Identifying Threats in Diplomatic Correspondence

CS 229 Final Project

Azusa Katagiri and Eric Min

December 11, 2013

1 Introduction

Since the inception of the field, International Relations (hereafter IR) scholars have placed substantial emphasis on the importance of threats and credibility in explaining interstate behavior—especially in the midst of crises that may precipitate armed conflict. Countless studies of theoretical, historical, and formal flavors have been devoted to the study of threat perception (Jervis 1976; Sartori 2002; Stein 2013). Empirical studies that actually measure and evaluate perceived threat, on the other hand, have been lacking. The paucity of micro-level data has hampered analysis on the use and effectiveness of threats. IR academics have long recognized this shortcoming but been unable to address it. Some recent studies have used laboratory experiments to assess threat perception within the general public (Rousseau and Garcia-Retamero 2007), but much of this research fails to explain actual policy decisions by governmental elites. To better understand threat perception, the IR field requires new empirical approaches that capture this elusive but crucial concept in a large-scale and systematic way. Machine learning may resolve this.

This project seeks to systematically identify, quantify, and predict perceived threats in international diplomacy. Not only does such data finally allow for empirical tests of long-standing academic theories of credible threat-making, but may also help to forecast the credibility of threats made in current and future correspondences beyond the international/political realm.

2 Data

Several organizations including the United States government have initiated projects to digitize declassified diplomatic documents.[1] For this project, we scrape the United States Department of State Office of the Historian's *Foreign Relations of the United States* (*FRUS*) collection, which archives significant intra-government documents from 1945 to 1980.[2] The vast majority of the documents are internal memos sent within the cabinet and with American diplomatic outposts.

The *FRUS* in total is too large and varied in topics to cover at once. As such, we focus on a couple important issue areas: Cold War Germany and Cuba (including the Cuban Missile Crisis). Each topic covers several volumes of the FRUS. Germany/Berlin involves 6,775 documents, while 4,931 exist on Cuba.

In order to perform the desired analysis, we split each memo into tokens and sentences (using Python's `nltk` package; specifically, `corpus` for removing stop words and `PorterStemmer` for lemmatizing). Stop words were also removed from the dictionary. A general summary of four collections of memos we scraped is in Table 1. For this project, and for historical reasons beyond the scope of this short update, we focus on the Germany/Berlin collection.

Collection	Documents	Sentences	Tokens
Germany/Berlin	6,775	260,283	30,712
Cuba	4,931	148,859	28,969
"Foreign Policy"	2,653	71,992	18,433
China	6,990	270,939	32,258

Table 1: Overall figures on collection sizes.

This data was substantially difficult to gather and clean for useful analysis. Moreover, no predetermined or recorded classifications exist for this data. The classifications necessary for this study—in both the training and testing phases—must be created from scratch. We perform brute-force bi-

[1] WikiLeaks has been actively divulging classified documents in digitized formats, as well.

[2] The collection is available at `http://history.state.gov/historicaldocuments`. The FRUS actually begins in 1861, but digitiziation begins with documents in 1945.

nary categorizations of whether or not a sentence contains threatening language (1 or 0, respectively).[3] For our purposes, and especially given the internal nature of the memos, here we focus on language that reveals *perception* of external threat. We provide a couple examples below. While our classifications were done using randomly chosen, unconnected sentences, these examples use whole paragraphs in order to better show the contrast between threatening and non-threatening language.

- "*The military are worried about the general attitude in Europe.* / There has been a let up on the ground that the threat is not as great as it was assumed to be. / *From our viewpoint the USSR's capability is still there. We have no way of knowing what Soviet intentions are.* / Initially, our problem was one of encouragement and of laying out the general outline of a plan for European defense."[4]
- "*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.*"[5]

For the purposes of this report, we classified 7,500 unique and randomly drawn sentences. 438 (or about 6%) were found to expressed perceived threat. This data was then converted into the appropriate token frequency matrix. However, to limit some noise and to slightly reduce the size of the data, we only limit ourselves to tokens that occur at least twice in the corpus of documents.[6] This sample hence contained 4,012 unique non-stop tokens.

3 Analysis

3.1 Naive Bayes and SVM

A "bag-of-words" approach is taken throughout. We first implement Bernoulli Naive Bayes (NB) with Laplace smoothing on the data. k-fold (with $k = 10$) cross-validation is used to leverage the data. The model yields an average accuracy rate of 91.25% with standard deviation 0.59%.

A support vector machine (SVM) approach slightly improved these results. ℓ-1 regularization was used with $C = 1$, where C is the penalization parameter for the primal problem

$$\min_{\gamma,w,b} \frac{1}{2}||w||^2 + C\sum_{i=1}^{m}\xi_i$$

subject to $y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i$ and $\xi_i \geq 0$, $i = 1, ..., m$. Due to the low proportion of threats in the data, we immediately address the issue of unbalanced data by weighting the data by class. Specifically, weights are such that $C' = C \cdot t$, where $t \in \{0, 1\}$.[7]

10-fold cross-validation shows an average accuracy rate of 94.24% with standard deviation 0.11%. As such, SVM seems to perform better than the NB implementation.

Table 2 displays the tokens most informative in identifying threatening sentences. Even at a cursory glance, SVM appears to choose far more substantively useful tokens than NB, which (based on tokens like `soviet`, `would` , `the`, `could`, `said`, and the like) appears to gauge informativeness based on frequency or basic existence in the sentence; those tokens are fourth, first, second, tenth, and

[3] We considered doing a multinomial classification based on varying degrees of threats. However, given that this is the first pass at a novel research agenda, we erred on the side of conservatism and decided on a binary approach, instead.

[4] Memorandum of Discussion of State–Mutual Security Agency–Joint Chiefs of Staff Meeting, Held at the Pentagon Building, January 28, 1953, 10:30 p.m. Available at `http://history.state.gov/historicaldocuments/frus1952-54v05p1/d374`.

[5] Telegram From the Department of State to the Mission at Berlin, October 3, 1961, 8:53 p.m. Available at `http://history.state.gov/historicaldocuments/frus1961-63v14/d167`.

[6] Cross-validation indicated that this threshold at 2 performs better than other potential values.

[7] Results that do not account for unbalanced data are markedly worse and not worth reporting if they are to be immediately discarded.

fifth most common tokens in the corpus, respectively.

NB		SVM	
soviet	nato	danger	induc
would	point	true	bit
berlin	east	seriou	battalion
the	access	creat	acceler
us	weapon	sovereignti	russian
forc	may	weak	fear
could	action	tension	doctrin
west	problem	attack	outstand
europ	posit	violenc	exploit
german	militari	threat	split
western	alli	graviti	just
germani	war	grow	soviet
said	state	resourc	divid
might	he	hostil	weapon
it	believ	road	violat
situat	use	greatest	war
danger	possibl	invas	happen
nuclear		competit	

Table 2: 35 most informative tokens by each method. (Top 18 are in the first column; next 17 are in the second.)

3.1.1 Sources of Error

Figure 1 displays training and test errors for the NB and SVM implementations above. Training and test error do not greatly change as the number of sentences in the training data (m) grows. One potential exception is SVM training error, which rises as a consequence of the model perfectly classifying threats in the much smaller training sets.

Given the small percentage of threats in the sample data, there may be some concern that the model's high accuracy comes from blindly classifying all test data as being "0"—that is, predominantly from false negatives. We therefore examine precision and recall for both NB and SVM. An attempt to graphically represent these changes over training data size is presented in Figure 2. The proportion of blue to red in each bar represents precision, while the proportion of blue to the horizontal line at 136 represents recall.

As such, the slight increase in test errors as m

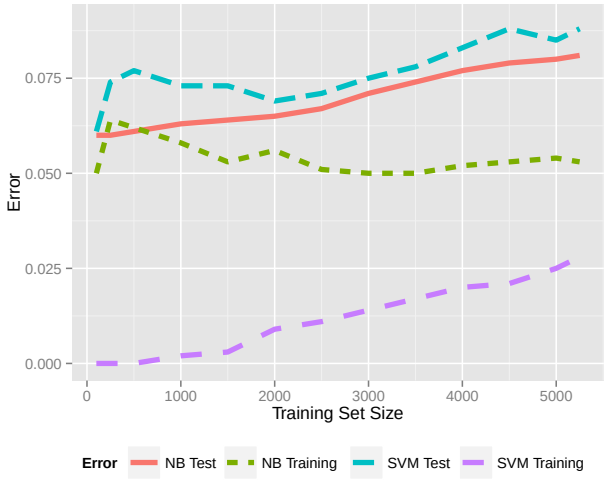
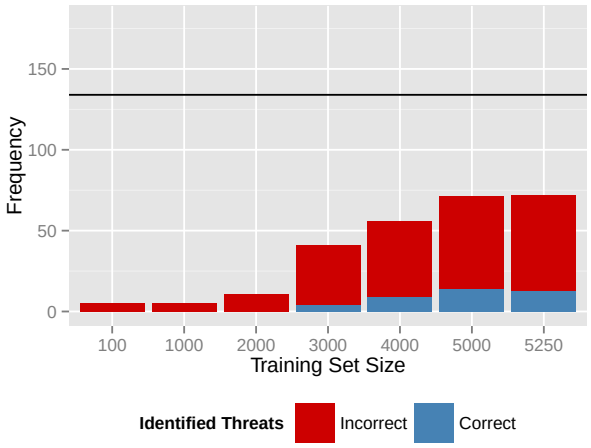


Figure 1: Training and test errors on the sample Berlin/Germany data using SVM. Calculations done using holdout cross-validation with a standard 70-30 split; k-fold cross-validation is not used due to computational/time constraints.

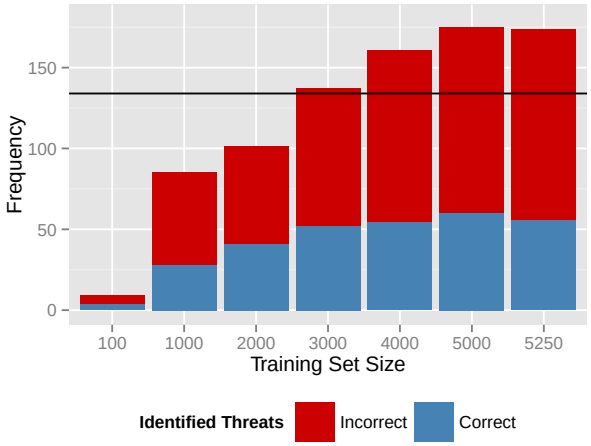
increases in Figure 1 may not be as troubling as it first seems. The low test error at small m is due to high numbers of false negatives. Indeed, when $m = 100$, the NB classifier predicts no perceived threats in the test data, and all test error comes from failing to identify "1"s. Both NB and SVM have improved recall with more data, though NB has noticeably worse performance. In our trials, NB's recall never exceeds 0.10, while SVM's reaches $0.40 - 0.44$ with large training sets and would likely continue to improve with more data. SVM also has higher precision than NB: the former wanders between 0.33 and 0.39, while the latter never exceeds 0.2. Overall, NB underidentifies threats and is often incorrect when doing so (probably related to its less convincing identification of influential tokens); SVM appears more effective, finding both a reasonable number of sentences as threats while also being more effective. We further summarize these results by calculating the balanced F-score for each test[8] and present these statistics in Figure 3. We openly admit that while SVM's results are better, they are not "ob-

[8]Recall that for the traditional F_1 measure,

$$F = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$



(a) Naive Bayes



(b) SVM

Figure 2: Threats identified by each model. The black horizontal line represents the 136 threats in the test data. (This was done via holdout cross-validation on a standard 70-30 split; k-fold cross-validation was avoided due to computational constraints.)

jectively" strong. Further refinements to the data, not feasible in our time frame, will likely improve this. (See thoughts in the "Conclusions.")

3.2 tf-idf

Data preprocessing already involved the removal of stop words. However, we may seek to take into account the relatively frequency of another subset of tokens that are not stop words yet also not highly informative in classification. We therefore

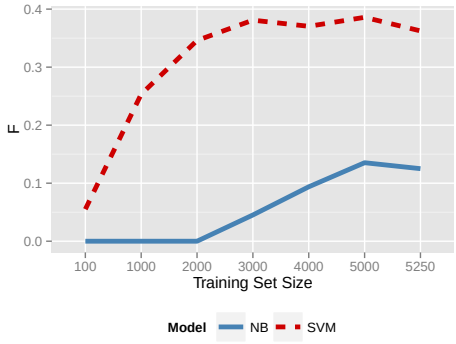


Figure 3: F_1 measures for both NB and SVM across varying training set sizes.

redo the analysis using term frequency-inverse document frequency (tf-idf) algorithm in order to account for tokens that show up frequently across documents (and therefore may not be useful or unique identifiers).

Given a collection of documents D where we denote an individual document as $d \in D$ and some token of interest t, tf-idf is calculated as

$$t_d = f_{t,d} \times \log\left(\frac{|D|}{f_{t,D}}\right)$$

where $f_{t,d}$ is the number of times t occurs in d, and $f_{t,D}$ is the number of times t occurs in D (Ramos 2003).

Given the nature of our diplomatic and Germany-centric documents, some very frequent tokens, both within and across sentences, include terms like *President*, *Secretary*, *Chancellor*, *Soviet*, *German*, and the like. Nonetheless, the refinement of the token frequency matrix using tf-idf has a negligible effect on the model's accuracy. The sparsity of the matrices may account for this. Due to high similarity, those redundant results are not presented here.[9]

4 Conclusions

The results of this paper suggest that machine learning techniques can indeed help us systemati-

[9] Another extension considered was the use of n-grams (specifically bi-grams and tri-grams). Neither approach improved results.

cally identify a key concept in many political environments, including international relations. The results are also admittedly preliminary and leave a substantial amount of room for improvement. Additional training data remains a vital ingredient toward that end.

The sparsity of our token matrices likely led to somewhat non-robust results. Our decision to use sentences as the unit of analysis here was primarily motivated by time constraints since all classifications were done manually. In the future, we would opt to classify and predict using memos, and then use principal component analysis to further narrow down the predictors of interest.

The models presented above run almost exclusively on the content of the sentences, without background context. We intend to continue with the SVM approach, which is more amenable to the addition of covariates that could increase accuracy. Covariates such as the sender or recipient of the memo may be important; correspondence involving more senior officials may be more prone to substantive and serious discussion of threats.

We are currently limited to studying perceived threats, rather than *use* of threats, due to data limitations. We require two-way communication between the United States and a second party to investigate how threats are used and recognized in a full bargaining and/or crisis setting. Disclosed documents on disarmament agreements or bilateral trade talks between the United States and another country (the United Kingdom would be an easy choice for linguistic reasons, though perhaps not useful for finding many threats) can be helpful on this front.

Hidden Markov Models (HMM) would be a good framework for understanding fluctuations in the existence of threats. Language in memorandums is only an indirect manifestation of threats; the actual "state of threat" that elites perceived is unobserved. We hoped to present HMM results in this paper, but the random sampling of our data was not amenable to HMM; we require a large quantity of sequential data, and at the document level, to test this effectively. Recent primatology work, however, does show promise in using HMM to examine changes in threats (Etting et al. 2013).

Finally, a clearer and more systematic definition of "threats" is fundamental to this undertaking's success. Our brute-force classifications appear to have some utility but could be more theoretically motivated. Using a more rigorous set of guidelines could easily improve the models more than statistical refinements could. The current iteration of this project quietly sidesteps a fundamental question: What is a (perceived) threat? This must be confronted in the near future.

References

[1] Etting, Stephanie E., Lynne A. Isbell, and Mark N. Grote. 2013. "Factors Increasing Snake Detection and Perceived Threat in Captive Rhesus Macaques (*Macaca mulatta*)." *American Journal of Primatology.* Published online; print forthcoming.

[2] Jervis. Robert. 1976. *Perception and Misperception in International Politics.* Princeton: Princeton University Press.

[3] Ramos, Juan. 2003. "Using tf-idf to determine word relevance in document queries." In First International Conference on Machine Learning, New Brunswick, New Jersey. Rutgers University.

[4] Rousseau, David L. and Rocio Garcia-Retamero. 2007. "Identity, Power, and Threat Perception: A Cross-National Experimental Study. *Journal of Conflict Resolution* 51(5): 744-771.

[5] Sartori, Anne E. 2002. "The Might of the Pen: A Reputational Theory of Communication in International Disputes." *International Organization* 56(1): 126-149.

[6] Stein, Janice G. 2013. "Threat Perception in International Relations." In *Oxford Handbook of Political Psychology, Second Edition*, eds. Leonie Huddy, David O. Sears, and Jack S. Levy. Oxford: Oxford University Press.