Forecasting Political Conflict in Asia using Latent Dirichlet Allocation Models *

Philip A. Schrodt
Department of Political Science
Pennsylvania State University
University Park, PA 16802
USA
schrodt@psu.edu

Version 1.0 : June 10, 2011

*Paper prepared for delivery at the European Political Science Association, Dublin, 16-18 June 2011. This project was funded in part by the National Science Foundation (SES-1004414). Data were provided by the Defense Advanced Research Projects Agency under the Integrated Crisis Early Warning System (ICEWS) program. The views, opinions, and/or findings contained in this paper are those of the author and should not be interpreted as representing the official views or policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the Department of Defense, nor those of the National Science Foundation. The original open source software discussed in this paper can be downloaded from the web site: `http://eventdata.psu.edu/` or from the author.

Abstract

Latent Dirichlet allocation models are a relatively new computational classification algorithm. In its standard application to document classification, the model assumes each document to be composed of a mixture of multiple, overlapping topics each with a typical set of words, and classification is done by associating words in a document with the latent topics most likely to have generated the observed distribution of those words. I apply this technique to the problem of political forecasting by assuming that the stream of events observed between a dyad of actors is a mixture of a variety of different political strategies and standard operating procedures (for example escalation of repressive measures against a minority group while simultaneously making efforts to co-opt the elites of that group). By identifying the dominant strategies being pursued at time t, one gets information that can be used to forecast likely patterns of interaction at a later time $t + k$. This approach is applied to event data generated for 29 Asian countries in the Integrated Conflict Early Warning System project for 1998-2010 to forecast the ICEWS conflict measures for rebellion, insurgency, ethno-religious violence, domestic political conflict and international conflict at a six month lead time. In random samples balancing the occurrence of negative and positive outcomes on the dependent variable, LDA combined with a logistic model predicts with around 60% to 70% accuracy in in-sample evaluation, and improves very substantially on the sensitivity of the classification compared with simple logistic models in full samples. A supervised version of LDA, however, does not provide much improvement over the unsupervised version, and shows some pathological behaviors. Some structure can be found in the factors, though more work is needed on this.

1 Introduction

Political event data have long been used in the quantitative study of international politics, dating back to the early efforts of Edward Azar's COPDAB [Azar, 1980] and Charles McClelland's WEIS [McClelland, 1976] as well as a variety of more specialized efforts such as Leng's BCOW [Leng, 1987]. By the late 1980s, the NSF-funded *Data Development in International Relations* project [Merritt et al., 1993] had identified event data as the second most common form of data—behind the various Correlates of War data sets—used in quantitative studies (McGowan et al 1988). The 1990s saw the development of two practical automated event data coding systems, the NSF-funded KEDS [Gerner et al., 1994, Schrodt and Gerner, 1994] and the proprietary VRA-Reader (`http://vranet.com`; [King and Lowe, 2004] and in the 2000s, the development of two new political event coding ontologies—CAMEO [Gerner et al., 2009] and IDEA [Bond et al., 2003]—designed for implementation in automated coding systems.

Much of the work with event data has focused on forecasting political conflict. Within the early warning literature, three primary methodological approaches exist: time series [Pevehouse and Goldstein, 1999, Shellman, 2004, 2000, Harff and Gurr, 2001], vector auto regression (VAR) [Goldstein, 1992, Freeman, 1989], and hidden Markov models (HMM) [Bond et al., 2004, Shearer, 2006, Schrodt, 2000, 2006]. This paper—like that of the HMM work—will look at event data as patterns since patterns are one of the most common modes of political analysis found in qualitative studies. In particular, various forms of qualitative "case-based reasoning"—see for example May [1973], Neustadt and May [1986], Khong [1992]—essentially match patterns of events from past cases to the events observed in a current situation (with some substitutions for equivalent events), and then use the best historical fit to predict the likely outcome of the current situation.[1] Instead of analyzing the effects of specific events in a vacuum (like Harff [1998] and her focus on specific "triggers" and "accelerators") a pattern-recognition approach allows discrete events or event counts to determine the likelihood of future events. This general concept can be implemented in a variety of different ways—see for example the various "artificial intelligence" approaches in Hudson [1991], Schrodt [1990], Bakeman and Quera [1995], Hudson et al. [2008] and the HMM studies cited earlier.

In this study, I will using the latent Dirchlet allocation (LDA) algorithm— a recently-developed classification method usually applied to document classification—to try to discriminate between patterns that do and do not precede various type of conflicts. The event data are from the recently-developed DARPA-funded Integrated Conflict Early Warning System data set (ICEWS; O'Brien 2010), described in more detail below. This paper is largely a "proof-of-exercise" exercise to determine whether LDAs are even plausible as an approach for event data analysis; the initial results do appear promising.

[1] See [Schrodt, 2004, chapter 6] for a much more extended discussion of this approach

2 Method

Latent Dirichlet allocation (LDA) models were introduced by Blei et al. [2003] and briefly described in the abstract of that article as:

> LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities.

In the typical LDA application to document classification, each document is assumed to be a mixture of multiple, overlapping *latent topics*, each with a characteristic set of words. Classification is done by associating words in a document with the topics most likely to have generated the observed distribution of words in the document. The purpose of LDA is to determine those latent topics from patterns in the data.

The latent topics are useful for two purposes. First, to the extent that the words associated with a topic suggest a plausible category, they are intrinsically interesting in determining the issues found in the set of documents. For example, one of the sample data sets in the *R* `lda` package [Chang, 2010] determines the set of issues discussed in a series of political blogs. Second, the topics can be used with other classification algorithms such as logistic regression, support vector machines or discriminant analysis to classify new documents. The full mathematical details of LDA estimation can be obtained from that paper or the other usual suspects on the web and will not be repeated here, as I am simply using this off-the-shelf (or off-the-`CRAN`, as the case may be.)

Despite the surface differences between the domains, the application of this technique to the problem of political forecasting is straightforward: It is reasonable to assume that the stream of events observed between a set of actors is a mixture of a variety political strategies and standard operating procedures (for example escalation of repressive measures against a minority group while simultaneously making efforts to co-opt the elites of that group). This is essentially identical to the process by which a collection of words in a document is a composite of the various themes and topics, the problem LDA is designed to solve. As before, the objective of LDA will be to find those latent strategies that are mixed to produce the observed event stream. These latent factors can then be used to convert full event stream to a much simpler set of measures.

The importance of latent dimensions in event data—rather than specifying the dimensions *a priori* based on some theory—is due to issues of measurement. As I noted in Schrodt (1994), if one is using event data in forecasting models—the objective of ICEWS—coding error is only one potential source of error that lies between "events on the ground" and the predictions of the forecasting model. These include

- News reports are only a tiny, tiny fraction of all of the events that occur daily, and are non-randomly selected by reporters and editors;

- Event ontologies such as WEIS, CAMEO and IDEA are very generic and bin together events that may not always belong together in all contexts;
- Forecasting models always contain specification error and cannot consider everything; for example few if any political forecasting models contain a full economic forecasting component;
- Political systems have a degree of intrinsic randomness due to their inherent complexity, chaotic factors even in the deterministic components of those systems, the impact of effectively random natural phenomena such as earthquakes and weather, and finally the effects of free will, so the error intrinsic to a forecasting model will never reduce to zero.

Because of these sources of error, the ability to determine latent dimension in event data is important in the overall scientific exercise of improving instrumentation for conflict forecasting. The latent dimensions of event data will never be not self-evident (or purely derivable from theory) because of the measurement factors noted above. We do not have a "god's-eye view" of political interactions—we have the highly (and non-randomly) selected view provided by the international media. Consequently determining methods that will allow these to be more effectively used to move the field forward more generally.

The LDA approach is similar in many ways to the hidden Markov approach. In both models, the observed event stream is produced by a set of events randomly drawn from a mixture of distributions. In an HMM, however, these distributions are determined by the state of a Markov chain, whose transition probabilities must be estimated but which consequently also explicitly provides a formal sequence. An LDA, in contrast, allows any combination of mixtures, without explicit sequencing except to the extent—as in this paper—that sequencing information is provided by the events in the model. The HMMs uses in political forecasting also tend to have a relatively small (typically about 5) set of states, and hence distributions, whereas LDA's typically use a larger number.

The forecasting component will use two different methods. First, a simple logistic regression will be used with the latent factor (rather than event) counts as the independent variables. Second, I will test a supervised-learning version of LDA, sLDA [Blei and McAuliffe, 2007]. Rather than determining arbitrary latent factors, which may or may not have any utility in classification, sLDA starts with the known classifications of cases, and derives factors that can be used in a logit or regression model to predict those values.

This paper, as an initial evaluation of the method, will use in-sample evaluation rather than the preferred split-sample approach used to evaluate predictive models. The basic scheme will be to use two months of data to predict the dependent variable six months later, i.e. a model of the form

$$M(lf_t, lf_{t-1}) \rightarrow conf_{t+6} \tag{1}$$

where lf_t are the latent event factors at time t and $conf_{t+6}$ is a measure of conflict six months later. The choice of six months is arbitrary—the method would work at any time

horizon—but is a "policy-relevant lead time" consistent with other forecasting work; that is, a period of time sufficiently long that there could be a policy response.

I used the R package `lda` [Chang, 2010] and the routines

```
library(lda):lda.collapsed.gibbs.sampler
```

and

```
library(lda):slda.em
```

to implement the basic LDA and sLDA respectively. The R logistic estimator

```
glm(...,family=binomial())
```

was used for the logistic model, and routines from the `ROCR` package [Sing et al., 2009] were used to produce the ROC curves and AUC estimates. Estimation of the simple LDA model was more or less immediate; the sLDA models took a minute or two to estimate. The resampling estimation was done using the Penn State Research Computing and Cyberinfrastructure high-performance computing facility `http://rcc.its.psu.edu/`. The event data counts were initially aggregated using a *Stata* script into a rectangular data set based on the dyads and categories discussed below ; this was converted to the sparse-count `LDA-C` format used in `lda` with a custom Python program. This software as well as the LDA/sLDA estimation scripts are available from the author.

3 Data

The basic data set used in the analysis is the DARPA-funded Integrated Conflict Early Warning System (ICEWS; O'Brien [2010], Schrodt [2010]) Asian data set, which covers the period 1997- 2010 and contains over 2,000,000 events for 29 Asian countries. The ICEWS dataset is produced using a proprietary automated event data coding program, JABARI, based on the open-source TABARI program but incorporating a number of new features, particularly pre-processing with open-source natural language processing software, which increases coding accuracy substantially over TABARI. ICEWS also uses multiple regional news sources to provide more comprehensive coverage of countries that tend to receive little media attention from western outlets (Fiji and the Comoros, for example). JABARI uses a 15,000-item actor dictionary to code for a broad range of domestic actors, including but not limited to military, police, rebel groups, and civilians, and this allows for detailed analyses of domestic events.

The key difference between the ICEWS event data coding efforts and those of earlier event data efforts is the scale. As O'Brien—the ICEWS project director—notes,

> ...the ICEWS performers used input data from a variety of sources. Notably, they collected 6.5 million news stories about countries in the Pacific Command (PACOM) AOR [area of responsibility] for the period 1998-2006. This resulted in a dataset about two orders of magnitude greater than any other with which

we are aware. These stories comprise 253 million lines of text and came from over 75 international sources (AP, UPI, and BBC Monitor) as well as regional sources (*India Today, Jakarta Post, Pakistan Newswire,* and *Saigon Times*).

3.1 Independent Variables

3.1.1 Actors

Most of the earlier event data analysis has been at the nation-state level, and in those contexts, the obvious unit of actor aggregation is the state-dyad, e.g. $USA - USSR$. Once substate actors are coded, the problem becomes more complex; D'Orazio et al. [2011] discuss this issue in some detail. While CAMEO scheme currently codes about three dozen distinct substate actor types, comparable work that has used event data to study substate behavior generally aggregates these into more general categories: for example in the VRA scheme, Bond et al. [1997] discuss "mass" and "state" actors; in the GEDS scheme, Harff and Gurr [2001] discuss "governing elites", "mass followship", "disadvantaged groups", etc; Davies et al. [1998] address "kindred groups", "communal groups", etc; Shellman [2000] discusses "government" and "dissidents".

Following earlier work on ICEWS, the actors in this analysis are aggregated into the following general categories

- gov: government agents such as the executive, police, and military
- par: political parties
- opp: armed opposition—rebels and military groups
- soc: society in general—civilians, businesses, professional groups
- ios: international actors
- usa: United States

3.1.2 Events

The majority of extant event data literature either scales all events, assigning them a score on a conflict-cooperation continuum or generates event counts reflecting the number of events that occur within conceptually unique categories. The Goldstein Scale (Goldstein [1992]), which is the most commonly used scaling technique within the event data literature (see Goldstein [1992], Schrodt [2007], Schrodt and Gerner [1994], Pevehouse and Goldstein [1999], Hämmerli et al. [2006], for sample uses) assigns a value to all events coded under the World Event Interaction Survey (WEIS) scheme on a -10 to 10 scale conflict/cooperation scale, with -10 reflecting the most conflictual events and 10 indicating the most cooperative is the most commonly used.

Despite its dominance within the event data literature, the Goldstein scale requires additional levels of aggregation beyond the initial scaling, which leads to a number of operational difficulties. For example, consider a day on which an armed killing (which receives a -10 score) and a peace-treaty signing (which receives a +10 score) occur on the same day between the same actors. Summing Goldstein scores would result in a net score of 0 in the previous example, which is the same score that days with no activity receive. While this example of two events exactly canceling is hypothetical, the problem of violent events masking the concurrent presence of cooperative actions—notably negotiations occurring during periods of on-going violence—is very real, and occurs frequently during such periods when the KEDS Levant and Balkans data are aggregated using Goldstein scores.[2] This is further complicated by the fact that comments and meetings have Goldstein scores that are small in magnitude, whereas violent events have a scale score of -10. Consequently a small amount of violence can mask a lot of talking. A similar problem plagued the scaled scores of the COPDAB data set, where the quip was made that "In COPDAB, three riots equals a thermonuclear war."

Due to this problem, we have shifted to utilizing count measures [Schrodt et al., 2001, Schrodt and Gerner, 2004, Shearer, 2006, D'Orazio et al., 2011], with good results. The approach we have used is similar to earlier Duval and Thompson [1980] event data count model which places all events into one of the four conceptually unique, mutually exclusive categories, and these can be readily translated from the WEIS system used in the original article to CAMEO, which in contrast to WEIS was deliberately structured so that these aggregations occurs in continguous categories:

- *Verbal Cooperation*: The occurrence of dialogue-based meetings (i.e. negotiations, peace talks), statements that express a desire to cooperate or appeal for assistance (other than material aid) from other actors. CAMEO categories 01 to 05.
- *Material Cooperation*: Physical acts of collaboration or assistance, including receiving or sending aid, reducing bans and sentencing, etc. CAMEO categories 06 to 09.
- *Verbal Conflict*: A spoken criticism, threat, or accusation, often related to past or future potential acts of material conflict. CAMEO categories 10 to 14.
- *Material Conflict*: Physical acts of a conflictual nature, including armed attacks, destruction of property, assassination, etc. CAMEO categories 15 to 20.

3.1.3 Time

The data have been aggregated at a monthly level: this is in keeping with the monthly coding of the ICEWS GTDS indicators described below. Since event data are coded to a precision

[2] Another alternative to the Goldstein scale is the Bond et al. [1997] and Jenkins and Bond [2001] utilize a different type of event count structure, which places all events into one of eight boxes which reflect whether an event is violent or non-violent and direct or indirect. The VRA "Conflict Carrying Capacity" approach differs from the Goldstein scale in its use of a ratio of counts, and more generally the "Cambridge" approach of VRA and various Harvard-based studies such as King and Lowe [2004] generally employs ratios and average scaled values rather than the counts and total scaled values used in most of the KEDS project studies.

of a day, we could use a higher level of resolution: for example studies have been done at the daily level (Pevehouse and Goldstein [1999], Shearer [2006], Schrodt [2006]), though weekly (Brandt and Freeman [2005], Shellman and Stewart [2007]), monthly (Schrodt [2007], Ward et al. [2010]), quarterly (Jenkins and Bond [2001]), and annual level aggregations are present within the literature. A number of studies find that different temporal aggregations on the same data can affect empirical results (Alt et al. [2001], Dale [2002], Shellman [2004]), so the use of a monthly aggregation probably has some effect on the results.

3.2 Dependent Variables

The dependent variables that will be forecast are the political conflict measures the ICEWS Ground Truth Dataset (GTDS), which provides a monthly, state-level, binary measure of whether or not each of the five types of political conflict "events-of-interest" (EOI) described below occur during each state-month.

- Rebellion: Organized opposition where the objective is to seek autonomy or independence; [REBELL]
- Insurgency: Organized opposition where the objective is to overthrow the central government; [INSURG]
- Ethnic Religious Violence: Violence between ethnic or religious groups that is not specifically directed against the government; [ETHREL]
- Domestic Political Crisis: Significant opposition to the government, but not to the level of rebellion or insurgency (for example, power struggle between two political factions involving disruptive strikes or violent clashes between supporters); [DOMCRI]
- International Crisis: Conflict between two or more states or elevated tensions between two or more states that could lead to conflict. [INTCRI][3]

The GTDS indicators were originally developed for 1998-2006 using human coding from a variety of sources; the 2007-2010 indicators have been coded using a combination of event-data indicators and machine-assisted coding.

4 Results

The LDA estimation was implemented in a set of *R* scripts. As usual, some initial experimentation was required to get the method to perform reasonably well. Most importantly, because positive instances of the INSURG, ETHREL and DOMCRI indicators are relatively

[3]Source for EOI descriptions: O'Brien [2010, p. 90]

rare in the full data set—only around 5% to 10%—classification algorithms will tend to simply predict the modal (negative) category. When the full data set was estimated, only the REBELL and INTCRI indicators had non-trivial predictors; the remaining models simply predicted the negative for all cases. This was corrected by taking a roughly balanced random sample of the negative cases.[4] This random sample, however, affects the results, so these will initially be presented as distributions rather than point estimates.

The LDA estimation uses a Gibbs sampler and consequently is a random process itself. I did not systematically estimate this variation, but in some small-sample experiments the AUC estimates for INSURG varied by around ±0.01 when the number of iterations in the estimator was set to either 64 or to the much more time-consuming 128. This is smaller than the variation induced by the random sampling of the negative cases, but still is an issue.

Table 1 shows the results of multiple random samples for the five EOIs; $LDAAcc$ is the accuracy—defined below—for the unsupervised LDA with 10 factors followed by classification using a logistic classification based on those factors. $LDAAUC$ and $sLDAAUC$ are the ROC "area under curve" measures for the unsupervised and supervised LDA respectively. As Sing et al. [2009, p. 3] notes, "[AUC] is equal to the value of the Wilcoxon-Mann-Whitney test statistic and also the probability that the classifier will score a randomly drawn positive sample higher than a randomly drawn negative sample." AUC is a widely-used measure of overall predictive accuracy; an AUC of 0.5 indicates that the model is only performing as well as chance. Ulfelder [2011] observes that in political forecasting,"An AUC of 0.5 is what youd expect to get from coin-flipping. A score in the 0.70s is good; a score in the 0.80s is very good; and a score in the 0.90s is excellent."[5] N refers to the number of random samples that the distribution is based on; this differs due to the tendency of sLDA to crash, particularly on the INSURG and ETHREL cases.

Three general conclusions can be drawn from Table 1. First, the unsupervised LDA does better than chance on all of the indicators with a reasonably high positive frequency, usually about 10% better as measured by accuracy, but almost 20% better for INTCRI. REBELL and INTCRI also have fairly high average AUC measures, whereas the AUC for the other three indicators is in line with the accuracy measure. Second, the variation across the random samples is quite wide, particularly for the three rare positive value indicators, where the range is around ±0.05. This is considerably higher than the variation due to the Gibbs sampler estimation procedures. Third, the supervised LDA generally does not perform noticeably better—or in some instance not at all better—than the unsupervised LDA. The average AUC is actually lower for the higher-positive-frequency REBELL and INTCRI, dropping substantially below 0.5 for INTCRI. It is roughly equal for ETHREL, somewhat higher for INSURG and only for DOMCRI is there a major difference, and even here the ranges of the estimates overlap.

Figures 1, 2 and 3 show the full distribution of the accuracy measures for REBELL. Figures

[4] All cases where the indicator was positive were included in each sample. The proportion of negative cases used were REBELL 30%, INSURG 12.5%, ETHREL 8%, DOMCRI 15%, INTCRI 30%

[5] `http://dartthrowingchimp.wordpress.com/2011/06/09/forecasting-popular-uprisings-in-2011-how-are-we-doing/`. Accessed 10-Jun-2011.

1 and 2 are roughly normally distributed and a spot-check on some other distributions confirmed this pattern. Figure 3, on the other hand, is decidedly skewed, more like a chi-squared distribution, with a few exceptionally high AUCs but most of these barely above chance.

These results suggest that little is to be gained from the sLDA estimation. This is compounded by the fact is a bug somewhere in the `lda:slda.em.mmsb.collapsed.gibbs.sampler` routine that causes R to crash. Both the reported cause of the crash—that is, the resulting error message—and the timing are unpredictable, so presumably some routine is mucking about somewhere in memory where it shouldn't be, and those changes eventually prove computationally fatal.[6] But randomly: in the runs on the HPC machines, some submissions would crash after a couple of iterations; some would run for the entire four hours I had allocated. Chang [2010, p. 9] alludes to a potential problem in the routine with "WARNING: This function does not compute precisely...when the count associated with a word in a document is not 1" and that is definitely the situation here (though it is also the case for the `poliblog` data set included with the package) and this problem may be related to that issue.

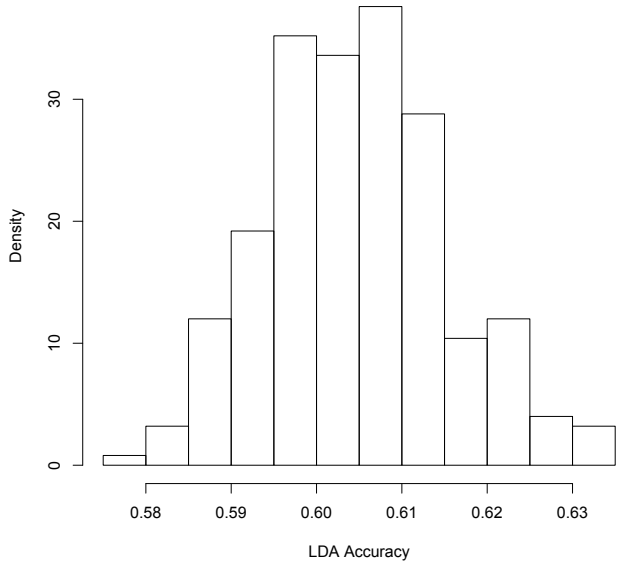

Figure 1: Distribution of LDA Accuracy for REBELL

[6]Though if someone would like to try to track down this bug, I have an ETHREL subset that reliably causes a crash.

Table 1: Distribution of LDA Accuracy and AUC, sLDA AUC

	LDA Acc.	LDA AUC	sLDA AUC
REBELL			
Mean	0.604	0.727	0.579
Min.	0.577	0.704	0.510
Max.	0.632	0.753	0.775
StDev	0.010	0.008	0.039
N = 250			
INSURG			
Mean	0.608	0.649	0.689
Min.	0.521	0.536	0.631
Max.	0.686	0.722	0.751
StDev	0.048	0.032	0.023
N = 125			
ETHREL			
Mean	0.577	0.620	0.615
Min.	0.493	0.490	0.559
Max.	0.638	0.715	0.707
StDev	0.025	0.043	0.024
N = 125			
DOMCRI			
Mean	0.621	0.590	0.666
Min.	0.574	0.530	0.592
Max.	0.671	0.636	0.705
StDev	0.018	0.022	0.018
N = 340			
INTCRI			
Mean	0.678	0.820	0.442
Min.	0.648	0.801	0.374
Max.	0.709	0.837	0.630
StDev	0.011	0.007	0.048
N = 235			

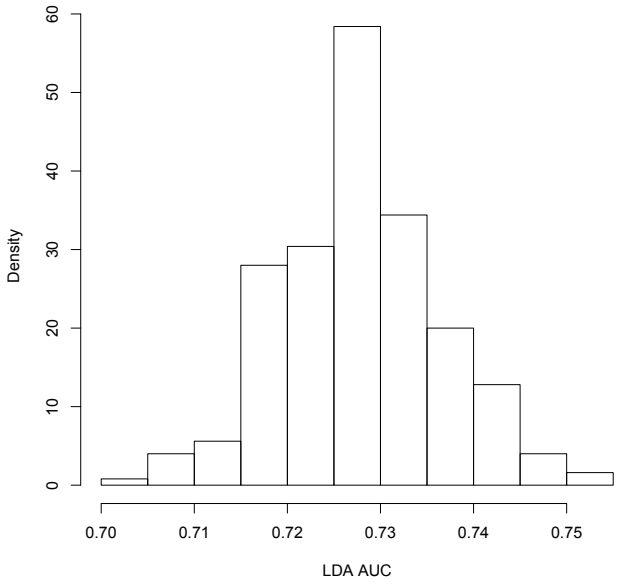

Figure 2: Distribution of LDA AUC for REBELL

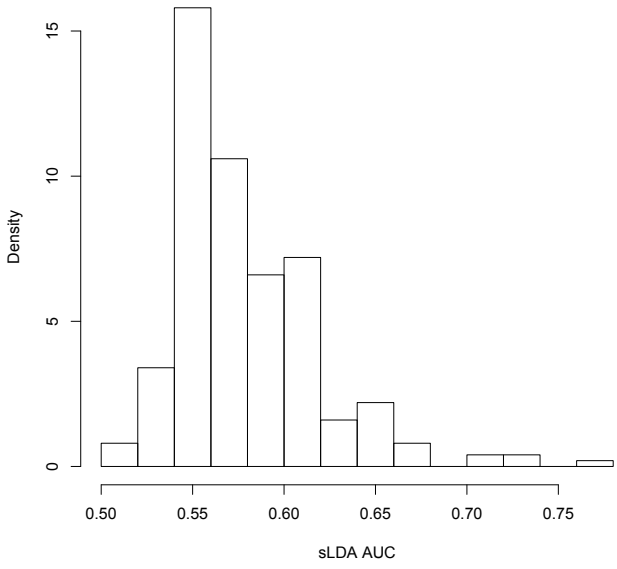

Figure 3: Distribution of sLDA AUC for REBELL

Tables 3 through 7 and Figures 4 through 13 show detailed classification results for each EOI. These are based on a single random subsample for each EOI, though they seem fairly typical. In the presentation of the results, each table is configured as shown in Table 2.

Table 2: EOI Classification Table Scheme

	true	
pred	0	1
0	TN	FN
1	FP	TP

The indicators are the usual

$$Accuracy = \frac{TP+TN}{TN+FN+FP+TP}$$

$$Specificity/Recall = \frac{TN}{TN+FP}$$

$$Sensitivity = \frac{TP}{TP+FN}$$

$$Precision = \frac{TP}{TP+FP}$$

$$F1 = 2 \cdot \frac{Precision \times Recall}{Precision+Recall}$$

The "ICEWS Reference Model" is a model that estimates a step-wise logistic using the same independent variables in the full-sample. This is not terribly useful since the accuracy and specificity in the full-sample is exaggerated by the low frequency of the positive cases, with a corresponding hit to sensitivity; I intend to do a balanced-sample estimation of these models in a future iteration of the paper. Nonetheless, the information is useful in showing that the LDA improves substantially on the sensitivity, typically by about a factor of at least 2, and in the case of ETHREL and DOMCRI, more than a factor of 5. How much of this is due to the method and how much to the balanced sample remains to be determined.

These individual results generally reinforce the analysis in Table 1, though with some additional detail. The ROC curves, assuming these are representative, show no particular pathologies, and in those instances where the AUC is near 0.5, the ROC generally follows the expected pattern of simply tracing the 45° line, though in the case of INTCRI it falls substantially below this. Consistent with the results in Table 1, in the illustrated case for DOMCRI, the sLDA ROC is substantially better than the LDA ROC, but for most of the EOIs it does worse, sometimes quite a bit worse.

Table 3: Classification Table: REBELL

		true		
pred	0	1	row N	
0	904	592	1496	
1	126	283	409	
col N	1030	875	1905	
	Acc	0.623	AUC	0.732
	Spec	0.604	Sens	0.691
	Prec	0.323	F1	0.421
sLDA AUC		0.527		
ICEWS Reference Model				
	Acc	0.852		
	Spec	0.996		
	Sens	0.387		
	N	4437		

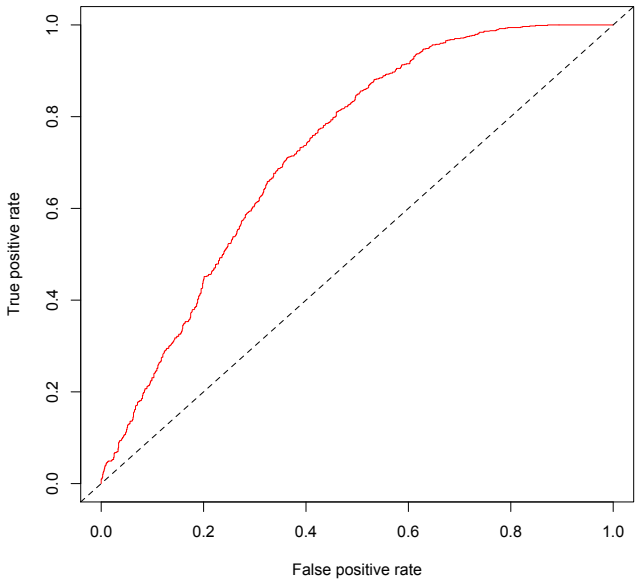

Figure 4: LDA ROC Curve: REBELL

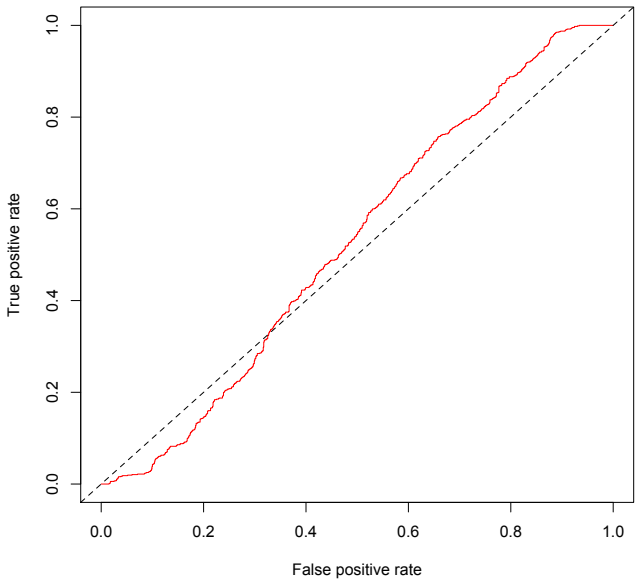

Figure 5: sLDA ROC Curve: REBELL

Table 4: Classification Table: INSURG

		true		
pred	0	1	row N	
0	375	231	606	
1	90	260	350	
col N	465	491	956	
	Acc	0.664	AUC	0.677
	Spec	0.619	Sens	0.743
	Prec	0.529	F1	0.571
sLDA AUC		0.716		
ICEWS Reference Model				
	Acc	0.915		
	Spec	0.987		
	Sens	0.334		
	N	4437		

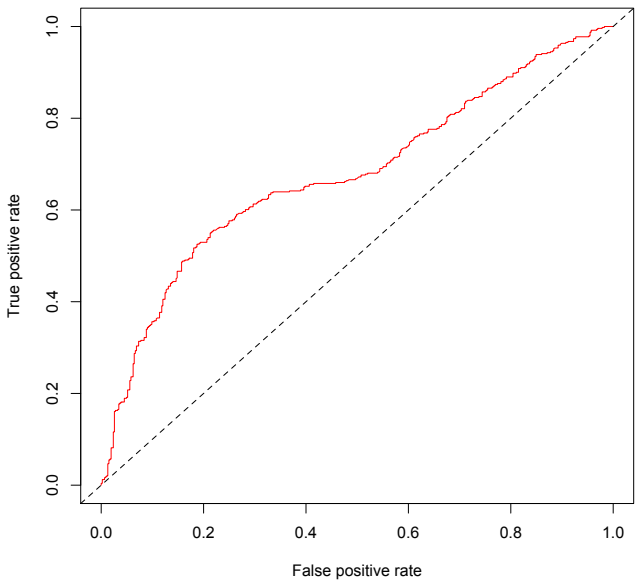

Figure 6: LDA ROC Curve: INSURG

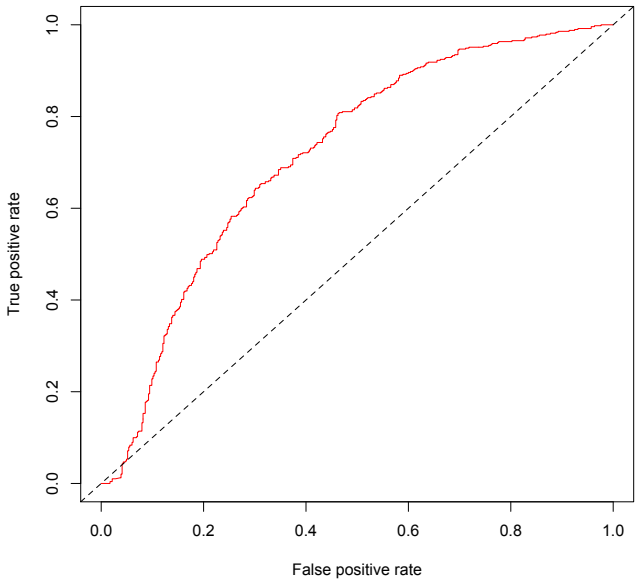

Figure 7: sLDA ROC Curve: INSURG

Table 5: Classification Table: ETHREL

	true			
pred	0	1	row N	
0	262	222	484	
1	51	71	122	
col N	313	293	606	
	Acc	0.550	AUC	0.618
	Spec	0.541	Sens	0.582
	Prec	0.242	F1	0.335
sLDA AUC		0.556		
ICEWS Reference Model				
	Acc	0.932		
	Spec	0.996		
	Sens	0.038		
	N	4403		

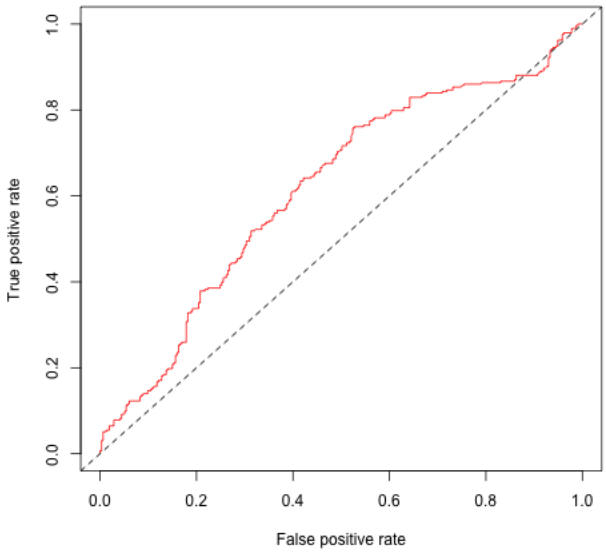

Figure 8: LDA ROC Curve: ETHREL

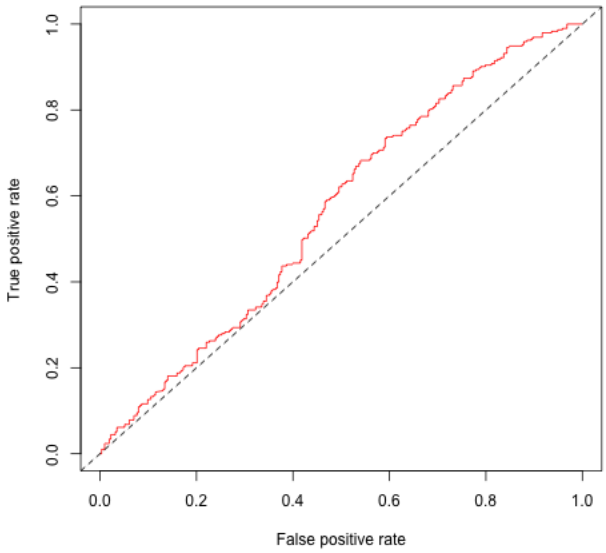

Figure 9: sLDA ROC Curve: ETHREL

Table 6: Classification Table: DOMCRI

		true		
pred	0	1	row N	
0	574	345	919	
1	17	46	63	
col N	591	391	982	
	Acc	0.631	AUC	0.572
	Spec	0.624	Sens	0.730
	Prec	0.118	F1	0.198
sLDA AUC		0.670		
ICEWS Reference Model				
	Acc	0.914		
	Spec	0.993		
	Sens	0.102		
	N	4433		

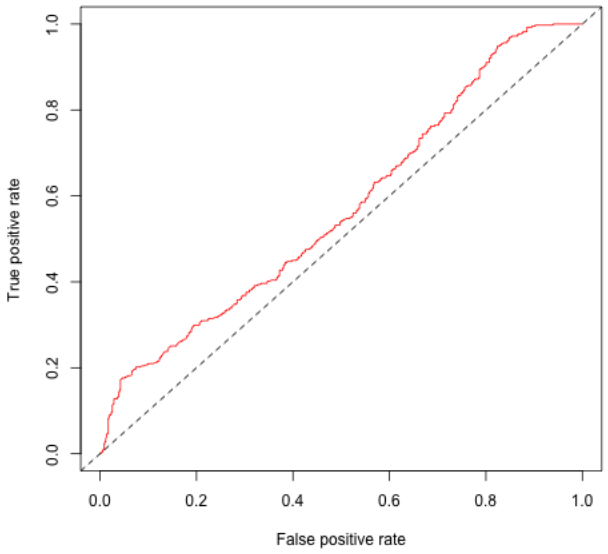

Figure 10: LDA ROC Curve: DOMCRI

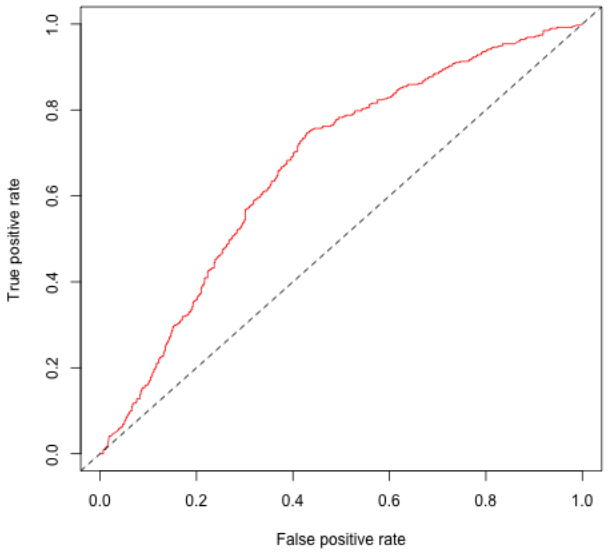

Figure 11: sLDA ROC Curve: DOMCRI

Table 7: Classification Table: INTCRI

		true		
pred	0	1	row N	
0	826	470	1296	
1	154	414	568	
col N	980	884	1864	
	Acc	0.665	AUC	0.803
	Spec	0.637	Sens	0.728
	Prec	0.468	F1	0.540
sLDA AUC		0.447		
ICEWS Reference Model				
	Acc	0.820		
	Spec	0.965		
	Sens	0.236		
	N	4437		

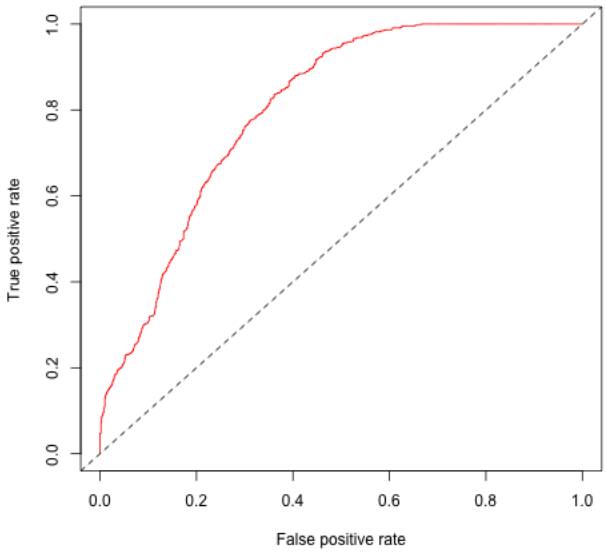

Figure 12: LDA ROC Curve: INTCRI

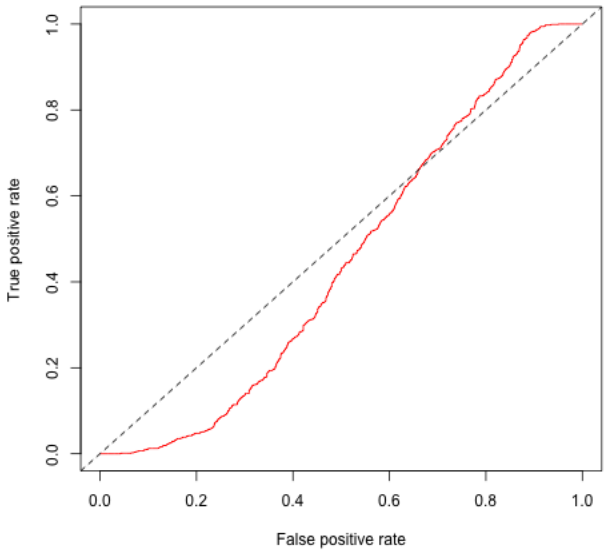

Figure 13: sLDA ROC Curve: INTCRI

Table 8 shows the most common eight events in each of the ten factors; this list was produced using the `lda:top.words()` procedure. The counts are in the form $< source > \_ < target > \_ < event-type >$ where the components of the variable correspond to the actor aggregation labels in Section 3.1 and the prefix $L1$ indicates the $t-1$ lag. Note that event counts can occur in multiple factors and, for example, *gov_gov_vercp*—the most common event in the data—occurs in most (but not all) of the factors in either its current or lagged form.

The illustrated set is not, shall we say, particularly transparent, though there is some clustering: for example $F7$ contains a disproportionate number of *gov_ios* counts, $F6$ focuses on *gov_opp* and $F4$ is mostly *gov_gov*. Additional analysis of these, particularly across a number of samples, might reveal further patterns.

I ran a principal components analysis (PCA) on the factor scores across the set of cases for several combinations of random sample and EOIs, and quite consistently the graph of the eigenvalues on these components is very flat, albeit usually with a noticeable drop-off for the last two components. This is quite different than the expected "skree slope" pattern one typically finds with a PCA, and would normally suggest that there are more factors than the 10 designated here, . This is also somewhat puzzling given that usually dyadic event data shows a very strong loading on a single conflict-cooperation dimension (hence the widespread success of the unidimensional Goldstein scale), though this is presumably complicated by the mixture of substate dyads used in this analysis. That result, in turn, would be consistent with the assumption that the event stream is a composite of multiple strategies, the assumption for using LDA in the first place. However, in contrast to application of LDA to text, event data—both at the event-category level and the dyad-level—are already highly structured, and consequently the flat PCA loadings may imply that there is not much additional structure left for the LDA to find.

4.1 Quirks in the analysis

The following notes are information that will otherwise disappear into the laboratory notebook but may be of interest to anyone trying to replicate/extend this work.

- One of the [many] free parameters is the number of iterations used in the Gibbs sampler and EM: I experimented with 8, 16, 32, 64 and 128 and the value of 64 appears to produce stable results as well as being reasonably fast. The LDA estimation is fairly quick under any choice of iterations; the sLDA takes a very long time when both the Gibbs and EM iterations are set to 128

- There are a variety of additional free parameters in the routines: these were mostly set to the defaults or to default-like values, e.g. the vector of initial regression parameters was set to 1.0.

- Using proportions rather than counts—a common approach when analyzing text *corpa*—made no difference in the classification accuracy on the full sample; it could still have an effect when balanced samples are used.

Table 8: Top Event Categories by Factor for REBELL

F1	F2	F3	F4	F5
L1gov_sta_vercf	L1gov_gov_vercp	L1gov_gov_vercp	gov_gov_vercf	L1gov_gov_vercp
gov_sta_vercf	gov_gov_matcf	gov_gov_vercp	gov_gov_vercp	L1gov_sta_matcp
L1gov_gov_vercp	gov_ios_vercp	gov_sta_matcp	gov_gov_matcf	L1gov_soc_vercf
L1gov_ios_vercf	L1gov_soc_vercf	L1gov_par_vercf	gov_soc_vercf	L1gov_ios_vercp
L1gov_gov_matcf	gov_sta_matcp	L1gov_gov_matcp	L1gov_soc_vercf	L1gov_gov_vercf
L1gov_sta_matcp	L1gov_gov_matcp	L1gov_gov_matcf	L1gov_gov_matcf	gov_soc_vercf
opp_sta_matcf	L1gov_soc_matcp	L1gov_par_matcp	L1gov_gov_vercf	gov_soc_matcp
F6	**F7**	**F8**	**F9**	**F10**
L1gov_ios_vercp	gov_gov_vercp	L1gov_sta_vercf	gov_gov_vercp	gov_sta_vercf
gov_ios_vercp	L1gov_gov_vercp	gov_ios_vercp	L1gov_gov_vercp	L1gov_sta_matcp
L1gov_opp_vercp	gov_ios_matcf	gov_sta_vercf	gov_sta_vercf	L1gov_sta_vercf
L1gov_opp_matcf	gov_ios_vercp	gov_par_vercf	gov_sta_matcp	gov_gov_vercf
L1gov_opp_vercf	gov_ios_vercf	gov_sta_matcp	gov_gov_vercf	gov_ios_vercp
gov_opp_matcf	L1gov_ios_vercf	L1gov_soc_matcp	L1gov_ios_matcp	gov_gov_matcf
opp_sta_vercp	gov_gov_vercf	gov_soc_vercf	gov_gov_matcp	L1gov_gov_vercf
L1opp_sta_vercp	gov_ios_matcp	L1gov_sta_matcp	L1gov_sta_matcp	L1gov_ios_vercp

- An experiment with reducing the number of variables by eliminating the disproportionately high frequency *gov_gov_vercp* variable and the very low frequency variables that involve interactions with the USA produced no discernible improvements.

5 Conclusion

As noted in the introduction, this is an exploratory proof-of-concept for the use of LDA, rather than the final word on the subject. Based on these preliminary results, is the approach worth pursing further?

Two features suggest that it may be. First, the accuracy and AUC measures are clearly doing better than chance, though not dramatically better. In addition, the *failure* of the sLDA—which actually does worse than chance in some instances—is in some ways reassuring, as it demonstrates that the technique will not fit anything. Second, the dramatic increase in sensitivity compared to the ICEWS reference model is very promising, since sensitivity is a critical issue on rare-events models, though this will have to be confirmed against balanced-sample tests of the reference model, and in out-of-sample testing.

There appears to be little evidence to suggest that the sLDA is worth pursuing further, both with respect to the basic results, the skewed distribution across the random samples, and

the bug, somewhere, in the code.[7]

There are several possible extensions to this approach. First, while the 4-category event aggregation system used here has generally produced good results in forecasting work, LDA is particularly well suited to using completely disaggregated event counts. Document classification problems typically involve vocabularies in the thousands or tens-of-thousands of distinct words, so unlike linear methods, LDA could accommodate a very large set of independent variables. This would also provide a test of whether anything is gained by disaggregation into the detailed categories found in existing event data sets, which in turn has implications for the importance of detailed coding accuracy in automated systems.

Second, logistic regression is not the only available classification algorithm. Support vector machines are commonly combined with LDA in out-of-sample classification; discriminant analysis is another available method and there are a variety of others such as neural networks. My guess is that these will probably produce fairly similar results but some experiments would be in order.

Third, the "events-only" models are a very hard empircal test compared to conflict prediction models generally—notably those of ICEWS and PITF—which typically combine structural information such as GDP/capita, infant mortality rate, democratization and ethnic fractionalization scores with event data, and these indicators are usually necessary to produce $AUC > .80$. In addition, the current scheme combines all of the disparate countries of Asia—from Australia and Japan to Myanmar and Fiji—into a single model, whereas hierarchical or random effects models would probably substantially improve the accuracy. In such approaches, the LDA factors are just one set of information going into the model, rather than the only set of information.

Fourth, Table 8 shows only one set of raw factors, which may or may not be typical, and may or may not be representative of the effect of the method for either data reduction or generating meaningful latent vectors. Two things need to be done to extend this: first, a composite of factors based on multiple balanced samples, and second, some weighting of the factors by the coefficients of the logistic model used for classification. In other words, look at the factors that are actually doing the work of classification, rather than all of the factors.

Finally, the ICEWS data, while extensive, is somewhat idiosyncratic and covers only 14 years. I plan to extend this both to predicting conflict in the 30-year KEDS Levant data set—specifically Israel's conflicts in Palestine and Lebanon—and to looking at the more difficult issue of forecasting onset-cessation models within the ICEWS data, following D'Orazio et al. [2011]

In a history of the first fifteen years of the KEDS/TABARI project (Schrodt 2006), the final section—titled "Mama don't let your babies grow up to be event data analysts" lamented the low visibility of event data analysis in the political science literature despite major advances in automated coding and the acceptance of analyses resulting from that data in every one of the major refereed political science journals.

[7] I would also note that Blei and McAuliffe [2007] has relatively few citations compared to Blei et al. [2003], so sLDA may not be working for much of anything.

The situation at the present is very different, largely due to ICEWS, which emerged about six months after I wrote that history. As far as I know, all three of the teams involved in the first phase of ICEWS used some form of event data in their models, and LM-ATL, the prime contractor for the only team whose models cleared the out-of-sample benchmarks set by ICEWS, invested substantial efforts in TABARI. Lockheed and various subcontractors have continued to invest in additional developments, both for ICEWS and potentially for other projects, and as noted in the previous section, there are now a number of proprietary systems in active development, in contrast to the previous fifteen years which saw only KEDS/TABARI and VRA-Reader. Furthermore, with the experimental extension of the ICEWS event data set to a global level, and the emergence of a number of systems that will be generating event data in real time based on Web sources, the amount and scope of available data sets will be changing substantially

References

James Alt, Gary King, and Curt S. Signorino. Aggregation among binary, count, and duration models: Estimating the same quantities from different levels of data. *Political Analysis*, 9:21–44, 2001.

Edward E. Azar. The conflict and peace data bank (COPDAB) project. *Journal of Conflict Resolution*, 24:143–152, 1980.

R. Bakeman and V. Quera. *Analyzing Interaction: Sequential Analysis with SDIS and GSEQ*. Cambridge University Press, New York, 1995.

David M. Blei and Jon D. McAuliffe. Supervised topic models. *Advances in Neural Information Processing Systems (NIPS 2007)*, 20, 2007. `http://www.cs.princeton.edu/~blei/papers/BleiMcAuliffe2007.pdf`.

David M. Blei, Andrew Ng, and Michael Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003. `http://www.cs.princeton.edu/~blei/papers/BleiNgJordan2003.pdf`.

Doug Bond, J. Craig Jenkins, Charles L. Taylor Taylor, and Kurt Schock. Mapping mass political conflict and civil society: Issues and prospects for the automated development of event data. *Journal of Conflict Resolution*, 41(4):553–579, 1997.

Doug Bond, Joe Bond, Churl Oh, J. Craig Jenkins, and Charles L. Taylor. Integrated data for events analysis (IDEA): An event typology for automated events data development. *Journal of Peace Research*, 40(6):733–745, 2003.

Joe Bond, Vladimir Petroff, Sean O'Brien, and Doug Bond. Forecasting turmoil in Indonesia: An application of hidden Markov models. Presented at the International Studies Association Meetings, Montréal, 2004.

Patrick T. Brandt and John R. Freeman. Advances in Baysian time series modeling and the study of politics: Theory testing, forecasting, and policy analysis. *Political Analysis*, 14: 1–36, 2005.

Jonathan. Chang. Package lda: Collapsed Gibbs sampling methods for topic models. `http://cran.r-project.org/web/packages/lda/`; Version date: 24-October-2010, 2010.

Thomas G.. Dale. Event data analysis and threats from temporal aggregation. Presented at the Florida Political Science Association Meeting, Sarasota, 2002.

John L. Davies, Barbara Harff, and Anne L. Speca. Dynamic data for conflict early warning. In John L. Davies and Ted Robert Gurr, editors, *Preventive Measures: Building Risk Assessment and Crisis Early Warning*, pages 79–94. Rowman and Littlefield, Lanham, MD, 1998.

Vito D'Orazio, Jay Yonamine, and Philip A. Schrodt. Ipredicting intra-state conflict onset: An events data approach using Euclidean and Levenshtein distance measures. Presented at the Midwest Political Science Association, Chicago. Available at `http://eventdata.psu.edu`, 2011.

Robert D. Duval and William R. Thompson. Reconsidering the aggregate relationship between size, economic development, and some types of foreign policy behavior. *American Journal of Political Science*, 24(3):511–525, 1980.

John R. Freeman. Systematic sampling, temporal aggregation, and the study of political relationships. *Political Analysis*, 1:61–98, 1989.

Deborah J. Gerner, Philip A. Schrodt, Ronald A. Francisco, and Judith L. Weddle. The machine coding of events from regional and international sources. *International Studies Quarterly*, 38:91–119, 1994.

Deborah J. Gerner, Philip A. Schrodt, and Ömür Yilmaz. *Conflict and Mediation Event Observations (CAMEO) Codebook*. `http://eventdata.psu.edu/data.dir/cameo.html`, 2009.

Joshua S. Goldstein. A conflict-cooperation scale for WEIS events data. *Journal of Conflict Resolution*, 36:369–385, 1992.

August Hämmerli, Regula Gattiker, and Reto Weyermann. Conflict and cooperation in an actor's network of Chechnya based on event data. *Journal of Conflict Resolution*, 50(159): 159–175, 2006.

Barbara Harff. Early warning of humanitarian crises: Sequential models and the role of accelerators. In John L. Davies and Ted Robert. Gurr, editors, *Preventive Measures: Building Risk Assessment and Crisis Early Warning*, pages 79–94. Rowman and Littlefield, Lanham, MD, 1998.

Barbara Harff and Ted Robert Gurr. Systematic early warning of humanitarian emergencies. *Journal of Peace Research*, 35(5):359–371, 2001.

Valerie Hudson, editor. *Artificial Intelligence and International Politics*. Westview, Boulder, 1991.

Valerie M. Hudson, Philip A. Schrodt, and Ray D. Whitmer. Discrete sequence rule models as a social science methodology: An exploratory analysis of foreign policy rule enactment within Palestinian-Israeli event data. *Foreign Policy Analysis*, 4(2):105–126, 2008.

Craig J. Jenkins and Doug Bond. Conflict carrying capacity, political crisis, and reconstruction. *Journal of Conflict Resolution*, 45(1):3–31, 2001.

Yuen Foong Khong. *Analogies at War*. Princeton University Press, Princeton, 1992.

Gary King and Will Lowe. An automated information extraction tool for international conflict data with performance as good as human coders: A rare events evaluation design. *International Organization*, 57(3):617–642, 2004.

Russell J Leng. *Behavioral Correlates of War, 1816-1975. (ICPSR 8606)*. Inter-University Consortium for Political and Social Research, Ann Arbor, 1987.

E. R. May. *"Lessons" of the Past: The Use and Misuse of History in American Foreign Policy*. Oxford University Press, New York, 1973.

Charles A. McClelland. *World Event/Interaction Survey Codebook (ICPSR 5211)*. Inter-University Consortium for Political and Social Research, Ann Arbor, 1976.

Richard L. Merritt, Robert G. Muncaster, and Dina A. Zinnes, editors. *International Event Data Developments: DDIR Phase II*. University of Michigan Press, Ann Arbor, 1993.

R. E. Neustadt and E. R. May. *Thinking in Time: The Uses of History for Decision Makers*. Free Press, New York, 1986.

Sean O'Brien. Crisis early warning and decision support: Contemporary approaches and thoughts on future research. *International Studies Review*, 12(1):87–104, 2010.

Jon C. Pevehouse and Joshua S. Goldstein. Serbian compliance or defiance in Kosovo? statistical analysis and real-time predictions. *The Journal of Conflict Resolution*, 43(4): 538–546, 1999.

Philip A. Schrodt. Parallel event sequences in international relations. *Political Behavior*, 12 (2):97–123, 1990.

Philip A. Schrodt. Pattern recognition of international crises using hidden Markov models. In Diana Richards, editor, *Political Complexity: Nonlinear Models of Politics*, pages 296–328. University of Michigan Press, Ann Arbor, 2000.

Philip A. Schrodt. *Patterns, Rules and Learning: Computational Models of International Behavior*. `http://eventdata.psu.edu/papers.dir/Schrodt.PRL.2.0.pdf`, 2nd edition, 2004.

Philip A. Schrodt. Forecasting conflict in the Balkans using hidden Markov models. In Robert Trappl, editor, *Programming for Peace: Computer-Aided Methods for International Conflict Resolution and Prevention*, pages 161–184. Kluwer Academic Publishers, Dordrecht, Netherlands, 2006.

Philip A. Schrodt. Inductive event data scaling using item response theory. Presented at the Summer Meeting of the Society of Political Methodology.Available at `http://eventdata.psu.edu`, 2007.

Philip A. Schrodt. Automated production of high-volume, near-real-time political event data. Presented at the American Political Science Association meetings, Washington, 2010.

Philip A. Schrodt and Deborah J. Gerner. Validity assessment of a machine-coded event data set for the Middle East, 1982-1992. *American Journal of Political Science*, 38:825–854, 1994.

Philip A. Schrodt and Deborah J. Gerner. An event data analysis of third-party mediation. *Journal of Conflict Resolution*, 48(3):310–330, 2004.

Philip A. Schrodt, Deborah J. Gerner, Rajaa Abu-Jabr, Ömür Yilmaz, and Erin M. Simpson. Analyzing the dynamics of international mediation processes in the Middle East and Balkans. Presented at the American Political Science Association meetings, San Francisco, 2001.

Robert Shearer. Forecasting Israeli-Palestinian conflict with hidden Markov models. Available at `http://eventdata.psu.edu/papers.dir/Shearer.IP.pdf`, 2006.

Stephen Shellman. Process matters: Conflict and cooperation in sequential government-dissident interactions. *Journal of Conflict Resolution*, 15(4):563–599, 2000.

Stephen Shellman. Time series intervals and statistical inference: The effects of temporal aggregation on event data analysis. *Security Studies*, 12(1):97–104, 2004.

Stephen Shellman and Brandon Stewart. Predicting risk factors associated with forced migration: An early warning model of Haitian flight. *Civil Wars*, 9(2):174–199, 2007.

Tobias Sing, Oliver Sander, Niko Beerenwinkel, and Thomas Lengauer. Package ROCR: Visualizing the performance of scoring classifiers. Available at `http://cran.r-project.org/web/packages/ROCR/`; Version dated 08-Dec-2009, 2009.

Michael D. Ward, Brian D. Greenhill, and Kristin M. Bakke. The perils of policy by p-value: Predicting civil conflicts. *Journal of Peace Research*, 47(5), 2010.