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

ABSTRACT

The ultimate goal of this project was to predict the outcomes of civil conflicts using the model we created and by doing that we were trying to determine the influence of the media on the outcome of the civil conflicts.

Using the datasets from the GDELT Project, we were able to create a model that classified conflict and non-conflict countries with ~73% accuracy.

Introduction

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.

RELATED WORKS

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 provides "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].

DATASET

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 The Attributes we used in the project are listed below:

1. *GDELT*
   a. *SQLDate*
   b. *Actor1CountryCode*
   c. *EventRootCode*
2. *GKG (Global Knowledge Graph)*
   a. *Date*
   b. *V2Locations*
   c. *V2Themes*
   d. *V2Tone*

TECHNIQUES APPLIED

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

There are over hundreds of themes in GKG datasets of the Gdelt project. Histogram is used to visualize the keywords of the results from Google BigQuery and reduces the amount of themes. 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.

Classification

In the classification, we would analyze different attributes statistically, and applying the results into the training matrix of the machine learning algorithm with supervisor approach or to the process of constructing the training set.

KEY RESULTS

First Approach

In the first approach, we 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 A1 and Table A1 in Appendix A). We used this data to analyze and find out the factor(s) of the event that the UK withdrew from the European Union. 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.

As we can see, the green line for topics safety and crisis fluctuated at the left most side; these two topics might be the main factor for the predict event of UK. The average tone in the same time period about same topics of the EU countries, Germany and France, were queried and plotted and are compared to that of UK. However, the plot of German and France show the similar pattern. The tones of the media were shifting from day to day at the beginning of 2015 through the mid of 2015. We concluded that the average tones could not be helpful for predict the specific event in every country by comparing the tones. So we moved forward to the tone approach and classification for figuring out how the tones of the media are related to the civil conflicts and predicting whether the conflict would happen within the period when the attributes have the similar pattern.

Figure 1 show the average tone of the top 3 themes in UK

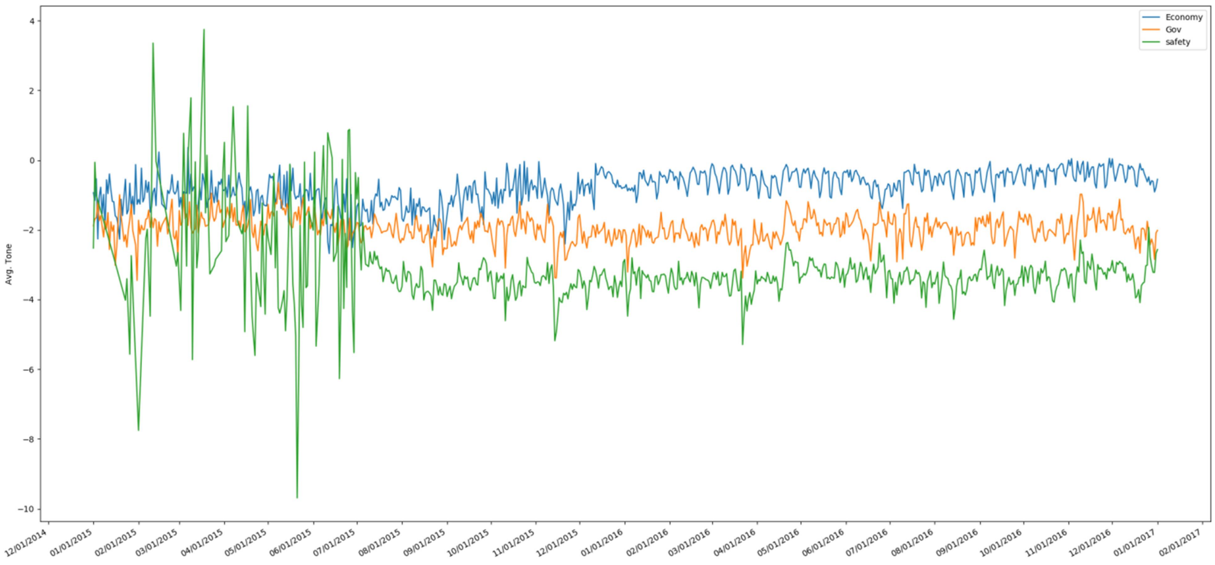
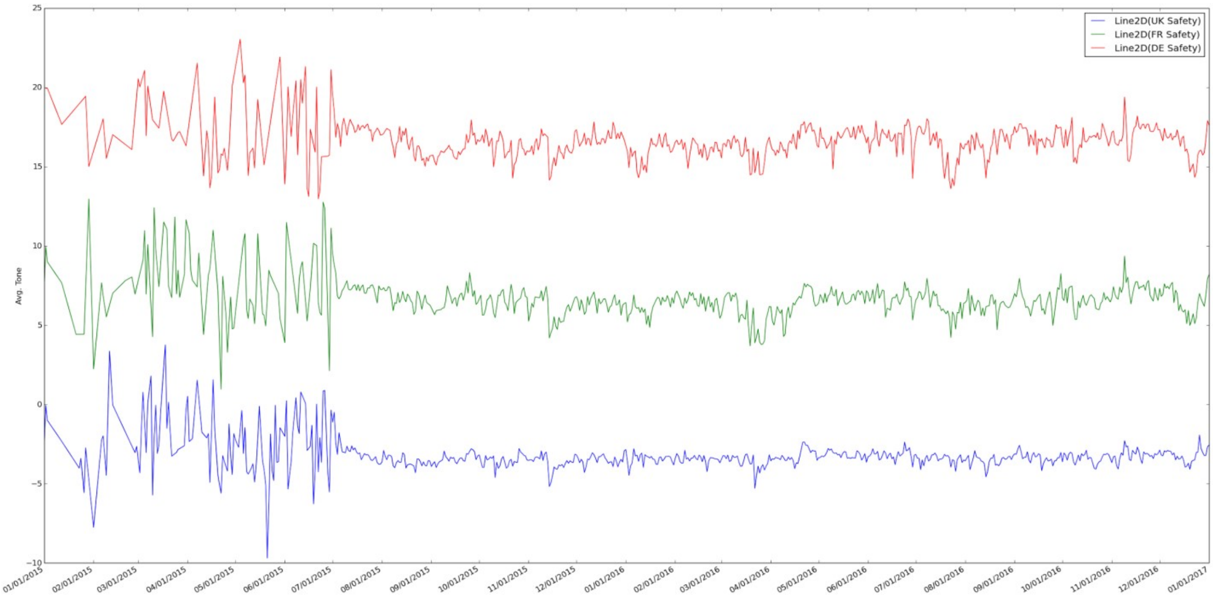


Figure 2 show the average tone of the U.K., Germany, and France about the topics safety and crisis from 01/01/2015 through 01/01/2017



Tone Approach

In order to examine whether the dataset of Gdelt project is helpful and useful to predict the stability of the society, we analyze the different attributes of GKG and EVENT databases. This approach focuses on the analysis of tone of the media. In this approach we analyzed the average tones from all source about the topics, safety and crisis, in peaceful countries and the countries that was in civil conflicts or unstable society. Because the Limitation of the free Google BigQuery account, we weren't able to query the data successfully. Alternatively, we use the Google Analytics Service to get the average tone of the countries with keywords, safety and crisis, and the specific location.

The training set in the Table 1 would use the marks as following. The marks used for difference are Small(S), Large (L), and Medium (M). The marks used for outliers are Plenty (P), None (N), and Few (F). The marks used for median position are Low (L), Mid (M), and High (H). The last mark set is Peace (P) and Conflict(C).

After the statistical analysis, we can see that the peaceful countries on the left have small interquartile ranges (IQR), meanwhile there are lots of outlier, except Canada. On the other hand the countries in unstable events have larger IQRs and fewer outliers. With this information, we can create the training set as Table 1. Since the number of dataset is limited, we can simplify the process of creating the training set. After the number of analysis increases, the table can be filled out in detail in future development. However, the extra classification would be helpful to increase the success rate of the prediction.

Classification

The classification model we chose to create was based around the event attribute in the GDELT dataset. The GDELT dataset uses twenty different base codes, shown in TableA2, used to categorize the events in a news report. Using Google BigQuery, we were able to collect the number of times each event was mentioned in a news report about a country the year before their conflict.

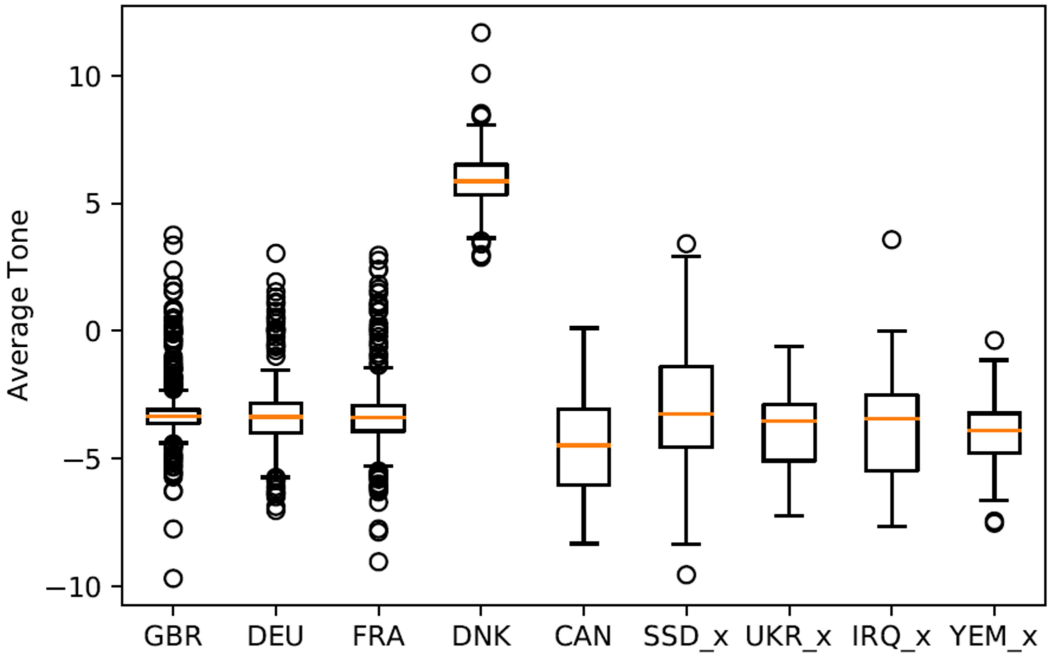


Figure 3 - shows the average tones statistics of the peaceful countries and the countries in civil conflicts or unstable events

Table1- shows the training set

Country Code	ΔIQR	Δ(max.-min.)	Outliers	Median Pos	Peace In Past
GBR	S	S	P	M	P
DEU	S	S	P	M	P
FRA	S	S	P	M	P
DNK	S	S	P	M	P
CAN	L	L	N	M	P
SSD	L	L	F	L	C
UKR	L	M	N	H	C
IRQ	L	M	F	H	C
YEM	L	M	F	M	C

Once the necessary data was extracted for each of the countries shown in TableA3, we began constructing the model. The first step was dividing the data into a training set and testing set. The training set was constructed with 12 peaceful and 12 conflict countries, while the testing set was made up of 6 peaceful and 6 conflict countries. We initially believed that there may be significant difference between which events were represented the most in conflict states and peaceful countries. this is not the case, since there are event categories that typically dominate the news cycle in any country e.g. consult, make public statement, while some events make small footprints in the media landscape, but are expected to still be significantly different between peaceful and conflict regions like engage in unconvential mass violence.

To make this data more useful, we began by calculating the normalized difference between the average event values for conflict countries and peaceful countries in the training data. This allows events with far lower counts to produce data that can compare with the rest of the data. This array shows which events are proportionately being reported on more in conflict areas, and vice-versa.

These first steps of the model set up the foundation of the rest of the algorithm. Classification occurs many matrix transformations later. Two simplistic classification techniques were used: correlation and normalized difference.

Two different correlation measures are created for the first technique, one that measures the correlation of the testing piece of data and the average values of the top conflict events, while the other does the same, but with peaceful events. The country being tested would then be classified based on the stronger correlation.

For the second classification technique, a normalized difference array was calculated from the average top event values and the test piece of data. If the sum of this array was positive, the country would be classified as conflict, otherwise it would be labeled as peaceful.

Model Evaluation

Confusion matrices were created after the classification process to evaluate each model. Using values from this matrix, we were able to calculate the accuracy, sensitivity, specificity, and precision of the model. Figure 4 shows the outcome of this below.

3445	671
2555	5329

	Correlation Model	Normalized Difference Model
Accuracy	0.5337	0.7312
Sensitivity	0.5547	0.5742
Specificity	0.5127	0.8882
Precision	0.5323	0.8370

Figure 4 – results of the classification

By every metric, the normalized difference classification wins. The accuracy of the correlation technique is only slightly better than guessing, but the difference classification was accurate 73% of the time.

APPLICATION

The results of our classification model show that it may be possible to predict if a country was going to engage in civil conflict in the next year. That could help local governments identify and address public issues quicker or it could be and indicator of when foreign aid should be sent out.

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.

Appendix A

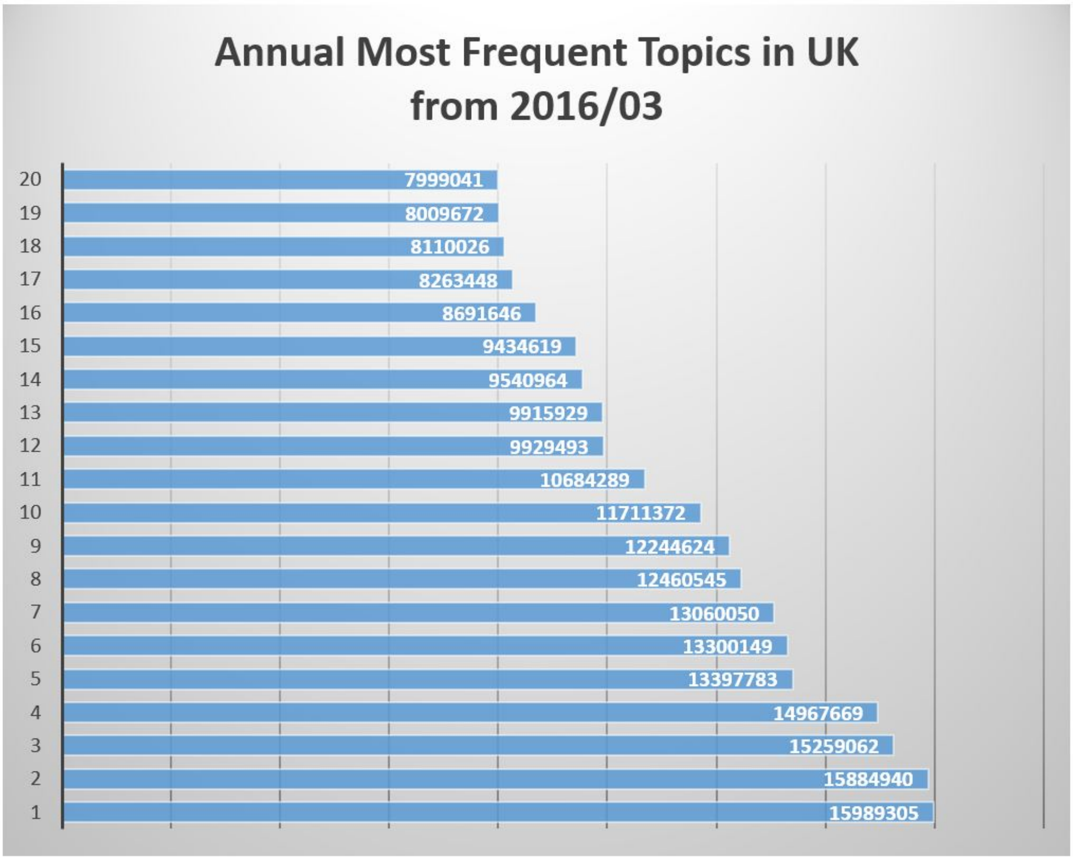


Figure A1 - shows the 20 most-frequent topics in 2015-01-01 through 2017-01-01 in the U.K.

Table A1 - matches the index to the number of bin in Figure A1.

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

Table A2 - Code Labels for Event Categories

Code	Event
1	Make Public Statement
2	Appeal
3	Express Intent to Cooperate
4	Consult
5	Engage in Diplomatic Cooperation
6	Engage in Material Cooperation
7	Provide Aid
8	Yield
9	Investigate
10	Damand
11	Disapprove
12	Reject
13	Threaten
14	Protest
15	Exhibit Military Posture
16	Reduce Relations
17	Coerce
18	Assault
19	Fight
20	Engage in Unconventional Mass Violence

Table A3: https://github.com/wwcao/DataMiningProject/blob/master/code_data/DataMiningData.xlsx