Blue Team

ISE 599 Final Project

Bringing Advance Clarity to Political Crises

Sean Eskew

Luke Haravitch

3 December, 2019

GitHub repository: https://github.com/lharavit/GDELTProject/blob/master/README.md

Link to Final Presentation

Executive Summary

Can a machine model be trained to beat human forecasters and accurately predict the intensity and causes of future political crises by using a database of near-real-time media reports?

Our research seeks to identify candidate methods for such a model by assessing their accuracy, simplicity, and explainability- features desired by those who would plan or assess necessary interventions to crises. Contingency planners will be particularly interested in the ability to gain useful information about the predicted nature and intensity of a current or future conflict. Policy makers want to know about the potential and realized effectiveness of historic, current, and/or future interventions. This prediction and feature selection problem is difficult because we are dealing with event data that are often sparse (low incidence rates) and may not capture the true event occurrences, but rather the media reactions to them. The useful signal in our data is therefore masked and requires thoughtful analysis to uncover. We explore linear, ensemble, and time-series prediction models and compare our performance against human forecasters from a recent geo-political forecasting competition. We recommend a Bayesian Ridge regression model because of its inherent feature reduction, updating from available data, and demonstrated accuracy. This model beat 61% of human forecasters and is capable of automatic predictions from a near-real-time data source. The model maintains explainability and can be used to inform causal assessments of future violence, but human experts are required to make this model's predictions most useful.

1. Proposed Approach

A. **Background** Societal factors; including political sentiment, economic status, religious views, and environmental conditions, likely have impacts on political based violence events. These violence events often require both diplomatic and security force intervention from local and outside entities. Both types of intervention would be improved with additional warning time and a clearer understanding of the factors causing unrest. If journalism is a reflection of current society, then within media data there exists information on the societal factors that cause this violence. If this data can be exploited to provide such advance clarity; decision makers can then ensure they have the right forces on hand and use the right diplomatic approaches to alleviate the situation/ crisis.

B. **Problem Statement** Given a media database and a country/region of interest, we seek to predict future counts of political violence events and identify the factors most relevant to the prediction. The counts will inform security intervention requirements, while the important factors will aid policy responses.

Figure 1 - Visual depiction of project problem statement.

C. **Data Selection** We use the Global Datasets of Events, Language, and Tone (GDELT) as a source of independent variables that describe the state of affairs in a country of interest. GDELT uses media reports from all over the world to create a database of events, each with a time, place, and coded description of the event. GDELT updates every 15 minutes, so the methodologies we use are capable of providing near real time insights.

We use the The Armed Conflict Location & Event Data Project (ACLED) as our response data. ACLED records the dates, actors, types of violence, locations, and fatalities of all reported political violence and protest events across Africa, South Asia, Southeast Asia, the Middle East, Europe, and Latin America. Political violence and protest activity includes events that occur within civil wars and periods of instability, public demonstrations, and regime breakdown. Specifically, we use the human-curated, battle death counts which are recorded at the country level, so they are available for download about a week after the end of the target month.

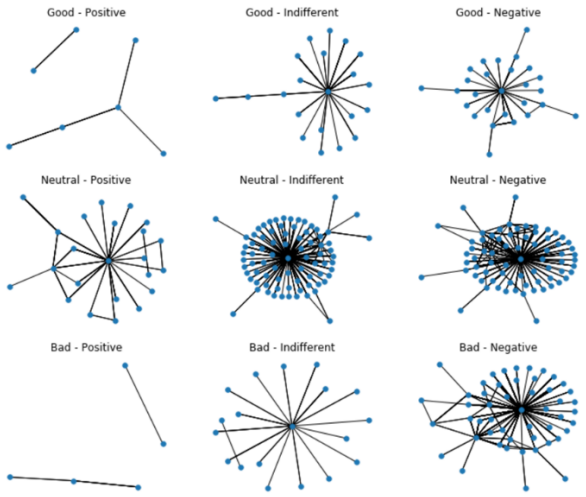
The methodologies we explore in this research use GDELT data as of the end of one month to predict the ACLED counts for the following month. For example, we use GDELT data (and historic ACLED data) up to 31 July, 2018 to predict the August, 2018 ACLED counts, which would be available in the first week of September, 2018.

D. **Data Pre-processing** In order to turn the GDELT data we pull from the web into a useable and relevant data set for this project, we aggregate events at the month level to produce a single value for each feature per month. We take two main approaches for these aggregations: network-based and event-based.

Network-based data approach: Each GDELT event has fields labelled as the actor and object (labelled “actor1” and “actor2,” respectively). In order to gain experience with the network analysis techniques we learned in class and to try a novel approach, we build networks of the actors involved in sub selections of each month’s events. We begin by segmenting a month’s events by Goldstein scale (a numeric score from -10 to 10 that captures the theoretical potential impact that an event will have on the stability of the country) and average tone (the average GDELT calculated tone of all documents containing one or more mentions of an event during the 15-minute update in which it was first recorded, scored from -100 to 100). We create nine segments as shown in Table 1. For each segment, we use each actor/object pair to build a network of the events, then calculate the number of nodes and average degree of each network. For each network measure, we now have nine new factors with which to build our predictive models. Figure 2 depicts the networks we create for each segment and the resulting predictor variables for a sample month of data from South Sudan.

	Goldstein Scale			
AvgTone		GS < -6 (Good)	-6 < GS < 6 (Neutral)	GS > 6 (Bad)
	AT > 3 (Positive)	Good – Positive	Neutral – Positive	Bad – Positive
	3 > AT > -3 (Indifferent)	Good – Indifferent	Neutral – Indifferent	Bad – Indifferent
	AT < -3 (Negative)	Good – Negative	Neutral – Negative	Bad – Negative

Table 1 - GDELT Data Network approach segments



	Network	Num_Nodes	Num_Edges	Avg_Degree
0	G-P	7	14	2.000000
1	G-I	21	67	3.190476
2	G-N	27	81	3.000000
3	N-P	23	101	4.391304
4	N-I	80	531	6.637500
5	N-N	71	611	8.605634
6	B-P	5	7	1.400000
7	B-I	19	35	1.842105
8	B-N	50	421	8.420000

Figure 2 - GDELT Data Network approach transformation

Event-based data approach: GDELT events are categorized by one of 20 different event codes that fall into four categories: engagement, action, posturing, and conflict as depicted in Figure 3. In this approach we simply tally the number of events having each code. We may also combine the tallies over each of the four categories for a coarser measure.

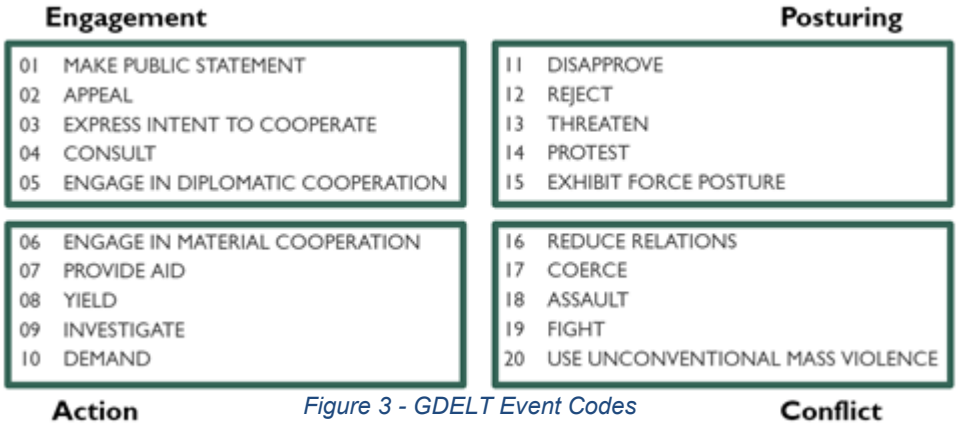


Figure 3 - GDELT Event Codes

Scaling: Continuing with our pre-processing, we performed feature scaling on our data through the use of the Scikit-Learn's StandardScaler function. This function centers and scales each individual feature around zero based on its mean and standard deviation. This approach, opposed to other scaler systems, allows for more robust dealings with outlying data as it is not forced into artificial boundaries (i.e. MinMaxScaler forcing all data between 0 and 1); as well as, easily incorporates future data into the set.

E. Data Modelling

Linear models: We use various multivariate linear, support vector, and tree models from the sklearn Python package to predict next month's ACLED battle death counts from current month GDELT Data as explained in the data pre-processing section. The simplest models use regression (Tree) along with some form of regularization (Lasso, Ridge, Bayesian Ridge) to penalize models that may overfit the data. Most models worked best when training on data from all countries (rather than building a specific model for each country).

Ensemble models: The sklearn package also contains models that are comprised of many sub-models. Random Forest (RF) builds many tree models each using a subset of the available variables and aggregates their predictions. Boosted methods (ADA and XGB) work similarly by increasing reliance on the most informative data from the training set as the model learns which data contribute most to model performance. We perform 10-fold cross validation on linear and ensemble models by training on 90% of the training data and testing on the remaining 10%. We repeat this 10 times for each 10% of test data slices to arrive at metrics of performance shown in Table 2. We refined each model by selecting features based on the contribution of each feature in previous model builds.

Model	MSE	R^2
RF	6344.424574	0.194006
Ada	6452.808832	0.214721
XGB	6850.908235	0.196834
Tree	9670.001076	-0.267918
Ridge	6512.153255	0.145782
Lasso	6365.155980	0.179827
SVR	9785.614184	-0.195104
NB	6160.908765	0.210215

Table 2 - Model performance of linear and ensemble models.

The Bayesian ridge regression model (labelled NB above) had the lowest mean square error (MSE).

Time-series models: In addition to the traditional multivariate statistical models above, we explored the efficacy of autoregressive models. Autoregressive integrated moving average (ARIMA) models used a univariate series (intensity levels for a particular country measured in ACLED monthly battle deaths) to forecast future values of the series. We used only one country at a time, tried both actual and log values, and did not experiment with any hierarchical (neither top-down or bottom-up) methods. We differenced these series twice to eliminate non-stationarity concerns. The performance measures shown in Table 3 are for the model that best fit the training data (2/1/2 ARIMA on actual values). Because of the suspected inter-relatedness of the GDELT and ACLED data, we also experimented with vector autoregressive (VAR) models that forecast multiple time series. The best VAR models also required differencing twice and worked best when using a small set (4 or 5) of exogenous variables. We tried using the battle death counts of other central-African countries, the network-based features (in this case the number of nodes for the following quadrants: G-P, G-I, B-N, N-P), and the event-based features (counts of events: 1, 9, 10, 15). For all time series models, we used functions from the statsmodels package.

In an effort to build models that produce insights and not just forecasts, we tried another time-series approach that tracks variable importance over time. We call this method ARIMA-Regression. We build a linear model with an early subset of our data (the best model used only the network-based node counts: G-P, G-I, B-N, N-P) with the ACLED count of the next month as the response variable (so if the features are from January, the response would be the February count as in the other statistical models above). Then we slide our subsetting window forward to include one more month of data and remove the oldest month of data. For each of these sub-setted models, we record the coefficients of our regression model and build a time series for each of them. Because we only change one data point in each model, the coefficients do not change as rapidly and can be used to show when variables change in importance over time. To make forecasts with this method, we forecast one month into the future and use the forecasted coefficients with the current month's GDELT data to predict the next month's ACLED count. Overall, this method is a relatively poor forecaster (it happens to perform well on the August 2018 questions, but that was an anomaly). The best ARIMA models for this method were also 2/1/2 models using the actual count values (not log of counts).

None of the time series methods we tried performed very well on these series because they are dominated by their stochastic components. Therefore, we do not recommend using any time-series methods to produce accurate forecasts for these problems, but they may be helpful to show how far back the predictor variables are useful (usually two months) and how variable importance changes over time (the ARIMA-Regression method).

Naive models: In order to provide relevant baselines of model performance, we experimented with some very simple naive models. The “last month” method simply uses the most recent available count as the forecast for the next time period. For example, we would assume that a country’s battle death counts for August, 2018 would be the same as their July counts. The “average” method forecasts that next month’s count will equal an average of the previous year’s counts. For example, we would assume that a country’s battle death counts for August, 2018 would equal the average of the monthly counts from August, 2017 through July, 2018. The “random” method assigns equal probability to each outcome. For example, we would assign a 20% probability to each of the five bins in the forecasting competition questions.

Experimental Results

A. **Metrics** In order to demonstrate the applicability of our approach to real problems, we compare our forecast performance (on intensity of future violence) to several hundred human forecasters who participated in a forecasting competition in 2018. We evaluate our performance using ordered Brier score (a measure of mean squared error) of a probabilistic forecast. There were four questions in the tournament that asked about ACLED counts in Central African countries in August 2018. An example is depicted in Figure 4.

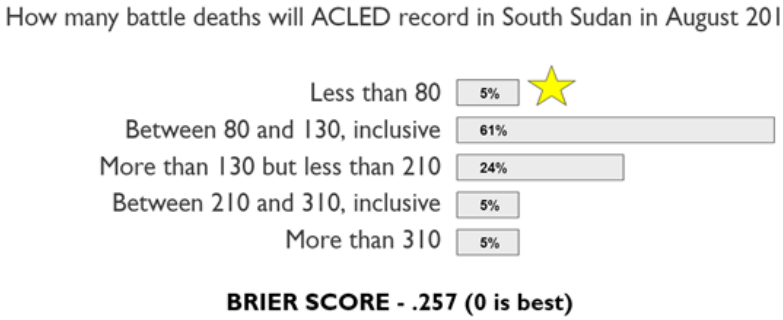


Figure 4 - An example forecasting competition question. The gray bars depict a respondents’ predicted probabilities that the result will fall in each of the five bins. The correct answer was 64 (depicted by the star). The respondent would have received a Brier score of 0.257 for this response.

We used data available up to the end of July 2018 to train and select models and compared their Brier scores to the roughly 400 humans who made forecasts on each question at or before 31 July 2018. The results are displayed in the table below along with the average percent of human forecasters that the model outperformed. Average Brier score is calculated as the average ordinal Brier score across the four questions (lower Brier score is better). Percent Forecasters beat is calculated as the average of the percent of human forecasters that were outperformed by the model over all four questions (higher percentage is better).

B. **Model Evaluation** We evaluate each model by its accuracy in predicting intensity (measured in number of ACLED recorded battle deaths), its simplicity, and its explainability. Accuracy results are shown and discussed in the previous section. Simplicity is handled during model selection by using measures like AIC to prune models to ensure they do not overfit the data and is included in our explainability assessment. For explainability, we evaluated models on our ability to make a coherent sentence about how the predictor variables explain the response variable and how the model works. If the model workings are easy to explain (naive and most linear models) or if the relationships are easy to explain, we call the model explainable. If both are easy to explain, we call it very explainable. It is another issue altogether to actually verify that these models' explanatory sentences were accurate. Verifying the causes of political crises is best performed by human curators, not by model evaluation metrics.

3. **Discussion** The results of our model evaluations are shown in Table 3 below.

Method		Average Brier Score	% Forecasters Beat	Explainability
TS	ARIMA-Regression	0.096	68	Yes
L	Ridge	0.108	67	Yes
L	Bayes Ridge	0.139	61	Yes
L	SVR	0.238	46	No
N	Random	0.247	45	No
L	Lasso	0.235	44	Yes
E	ADA	0.259	38	No
E	XGB	0.274	37	No
L	Tree	0.328	36	Yes
E	RF	0.358	26	No
TS	ARIMA	0.498	13	Yes
TS	VAR	0.588	9	No
N	Last month	0.563	7	Very
N	Average	0.563	7	Very

Table 3 - Model performance. The category of each model is listed on the left of the chart (L for linear models, E for ensembles, TS for time series, and N for naive).

Table 3 is ordered by the average percent of humans each model out-performed. Although the ARIMA-Regression model performed exceptionally well on these specific competition questions (August 2018), it did not generally produce accurate forecasts. The ridge models, however, produced consistently accurate forecasts during cross-validation (see Table 2) and select the most consistently informative features thus enhancing their relevance for policy makers. The time series and naive models did not fare well on these test forecasts because these battle death count series are governed by their residuals. They are highly variable and previous values are often not helpful. The ensemble models did not perform very well and are generally difficult to explain to policy makers (both in terms of how the models work and which variables are most informative).

Conclusion

A. **Lessons Learned**

1. Linear models are the best for predicting intensity levels and identifying important features that may help policy makers decide on appropriate interventions.
2. Time series of intensity levels are governed by residuals rendering auto-regressive methods ineffective.
3. Models trained on data from multiple countries out-performed those trained on individual countries because Central African countries tend to influence one another's political dynamics.
4. Shifts in political dynamics may be detected by how linear model coefficients (or feature selections) change over time.

B. **Future Work** More research must be done on the thresholds of battle death (or other count) levels that result in specific interventions. This way, we may turn the above forecasting problem into a classification problem: given the current (and recent) conditions (gleaned from analysis of GDELT media data), what is the appropriate intervention response? The same experts that may help inform this type of solution should also be able to help verify the causal identification aspect of the models. Also, research should be conducted into how best to combine the results of several models (and even naive and expert human forecasters too) to arrive at more accurate and contextual forecasts. For example, we may want to use an average of some time series and some linear models to create a more accurate forecast of intensity. And those models may be improved by what we learn from the feature selection and data weighting of the ensemble models. The change in regression coefficients over time and across countries may help inform policy makers to changing political dynamics. This approach allow for contributions from multiple models that on their own were not capable of producing accurate or contextually appropriate forecasts.