Data Analysis on the GDELT Project

Global Database of Events, Language, and Tone

Curry Buscher
University of Colorado, Boulder
Curry.Buscher@Colorado.edu

Rohan Baishya
University of Colorado, Boulder
Rohan.Baishya@Colorado.edu

Benjamin King
University of Colorado, Boulder
Benjamin.King@Colorado.edu

Gayathri Gude
University of Colorado, Boulder
Gayathri.Gude@Colorado.edu

1 PROBLEM STATEMENT/MOTIVATION

The data being obtained from the GDELT project, archives a span of 45 years of "emotional snapshots". This data is essentially known to be a study of the global human society. The attributes within the dataset include information regarding societal influencers (people), risk assessment & global trends, policy reactions, and humanitarian & crisis responses. The purpose to analyzing this data is to spot trends and recognize patterns in common global events, specific to locations and human interactions. It will be investigated what events could be predicted given prior event patterns along with uncovering unexpected correlations within the data.

2 LITERATURE SURVEY

The GDELT database has been used in several prior studies. One of the ways the GDELT foundation has studied this data is to generate maps of influence in relation to 'influencers' around the world. This map connects each influencer to another based on their geographic region. This study allows researchers to determine the flow of information from a primary source throughout a region, and which influencers' input is essential to disseminate information worldwide [1]. One of the ways this was applied was to study Edward Snowden and the global effects of the continual efforts of the Wikileaks organization [2]. Next, the founder of GDELT has used this database to study wildlife trafficking in relation to global events. This study won the 2016 Wildlife Crime Tech Challenge.[3] GDELT's 'Television Explorer' system has been used by a variety of reporting organizations to analyze how the media responds to certain events. One such use analyzed Fox news' coverage of mass shootings after Robert Mueller indicted Russians in the US.[4] Another study analyzed the trend of discussing gun violence and gun control in the media, and found that after mass shootings, these conversations spiked, then returned to a normalized level.[5] The GDELT database has also been used to track interest in various topics over time, including cryptocurrency and blockchain technologies. This summary provides an in-depth overview of these emerging technologies, and how they have come to be popularized among a general audience.[6]

3 PROPOSED WORK

To begin analyzing the data, a lot of data preprocessing will need to be done due to the sheer size and variety of data included in the 6.4 GB of GDELT data. To do this, the first step will be to pinpoint and narrow down the data to the attributes that correlations will be made for, and make sure to replace holes with projections, as well as make sure each value per attribute is homogenous based on variable type. The data will be sorted as well as deemed appropriate and specific sizes of data will need to be determined, where a proper span of information is being represented with respect to the comparisons and correlations.

The focus will be to connect historical geographic data to various attributes to make conclusions and predictions on what can happen, where it can happen, and the likelihood of these events happening at those times and places. Furthermore, we will analyze how policy changes over time have affected the various aspects of these events, as well as how different regimes and political circumstances in different countries affect social distress and protests in those areas over time.

4 DATA SET

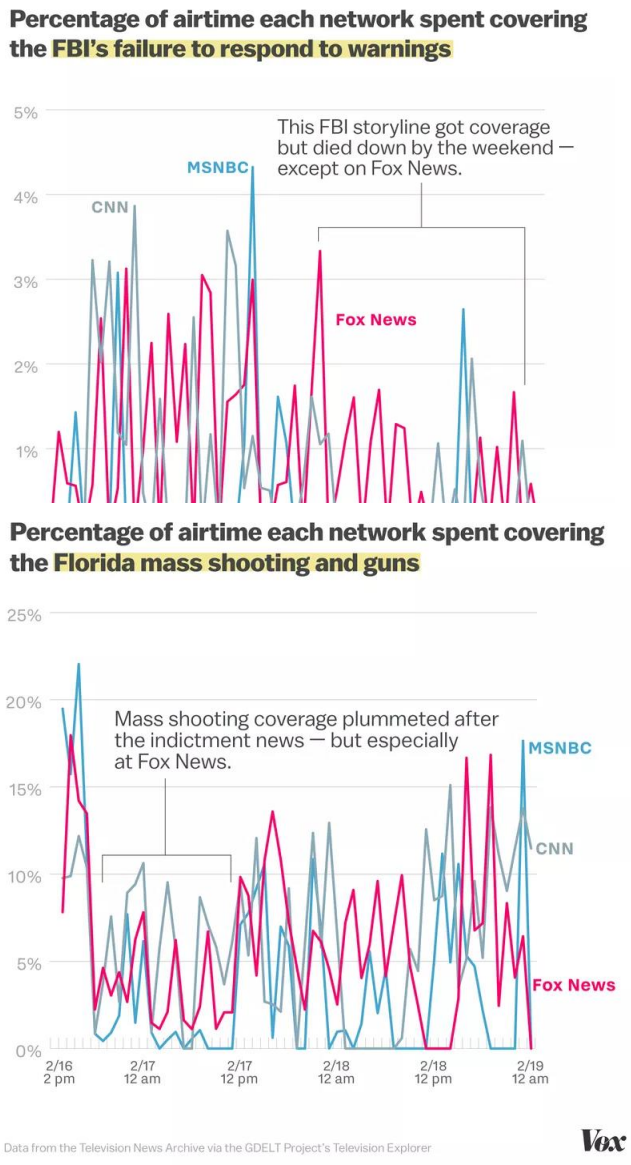
Our dataset spans approximately 50 years of global data, and is sized at more than 6.4 GB. This is a summary of the original dataset which goes back 215 years and has an average size of 2.5 TB per year. The dataset includes numerous different attributes: **Latitude**-Quantitative/Discrete, **Longitude**-Quantitative/Discrete, **Country Code**-Qualitative/Nominal, **Country Abbreviation**-Qualitative/Nominal, **Number of Events**-Quantitative/Discrete, and **Event Scale/Magnitude**-Quantitative/Continuum. The dataset

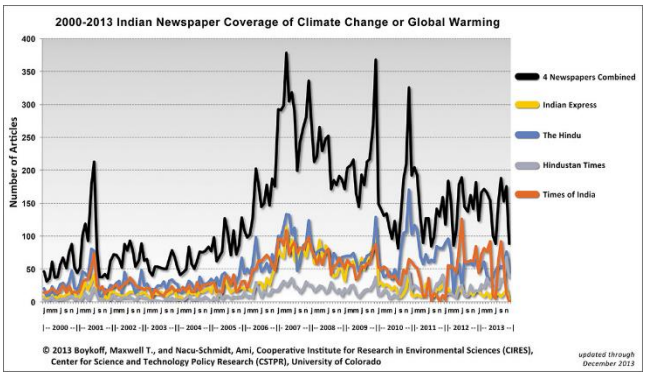
can be found on the GDELT website, and is currently stored on Ben King's computer.[7]

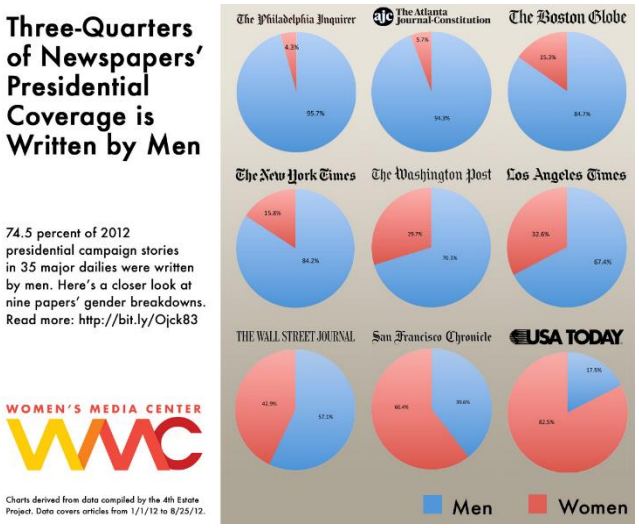
This dataset is a csv formatted package, containing all relevant datasets. This subset will then be cleansed and narrowed to specific topics, times, and attributes. From here, specific files will be mined to obtain the ultimate correlations and conclusions.

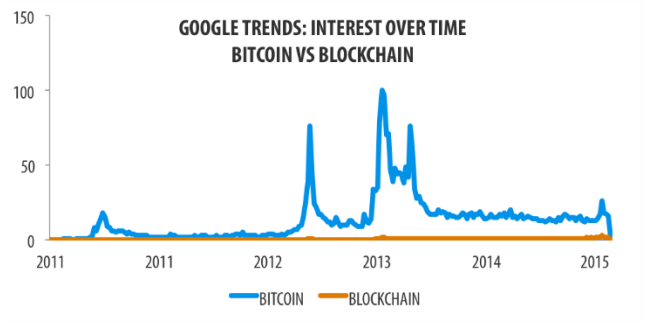
5 EVALUATION METHODS

The goal of the data analysis is to map the correlations found year by year as well as group by time periods of significance. The data that will be collected and analyzed will further be compared to existing data analytics and see how they match up. Below are some data summaries in the form of visuals that comparisons will be made to and conclusions drawn with respect to our processed data:









In order to evaluate the above data and form conclusions, trends and data clustering and classification will be examined and data represented with plenty of visuals implementing MatLab plots, numpy, and pandas, after which correlations of incidents between locations and dates will be found. The analysis done will hopefully result in applicable uses such as analyzing risk assessments and crisis response times and methods to pinpoint where high risk zones are and what zones need more help and faster help. The effectiveness of policy changes will be tested as well, and predictions will be able to be made based on countless factors e.g. industry, geographical location, government, economic conditions, time, etc.

6 TOOLS

Tools that will be used include Python 3 (with a lot from the Pandas and Numpy Libraries) for preprocessing, processing, sorting, and analyzing the data throughout our project. VIsualization of data in the form of graphs and charts will be made primarily with Matlab Plots.

Tensorflow will be implemented for numerical computations and graphing our data with mathematical functions and operations as well as representing the multidimensional data arrays that connect them. Keras will be used on top of Tensorflow to enable fast experimentation and deal with configurable modules that can be plugged together with little restriction.

Git/GitHub will be used version control and to monitor/update our progress among the group. Our project link is https://github.com/cubu0178/WeLikeBigData

Finally, WEKA will be incorporated for specific, detailed data analytics including data clustering and classification as well as visualization.

7 MILESTONES

The following table shows the timeline in correlation to class deadlines for each milestone. This project will be aimed to follow these given deadlines, as shown below.

Milestone	**Due Date**
Proposal Paper	March 6, 2018
Data Cleaned	March 12, 2018
Numpy & Pandas Data Exploration	March 14, 2018
Keras & Tensorflow Analysis	March 19, 2018
Weka Analysis	March 26, 2018
Graphing Results	April 2, 2018
Progress Report	April 10, 2018
Remaining work	April 16, 2018
Final Presentation	April 23, 2018
Final Paper	May 1, 2018

8 SUMMARY OF PEER REVIEW SESSION

From the presentation peer review, the current approach of finding unique trends and building from there is an appropriate path forward. The feedback received from the course coordinator was generally positive, as it referred to the nature and breadth of the current analyses, and will serve as a starting point for all future data analysis.

REFERENCES

[1] The GDELT Project. (n.d.). Retrieved March 06, 2018, from https://www.gdeltproject.org/solutions.html#influencers

[2] Leetaru, K. (2013, December 31). King Snowden and the Fall of Wikileaks. Retrieved March 06, 2018, from http://foreignpolicy.com/2013/12/31/king-snowden-and-the-fall-of-wikileaks/

[3] Wildlife Trade News - Winners of Wildlife Crime Tech Challenge announced. (n.d.). Retrieved March 06, 2018, from http://www.traffic.org/home/2016/20/winners-of-wildlife-crime-tech-challenge-announced.html

[4] Chang, A. (2018, February 19). Fox Newss appalling past 72 hours, analyzed. Retrieved March 06, 2018, from https://www.vox.com/2018/2/19/17027456/fox-news-mueller-indictment-trump

[5] Qureshi, H. (2018, February 21). Conversations about gun violence spike when shootings occur, then decline soon after. Retrieved March 06, 2018, from http://www.dailyhelmsman.com/news/conversations-about-gun-violence-spike-when-shootings-occur-then-decline/article_d938d98c-1754-11e8-9c61-4b1b27430f24.html

[6] Tracking Bitcoin/Cryptocurrencies And Blockchain Using GDELT Summary. (2018, January 08). Retrieved March 06, 2018, from https://blog.gdeltproject.org/tracking-bitcoin-cryptocurrencies-blockchain-using-gdelt-summary/

[7] The GDELT Project. (n.d.). Retrieved March 06, 2018, from https://www.gdeltproject.org/data.html#rawdatafiles

[8] Who's Writing Nine Newspapers' Presidential Election Coverage? (n.d.). Retrieved March 06, 2018, from http://vitaminw.co/politics/whos-writing-nine-newspapers-presidential-election-coverage

[9] Inside the Greenhouse. (n.d.). Retrieved March 06, 2018, from http://sciencepolicy.colorado.edu/icecaps/research/media_coverage/india/index.html