Case Study Title: Bringing Advance Clarity to Political Crises

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?

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.

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.

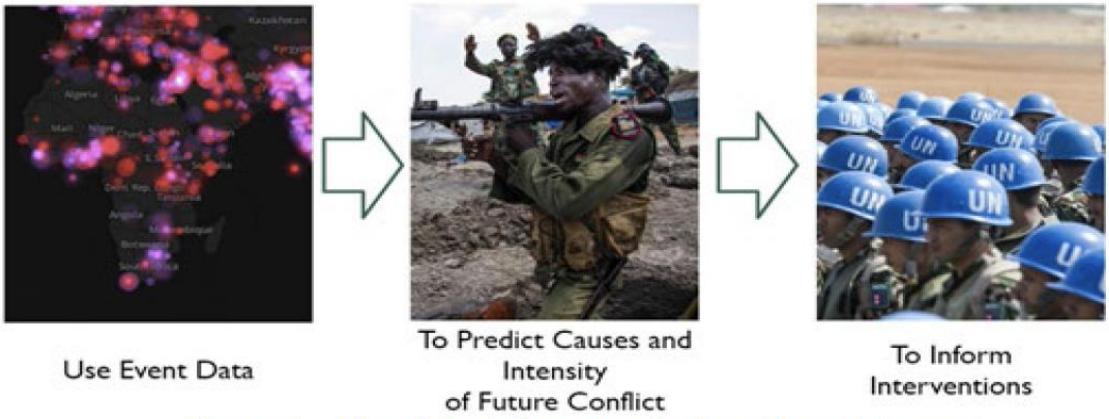


Figure 1 - Visual depiction of project problem statement.

Data Selection (GDELT and ACLED):

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.

Data Pre-Processing:

In order to turn the GDELT data we pull from the web into a useable and relevant data set for this case study, 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.

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.

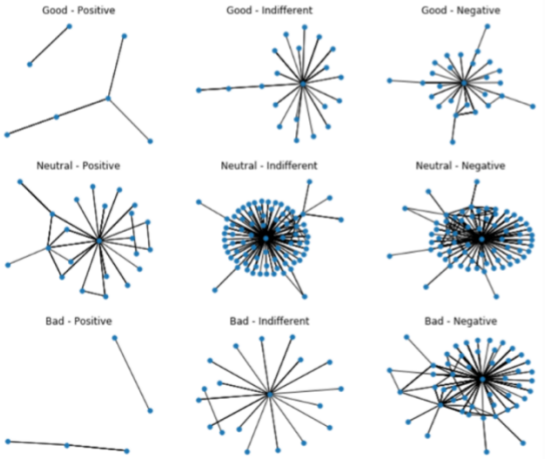
Scaling: Continuing with our pre-processing, we performed feature scaling on our data through the use of the Scikit-Learn's StandardScaler function. This function centers and scales each individual feature around zero based on its mean and standard deviation. This approach, opposed to other scaler systems, 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.

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.

		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

Engagement

01 MAKE PUBLIC STATEMENT
02 APPEAL
03 EXPRESS INTENT TO COOPERATE
04 CONSULT
05 ENGAGE IN DIPLOMATIC COOPERATION

06 ENGAGE IN MATERIAL COOPERATION
07 PROVIDE AID
08 YIELD
09 INVESTIGATE
10 DEMAND

Action

Posturing

11 DISAPPROVE
12 REJECT
13 THREATEN
14 PROTEST
15 EXHIBIT FORCE POSTURE

16 REDUCE RELATIONS
17 COERCE
18 ASSAULT
19 FIGHT
20 USE UNCONVENTIONAL MASS VIOLENCE

Conflict

Figure 3 - GDELT Event Codes

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.

Questions

i) After data pre-preprocessing, data modeling is the next applicable step. Describe in detail how you could use linear models to do data modeling for this case study.

ii) Does it make sense to use ensemble models for doing data modeling? Why or why not?

iii) What about time-series models? Would you want to use auto-regressive models for this problem? Why or why not?

iv) Suppose you tried time series models but they did not work well (indeed, this is what we found in this project). What could be some reasons?

v) We mentioned the ordered Brier score at the end of the case study. Look up this metric and discuss in detail why it is the right metric for this problem. What are its possible shortcomings?

vi) Discuss possibilities for future work based on some of the lessons we enumerated.

vii) Although this project was about geopolitical events, could a similar approach be used for 'other kinds' of events? What other kinds, and why (or why not)?