography department of one of the universities. While we looked at international relationships, we did not take into account distance between countries specifically. They also used a longer time frame by employing GDELT 1.0 data, while we only used the more recent GDELT 2.0 data set.

They had promising results depicting international relationships with GDELT, and were able to validate their findings with other data sources. These sources included more traditional economic and military data. Generally, we were very pleased with what the literature indicated about GDELT and were optimistic about using this data for our project.

3 Methods

Our initial exploratory data analysis was done using Google BigQuery. Google BigQuery is a tool that was showcased on the GDELT website that allows you to query the whole data set with SQL. The data is updated constantly and the query engine is free to use under certain rate limits. It let us get early insights without managing any infrastructure of our own. Alone, it did not let us create any visualizations.

We decided to pursue creating our own scraper that would download the entire dataset to S3, where we could use tools like Spark to analyze it. We were successful in creating this scraper, it would parse the list of available files and determine which ones were not yet downloaded then download them to the cluster and upload them to an S3 bucket. Even though we could successfully set up this scraper, we ran into some practical issues. First, we had to run an EMR cluster all the



	Pros	Cons
1. Set up our own scrape	● Updated every 15 min ● Control the ETL ● Can Incorporate other datasets besides events	● You have to pay for everything ● You have to bring cluster up for data updates
2. Use Amazon's copy of the Data	● Already in S3 ● Collected for you, you don't have to worry about it	● Only updated once per month ● Files are to small for hadoop to perform optimally
3. Use BigQuery	● Collected for you ● Free to use/store ● Can query in web interface ● Connection to Tableau	● Cannot access the row level data ● Can only analyze with SQL ● 1 TB monthly usage limit

Figure 2: Pros and cons of each approach.

time, as the GDELT data is constantly being updated. Additional, each file that would be put into S3 was relatively small, running spark jobs against the bucket was problematic because the files were far less than the block size.

We briefly pursued using a copy of the data on a public S3 bucket hosted by Amazon. We ran into the same issue of small file sizes. Additionally we found that this version of the data set was not updated more than once monthly.

So after investigating these approaches we returned to BigQuery. We found that when used in conjunction with Tableau, a data visualization software, all of our analysis needs will be met. Tableau can connect to almost all types of databases, and includes a native connection capability for BigQuery. Visualizations and analyses of the full 90 GB data set can be done in a matter of 20 to 30 seconds. The full stack could be used for free, and we had no issues with rate limiting.

4 Results

We took a case study approach to figuring out what insight could be gained on major world events using the GDELT data. We pursued three international events or relationships, the 2016 United States election, terrorist attacks in France, and the relationship between North Korea and the United States.

You might notice in this section's figures that we used fractional dates throughout our analyses because they provide more procession then discrete dates and worked better in Tableau. In this scheme date times are floating point decimals where each year is one unit. So a few minutes after midnight on New Years Eve at the start of 2017 would be something like 2017.0005.

4.1 2016 U.S. Election

The 2016 United States election was an event that garnered extreme amounts of media coverage. We examined Hillary Clinton's interactions with other actors over the course of 2015, 2016 and 2017. As detailed later in this paper, GDELT was unable to specifically identify Donald Trump as an actor at this time.

We looked at the time series of events tagged with Hillary Clinton as an actor. In the figure below, the aggregate of all those events is graphed. We also examined similar time series data split between events where she was Actor 1 or 2 alone. For the purposes of this analysis, we normalized the average tone scores so that -2 was neutral, higher scores were positive sentiment and lower scores were negative sentiment. We discovered two main findings.

First, Hillary Clinton's disappearance after the election, which has been anecdotally showcased in media coverage, is empirically noted through this analysis. There is a drastic increase in articles published about her in the lead up to the election. Almost immediately afterwards, this coverage plummets to levels lower than since the start of 2015. The story of the election turned from a battle between Clinton and Trump to a single actor story of Trump alone after his victory.

Secondly, we found that the average tone scores were easily tripped up by some of the more complex interactions that Clinton was involved with over the course of the election. Most notably was the day after the election when results were being published and analyzed. The highest magnitude point in the below figure, it is noticeably blue. Interestingly, this measure of tone was positive for cases when Clinton was either Actor 1 or Actor 2. Obviously, this was not a positive event for her. The sentiment scoring was unable to parse out the true directional sentiment, which in this case would be markedly negative for Clinton.

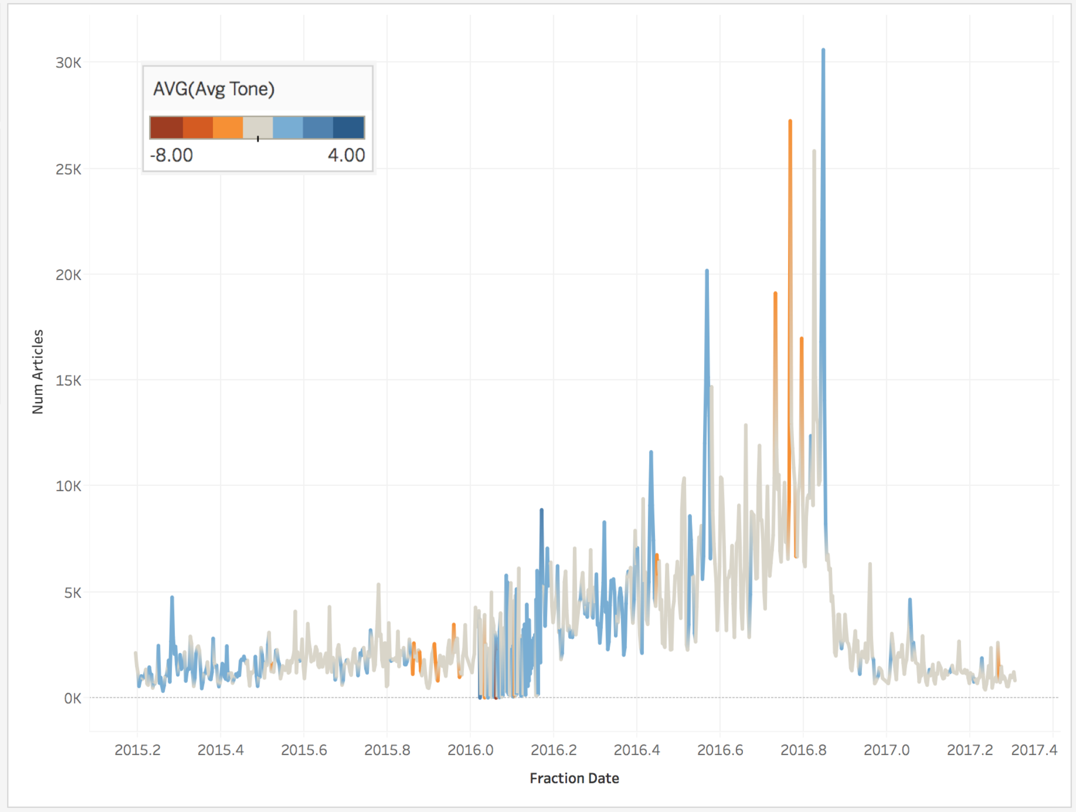


Figure 3: Mentions of Hilary Clinton as Actor 1 or Actor 2 Over Time. The color is the average of the ’Avg Tone’, or sentiment score, of events for that day

4.2 November 2015 Paris attacks

On November 13, 2015, Islamic State of Iraq and the Levant (ISIL) militants carried out a series of coordinated attacks in Paris, France. The events received extensive media coverage, the effects of which can be witnessed in GDELT. On the day of the attacks, there were $8,075$ events involving a French actor[1]. The following day day, there were $32,848$ such events, a 307% increase on the day prior. The additional media coverage brought with it a significant increase in negative sentiment. The AvgTone measure dropped from -2.510 to -5.834 the day after the attacks. Interestingly, it took until November 29 for sentiment to return to its previous levels. Other European countries were also affected as seen in Figure 4, but their Avg Tone rebounded more quickly while France’s continued to languish.

[1] Actor1 or Actor2 has country code France

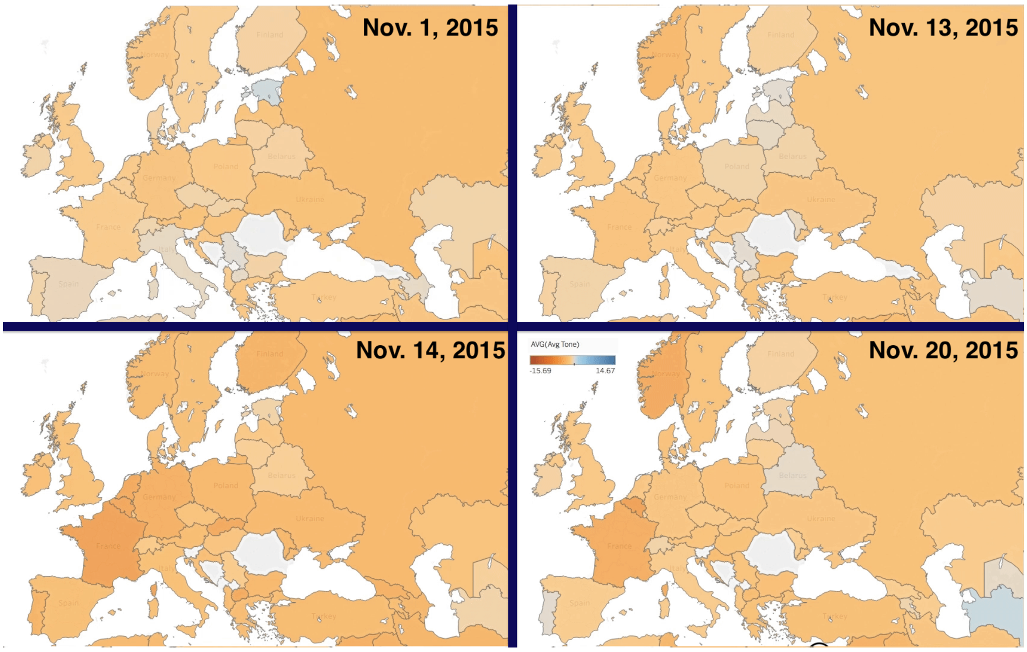


Figure 4: Changes in sentiment before and after the Paris terrorist attacks

4.3 U.S. - North Korea relationship

Our final case study examined the relationship between the United States and North Korea. The Actor 1/Actor 2 model lends itself very well to modeling this kind of two way international relationship. The main events in this contentious relationship have in the past few years centered around nuclear test launches by North Korea, also known as PRK - The Peoples Republic of Korea. As seen in the figure below, there are spikes in the number of articles published around each launch. Actions done by PRK to the USA are drawn in red, while actions by the USA to PRK are in blue.

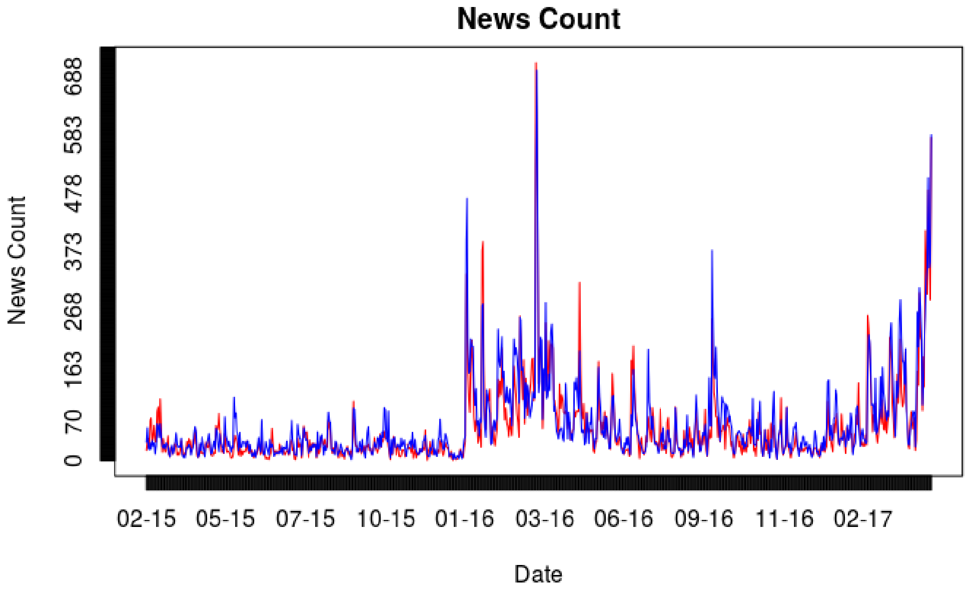


Figure 5: Number of Articles About U.S. - North Korea Relationship

We see that after the first nuclear test, which occurred in early 2016, there was a clear decline in sentiment score. The sentiment of news articles stabilized after the initial reactions. The tension between the countries was already apparent - the average sentiment score even before the nuclear test was about 1 unit below the global average.

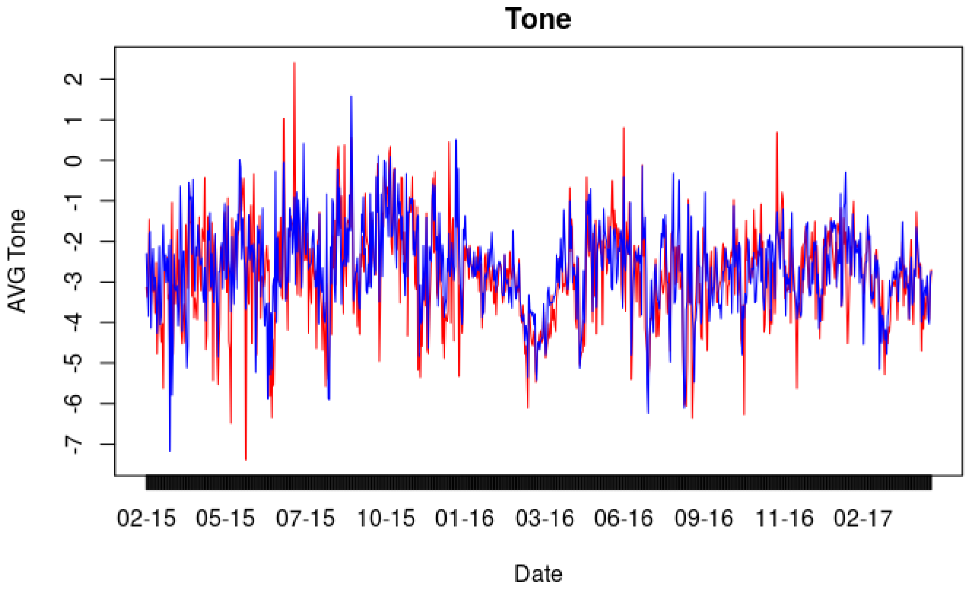


Figure 6: Drop in Tone After First Incident

To further examine the relationship, we determined the most frequent actions in each direction. The most common action covered by the media done by North Korea to the United States was ’Arrest, detain’. This is an obviously hostel action and was due with U.S. citizens being arrested in North Korea which resulted in media outrage. The United States does not take such direct action, and there most frequent action is to ’Make Statement’.

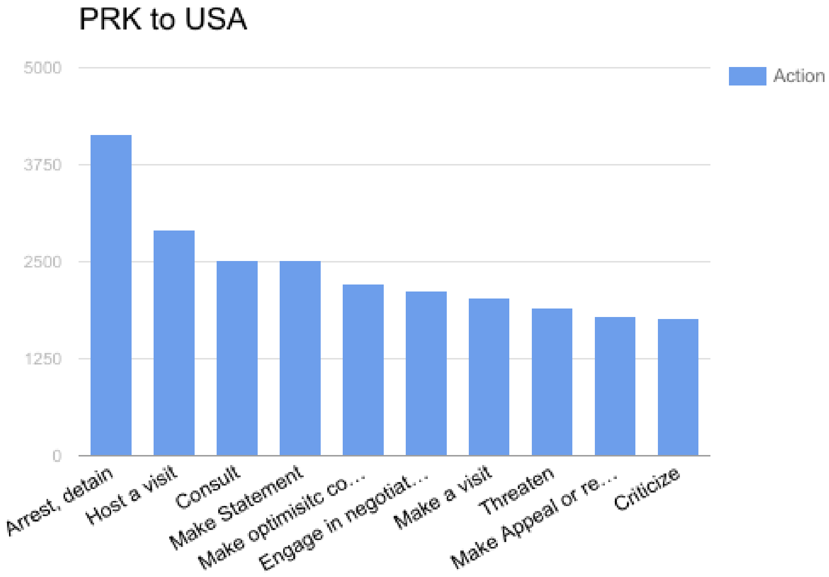


Figure 7: Common Directional Actions From PRK to USA

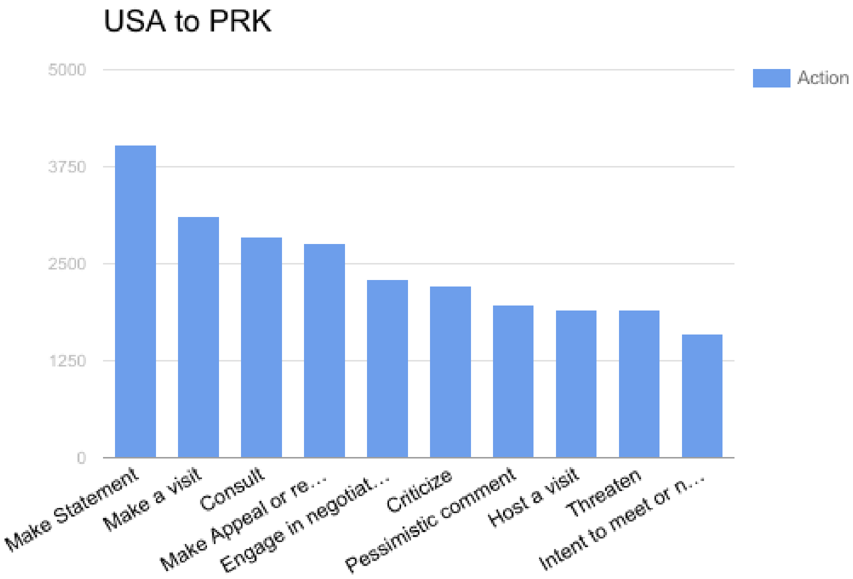


Figure 8: Common Directional Actions From USA to PRK

5 Conclusion

Over the course of this project, we were able to meet the goals we set out to accomplish. First, we developed a working knowledge and best practices around the GDELT data set. We spent significant time pursuing multiple solutions for setting up an analysis environment for this data set. As detailed in the method section, we found that a combination of Tableau and BigQuery was better for the purposes of this project then a more capable, but also harder to manage solution that would rely on Spark. While this solution is not workable for all analytics, specifically any models that require use of the data at an atomic level, it was appropriate for time series analysis and other aggregations.

After building our analytics capabilities we moved towards developing some case studies for the effectiveness of the GDELT data to gain insight on world events. Generally, we were able to quantify complex interactions between global actors. This work could serve as a proof of concept for a model that would detect or predict sentiment, actions or other features relevant to the data set. While developing these findings, we did find some shortcomings of the data set. Detailed in the following paragraphs, these shortcomings would have to be taken into account heavily when accessing the viability of a particular model or technique.

First, we ran into some issues with the sentiment score. On average the baseline score for an article was negative. This is not inherently an issue, but we did find that this bias made it more difficult to detect changes to sentiment. Truly positive events were harder to parse out and negative delta in sentiment was less noticeable since almost all articles already appeared negative. We also found that because sentiment was not directionally related to the actors it was hard to tell the correct story. For example, election day for events in which Hillary Clinton was actor 2 (the target, as opposed to the subject) appeared to be of positive sentiment. This was obviously not the case, and the sentiment algorithm was tripped up on words like triumph and victory. These words were actually describing her opponent, it was a negative event for her.

Secondly, there were some issues with the magnitude of events tracked by GDELT during certain time periods. Some days had to be excluded from analysis because they had very few event recorded. This caused the variance of metrics like tone and Goldstein score to become too large to be included in any sort of fair analysis.

Finally, it appears that the actor names are generated from a predefined set, which means that new actors are not correctly tagged. The most glaring example of this came in the election case study. While both Hillary and Bill Clinton were tagged as actors in various events, Trump was never mentioned explicitly. This is probably because the Clintons have long been involved in world events and have made it into a set of keywords somewhere in the GDELT system. Trump, a newcomer, was not in this set even though he is now a major actor. This caused some trouble in our analysis of the election, resulting in our focus on Clinton alone.

References

[Iul14] A. M. Iuliano. Gdelt: The global database of events, language, and tone. *Choice*, 52(1):53, 2014.

[KA16] Haewoon Kwak and Jisun An. Two tales of the world: Comparison of widely used world news datasets gdelt and eventregistry. 2016.

[YLW16] Yihong Yuan, Yu Liu, and Guixing Wei. Exploring inter-country connection in mass media: A case study of china. *Computers, Environment and Urban Systems*, 62:86–96, 2016.