Identifying Threats: Using Machine Learning in International Relations

Azusa Katagiri and Eric Min
Stanford University

Motivation

What role do threats really play in international relations (IR)? IR scholars have long highlighted the importance of threats in explaining interstate behavior, especially in crises.

Many theoretical, historical, and formal studies have analyzed threats at a small-*n* and unsystematic manner, leaving most of these prominent theories untested. (See Schelling 1960; Jervis 1976; Fearon 1994; Sartori 2002; Stein 2013.)

Our project attempts to use machine learning techniques to empirically identify, classify, and quantify (perceived) threats. Doing so will finally allow for rigorous, systematic, and time-series testing of long-standing theories on the place of threat-making in international politics.

General Approach

Our focus is on threats perceived by elite policymakers, as opposed to the public and/or media. We also seek fine-grained data that allows intra-event analysis.

To that end, we find data of intra-elite communications and scrape them using text-based methods. We then use machine learning techniques to classify all documents as either expressing perceived threat or not. Lastly, we use time-series analysis to assess the validity of this new data.

Data

We scrape the *Foreign Relations of the United States 1945-1980* (FRUS) collections, which are a compilation of memoranda, telegrams, and reports sent within the federal government and with American embassies abroad.

To ensure tractable and consistent results, we examine all volumes on the Soviet Union and the Eastern Bloc between 1952 and 1977—the height of the Cold War.

This provides a total of 7,842 documents and 29,607 unique stemmed tokens.

The document-term matrix was reweighted using tf-idf. Sparse terms not present in at least 80% of documents were dropped, leaving 312 tokens of interest.

Sample Classification

664 memos were randomly drawn and hand-classified for training and testing.

"Threat" is an elusive concept to measure, much less perceive. We classified memos as expressing threat if they conveyed a real and/or imminent potential for either (1) armed hostilities involving the United States and/or (2) a substantial risk in the bipolar balance of power, usually via allied nations. An example of a positive case is below:

> *We are alive to problems presented by GDR harassment at crossing point and on autobahn. Latter particularly disturbing since touches Allied vital interest, and manner dealing with possible pattern such encroachments being considered in Quadripartite contingency planning. We shall look to you and our Mission in Berlin to detect any such pattern as it begins to emerge. We shall also welcome any suggestions for particular measures to forestall or retaliate for harassment.*

The sampled data had the following distribution:

Perceived	Not Perceived
92 (13.9%)	572 (86.1%)

Classification Model Selection

Of seven techniques optimized and evaluated through 10-fold cross validation, *random forests* (RF) gave the best overall performance. Lasso was the next best. With random forests, performance further improved via down-sampling (equalizing the number of positive and negative cases in the training set).

Model Diagnostics

Model	MCC	Kappa	F_1	Acc.	Spec.	Sens.	AUC
RF Down-Sampled	0.367	0.367	0.908	0.843	0.478	0.902	0.815
RF	0.116	0.059	0.925	0.861	0.043	0.993	0.728
Lasso	0.280	0.279	0.907	0.837	0.348	0.916	0.724
kNN (k=9)	-0.063	-0.043	0.911	0.837	0.000	0.972	0.717
SVM	0.207	0.126	0.928	0.867	0.087	0.993	0.700
PLS	0.273	0.259	0.919	0.855	0.261	0.951	0.779
Bagged	0.275	0.141	0.932	0.873	0.087	1.000	0.457

Table 1 : Summary statistics for seven classification techniques.

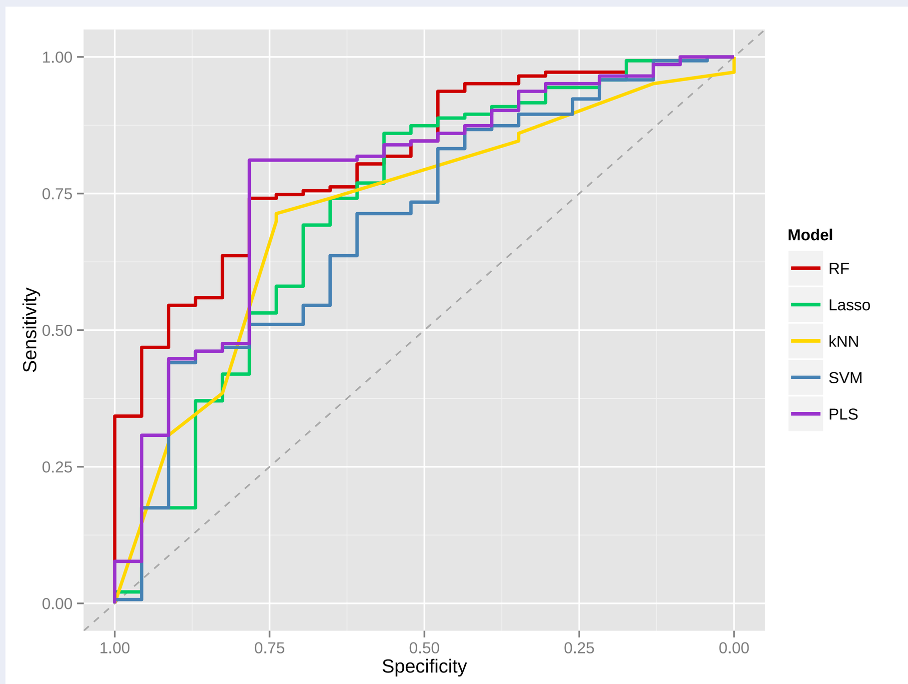


Figure 1 : ROC plots for five best classification models.

RF Down-Sampled	Lasso
soviet	call
immedi	arrang
berlin	measur
war	communist
strong	defens
seem	serious
militari	side
danger	concern
told	war
call	remain
east	danger
concern	determin

Table 2 : 12 most informative tokens.

Predicted Data

Predicted data aligns quite well with historical events between the US and USSR.

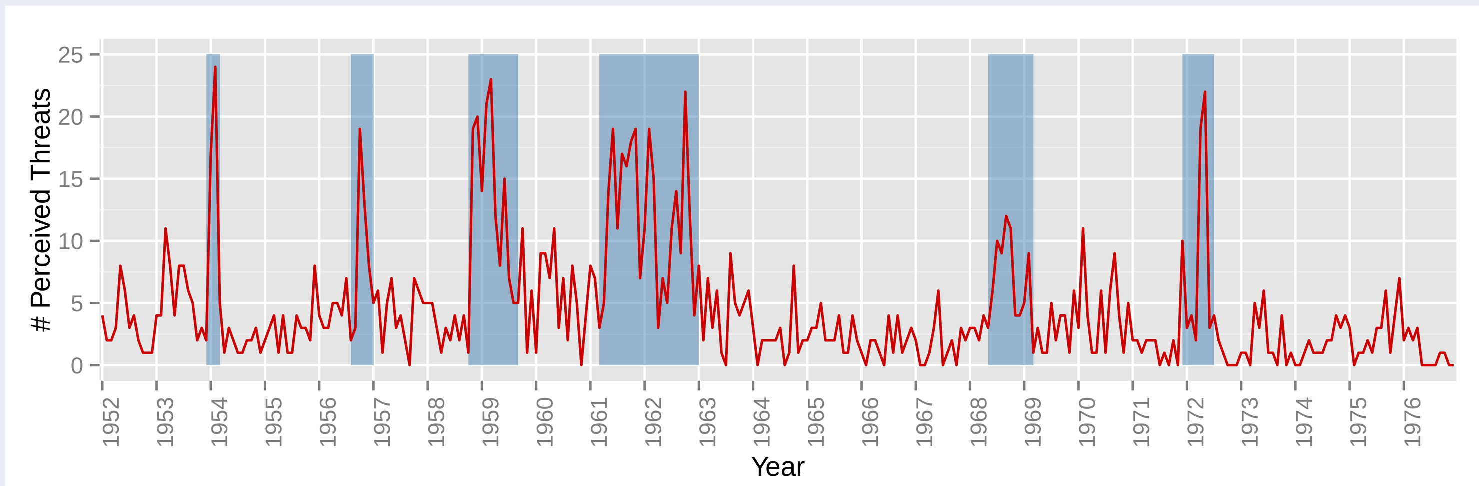


Figure 2 : Predicted monthly data on FRUS using down-sampled random forest model. Highlighted spikes align with following events, respectively: (1) the Berlin Conference of 1954; (2) the Hungarian Revolution of 1956; (3) Khrushchev's ultimatum in November 1958; (4) the Berlin Crisis of 1961 to the Cuban Missile crisis in 1962; (5) the Prague Spring, and the Warsaw Pact Invasion of Czechoslovakia in August 1968; and (6) the Easter Offensive in Vietnam, and SALT negotiations in April-May 1972.

Evaluating the Data Using COPDAB

In order to assess our new threat perception data's utility, we examine the Conflict and Peace Data Bank (COPDAB; see Azar 1980), which includes a directional monthly measure of cooperative/conflictual events between the United States and the Soviet Union. The higher the measure, the more conflict. We see whether our threat perception data can help explain variation in hostile events between these two countries from 1952 to 1977, or vice versa.

Time-Series Model Selection

According to Augmented Dickey-Fuller and Phillips-Perron tests, the data satisfies the stationary assumption, allowing for use of a simple vector auto-regressive (VAR) model. Several tests and measures indicated $p = 2$ to be the most appropriate lag. The main models are thus specified using the generated threat perception (TP) and COPDAB (CD) data:

$$CD_{a\to b,t} = \alpha_0 + \alpha_1 TP_{t-1} + \alpha_2 TP_{t-2} + \alpha_3 CD_{a\to b,t-1} + \alpha_4 CD_{a\to b,t-2} + \alpha_5 t + \varepsilon_t$$

where $a \to b$ may be either actions the US takes against the USSR ($US \to USSR$), or the USSR against the US ($USSR \to US$), depending on the model.

VAR Results

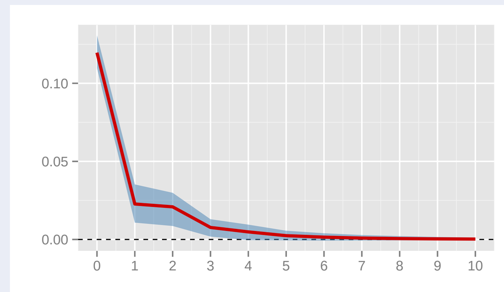
$CD_{a\to b,t}$	TP_{t-1}	TP_{t-2}	$CD_{a\to b,t-1}$	$CD_{a\to b,t-2}$	Const.	Trend
USSR → US (RF)	64.07**	15.86	0.27***	0.30***	11.52	-0.07
US → USSR (RF)	5.61	72.64***	0.27***	0.04	22.50***	-0.18***
USSR → US (L)	146.13***	47.96	0.27***	0.31***	12.2	-0.07*
US → USSR (L)	-25.45	125.07***	0.28***	0.06	28.14***	-0.18***

RF = random forest; L = lasso. Significance at 0.1 (*), 0.05 (**), and 0.01 (***) levels.

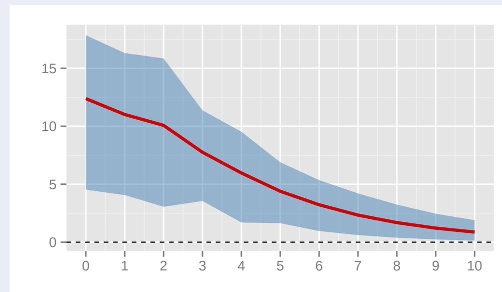
Similar models using TP as the DV only indicated strong autocorrelation within TP.

Impulse Response Functions

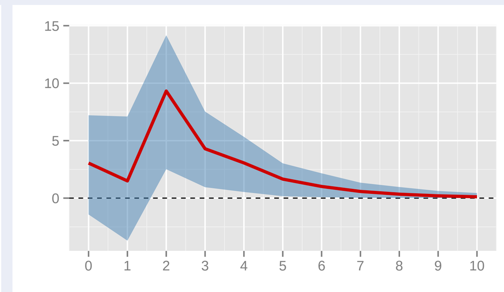
A one standard deviation shock in perceived threat has the following effects:



(a) Threats



(b) USSR → US



(c) US → USSR

Conclusions

Using a novel source of data, this project indicates that policymakers aptly foresee threats as embodied by conflicts and antagonistic events. *Threat perception precedes actual aggressive behavior.* While this may seem intuitive, such a process has never (to our knowledge) been empirically evaluated using quantitative methods. It also has substantial implications on IR theory, where threats are operationalized through explicit threat-making and major events.

The classifiers could be improved. Future steps include employing automated text segmentation processes (e.g. TextTiling) to create more consistent data.

Contact: **azusak@stanford.edu** and **ericmin@stanford.edu**

Eric Min gratefully acknowledges the financial support of the National Science Foundation's Graduate Research Fellowship #DGE-114747.