Data Mining to Explore the Relationships between Media and Society

Brock Bylovas
Electrical and Computer Engineering
University of Colorado-Boulder
Brock.Bylovas@colorado.edu

Weipeng Cao
Computer Science
University of Colorado-Boulder
wwcao@colorado.edu

Henok Hailenariam
Computer Science
University of Colorado-Boulder
heha8905@colorado.edu

Trevor Stephens
Computer Science
University of Colorado-Boulder
Trevor.Stephens@colorado.edu

PROBLEM STATEMENT AND MOTIVATION

The significance of the media within modern societies should not be underplayed. Media outlets reflect the present states of their respective regions all around the world. Political elections, international and domestic affairs, and the general mood of a population can be encapsulated, and even strongly influenced, by what's being reported in the news. The general purpose of this project is to examine the relationship the media has with society. We will be examining the patterns of style the news has throughout time, especially during major political cycles, determining if there are any markings in reporting that could predict domestic and international conflict, and monitor the way media has changed throughout time.

PREVIOUS WORK

Many researchers involved in the field of media and politics have previously worked on projects similar to this one. Two of the articles that capture common themes with previous research will be briefly discussed here. Each of these articles uses the same database, Global Data on Events, Location and Tone (GDELT), that our project will be using. GDELT will be described in more detail in a following section. In the article "Did the Arab Spring Really Spark a Wave of Global Protests?" from the journal *Foreign Policy*, Kalev Leetaru used GDELT to examine the potential influence the Arab Spring had on other domestic conflicts. Monitoring the number of news articles that discussed protest activity, Leetaru found the Arab Spring was associated with a 25 percent increase in protest activity around the world [1]. Leetaru concluded his article saying that information provided by databases like GDELT provide "the first glimpse of what the future of data-driven diplomacy may look like, moving from anecdote to actuality" [1].

This same sentiment was expressed by James E. Yonamine in his Political Science dissertation, "A Nuanced Study of Political Conflict Using the Global Datasets of Events Location and Tone (GDELT) Dataset". With the information provided in the dataset, Yonamine analyzed the effects political violence had on the Tel Aviv Stock exchange, how much civil war affects

interstate war, and attempted to build an empirical model to generate predictions of future conflict in Afghanistan. For the first section of his dissertation, he found that the stock exchange did not significantly vary when attacks against Israel occurred, but many specific companies within the stock exchange do [2]. With regards to the second part of his dissertation, Yonamine found strong support for the idea that there is an increased likelihood of interstate conflict after an onset of domestic conflict by examining features of the GDELT dataset [2]. Using the dataset, he was able to create a model capable of providing forecasts of conflict at a significant rate. He describes his model as imperfect, but says that his work shows the feasible use of media data to predict societal behavior [2].

PROPOSED WORK

Data Reduction

Attribute Subset Selection

The original datasets from GDELT project database that is chronological with the additions of current news occurring every 15 minutes.. The first step for data reduction would be to reduce the number of attributes since the elimination of actual objects is not ideal for this project. For example, the geographical information and link of the news would be removed from the original database. This task is archived with Google BigQuery.

Histogram

Histogram is used to visualize the the keywords of the results from Google BigQuery. Depending on the frequency of the keywords we would construct the queries of Google BigQuery with the keywords that we are interested for analysis further.

DATA SET

The information being gathered for this project comes from the Global Database of Events, Language, and Tone (GDELT). The GDELT Project describes itself as a global database of society. GDELT monitors broadcast, print, and web news from every country in over 100 languages, capturing what media outlets all over the world are saying. This database collects data every 15 minutes, ensuring that our team will have plenty of data points to work with. At its current size in April 2017, the total uncompressed size of GDELT is over 340GB, which makes BigQuery so important for this project. Within the dataset, there are 46 different features encoded in the CAMEO format which fall under the following categories: EventID and Date Attributes, Actor Attributes, Event Action Attributes, Event Geography Attributes, and Data Management Fields.

The features in the dataset are many, but some of them are unnecessary for our project. In addition, there are repeated fields which are for the backward compatible. Thus, we would need to select those might useful in our project.

- Date
- Actions
- Average Tone (using Goldstein Scale)
- Global Content Analysis Measures(for sentiment analysis)
- Location (country code, latitude, altitude)
- Actors

EVALUATION METHODS

We will be collecting a series of features from GDELT to evaluate in order to reach our project's objective. Data relating to the tone of a piece, key actors of articles, geographical and temporal information, potential impact, and the types of articles will be collected. To evaluate the patterns that news has throughout time, we will look at the variation and standard deviation these features have over specified time periods. To measure how media has changed throughout time, we will look at

the variation and derivatives these features have over the course of many years. If there are any significant changes, additional background research will be conducted to see if there are any plausible events around that time that can be used to explain these differences. For domestic and international conflict and their relationship with the media, we will look at the stability of the average society, quantity of positive and negative news, and the stability factors over time. In this project we will consider the tone of the media which is analyzed by google with some power algorithms as the stability measurement. This data will be used to determine whether or not these features can be used to predict conflict in the future by comparing the data patterns to the expected pattern, such as negative or positive correlation and/or tendency.

TOOLS

This section provides the major tools we will be using throughout the project.

Google BigQuery

· Google BigQuery is online data warehouse and offers scalability and Flexibility for data processing. It allows the project team to leverage Google's hardware infrastructure to complete some of the data cleaning, data preprocessing and data storage.

Gephi

· Network Visualization can be created with Gephi

GDELT Analysis Service

- Tools used to visualize and explore GDELT
- Python with scientific packages are used for general computation

MILESTONES

The milestones this project will follow are outlined below.

Learn Doc and Setup Tools

· At this step we setup the environment for the next steps, e.g. coding python to load the sample data after we will learn the documentation of the data sets and how to use the tools.

Download and Load Data Sets

· At this step we can run the application resulting in the extraction of the most ideal data from Google BigQuery to our local machines.

Manipulate Data

· At this step we can reformat the data as expected. We as well should be able to extract the data in desired column and output to desire format for other tools.

Plot Graphs/Create Model

· This is the final step and we should be able to plot effective graphs as described in EVALUATION section in addition to having a reliable model for event prediction based off the aforementioned data.

Results So Far

Factors for UK leaving EU

So for, we have queried the most frequent themes of the global media about United Kingdom on from 03/01/2016 to 03/01/2017 with Google BigQuery. The theme are sorted and plotted with histogram (Figure 1 and Table 1). We used this data to analyze and find out the factor(s) of the event that UK leaves EU. First of all, we assume the decision of this event is due to the negative of the society of UK about being part of EU, so that we come up with a model that the stability of the society is negative proportional to the time. In addition, the model should trend to be positive after the decision of UK. We would select the few of the most frequent themes and query the data with the theme individually. We plot the data and try to compare them to our model.

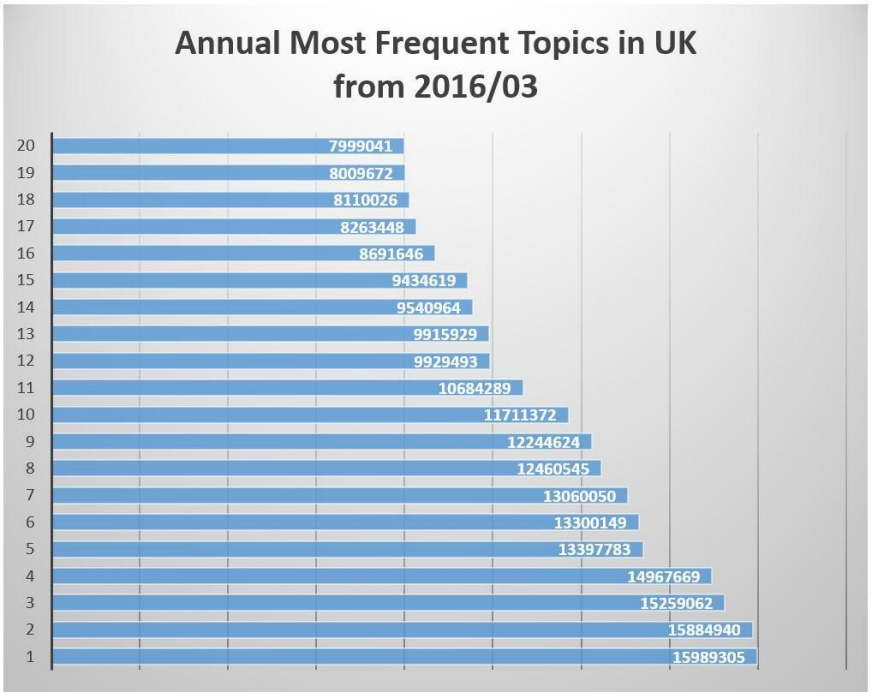

Figure 1

Table 1 matches the index on Figure 1 to the themes

Index	Themes
1	LEADER
2	USPEC_POLITICS_GENERAL1
3	EPU_ECONOMY_HISTORIC
4	GENERAL_GOVERNMENT
5	CRISISLEX_CRISISLEXREC
6	CRISISLEX_C07_SAFETY
7	USPEC_POLICY1
8	MANMADE_DISASTER_IMPLIED
9	EPU_POLICY_GOVERNMENT
10	UNGP_FORESTS_RIVERS_OCEANS
11	EDUCATION
12	TAX_FNCACT_PRESIDENT
13	WB_2432_FRAGILITY_CONFLICT_AND_VIOLENCE
14	WB_696_PUBLIC_SECTOR_MANAGEMENT
15	TAX_FNCACT_MINISTER
16	GENERAL_HEALTH
17	TAX_FNCACT_POLICE
18	TAX_ECON_PRICE
19	SECURITY_SERVICES

20	MEDIA_MSM

REFERENCES

[1] Leetaru, Kalev (May 29, 2014). "Did the Arab Spring Really Spark a Wave of Global Protests? The world may look like it's roiling now, but the 1980s were far worse.". Foreign Policy. Retrieved June 2, 2014.

[2] Yonamine, James E. "A nuanced study of political conflict using the Global Datasets of Events Location and Tone (GDELT) dataset". Retrieved June 2, 2014.