Reading Between the Lines: Prediction of Political Violence Using Newspaper Text

Hannes Mueller
Christopher Rauh

September 2017

Barcelona GSE Working Paper Series

Working Paper nº 990

Reading Between the Lines: Prediction of Political Violence Using Newspaper Text*

Hannes Mueller Christopher Rauh

September 19, 2017

Abstract

This article provides a new methodology to predict armed conflict by using newspaper text. Through machine learning, vast quantities of newspaper text are reduced to interpretable topics. These topics are then used in panel regressions to predict the onset of conflict. We propose the use of the within-country variation of these topics to predict the timing of conflict. This allows us to avoid the tendency of predicting conflict only in countries where it occurred before. We show that the within-country variation of topics is a good predictor of conflict and becomes particularly useful when risk in previously peaceful countries arises. Two aspects seem to be responsible for these features. Topics provide depth because they consist of changing, long lists of terms which makes them able to capture the changing context of conflict. At the same time topics provide width because they are summaries of the full text, including stabilizing factors.

*Mueller (corresponding author): Ramon y Cajal Researcher at IAE (CSIC), Barcelona GSE (email: h.mueller.uni@gmail.com); Rauh: Assistant Professor at University of Montreal (email: christopher.raphael.rauh@umontreal.ca). We thank Tim Besley, Melissa Dell, Vincenzo Galasso, Hector Galindo, Matt Gentzkow, Stephen Hansen, Ethan Kapstein, Daniel Ohayon, Akash Raja, Bernhard Reinsberg, Anand Shrivastava, Ron Smith, Jack Willis, Stephane Wolton and the participants of the workshops and conferences ENCoRe Barcelona, Political Economy Cambridge (internal), EPCS Freiburg, ESOC in Washington, Barcelona GSE Calvo-Armengol, NBER SI Economics of National Security, Conflict at IGIER, and the seminars PSPE at LSE, BBE at WZB, and Macro Lunch Cambridge for valuable feedback. We are grateful to Alex Angelini, Lavinia Piemontese, and Bruno Conte Leite for excellent research assistance. We thank the Barcelona GSE under the Severo Ochoa Programme for financial assistance. All errors are ours.

1 Introduction

The conflict literature has made significant progress in understanding which countries are more at risk of suffering an armed conflict.[1] However, many factors that have been identified as leading to increased risk, like mountainous terrain or ethnic polarization, are time invariant or very slow-moving, and therefore not useful in predicting the timing of conflict. Other factors, like GDP levels or political institutions, still vary more between countries than within countries over time. This means it is easier to predict whether a country is at risk *in general* rather than *when* a country is particularly at risk. Yet, understanding the timing of conflict is critical for policy.

An additional problem of forecasting the timing of armed conflict is that it is rare and at the same time relatively concentrated in some countries. This is problematic because it implies that the variation between countries can dominate the analysis unless the between- and within-country variation are separated explicitly. Empirical models that are overall quite accurate can therefore be of little use on the time dimension. We show, using a simple panel regression model, that many variables commonly used in the literature indeed face this problem. This means they predict conflict where it occurred before, and therefore fail to predict conflicts in previously peaceful countries.

As a solution to this problem, we propose data generated from news sources. To this end, we implement an automated method to quantify the content of news using the latent Dirichlet allocation (LDA) model (Blei, Ng and Jordan 2003), which we apply to over 700,000 newspaper articles from English-speaking newspapers. There are two advantages that topics have over existing methods of analyzing text. First, topics provide depth because, by design, they put words into context. The context can be useful for forecasting. Second, topics provide width because they allow us to use the whole text, including stabilizing factors, when forecasting conflict. This means we can let the data speak without losing interpretability of the results.

At the prediction stage, we rely on a simple panel regression model, which uses all generated topics as explanatory variables. The result is a model able to forecast out-of-sample the onset of civil war, armed conflict, and even movements of refugees a year before they occur. It relies entirely on news text and can therefore provide forecasts without the need to extrapolate or wait for other data sources. Furthermore, the procedure can be implemented with only minimal personal judgement

[1] See, for example, Goldstone et al. (2010), Fearon and Laitin (2003), Esteban, Mayoral and Ray (2012), Besley and Persson (2011*a*). See Blattman and Miguel (2010) for a summary of the literature.

and appears to generate consistent summaries of text, which could be used in other applications as well.

Our empirical methodology proceeds in three steps. We first download newspaper articles and collect words and series of words, referred to as tokens, in one vector for each article. Newspapers have the advantage that they stretch several decades and report on events in all countries, which means that even rare events are sufficiently common to be analyzed with quantitative methods. We downloaded all articles on 185 countries from the New York Times, the Washington Post, and the Economist for all available years since 1975. This gives us a basis of 700,000 newspaper articles with a little less than one million unique word combinations, even after excluding stop words, rare words, and stemming.

As a second step, we develop a topic model tailored for the purpose of summarizing the content of news reports in a country and year. We use the LDA model to generate quantitative summaries of the articles. In this way, the high dimensionality of token vectors (0.9 million) can be decreased to as many topics as we choose. The main advantage of this methodology is that we do not need to impose any judgement on which part of the text is important when predicting conflict - we can let the data speak.

As the final step, we use the within-country variation, i.e. the emergence and disappearance of topics on the country level, to predict conflict out-of-sample. For this step, we calculate the share of words written on each topic in every country and year. We then use these topic shares in a country fixed effects regression to predict the onset of conflict in the following year. We show that reporting on specific topics increases before conflict, whereas reporting on other specific topics decreases. In this way the timing of conflict can be forecasted more accurately than through variables previously used by the literature. In addition, topics are meaningful and can therefore help us understand predictors of (in)stability.

We show that forecasts relying on the overall variation, even if they were estimated with fixed effects, introduce a bias against new onsets of conflict in previously peaceful countries. Methods which rely on the overall variation will therefore tend to attribute low risk to countries which have not experienced an onset in the past - even if the within-country variation would indicate high risk. Since our topics provide a lot of useful within-country variation, they can provide early warning for countries which did not experience a conflict onset in the sample used to train the model. This

is an important difference to standard methods.

We use a stepwise selection method to explore why the estimated topics provide such strong predictive power. First, topics rely on a long list of terms that co-occur. The algorithm that generates topics learns, for example, that specific non-conflict words are associated with conflict words. Topics with a conflict content can then add forecasting power beyond conflict indicators, conflict events, and conflict keyword counts. Second, the model uses several negative associations between topics and conflict. A lot of the forecasting power is maintained even if only these non-conflict topics are used for the forecast. The fact that our forecast relies so heavily on negative correlations means that (the absence of) stabilizing factors could be key to understanding the timing of conflict, even in new outbreaks. We find, for example, that news which describe judicial procedures systematically decrease before conflict occurrs.

We proceed as follows. We first discuss related literature in Section 2. In Section 3 we present a way to evaluate the ability of a model to forecast the timing of conflict. Section 4 presents our methodology of aggregating news text into topics. Section 5 presents the main results. Secton 6 demonstrates the close link between predicting the timing of conflict and predicting new, otherwise unforeseen conflicts. In Section 7 we explore why the topic model provides such useful forecasts of the timing. Section 8 concludes.

2 Related Literature

The academic literature has made large strides towards understanding the triggers of civil conflict. A part of the literature has focused on establishing links to specific factors such as ethnic cleavages (Reynal-Querol and Montalvo 2005, Esteban, Mayoral and Ray 2012, Caselli and Coleman 2013), climate (Miguel, Satyanath and Sergenti 2004, Dell, Jones and Olken 2012, Buhaug et al. 2014) or natural resources (Brückner and Ciccone 2010, Bazzi and Blattman 2014). This literature is more concerned with causal identification and less with forecasting power. Another part of the literature has looked at using a mix of political and economic indicators to explain conflict. Examples are Fearon and Laitin (2003), Collier and Hoeffler (2004), Collier et al. (2009), Gleditsch and Ruggeri (2010), or Besley and Persson (2011*a*). For a review of this literature see Blattman and Miguel (2010).

Naturally, the forecasting literature started with relying on structural factors in forecasting.[2] An exhaustive review of this growing literature is beyond the scope of this paper. For an overview see Schrodt, Yonamine and Bagozzi (2013), Ward et al. (2013), and Hegre et al. (2017). In what follows, we will therefore focus on the use of country fixed effects and newspaper text in forecasting.

Country fixed effects are typically absent in the forecasting literature which explains why slow-moving, structural variables typically play such an important role in forecasting. Rost, Schneider and Kleibl (2009), for example, use cross-sectional logit regressions on economic and political variables as well as proxies for violations of human rights to predict conflict onset within a 5-year window. They find substantial predictive power of their model within this time frame. Goldstone et al. (2010) provide predictions of political instability at the country level within a two-year horizon. Their statistical method compares country-years before instability to country-years in the same region that were not followed by onset. Their main finding is that the best predictors of instability are slow-moving variables such as political institutions or infant mortality. Hegre et al. (2013) forecast conflict for the period 2010-2050 using a combination of variables such as population, infant mortality, and education.

News text has been used extensively to predict conflict. Brandt, Freeman and Schrodt (2011), for example, use the Conflict and Mediation Event Observations (CAMEO) coding scheme in their analysis of news sources when defining conflict events in the Levant. CAMEO uses dictionaries of verbs and actors developed over a decade in several large research projects to identify events and the involved parties in these events.[3] Many modern applications such as the Global Data on Events, Location and Tone (GDELT) or Integrated Conflict Early Warning System (ICEWS) rely on such coding rules to automatically extract events from text in real-time. Most closely related to our paper is Ward et al. (2013) who use a combination of event data based on ICEWS. Their model has a striking degree of accuracy in predicting civil war incidence (occurrence) and performs well out-of-sample. A simpler way to forecast conflict with text is proposed by Chadefaux (2014) who relies on keyword counts of a list of predetermined words to construct an index of tension on a weekly basis for the period 1902 to 2001. He uses the constructed tension data to predict onset of conflict weeks before it occurs and shows that news data can contribute significantly to a standard

[2]However, Ward, Greenhill and Bakke (2010) demonstrate that focusing on statistically significant relationships does not necessarily contribute to the prediction of conflict.

[3]For an introduction see Gerner et al. (2002) and Schrodt, Gerner and Yilmaz (2009).

model. Most recently, Chiba and Gleditsch (2017) combine structural and event data to forecast civil war on the monthly level in a logit framework without fixed effects. A lot of their forecasting power comes from constant slow-moving variables such as ethnic fractionalization, GDP per capita, or population. The within- and out-of-sample gains of adding conflict events is relatively modest. This is an interesting contradiction to the findings in Ward et al. (2013) which might be explained by use of country fixed effects in the latter.

We add to the forecasting literature in two ways. First, an important conceptual contribution of this project is that we explicitly separate the within-country from the between-country variation before forecasting. We do this by running linear fixed effects regressions and then relying on the within model to forecast. To the best of our knowledge, no paper in the existing literature has tried to separate within and between variation this way.[4] Second, we use a topic model to automatically summarize all news text in a few variables. Our forecasting model can therefore rely on the complexity of the entire newspaper text written on each country and we demonstrate that this has some benefits.

However, our approach comes at the cost of requiring the entire text of articles instead of relying on search queries only. In addition, we try to forecast rare events which implies that we need news sources that are consistently available for decades. We use three newspapers which leave us with a little more than 700,000 articles and prevents us from trying to predict conflict at the quarterly or even monthly level. We therefore see our approach as complementary to Ward et al. (2013) and Chadefaux (2014). Ward et al. (2013) use event data constructed from more than 30 million news stories while Chadefaux (2014) searches keywords in over 60 million pages of news text.

Quinn et al. (2010) use a topic model, in which documents are assigned only a single topic, to categorize over 100,000 speeches in the US congress. They estimate their topic model of 42 topics to show that topics can be used to analyze democratic agenda dynamics over a long time period. Topics generated by LDA have also been used by Hansen, McMahon and Prat (2014) to quantify discussions in the central bank committee of the Bank of England. We contribute to this literature by applying the topic model to newspaper text which spans large cross-country panels.

[4] Chadefaux (2017*b*) differs from our approach but heads in a similar direction. He compares the forecasting power of a model with just country fixed effects to an augmented model with asset prices and fixed effects.

3 Forecasting the Timing of Conflict

In this section, we present a method to evaluate the ability of a model to forecast the timing of conflict. We do this in three steps. First, we explain the basic problem of forecasting the timing in a stylized example. Second, we present a method to circumvent this problem. Third, we use this method to evaluate five empirical models of conflict taken from the literature.

3.1 The Problem of Forecasting the Timing of Conflict

It is a well-known fact in the conflict literature that rich countries with strong political institutions are less likely to enter conflict than poor countries with weak institutions. A simple way to forecast conflict would therefore be to attribute a higher risk of conflict to poor countries with weak institutions. However, to make conclusions from such a model about how marginal changes in income or institutions affect stability requires a leap of faith. The required leap of faith is that the variation between countries, the *between variation*, is useful to forecast the variation within countries across time, the *within variation*. But this is not always true.

As a stylized example take the standard democracy score from Polity IV ($polity2$), which is meant to capture the level of democratization in a country. If one runs a regression of civil war onset in year $t+1$ on $polity2$ in year t one gets a significant, negative relationship. Countries that are more democratic tend to be less at risk of experiencing a conflict the following year. However, if one uses this estimated relationship to look *within* countries one gets a falling likelihood of conflict as conflict approaches. In other words, the overall model marks countries as relatively safe right before they experience an outbreak of armed conflict.[5]

This example only serves as an illustration. The literature typically adds controls and uses a non-monotonous relationship between civil war and democracy.[6] But the problem could be relevant in more sophisticated applications as well. For example, a commonly used fragility index, the Fragility Index of the Fund for Peace, relies on many more factors but was still falling in Syria, Libya and Egypt right before these countries experienced dramatic outbursts of violence in 2011.

A textbook solution to this problem is the use of country fixed effects but we will show in this

[5]The reason is that in simple regressions with one variable the democracy score and conflict have a negative relationship between countries but a positive relationship within countries.

[6]See, for example, Fearon and Laitin (2003) and Goldstone et al. (2010). Political institutions used this way can add useful forecasting power.

section that when forecasting out-of-sample, the problem goes further. Even if a fixed effect model is used, the within variation might not contribute to the forecast if the estimated fixed effects are included. Relative to the magnitude of the fixed effect, the within variation is relatively subtle and so the "signal" contained in it is not visible when forecasting out of sample. Accordingly, it might happen to be impossible to predict new developments, i.e. onsets in previously peaceful countries or stability in violent countries.[7]

3.2 Out-of-sample Evaluation of Forecasts

We now present a method to evaluate the ability of a model to forecast the onset of conflict. Our starting point is the perspective a policymaker could have. On December 31st of year T the policymaker is interested in where conflict might break out in the following year. This procedure has two steps; first, the policymaker trains a model using all information available up until year T; second, once the training is completed, the policymaker uses the trained model and data in T to produce forecasts for $T+1$.

In the first step we distinguish between the *overall* model, which contains the entire model including the estimated fixed effects, and the *within* model, which disregards the baseline risk contained in the fixed effects. For example, to test whether the democracy index $polity2_{it}$ contains useful within variation for a forecast we would run a regression of the form

$$y_{it+1} = \alpha + \beta_i + polity2_{it} \cdot \beta^{FE} + \varepsilon_{it},$$

where y_{it+1} is conflict onset in $t+1$ and β_i are a full set of country dummies.[8] The linear fixed effects model allows us to separate the between variation contained in the β_i from the useful time variation contained in the within fitted values, $polity2_{it} \cdot \hat{\beta}^{FE}$.

In the second step, the two sets of fitted values in T are used to produce forecasts for $T+1$. The fitted values are converted into a binary forecast, i.e. negative for peace or positive for onset of conflict, depending on whether the fitted value is above a cutoff c or not. These negatives and positives can then be evaluated with the actual realizations of onset.

[7] We will demonstrate this in Section 6.

[8] The newest onset in the training sample is in $t+1=T$. This means that the newest information on $polity2_{it}$ used in training is from $T-1$.

These two steps are repeated every year to get an impression of the ability of the model to forecast conflict. In our main application we let T go from 1995 to 2013 and implement the above steps for all years in between. This means we first predict the onset of conflict in 1996 with information available in 1995. We then predict onset in 1997 with all information available in 1996 and so forth. In order to evaluate the forecasting power of our model we collect all out-of-sample predictions and pool them for evaluation. This makes sure that the evaluation is conducted with a common cutoff which could therefore also be used to forecast an unknown future, based on the performance in the past.

The key trade-off that a policymaker faces when deciding on the right cutoff c is between false negatives, outbreaks of conflict with a negative forecast, and false positives, positive forecasts without onset. If a policymaker chooses a high cutoff there will be fewer positives and therefore also less false positives. However, the forecast will miss more actual outbreaks of conflict, i.e. it will generate more false negatives. If the policymaker instead chooses to warn in all countries then she will be able to have no false negatives but a lot of false positives.

This trade-off can be shown in receiver operating characteristic (ROC) curves which visualize the performance of the model for all possible cutoffs c. In the figures containing ROC curves (e.g., Figure 1), we report the true positive rate (TPR) on the y-axis. The true positive rate is given by the formula

$$TPR_c = \frac{TP_c}{FN_c + TP_c}$$

and is a measure of how many false negatives (FN_c) are generated for a given level of true positives (TP_c). We want this measure to be as close to 1 as possible, which would indicate that all conflict onsets have been spotted correctly without missing a single one, i.e. with $FN_c = 0$.

The false positive rate (FPR) is reported on the x-axis of Figure 1. It is given by the formula

$$FPR_c = \frac{FP_c}{FP_c + TN_c}$$

and is a measure of how many false positives (FP_c) are generated for a given level of true negatives (TN_c). We want this to be as low as possible. Optimally, we would want no false warnings, i.e. $FP_c = 0$. The 45-degree line in ROC curves is the benchmark that would be reached by random

forecasts. For each ROC curve we also report the area under the curve (AUC).

There are some problems associated with the linear fixed effects model and a broad academic debate has brought pro- and counter-arguments for its adoption.[9] We nonetheless choose the linear model for several reasons. First, because we are forecasting, we are not interested in the precision or even size of the estimated coefficients and due to our focus on ranking forecasts, we do not mind that the fitted values from the model are not bound between 0 and 1. Instead, we can simply rely on showing that our model is able to produce useful rankings when forecasting out-of-sample. This deflates both of the main arguments against the use of a linear model.

The key in our application is, however, that within and between variation are additive in the linear model so that we can use the within fitted values alone in forecasting. This is a crucial difference to the fixed effect logit model, for example. The logit model estimates a set of fixed effects but drops all time series which have no variation in y_{it+1}, i.e. all countries which were always or never in conflict. In addition, it is unclear how one would separate the within variation contained in a logit model to use them separately in forecasting.[10]

3.3 An Evaluation of Standard Models

We now illustrate that isolating the within variation from the between variation (which together form the overall model) can yield fundamental insights regarding how precisely a model predicts the timing of conflict out-of-sample. We do this by applying the method described in the previous section to five different models from recent publications in economics and political science shown to explain conflict onset and in some cases even forecast it. We discuss the details of the five models in the Appendix C.[11]

First, we use a model of rainfall shocks in Africa. Second, we use foreign aid shocks and income shocks interacted with the country's institutional environment. Third, we use a combination of standard economic and political variables. Fourth, we use a mixture of Integrated Crisis Early Warning System (ICEWS) event data and economic and political data. Fifth, we use a model of conflict word counts based on our articles together with measures of conflict history and political

[9] For a summary see Beck (2015).

[10] In Appendix Section E.7 we discuss the logit model and show in Figure E.18 that topics maintain their superior predictive power when using logit.

[11] The five models we recreate are published in Miguel and Satyanath (2011), Besley and Persson (2011*b*), Goldstone et al. (2010), Ward et al. (2013) and Chadefaux (2014). Wherever available, we use the respective replication datasets.

institutions. Finally, we add a model in which we regress conflict on country fixed effects only, i.e. we do not even try to predict the timing of conflict but rely instead on the simple logic that onset will occur where it occurred before. This model should, by definition, have no useful within variation.

In all six models, we predict armed conflict and civil war onset as measured by battle-related deaths from the Uppsala Conflict Data Program (UCDP/PRIO).[12] This includes all battle-related deaths which took place in armed conflict. The UCDP defines an armed conflict as a contested incompatibility that concerns government and/or territory over which the use of armed force between two parties, of which at least one is the government of a state, has resulted in at least 25 battle-related deaths in one calendar year. It also gives four types of conflict - we include battle-related deaths that occurred during internal and internationalized internal armed conflict.[13]

As discussed in the previous section we evaluate the forecasting performance of the different models using ROC curves. The blue ROC curves in Figure 1 show the performance of the respective overall model which is what is typically reported in the forecasting literature. To illustrate, take the forecast of civil war of the "Economics and Politics" overall model in panel (a). It reaches a true positive rate of around 70 percent for a false positive rate of 20 percent when predicting civil war. This means that the model correctly forecasts 70 percent of all civil war outbreaks while only marking 20 percent of peaceful years as under conflict risk. The area under the curve (AUC) is almost 0.8 which is comparable to the findings in the literature. When predicting armed conflict in panel (b) the model does a little worse. The blue ROC curve stretches out more in panel (a) than in panel (b) and the reported AUC is higher. Generally, forecasting civil wars in panel (a) seems to be easier than forecasting armed conflict in panel (b).

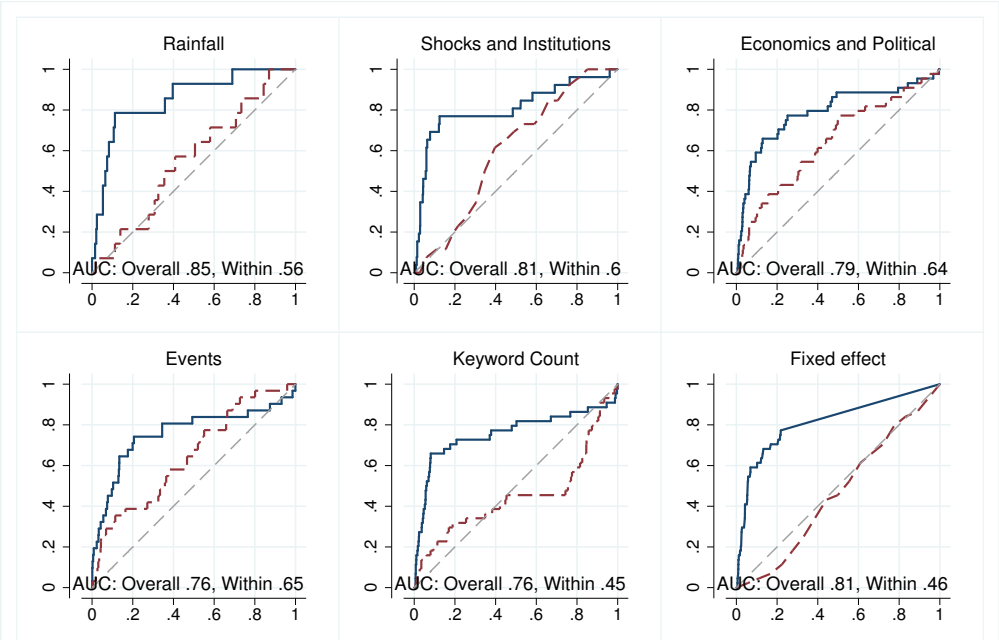
We then test how much of this forecasting power is maintained when looking at the within variation alone. The performance of this forecast is reported as a red dashed line in Figure 1. Almost all models show a dramatic decline in forecasting power from the overall model to the within model. This indicates that most overall models receive their forecasting power from the between variation. Notable exceptions are the model based on political and economic variables,

[12]Coding of this data is based on Pettersson and Wallensteen (2015) and Gleditsch et al. (2002). See Sambanis (2004) for a discussion of conflict data generally.

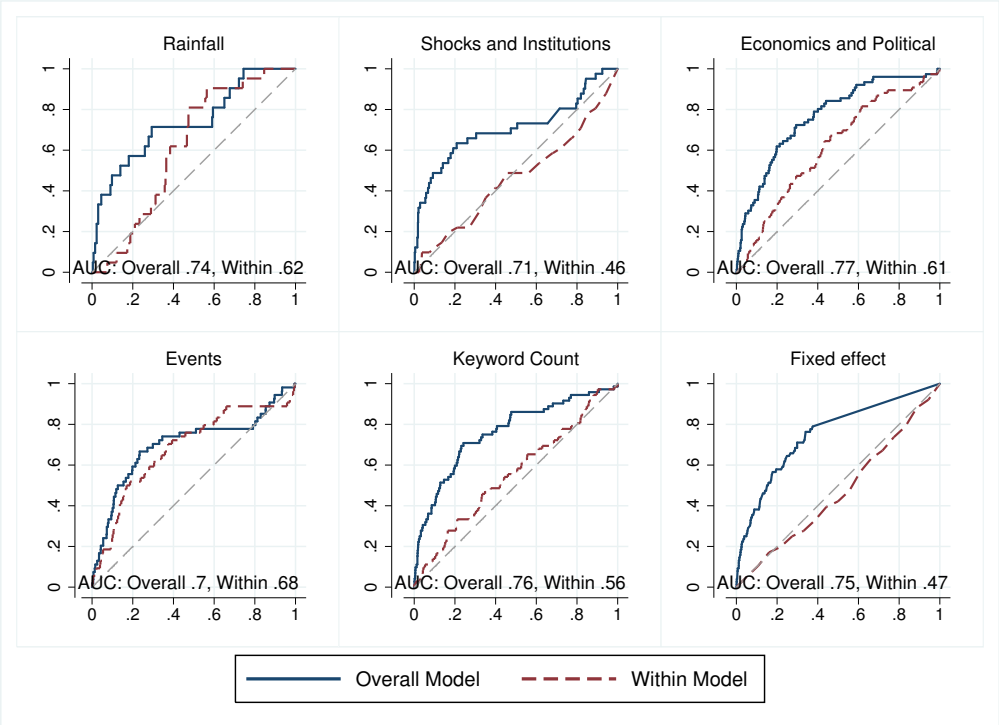
[13]All recent casualties in Afghanistan are, for example, coded as stemming from an internationalized internal conflict. We ran extensive robustness checks regarding our definition. For a detailed discussion see the Appendix.

Figure 1: ROC Curves for Onset (X-Axis: FPR, Y-Axis: TPR)

(a) Civil War



(b) Armed Conflict



Notes: Predictions result from a panel estimated as in equation (2). The variables included for each model as $\mathbf{x}_{it}$ are specified in Section 3.3 and Appendix Table C.1. The within model is the overall model net of country fixed effects as presented in equation (3).

which always retains at least some forecasting power, and the model based on events which retains almost all of its forecasting power when predicting armed conflict. The latter finding suggests that using news text can provide a useful source of variation when predicting the timing of conflict.

Note also that the model with only country fixed effects performs as well as most other overall models. This confirms that it is possible to gain fairly good forecasts by assuming that conflict breaks out where it broke out before. Howver, in many applications this is not satisfactory and leads to forecasting mistakes if peaceful countries destabilize or violent countries stabilize.

4 A Topic Model of Newspaper Text

This section discusses the news reports we rely on and how they can be summarized with the help of a topic model. We also report on the content of the estimated topics.

4.1 News Text

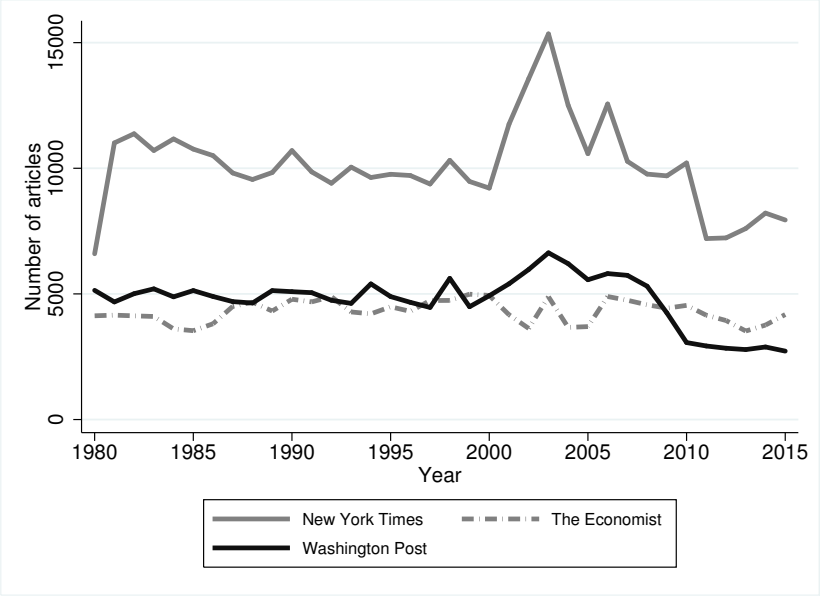
The first choice we face is the selection of our news sources. Due to their availability over a long time span and international coverage, we focus on three major newspapers published in English, namely the Economist (available from 1975), the New York Times (NYT) (available from 1980), and the Washington Post (WP) (available from 1977). From the database LexisNexis we downloaded all articles dating from January 1975 to December 2015 containing country names (or slight permutations thereof) or capital names in the title.[14] In total, we downloaded more than 700,000 articles, of which 174,450 are from the Economist, 363,275 from the NYT, and 185,523 from the WP.

On average about 120 articles are written on a country in a given year. However, the extent of coverage varies drastically so that we observe between 1 and more than 5,500 articles in a given year. As a general idea, more populous, richer and more democratic countries are covered more. In addition, coverage increases in and before conflict. On average, a conflict year is covered with about 100 articles more, while a pre-conflict year is covered with almost 70 articles more than the average year. However, our methodology accounts for changes in coverage by using topic shares, i.e.

[14] In the case of the Economist we also search in the leading paragraph as the title rarely contains a country or capital name. Searching and downloading articles was conducted manually in accordance with LexisNexis terms and conditions.

we disregard how much is written on a country and focus instead on what is written on a country, as it facilitates forecasting across countries.

Figure 2: Number of articles by news source over time



In order to improve the performance of our machine learning algorithm, we process the raw texts of articles of all three newspapers according to standard text mining procedures.[15] First, we remove a library of common words, which in text mining are referred to as stop words, such as "to" or "that". Second, we lemmatize and then stem words using the Snowball algorithm, which is an updated version of the algorithm from Porter (1980).[16] Lemmatizing groups distinct forms of the same word into one word, while stemming attempts to harmonize different usages of one word, such that, e.g. "running", "ran", and "run" all become "run". However, unlike the example, the outcome does not necessarily represent an English word. This leaves us with more than 5.5 million unique tokens, which are not only single words, but also tokens of sequences of two words and three words, referred to as bigrams and trigrams, respectively. Then as a final step, we remove overly frequent and rare tokens, and are left with around 0.9 million tokens. This high dimensionality makes it impossible to use the token vectors in standard regressions. Here is where the literature has typically reduced dimensionality by focusing on particular words.

[15] An example of the following steps is presented in Appendix G.1.

[16] The Python package for lemmatizing is available at `http://www.nltk.org/` and for stemming at `http://snowball.tartarus.org`.

4.2 LDA Topic Models

In order to reduce the high dimensionality of our data set, we use the latent Dirichlet allocation (LDA) to model topics, a method introduced by Blei, Ng and Jordan (2003). Topics are probability distributions over words. The LDA model in text analysis assumes that each document is a mixture of a small number of topics and that each word's creation is attributable to one of the document's topics.

The exercise consists in splitting each article into topics k. One can imagine a journalist writing about a topic will use a combination of words related to that topic. For instance, an article about sports might be more likely to contain words such as "football", "win", "fans", and "game", whereas an article about a conflict might be more likely to use words such as "violence", "casualties", and "soldier". Through Bayesian learning, the algorithm optimizes the weighted word lists, i.e. the topics, in order to discriminate between articles. For instance, the word "game" might be more of a sports-topic word and, therefore, indicates that an article is on sports. Ultimately, the mixed-membership model represents each document as a set of shares of topics. One could imagine that an article is classified as 70 percent sports and 30 percent conflict if a particularly violent soccer match took place. While the number of topics K is pre-specified, the content of the topics is not. The topics are identified by looking at which tokens co-occur in articles.

The LDA model requires just three parameter assumptions and can be implemented with a Gibbs sampling technique which we adapt from Phan and Nguyen (2007). In Appendix D we provide a technical discussion of the LDA and the Gibbs sampler. The parameters to choose are α, β, and the number of topics K. High values of α imply that each article is likely to consist of a mix of many topics. Analogously, a high value of β favours a topic to contain a mixture of most words, whereas low values allow topics to consist of a limited number of prominent words. Concerning α and β we follow the literature. Our preferred specifications, which we will be using for all of the baseline results presented in Section 5, is composed of 15 topics and hyperparameters $\alpha = 3.33$ and $\beta = 0.01$.[17]

[17] We estimate models for 5, 50, and 70 topics for robustness checks discussed in Appendix E. The estimated topics of course become more specific with the number of topics.

4.3 Topic Estimation Results

In order to be able to use the estimated topics for out-of-sample forecasting we need to estimate the model for each sample of text ending in year T. We start training our model on text until $T = 1995$ and apply it out-of-sample to predict onset in $T+1 = 1996$ so that the first topic model we estimate uses all articles between 1975 and 1995. We estimate one topic model for each consecutive year $T \in \{1995, 1996, ...2013\}$, where the last model uses all text from 1975 up to 2013 to predict the last conflict available from UCDP in 2014.[18]

Remember that our K topics are probability distributions over 0.9 million terms. Typically, around 90 percent of the probability mass in each topic is concentrated in the 10,000 most likely terms and in Appendix G.1 we show that, while some common terms are shared, each of the topics relies on a different part of the 0.9 million terms. When we look at the most common terms in each topic it is fairly easy to come up with a title for a topic, i.e. in our application to news content the K estimated topics seem natural and intuitive. For example, in all years topics appear which we can classify as conflict, sports, tourism, politics and the economy. By relying on different, large groups of terms, topics provide *width* for the forecasting exercise. At the same time, the topic model ensures that we can analyze which topics are most useful for the forecast.

In Figure 3, we present four topics when $T = 2013$, i.e. all text from 1975 to 2013, as word clouds of the top 50 terms of the topic. In these clouds, the size of each word is proportional to its likelihood within the corresponding topic. Notice that words are stemmed/lemmatized versions so that "armi", for example, stands for "army" and "armies". The two clouds in Figures 3a and 3b suggest (potential) violence. Terms like "force" and "military" indicate as much. We therefore call these conflict topics. Figure 3c seems to summarize processes in the judicial system, indicated by words such as "court" and "case". The topic in Figure 3d seems to describe economics. We refer to topics like these as non-conflict topics.

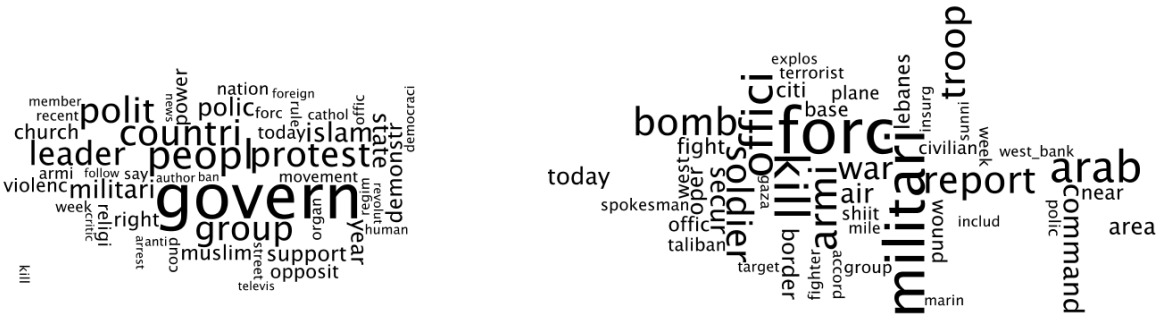
It is important to keep in mind that the tokens shown in these word clouds are only the tip of the iceberg. Topics are a probability distribution over thousands of tokens. The full list of terms associated with the topics in Figure 3, for example, could capture factors that trigger or at least anticipate conflict. In this sense topics have *depth* which could also be useful for forecasting.

[18]The restriction is actually the conflict data. We have text written until 2015 and used it to predict 2016/2017. Published working paper versions contain these forecasts.

Figure 3: Word Clouds of Topics

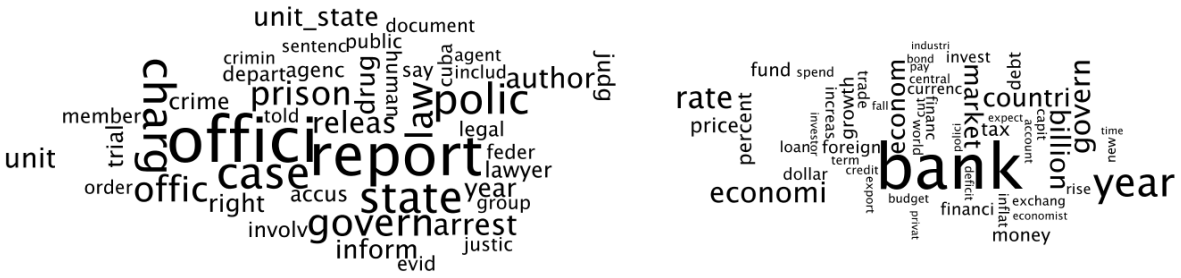
(a) Conflict 1

(b) Conflict 2



(c) Justice

(d) Economics



Notes: These are the top 50 words of four out of 15 topics computed using LDA with $\alpha = 3.33$ and $\beta = 0.01$ for the entire sample until 2013. The size of a term represents its probability within a given topic. The position conveys no information. A list of the 15 topics is exhibited in Appendix Table I.1.

The algorithm uses this depth by learning in each sample which terms are associated with each other and therefore form a topic. As we move through time, new aspects will be associated with, for example, *conflict2* through the co-occurrence that the Gibbs sampler uses to build the topic. In Table 1 we show the change in the top 50 terms for the *conflict2* topic between 1995 and 2015.[19] In column (1) we show the tokens which appear in the top 50 list in both years. These are mostly generic conflict tokens like "force", "attack", "army", "war", "soldier" or "guerrilla". In columns (2) and (3) we show the tokens that only appear in the years 1995 and 2015, respectively. In the year 1995 the terms "unit_nation", "serb", "libanes" and "gulf" were associated with conflict. In the year 2015 these terms are replaced by terms like "terrorist", "insurg", "shiit" and "sunni". From this change it becomes clear how the *conflict2* topic has adapted to the new international

[19]It is important to stress here that the preservation of the identity of topics is not important for the forecasting exercise as topics enter the regressions anonymously, i.e. as topics 1, 2, 3, ... 14. The name *conflict2* is simply a name we give to two similar probability distributions that appear both in 1995 and 2015.

context and surge in asymmetric armed conflicts.[20] This is an advantage of using a machine learning algorithm to summarize text. Topics allow the specific vocabulary of some countries and events to be put into a broader context.

Table 1: Word list of topic *conflict2* - 1995 vs 2015

Both years	Only 1995	Only 2015
forc	unit	bomb
militari	serb	american
attack	nation	group
armi	unit_nation	islam
kill	lebanes	secur
troop	defens	peopl
soldier	mile	polic
offici	gulf	citi
war	weapon	wound
report	aircraft	shiit
fight	missil	taliban
command	ship	milit
govern	plane	insurg
rebel	use	leader
guerrilla	christian	men
civilian	tank	terrorist
arm	town	violenc
border	western	northern
area	peac	capit
oper	say	sunni
offic		
base		
air		
near		
southern		
fighter		
muslim		
today		
control		
week		

Note: Table shows the 50 most likely words in the *conflict2* topic in 1995 and 2015. Words are ordered by prominence within topic.

[20] Kalyvas and Balcells (2010) stress how important the international context is for explaining the character and strategies used in armed conflict.

After estimating the topic model, we possess data on the composition of each article m in terms of the K topics, η_m. We aggregate the shares in each article to receive a topic distribution in a country-year, while taking into account the prior probability distribution of topics in the Dirichlet distribution. Call M_{it} the group of articles written in country i and year t. The $k \times 1$ vector of topic shares in country i in year t is then

$$\theta_{it} = (\sum\nolimits_{m \in M_{it}} \eta_m N_m + \alpha) / \left(\sum\nolimits_{m \in M_{it}} N_m + K\alpha \right) \tag{1}$$

where $\sum_{m \in M_{it}} N_m$ is simply the total number of articles. Note that α enters here as the strength of the prior. If only few words are written in a country-year then the deviation from this prior will be relatively weak. In order to use our estimates when forecasting out-of-sample, we estimate a full panel of topic shares θ_{it} for each sample ending in year $T \in \{1995, 1996, ...2013\}$ separately.

Our use of topic shares alleviates some of the critizism in the literature regarding the use of news as a source of data.[21] Indeed, as in most other studies of conflict, our left-hand-side variable is based on events which are partly reported by news agencies. However, our predictors rely on content rather than the quantity of reporting. We show in the next section that changes in content can predict changes in reported violence out-of-sample. Moreover, in order to illustrate that we are not merely picking up news biases, we also show that we can predict refugee movements, which are collected and reported by local agents directly to the United Nations High Commissioner for Refugees (UNHCR).

[21] See, for example, Woolley (2000) and Weidmann (2016).

5 Predicting Conflict with Newspaper Topics

In this section, we combine the forecast evaluation from Section 3.2 with the topic model from section 4.3. In each year T between 1995 and 2013 we use the text written up until year T to predict conflict in $T+1$. We then draw ROC curves to evaluate our forecasts.

In each step, we first estimate a topic model which uses the text written between year 1975 and year T. From the topic model we obtain the vector of 15 topic shares θ_{it} in country i at time t, which we calculate as in equation (1). We then use these shares to train our model with conflicts that happened before $T+1$ to predict outbreaks in $T+1$.

Formally, we use the trained parameters $\hat{\alpha}, \hat{\beta}_i$ and $\hat{\beta}^{topics}$ to calculate two sets of fitted values for year $T+1$. The overall fitted values

$$\hat{y}_{iT+1}^{overall} = \hat{\alpha} + \hat{\beta}_i + \theta_{iT}\hat{\beta}^{topics} \tag{2}$$

and the within fitted values

$$\hat{y}_{iT+1}^{within} = \hat{\alpha} + \theta_{iT}\hat{\beta}^{topics}. \tag{3}$$

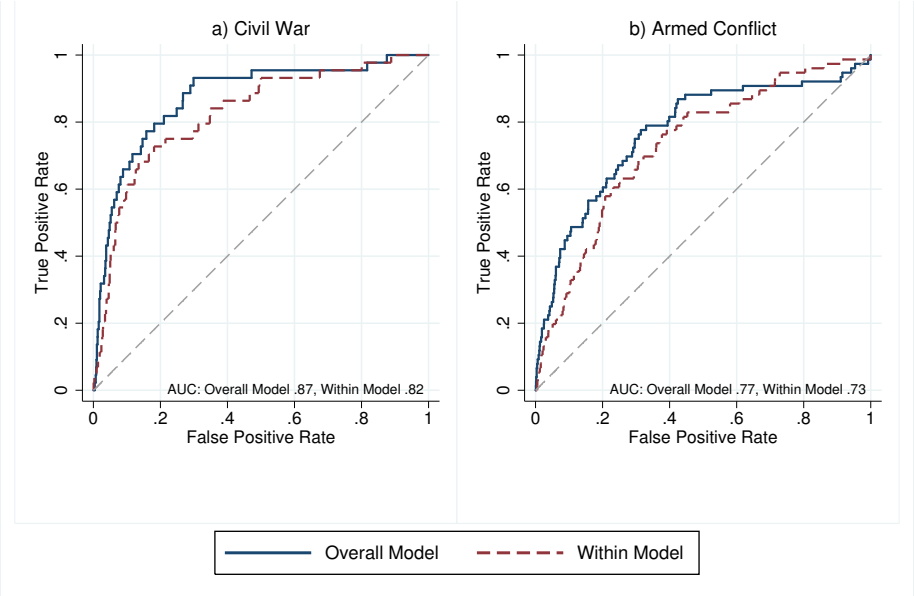
Using a set of varying cutoffs c and our estimates for $\hat{y}_{iT+1}^{within}$ and $\hat{y}_{iT+1}^{overall}$, we then calculate the TPR_c and FPR_c for each cutoff c, which we present in standard ROC curves as explained in Section 3.2.

5.1 Main Results

Our main results are shown in the two graphs in Figure 4, which show ROC curves for the onset of civil war (left) and armed conflict (right). The blue solid lines show the forecasting performance using the fitted values from the overall news model $\hat{y}_{it}^{overall}$ while the red dashed lines provide the ROC curve of the within model $\hat{y}_{it}^{within}$.

Figure 4 shows that news topics fare well at predicting onset of both civil war and armed conflict. When predicting civil war onset, the news model generates a TPR of about 80 percent for a FPR of 20 percent. Furthermore, the predictive power of the within model is very close to the predictive power of the overall model. This is quite a striking finding given the difficulty of forecasting the timing of such rare events. When predicting the onset of armed conflict the model

Figure 4: ROC Curves for Onset (Topics Model)



Notes: Predictions result from a panel estimated as in equation (2). The topic model contains 15 topics as θ_{it} derived using LDA with $\alpha = 3.33$ and $\beta = 0.01$ and are aggregated at the country-year level. The within model is the overall model net of country fixed effects as presented in equation (3).

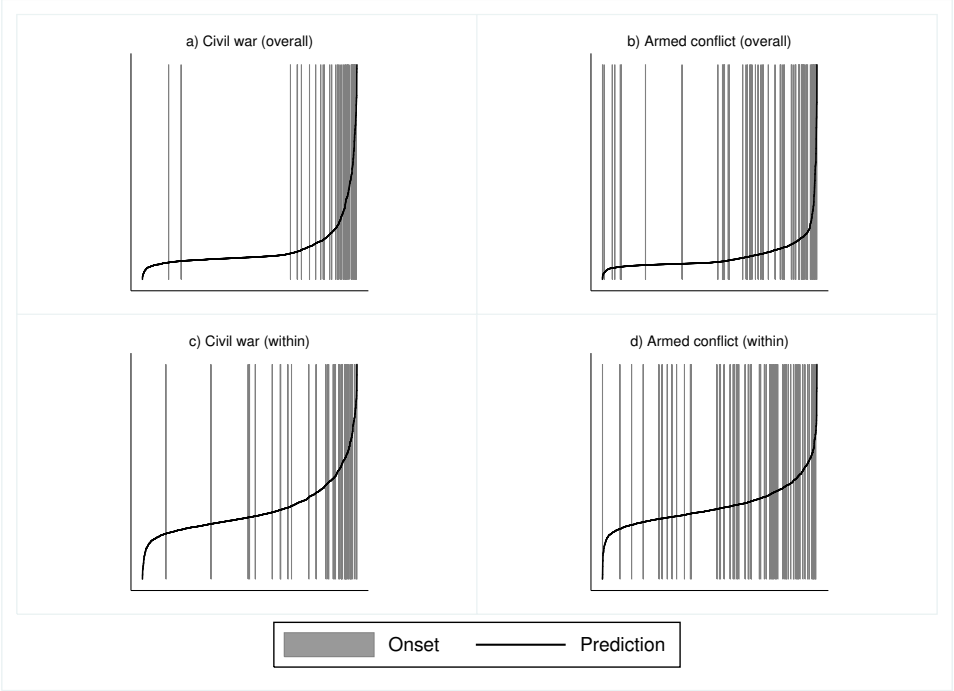
performs worse. Again, the within variation seems to be an important driver of the ability to forecast conflict. The AUC only drops from 0.87 and 0.77 in the overall model and to 0.82 and 0.73 in the within model for civil war and armed conflict, respectively. This is an important difference to the other models evaluated in Section 3.3, which suffer a drop of around 20 percentage points of the AUC, as summarized in Appendix Table I.2.

Appendix Figure I.1 draws all ROC curves in one single figure. Given that not all variables are available for the same time intervals, in Appendix Figures I.2 and I.3 we contrast the predictions of our topic model with other models only using overlapping predictions, i.e. country-years for which we have predictions for both models. This confirms the impression that similar AUCs of the overall model are consistent with significant differences in performance when predicting the timing of conflict.

Before we turn towards robustness checks, we illustrate the quality of our forecasts. We first show separation plots which order all observations according to the predicted values from the model and then compare them to the actual realizations of onset. Figure 5 reports four separation plots, one for each of our models. A good forecasting model has a stronger association of onsets (the gray

vertical lines) with high fitted values (black curves), i.e. we want the gray lines to bunch towards the right. Both in the within and the overall model there are relatively few observations to the left of the figure.

Figure 5: Separation Plots for Onset (Topics Model)

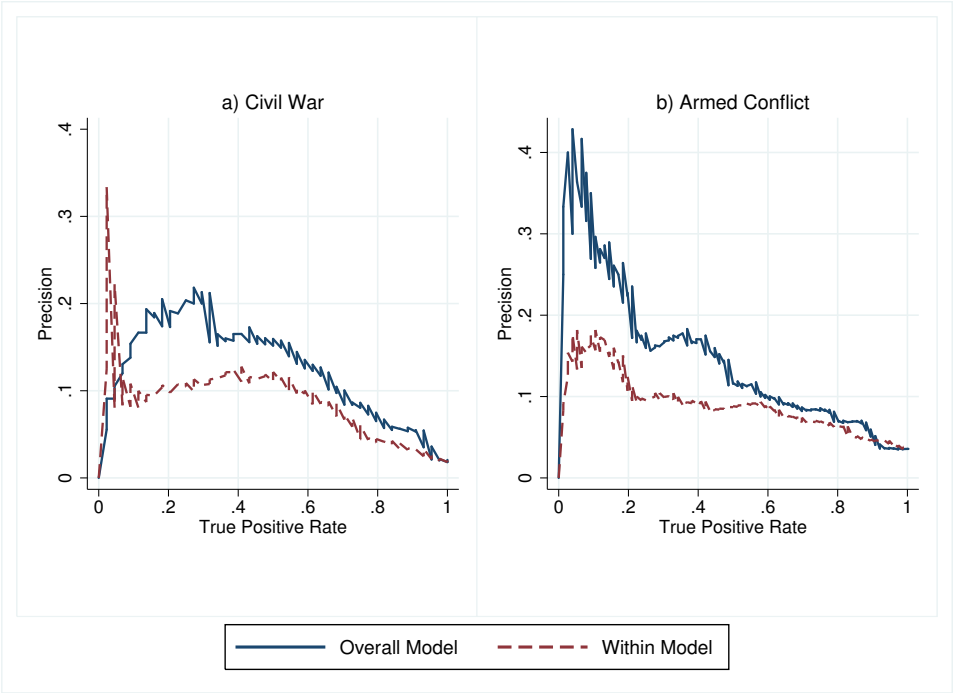


Notes: Predictions result from a panel estimated as in equation (2). The topic model contains 15 topics as θ_{it} derived using LDA with $\alpha = 3.33$ and $\beta = 0.01$ and are aggregated at the country-year level. The within model is the overall model net of country fixed effects as presented in equation (3). For a description of the plots see the text.

The problem in predicting rare events is that, even a low false positive rate can mean that many more false positives are generated than true positives. A way to look at this is precision, which relates the number of true positives to the number of false positives. In Figure 6 we report standard precision/recall curves for our forecasts of onset. Generally, the precision of the within model when forecasting onset is around 10 percent. This would imply that for every ten onsets that are forecasted one turns out to be true. In the overall model, the ratio is at times twice as high but, as argued before, this is often due to the fact that repeated onsets are forecasted with the fixed effect.

We run several robustness checks regarding these basic results. Details and corresponding Figures are all reported in Appendix E. First, in Figure E.1 we show that prediction results when using topics for conflict incidence, unsurprisingly, perform even better than for the prediction of

Figure 6: Precision Recall Curves (Topics Model)



Notes: Predictions result from a panel estimated as in equation (2). The topic model contains 15 topics as θ_{it} derived using LDA with $\alpha = 3.33$ and $\beta = 0.01$ and are aggregated at the country-year level. The within model is the overall model net of country fixed effects as presented in equation (3).

onset.

Second, we use more and less topics.[22] As shown in Figure E.2, with five topics the forecasting power drops only slightly. With 50 and 70 topics the AUC falls a bit more (see Figures E.3 and E.4). This could be explained by a closer fit of topics to specific situations, which then do not generalize out-of-sample.

Third, we combine topics with additional information. We show that, building on other existing models, the topic shares add forecasting power. This indicates that it is not simply reporting on basic economic and political facts that helps us forecast. We also test a model in which we augment our topics with standard variables from the literature. There seem to be few benefits from this (see Figure E.5). This result is reassuring for policy applications which need to rely on news data alone because of the delay in release of the standard data. Moreover, we add an indicator for contemporaneous armed conflict incidence to the onset model of civil war to see whether news topics add forecasting power (see Figure E.6). Indeed, topics add forecasting power beyond an

[22] We adjust α using the ratio $\alpha = 50/K$.

already very high benchmark in this model. This also confirms that news topics capture more than the risk of escalating violence. In Appendix Figure E.7 we also show that the model reaches a precision of 25 percent at a true positive rate of 50 percent.

Fourth, we look at a large number of different definitions of conflict finding similar results (see Figures E.8- E.10). We also use our topic model to predict refugee movements using data on refugees from the UNHCR. We predict the onset of a large number of refugees using two different cutoffs, 30,000 and 130,000 refugees, which makes these movements about as common as armed conflict and civil war, respectively. Again, as can be seen in Figure E.11, the within variation has a lot of predictive power, in this case as much or more than the overall model. This exercise underlines the usefulness of news text in providing early warning for events which are not themselves derived by news. We also analyze how well our model performs when forecasting conflict two years before onset. The within model performs only slightly worse (see Figure E.12). Interestingly, the overall model performs very similarly which confirms the idea that the between variation dominates the overall model. If a forecast is time invariant, it does not matter whether conflict breaks out one or two years later.

Fifth, we break results down by region (Figure E.13) or year (Figures E.14 and E.15) and find that topics consistently perform well, i.e. the predictive power of topics is not restricted to any particular geographical region or time period.

Sixth, we contrast our forecasting model with the events used in Ward et al. (2013) and conflict keyword counts used in Chadefaux (2014) but now without adding the additional, more standard, variables used in the original articles. If conflict keyword counts would forecast as well as the topics there would be no reason to go through the more demanding topic estimation. We find that the within variation from the keyword count variables are now useful for predicting the timing of civil war onset (Figures E.16 and E.17). This indicates that the common practice of mixing standard variables with news variables when adding country fixed effects might actually prevent news variables from providing more useful forecasts of the timing of onset. This is typically not visible because only the overall variation is used. We also find that topics still add forecasting power beyond keyword counts and news events. In Section 7 we return to discuss why this might be the case.

Seventh, in order to bridge the gap to the existing machine learning and forecasting literature

we also use a neural-networks technique to forecast conflict with topics. The method is described in more detail in the Appendix and results are in Figure E.19. The upshot is that there seem to be only moderate gains from using a neural network for the task at hand which is in line with the findings by Goldstone et al. (2010) who find little improvement when switching to neural networks from simpler regression models.

The main takeaway from all these tests is that the timing of conflict can be predicted using automated summaries of news reports. The topic model produces a relatively high true positive rate for relatively low rates of false positives. Our methodology further demonstrates that the within-country variation of the topic model has almost the same predictive power as the overall model. We show in the next section that this is key when one tries to predict in previously peaceful countries.

6 Why Predicting the Timing is Important

The literature has been criticized for failing to provide early warning when new instabilities emerge. Margolis (2012), for example, laments that *"Policymakers paying attention to the recent history of popular current stability indices, for example, could not have anticipated that instability would sweep across the Middle East."* [Margolis 2012, p. 14]. This has led to some soul-searching in the literature and, most recently, to claims that forecasting new civil wars might have reached a limit.[23]

Certainly, there are good reasons to believe that there is a natural boundary to the precision that can be reached in forecasting. However, our results suggest that two methodological shortcomings might contribute to the fact that new fragilities are more surprising than they need to be. The first is that standard methods without fixed effects will identify a lot of the risks from the between variation. We have argued in Section 3.1 that this can lead to mistakes.

Second, even when fixed effects are used, it can be of interest to drop them for forecasting because the use of the overall variation leads to a focus on cases which had previous onsets in the training sample. In Appendix F we explain this effect theoretically and also provide Monte Carlo simulations to show that there is indeed a bias in the overall model against cases without previous onsets.

[23] See Chadefaux (2017*a*) and Cederman and Weidmann (2017) who argue that violent conflict might occur exactly because it follows rule-breaking ways which are impossible to predict.

To illustrate this in the actual data, we calculate the AUC from the within and overall topic model but focus on new conflict onsets, i.e. the first onset in our sample. The results are reported in Table 2. In the full sample, the overall model provides a better forecast than the within model. However, when looking at cases of onsets in previously peaceful countries the within model gives a much better forecast than the overall model.

Table 2: AUC of topics model in previously peaceful countries vs full sample

Sample	Civil War Onsets			Armed Conflict Onsets		
	overall	within	dif.	overall	within	dif.
Full	0.87	0.82	**0.05**	0.77	0.73	**0.04**
Previously peaceful	0.76	0.83	**-0.07**	0.61	0.70	**-0.08**

Note: The topic model contains a set of 14 topic shares as estimated in 2013. The AUC is the surface under the ROC curve. A value of 1 implies that forecasts of the model are perfect. A value of 0.5 implies that forecasts are as good as random.

Table 2 demonstrates that the overall model induces a bias against new conflicts, whereas the AUC of the within model is relatively robust to the switch of sample. This is because the within model does not use any information on previous onsets to make the forecast. Instead, it relies entirely on the usefulness of the within variation to make risk evaluations. Of course, the fixed effect can contain information about risk but from the perspective of Table 2, it seems to be a good idea to develop models with more useful within variation. Optimally, one would want a model which is able to forecast the huge differences between countries entirely through the within variation in these countries.

The bias against new conflicts in the overall model is also clearly visible in the list of top-risk countries produced by the within and overall models.[24] For example, when $T = 2010$ the within model predicts the onset of civil war in 2011 to be most likely in four countries without previous onsets and in Yemen which had been relatively peaceful for over a decade. However, the overall model predicts civil war onsets in Chad, Sri Lanka, Uganda, Colombia and the Philippines - all of which had had several onsets in their recent past. Accordingly, high-profile onsets in 2011, like those of Libya and Syria, are ranked as much more likely by the within model than by the overall model. This is the key problem when relying on the overall model - it leads to a heavy focus on

[24] Examples are in Appendix Table H.1.

countries with previous, recent onsets.

Hence, some of the current frustration in forecasting conflict could stem from a methodological bias against spotting new conflicts in the existing literature. At the heart of this bias is the reliance on between variation when forecasting which makes models blind to new developments.

7 Reading Between the Lines

We now explore why topics provide such useful forecasting power on the time dimension.[25] To do this we first let a simple machine learning algorithm choose variables to predict conflict within-sample. We use the least absolute shrinkage and selection operator (LASSO) with country fixed effects to choose variables from a pool of over 30 variables, including our 15 topic shares.[26] We base our analysis on the topic model estimated in 2013 as this is the last year of text we can use for estimation. The other variables in the pool are all previously used variables based on Chadefaux (2014), Ward et al. (2013), and Goldstone et al. (2010). This includes a host of standard political and economic variables, word counts based on our text, and two event counts from ICEWS. To this we add the incidence of armed conflict when explaining the onset of civil war a year later. In order to get a more and less restrictive set of variables, we vary the parameter that captures the weight given to choosing few variables. We pick three levels to show how the chosen model evolves with increasing selectivity.[27]

Table 3 shows the six models selected by the LASSO. Columns (1) to (3) show variables selected when predicting the onset of civil war and Columns (4) to (6) when predicting the onset the onset of armed conflict. We report the share of the topic variables in these models in bold at the top of the table. The LASSO always chooses at least 50 percent of all variables from amongst the topics and this share is higher in the more selective models.

The first message from Table 3 is that conflict topics are chosen by the LASSO despite the fact

[25] In order to "harmonize" topics across samples we choose a baseline year (2013) relative to which we define all topics. We count the words that coincide across two topics within the top 50 keywords weighted by their prominence. Finally, we assign the same name to the most similar topic (e.g. *conflict1*) if it has at least 3 coinciding words amongst the top 50 keywords. The similarity between topics was confirmed through eyeballing which revealed high consistency.

[26] The LASSO minimizes the usual sum of squared errors but augments the mean squared error objective function with an additional penalty term that is a weighted sum of the absolute value of the regression coefficients. The resulting minimization leads to a small set of the most important predictors. Our analysis uses the lassoShooting algorithm provided on Christian Hansen's webpage as replication file for Belloni et al. (2011).

[27] The relevant parameter to do this is λ and we pick 100, 150, and 200.

that conflict keyword counts based on the same text, conflict events and armed conflict incidence are available. This must be because the conflict topics rely on tokens which add useful variation beyond the core conflict keywords, conflict events, and even lower-level conflict. This is not implausible as we know that the Gibbs sampler forms large word lists around key conflict words. This depth seems to help in the forecast.

The second message is that the large majority of coefficients on topic shares within-sample are negative. This suggests that stabilizing factors, captured by non-conflict topics, play a key role when explaining the timing of conflict. In order to evaluate the contribution of these stabilizers we run an additional robustness check in which we exclude all conflict topics (up to 3) in all years and try to forecast only with the remaining non-conflict topics.[28] The result of this attempt are in Figure 7. Strikingly, both armed conflict and civil war can still be forecasted fairly well without relying directly on the conflict topics. The within model retains almost all of its forecasting ability when predicting armed conflict (AUC of 0.72 when excluding the conflict topics compared to an AUC of 0.73) and suffers only a little more when predicting civil war (AUC of 0.79 compared to an AUC of 0.82). We conclude from this that non-conflict topics are important precursors of conflict. The forecasting power is much lower when using only the conflict topics.[29]

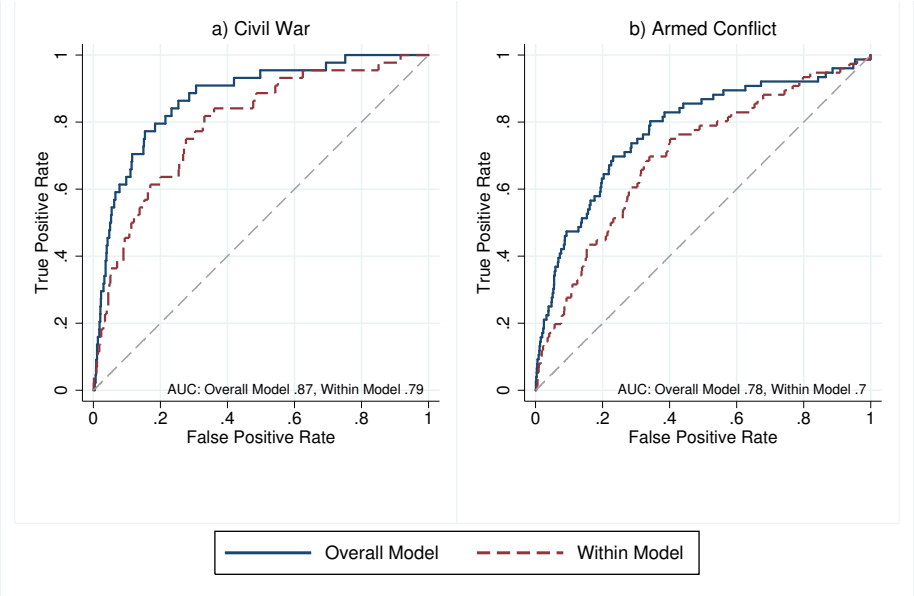
[28] We re-normalize the non-conflict topic shares so they add up to one. Results here are robust to not doing this.
[29] See Appendix Figure E.20 in which we present results from the analog exercise using only conflict topics.

Table 3: Lasso model

Selectivity level	mildly	regular	very	mildly	regular	very
	civil war onset next year			armed conflict onset next year		
	(1)	(2)	(3)	(4)	(5)	(6)
Topic shares						
conflict1	0.0366	0.0564		0.306**	0.259**	0.275***
	(0.0685)	(0.0599)		(0.121)	(0.103)	(0.0999)
conflict2	0.256**	0.300***	0.281***	0.304**		
	(0.104)	(0.103)	(0.0961)	(0.117)		
justice	-0.158**	-0.115*	-0.117**	-0.256***	-0.215***	-0.206***
	(0.0664)	(0.0617)	(0.0541)	(0.0826)	(0.0712)	(0.0705)
international relations2	-0.236**			-0.130	-0.0554	
	(0.102)			(0.0992)	(0.0909)	
civic life2	-0.0869*	-0.00783	-0.0247	-0.0196	-0.0679	
	(0.0518)	(0.0370)	(0.0298)	(0.0671)	(0.0520)	
asia	-0.180**	-0.151**	-0.142**			
	(0.0803)	(0.0734)	(0.0650)			
sports	-0.0490					
	(0.0365)					
politics	-0.141***					
	(0.0472)					
business	-0.136**					
	(0.0549)					
economics				-0.0256		
				(0.0891)		
Other variables						
25+ battle death	0.0699***	0.0713***	0.0749***			
	(0.0163)	(0.0164)	(0.0165)			
democracy score	4.81e-05					
	(0.000198)					
partial autocracy				0.0244	0.0270*	
				(0.0151)	(0.0145)	
partial dem. with factionalism				-0.00845	-0.00163	-0.00888
				(0.0124)	(0.0104)	(0.00981)
partial dem. w/o factionalism	0.0154					
	(0.0105)					
full democracy	0.0174*			0.00183	0.00442	
	(0.0102)			(0.0165)	(0.0118)	
4+ neighbouring conflicts	0.0247					
	(0.0396)					
child mortality rate				-3.86e-05		
				(0.000212)		
ln (child mortality rate)	0.00707			0.00376		
	(0.00531)			(0.00852)		
% pop. discriminated	0.111*	0.108*				
	(0.0604)	(0.0616)				
% pop. excluded from power				-0.0488		
				(0.0442)		
Country fixed effects	yes	yes	yes	yes	yes	yes
Observations	4,561	4,644	4,931	3,991	4,226	4,226
R-squared	0.039	0.034	0.030	0.012	0.008	0.006
Number of countries	140	141	143	138	139	139
% topics in model	**56%**	**71%**	**80%**	**50%**	**57%**	**67%**

Notes: Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. The table displays the selected variables using LASSO with parameter λ equal to 100 (columns 1 & 4), 150 (columns 2 & 5), 200 (columns 3 & 6) from 15 topics and 30 variables from other models. Topics are from the year 2013. The most prominent words of each topic in 2013 are displayed in Table I.1. Summary statistics of all variables are displayed in Table C.1.

Figure 7: ROC Curves for Onset (Only Non-Conflict Topics)



Notes: Predictions result from a panel estimated as in equation (2). The topic model contains the 12 non-conflict topics as θ_{it} derived using LDA with $\alpha = 3.33$ and $\beta = 0.01$ and are aggregated at the country-year level. We excluded the 3 conflict topics and re-normalized so that the 12 remaining topics sum to 1. The within model is the overall model net of country fixed effects as presented in equation (3).

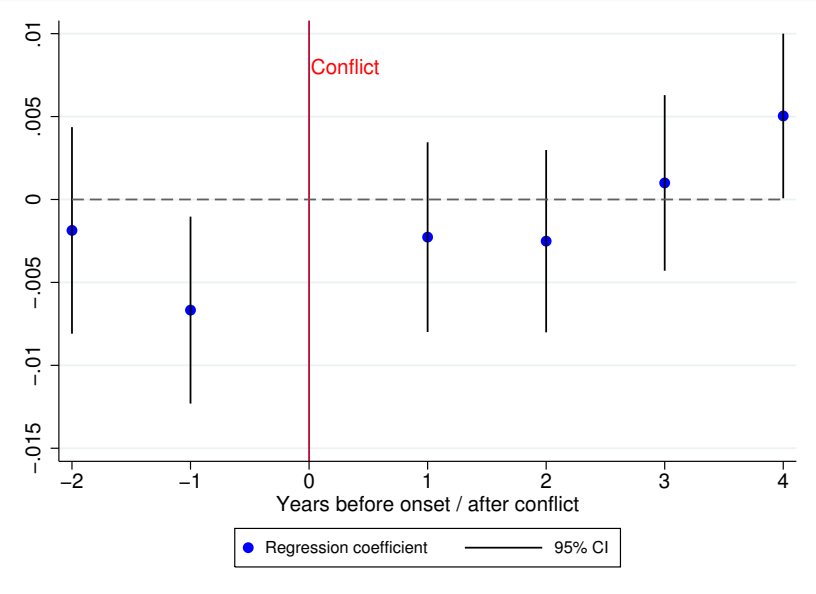
One example of the forecasting power inherent in non-conflict topics is the *justice* topic. Its topic share is the only variable that is picked in all models in Table 3. In Figure 8 we show the dynamics of the topic before the onset and after the end of conflict.[30] Writing on the *justice* topic decreases significantly one year prior to conflict and increases after conflict has ended. Importantly, this holds controlling for reporting on all other topics, i.e. it is not driven by a crowding out of the *justice* topic by the conflict topics.[31]

This relationship is correlational rather than causal. Nonetheless, the fact that newspaper reports on law enforcement and justice disappear before and after conflict is in line with research on the role of checks and balances for preventing conflict (e.g., Besley and Persson 2011*b*, Blattman, Hartman and Blair 2014) and the role of post-conflict justice (e.g., Meernik 2005, Olsen, Payne and

[30]To generate the figure we regress the *justice* topic share on dummies for the number of years before the onset of conflict and the number of years after conflict has ended. In this regression we control for country fixed effects and the remaining topic shares. The coefficients on the dummies then capture how much more (or less) reporting there is on justice is in the years around conflict relative to other years in peace and controlling for country-specific reporting and reporting on other topics.

[31]In addition, we show in Appendix Figure I.4 that shifts in topic shares of *justice* are not driven by shifts in which all journalists suddenly report much less on justice. This is in line with Nimark and Pitschner (2016) who show that small news stories are picked up by only some news sources while larger events unify reporting across sources.

Figure 8: The *Justice* Topic Before and After the Outbreak of Conflict



Notes: The coefficients and confidence intervals are obtained by regressing the *justice* topic share on the remaining topic shares, and dummies for the number of years before the onset of conflict and the number of years after conflict has ended. In this panel regression with country fixed effects, conflict years have been set to missing.

Reiter 2010) in ensuring stability. The latter literature is relevant as news stories on the *justice* topic increase significantly after conflict and reach particularly high levels in countries where peace is stable. However, the significant decrease before conflict is what lends the model its forecasting power.

8 Conclusions

In this paper, we present a new method of predicting conflict through news topics which are generated automatically from a topic model. Topic models have the ability to diminish the dimensionality of text from counts of close to one million expressions to, for example, 15 topics. These topics can then be used in simple linear regression models to predict the onset of conflict. We have used ROC curves to show that, aggregated this way, news text becomes a useful predictor. When predicting onset one year ahead, a method based entirely on topics is able to forecast the timing of conflict better than the main models used in the literature.

Three factors make topics particularly appealing for forecasting. First, the results can be easily interpreted because topics provide meaningful summaries of text. Second, the algorithm which

generates topics is able to learn from the changing association of terms. We have shown, for example, that new terms like "terrorist" or "insurgent" serve as key indicators of conflict risk in recent years, whereas they did not in 1995. Third, the topic model uses negative associations between topics and conflict risk in the prediction. In fact, large parts of the forecast seems to come from topics not directly related to conflict. The relationship between less reports on judicial procedures and law enforcement and higher conflict risk is particularly strong.

Our findings highlight that models need to be tested for whether their within variation is meaningful. If not, policymakers might rely on meaningless changes of risk across time. Furthermore, we have shown that relying on the overall variation of models, even if they contain useful within variation, can lead to a bias against onsets in previously peaceful countries. Ultimately, researchers and policymakers therefore face a trade-off between a better prediction overall and a model that is more useful in spotting new instabilities.

In addition, an implementation of the model presented here for policy purposes shares problems with other forecasts. First, forecasts do not provide a causal analysis of the underlying factors leading to high risk but only produce a warning of that risk. Additional analysis of the specific circumstances is needed to identify ways to address the conflict risk. Second, precision remains a problem despite the fact that our method improves upon what exists. Policymakers need to understand that, even in the best model, for five warnings made, four will be false warnings.

Topic models could provide a useful alley for research in political events more generally. We believe that it is possible to learn about the factors that influence these events from what we call the depth and width of topics. Technical extensions or refinements could include using more recently developed topic modelling techniques, such as dynamic topic models (Blei and Lafferty 2006) or a structural topic model (Roberts et al. 2013).

References

Banks, Arthur. 2005. "Cross-National Time-Series Data Archive." Binghamton: Databanks International.

Bazzi, Samuel and Christopher Blattman. 2014. "Economic shocks and conflict: Evidence from commodity prices." *American Economic Journal: Macroeconomics* 6(4):1–38.

Beck, Nathaniel. 2015. "Estimating grouped data models with a binary dependent variable and fixed effects: What are the issues?" Annual meeting of the Society for Political Methodology, July.

Bell, Sam R, David Cingranelli, Amanda Murdie and Alper Caglayan. 2013. "Coercion, capacity, and coordination: Predictors of political violence." *Conflict Management and Peace Science* 30(3):240–262.

Belloni, Alexandre, Victor Chernozhukov, Christian Hansen et al. 2011. "Inference for high-dimensional sparse econometric models." Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

Besley, Timothy and Torsten Persson. 2011*a*. "The Logic of Political Violence." *Quarterly Journal of Economics* 126(3):1411–1445.

Besley, Timothy and Torsten Persson. 2011*b*. *Pillars of prosperity: The political economics of development clusters*. Princeton University Press.

Blattman, Christopher, Alexandra C Hartman and Robert A Blair. 2014. "How to promote order and property rights under weak rule of law? An experiment in changing dispute resolution behavior through community education." *American Political Science Review* 108(01):100–120.

Blattman, Christopher and Edward Miguel. 2010. "Civil war." *Journal of Economic Literature* 48(1):3–57.

Blei, David M, Andrew Y Ng and Michael I Jordan. 2003. "Latent Dirichlet allocation." *The Journal of Machine Learning Research* 3:993–1022.

Blei, David M and John D Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*. ACM pp. 113–120.

Brandt, Patrick T, John R Freeman and Philip A Schrodt. 2011. "Real time, time series forecasting of inter-and intra-state political conflict." *Conflict Management and Peace Science* 28(1):41–64.

Brandt, Patrick T, John R Freeman and Philip A Schrodt. 2014. "Evaluating forecasts of political conflict dynamics." *International Journal of Forecasting* 30(4):944–962.

Brückner, Markus and Antonio Ciccone. 2010. "International Commodity Prices, Growth and the Outbreak of Civil War in Sub-Saharan Africa." *The Economic Journal* 120(544):519–534.

Buhaug, Halvard, J Nordkvelle, T Bernauer, T Böhmelt, M Brzoska, JW Busby, A Ciccone, Hanne Fjelde, E Gartzke, NP Gleditsch et al. 2014. "One effect to rule them all? A comment on climate and conflict." *Climatic Change* 127(3-4):391–397.

Caselli, Francesco and Wilbur John Coleman. 2013. "On the theory of ethnic conflict." *Journal of the European Economic Association* 11(s1):161–192.

Cederman, Lars-Erik and Nils B Weidmann. 2017. "Predicting armed conflict: Time to adjust our expectations?" *Science* 355(6324):474–476.

Chadefaux, Thomas. 2014. "Early warning signals for war in the news." *Journal of Peace Research* 51(1):5–18.

Chadefaux, Thomas. 2017*a*. "Conflict forecasting and its limits." *Data Science* Preprint(Preprint):1–11.

Chadefaux, Thomas. 2017*b*. "Market anticipations of conflict onsets." *Journal of Peace Research* 54(2):313–327.

Chiba, Daina and Kristian Skrede Gleditsch. 2017. "The shape of things to come? Expanding the inequality and grievance model for civil war forecasts with event data." *Journal of Peace Research* 54(2):275–297.

Colaresi, Michael and Zuhaib Mahmood. 2017. "Do the robot: Lessons from machine learning to improve conflict forecasting." *Journal of Peace Research* 54(2):193–214.

Collier, Paul and Anke Hoeffler. 2004. "Greed and grievance in civil war." *Oxford Economic Papers* 56(4):563–595.

Collier, Paul, Anke Hoeffler, Dominic Rohner et al. 2009. "Beyond greed and grievance: feasibility and civil war." *Oxford Economic Papers* 61(1):1–27.

Dell, Melissa, Benjamin F Jones and Benjamin A Olken. 2012. "Temperature shocks and economic growth: Evidence from the last half century." *American Economic Journal: Macroeconomics* 4(3):66–95.

Esteban, Joan, Laura Mayoral and Debraj Ray. 2012. "Ethnicity and conflict: An empirical study." *The American Economic Review* 102(4):1310–1342.

Fearon, James D and David D Laitin. 2003. "Ethnicity, insurgency, and civil war." *American Political Science Review* 97(01):75–90.

Gerner, Deborah J, Philip A Schrodt, Omur Yilmaz and Rajaa Abu-Jabr. 2002. "The creation of CAMEO (Conflict and Mediation Event Observations): An event data framework for a post cold war world." Annual meeting of the American Political Science Association.

Girardin, Luc, Philipp Hunziker, Lars-Erik Cederman, Nils-Christian Bormann and Manuel Vogt. 2015. "GROWup–Geographical Research on War, Unified Platform. ETH Zurich.".

Gleditsch, Kristian Skrede and Andrea Ruggeri. 2010. "Political opportunity structures, democracy, and civil war." *Journal of Peace Research* 47(3):299–310.

Gleditsch, Nils Petter, Peter Wallensteen, Mikael Eriksson, Margareta Sollenberg and Håvard Strand. 2002. "Armed conflict 1946-2001: A new dataset." *Journal of Peace Research* 39(5):615–637.

Goldsmith, Benjamin E, Charles R Butcher, Dimitri Semenovich and Arcot Sowmya. 2013. "Forecasting the onset of genocide and politicide: Annual out-of-sample forecasts on a global dataset, 1988–2003." *Journal of Peace Research* 50(4):437–452.

Goldstone, Jack A, Robert H Bates, David L Epstein, Ted Robert Gurr, Michael B Lustik, Monty G Marshall, Jay Ulfelder and Mark Woodward. 2010. "A global model for forecasting political instability." *American Journal of Political Science* 54(1):190–208.

Hansen, Stephen, Michael McMahon and Andrea Prat. 2014. "Transparency and deliberation within the FOMC: a computational linguistics approach." CEP Discussion Paper No 1276.

Hegre, Håvard, Joakim Karlsen, Håvard Mokleiv Nygård, Håvard Strand and Henrik Urdal. 2013. "Predicting Armed Conflict, 2010–20501." *International Studies Quarterly* 57(2):250–270.

Hegre, Håvard, Nils W Metternich, Håvard Mokleiv Nygård and Julian Wucherpfennig. 2017. "Introduction: Forecasting in peace research." *Journal of Peace Research* 54(2):113–124.

Heinrich, Gregor. 2009. A generic approach to topic models. In *Machine Learning and Knowledge Discovery in Databases*. Springer pp. 517–532.

Kalyvas, Stathis N and Laia Balcells. 2010. "International system and technologies of rebellion: How the end of the cold war shaped internal conflict." *American Political Science Review* 104(03):415–429.

Kennedy, Ryan. 2015. "Making useful conflict predictions: Methods for addressing skewed classes and implementing cost-sensitive learning in the study of state failure." *Journal of Peace Research* 52(5):649–664.

King, Gary and Langche Zeng. 2001. "Improving forecasts of state failure." *World Politics* 53(04):623–658.

Margolis, J Eli. 2012. "Estimating State Instability." *Studies in Intelligence* 56(1):13–24.

Meernik, James. 2005. "Justice and peace? How the International Criminal Tribunal affects societal peace in Bosnia." *Journal of Peace Research* 42(3):271–289.

Miguel, Edward and Shanker Satyanath. 2011. "Re-examining economic shocks and civil conflict." *American Economic Journal: Applied Economics* 3(4):228–232.

Miguel, Edward, Shanker Satyanath and Ernest Sergenti. 2004. "Economic shocks and civil conflict: An instrumental variables approach." *Journal of Political Economy* 112(4):725–753.

Muchlinski, David, David Siroky, Jingrui He and Matthew Kocher. 2016. "Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data." *Political Analysis* 24(1):87–103.

Mueller, Hannes. 2016. "Growth and violence: Argument for a per capita measure of civil war." *Economica* 83(331):473–497.

Nimark, Kristoffer P and Stefan Pitschner. 2016. "Delegated Information Choice." Mimeo.

Olsen, Tricia D, Leigh A Payne and Andrew G Reiter. 2010. "Transitional justice in the world, 1970-2007: Insights from a new dataset." *Journal of Peace Research* 47(6):803–809.

Pettersson, Therése and Peter Wallensteen. 2015. "Armed conflicts, 1946–2014." *Journal of Peace Research* 52(4):536–550.

Phan, Xuan-Hieu and Cam-Tu Nguyen. 2007. "GibbsLDA++: AC/C++ implementation of latent Dirichlet allocation (LDA).".

Porter, Martin F. 1980. "An algorithm for suffix stripping." *Program* 14(3):130–137.

Quinn, Kevin M, Burt L Monroe, Michael Colaresi, Michael H Crespin and Dragomir R Radev. 2010. "How to analyze political attention with minimal assumptions and costs." *American Journal of Political Science* 54(1):209–228.

Reynal-Querol, Marta and Jose G Montalvo. 2005. "Ethnic polarization, potential conflict and civil war." *American Economic Review* 95(3):796–816.

Roberts, Margaret E, Brandon M Stewart, Dustin Tingley, Edoardo M Airoldi et al. 2013. "The structural topic model and applied social science." Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation.

Rost, Nicolas, Gerald Schneider and Johannes Kleibl. 2009. "A global risk assessment model for civil wars." *Social Science Research* 38(4):921–933.

Sambanis, Nicholas. 2004. "What is civil war? Conceptual and empirical complexities of an operational definition." *Journal of Conflict Resolution* 48(6):814–858.

Schneider, Gerald, Maya Hadar and Naomi Bosler. 2017. "The oracle or the crowd? Experts versus the stock market in forecasting ceasefire success in the Levant." *Journal of Peace Research* 54(2):231–242.

Schrodt, PA, DJ Gerner and O Yilmaz. 2009. Conflict and Mediation Event Observations (CAMEO): An Event Data Framework for a Post Cold War World. In *International Conflict Mediation: New Approaches and Findings*, ed. Gartner S Bercovitch J. New York: Routledge.

Schrodt, Philip A, James Yonamine and Benjamin E Bagozzi. 2013. Data-based computational approaches to forecasting political violence. In *Handbook of computational approaches to counterterrorism*. Springer pp. 129–162.

Ward, Michael D, Brian D Greenhill and Kristin M Bakke. 2010. "The perils of policy by p-value: Predicting civil conflicts." *Journal of Peace Research* 47(4):363–375.

Ward, Michael D, Nils W Metternich, Cassy L Dorff, Max Gallop, Florian M Hollenbach, Anna Schultz and Simon Weschle. 2013. "Learning from the past and stepping into the future: Toward a new generation of conflict prediction." *International Studies Review* 15(4):473–490.

Ward, Michael D, Nils W Metternich, Christopher Carrington, Cassy Dorff, Max Gallop, Florian M Hollenbach, Anna Schultz and Simon Weschle. 2012. Geographical models of crises: Evidence from ICEWS. In *Advances in Design for Cross-Cultural Activities*, ed. Dylan D. Schmorrow and Denise M. Nicholson. Vol. 429 CRC Press.

Weidmann, Nils B. 2016. "A closer look at reporting bias in conflict event data." *American Journal of Political Science* 60(1):206–218.

Weidmann, Nils B and Michael D Ward. 2010. "Predicting conflict in space and time." *Journal of Conflict Resolution* 54(6):883–901.

Witmer, Frank DW, Andrew M Linke, John O'Loughlin, Andrew Gettelman and Arlene Laing. 2017. "Subnational violent conflict forecasts for sub-Saharan Africa, 2015–65, using climate-sensitive models." *Journal of Peace Research* 54(2):175–192.

Woolley, John T. 2000. "Using media-based data in studies of politics." *American Journal of Political Science* 44(1):156–173.

Appendix

A Machine Learning and Forecasting: An Overview

In Table A.1 we provide a brief literature summary of the forecasting literature since 2000. In this review we focus on articles that try to predict the time dimension of conflict. To provide a

better impression of the part of the literature using machine learning techinques we categorise the literature along these lines. A recent article using machine learning, Muchlinski et al. (2016) uses a random forest to select variables for its forecasting model. This paper reports relatively high AUC but uses fixed characteristics like mountainous terrain to produce the forecast so that the same set of variables cannot be used in our setting.

Table A.1: Related literature

No machine learning	King and Zeng (2001); Rost, Schneider and Kleibl (2009); Goldstone et al. (2010); Ward, Greenhill and Bakke (2010); Weidmann and Ward (2010); Goldsmith et al. (2013); Ward et al. (2013); Chadefaux (2014); Kennedy (2015); Chiba and Gleditsch (2017); Chadefaux (2017*b*); Schneider, Hadar and Bosler (2017); Witmer et al. (2017)
Machine learning	Brandt, Freeman and Schrodt (2011); Ward et al. (2012); Bell et al. (2013); Brandt, Freeman and Schrodt (2014); Muchlinski et al. (2016); Colaresi and Mahmood (2017)

While we use machine learning, our work relates most closely to work that does not. The reason is that we use machine learning techniques to summarize the text but then use these summaries in a way which is more akin to the standard structural part of the literature. We do this for three reasons. First, we want to compare our results to the existing work which uses large country panels. Second, because we want to show that text summaries in topics can, in fact, contribute to our understanding of which factors increase conflict risk. This is easier to illustrate in classic regression frameworks despite their known weaknesses in forecasting. The third reason is that the regression framework allows us to analyze which part of the variation (between or within) does the forecasting. This illustration is less straightforward when using machine learning techniques.

B Evaluation Method

In this appendix we discuss the evaluation method we use throughout the article. In the first step a model is trained by estimating all conflict onsets in $t \in \{1976, 1977, 1978, ..., T\}$ with the data available until year T.[32] Formally, we assume that the policymaker uses a regression of conflict

[32]In practice this means that the newest independent variables used are from $T-1$ because the idea is to train a model which is able to forecast T with $T-1$.

onset in $t+1$ on a set of variables from year $\mathbf{x}_{it}$ and fixed effects to train the model, i.e.

$$y_{it+1} = \alpha + \beta_i + \mathbf{x}_{it}\beta^{FE} + \varepsilon_{it}, \tag{4}$$

where α is a constant and β_i is a set of country fixed effects. Note that because only information until T can be used in this step, the most recent conflict that is used in the training of the model occurred in year T. To illustrate, assume that a policymaker wants to forecast conflict in 2016. She would then use conflict outbreaks that happened until 2015 to train her model.

Once the training is completed the policymaker then uses the trained model to produce fitted values for $T+1$. In this step, the respective estimates for $\hat{\alpha}$, $\hat{\beta}_i$, and $\hat{\beta}^{FE}$ are used together with the variable values in year T, $\mathbf{x}_{iT}$, to produce the fitted values in $T+1$. We call the fitted values that use the entire model the overall model. Formally, the fitted values here are given by

$$\hat{y}_{iT+1}^{overall} = \hat{\alpha} + \hat{\beta}_i + \mathbf{x}_{iT}\hat{\beta}^{FE}.$$

We will show that the variation contained in the estimated fixed effects $\hat{\beta}_i$ often dominates the model $\mathbf{x}_{iT}\hat{\beta}^{FE}$ in terms of magnitude. For this purpose we also isolate the within variation in what we call the within model. Formally, the fitted values here are given by

$$\hat{y}_{iT+1}^{within} = \hat{\alpha} + \mathbf{x}_{iT}\hat{\beta}^{FE}.$$

In the second step, these two sets of fitted values are used to produce forecasts. In this step, the fitted values are converted into a binary forecast (i.e. we predict peace or the onset of conflict) depending on whether the fitted value is above a cutoff c or not. For values below the cutoff no onset, a negative, is forecasted. For fitted values above this cutoff onset, a positive, is forecasted. These zeros and ones can then be evaluated with the actual realizations of onset.

These two steps are repeated in several years to get a good impression of the overall ability of the model to forecast conflict. In our main application, for example, we let T go from 1995 to 2013 and implement the above steps for all years in between. This means we first predict conflict in 1996 with information available in 1995. We then predict conflict in 1997 with all information available in 1996 and so forth.

The ROC curves we report evaluate all these predictions together. For this we pool all the fitted values from each iteration with the sample until T and compare them to the respective realization for $T + 1$.

C Discussion of Existing Models

We use five models to forecast conflict. The first model uses two variables provided in the replication dataset for Miguel and Satyanath (2011), contemporaneous rainfall growth and lagged rainfall growth for about 40 African countries.

The second model uses six variables from the replication dataset from Besley and Persson (2011*b*). Their model combines proxies for external economic shocks (membership of the security council as a measure for an aid shock and natural disasters) together with a dummy for the post-cold war period and a measure of political constraints in the country. We use their measure of good institutions which is based on executive constraints from Polity IV and generate interaction variables between this variable, the post-cold war variable and the security council membership. We also include the interaction between the good institutions dummy and natural disasters. In all of this we follow the replication do files provided with the data.

The third model we use is from Goldstone et al. (2010). We follow their generation of the political institutions dummies closely. These measures are based on various dimensions of the Polity IV data. We also add the newest data on child mortality from the World Bank and the share of population which is discriminated from the Geographical Research On War, Unified Platform (GROWup) Girardin et al. (2015). This is a slight deviation from the original model in Goldstone et al. (2010). However, the discrimination variable is extremely robust within-sample so that we doubt that this lowers the ability of this model to forecast. Finally, we add a dummy for armed conflict in more than four neighboring countries.

The final two models are based on news. The fourth model counts the keywords defined by Chadefaux (2014) in our 700,000 articles. These are the following: tension(s), crisis, conflict, antagonism, clash, contention, discord, dissent, disunion, disunity, feud, division, fight, hostility, rupture, strife, attack, combat, shell, struggle, fighting, confrontation, impasse. To the keyword count, the word count and their interaction, the model adds the number of years since the last

conflict, the number of years squared and cubed.

The fifth model is based on Ward et al. (2013) and includes events and a range of political and institutional indicators. We use the ICEWS database, which was assembled using the Raytheon/BBN Serif/ACCENT coder.[33] From this event data, we construct the number of high-intensity conflict events (for example, protests, fighting, killings) and low-intensity conflict events (for example, demands or threats) taking place between the government and opposition groups.[34] In addition, we add autocracy and democracy scores, the share of population excluded and the share squared, log child mortality, a dummy for armed conflict in more than four neighboring countries, and the share of population discriminated against. For the event data and the keyword counts, we test a model including only these variables, i.e. excluding institutional variables. For events we use counts of high intensity events of dissidents, ethnic groups, opposition, or any domestic group versus the government, as in Ward et al. (2012). For keywords and events we find that focusing on news data improves the predictive power of the within variation.

We merge these different data with data on battle-related deaths from UCDP/PRIO. In the construction of our internal war variable we tried to err on the inclusive side, i.e. within reasonable boundaries of doubt we want to code a year as a conflict year if some violence took place. We want to be inclusive because we want to consider as much ongoing violence as possible. We therefore count battle-related deaths in internal and internationalised internal conflicts. The latter includes, for example, casualties caused by international terrorism and violence in Afghanistan. We use the best estimates for battle-related deaths and define armed conflict as a year with at least 25 battle-related deaths and a year of civil war as a year with at least 1000 battle-related deaths. In addition, we code a year as being in conflict if the mean between the low and the high estimate crossed these thresholds. We have run robustness checks in which we both expand and restrict this definition - results always stay similar.

In order to improve comparability across models we only include countries with more than 1 million inhabitants, which causes us to drop only about 0.2 percent of the world population. Summary statistics for all variables in our sample are in Table C.1.

[33] Prior to 2012, ICEWS used the Kansas-based TABARI coder (and a Lockheed-produced derivative called JABARI).

[34] We also tested a model using dummies for whether there were high or low intensity events and find very similar results.

Table C.1: Sample summary statistics

	N	Mean	Median	Min	Max	*Shocks and Institutions*	N	Mean	Median	Min	Max
Main						*Shocks and Institutions*					
topic 1 share	5,228	0.053	0.036	0.010	0.334	natural disaster	5,975	0.306	0.461	0	1
topic 2 share	5,228	0.072	0.038	0.011	0.559	natural disaster * good institutions	5,975	0.054	0.226	0	1
topic 3 share	5,228	0.042	0.048	0.006	0.432	security council member	5,975	0.067	0.250	0	1
topic 4 share	5,228	0.056	0.054	0.009	0.597	sec. c. m. * good inst.	5,975	0.014	0.118	0	1
topic 5 share	5,228	0.070	0.043	0.006	0.384	sec. c. m. * cold war	5,975	0.045	0.207	0	1
topic 6 share	5,228	0.058	0.038	0.010	0.765	sec. c. m. * cold war * good inst.	5,975	0.009	0.096	0	1
topic 7 share	5,228	0.075	0.044	0.007	0.514	civil war	5,975	0.084	0.278	0	1
topic 8 share	5,228	0.073	0.051	0.009	0.426	armed conflict	5,975	0.172	0.377	0	1
topic 9 share	5,228	0.073	0.050	0.010	0.514						
topic 10 share	5,228	0.066	0.048	0.011	0.473	*Economics and Political Inst.*					
topic 11 share	5,228	0.066	0.046	0.009	0.407	partial autocracy	6,429	0.201	0.401	0	1
topic 12 share	5,228	0.076	0.068	0.010	0.653	partial democ. w factionalism	6,429	0.172	0.377	0	1
topic 13 share	5,228	0.091	0.092	0.011	0.550	partial democ. wo factionalism	6,429	0.117	0.322	0	1
topic 14 share	5,228	0.066	0.042	0.007	0.582	full democracy	6,429	0.212	0.409	0	1
topic 15 share	5,228	0.063	0.056	0.008	0.437	child mortality (in logs)	6,429	3.912	1.209	0.993	6.095
articles	5,228	124.187	248.286	1	5542	share of pop. discriminated	6,429	0.055	0.152	0	0.98
civil war	5,228	0.082	0.274	0	1	$\geq$4 armed conf. neighbours	6,429	0.021	0.143	0	1
armed conflict	5,228	0.189	0.391	0	1	civil war	6,429	0.077	0.267	0	1
						armed conflict	6,429	0.176	0.381	0	1
Rainfall											
precipitation growth	965	0.020	0.218	-0.609	1.677	*Events*					
precip. growth t-1	965	0.021	0.215	-0.550	1.677	high	2,556	11	38.718	0	712
civil war	965	0.122	0.328	0	1	low	2,556	39.317	146.678	0	2146
armed conflict	965	0.254	0.435	0	1	autocracy score	2,556	-1.930	17.031	-88	1.00E+01
						democracy score	2,556	1.604	18.008	-88	10
Keyword Count						share of pop. excluded	2,556	0.148	0.208	0	0.879
conflict word count	5,097	140.361	368.114	0	10657	share of pop. excluded ^{2}	2,556	0.065	0.152	0	0.773
total conflict words	5,097	19621.310	5680.796	3228	30790	child mortality (in logs)	2,556	3.384	1.207	0.993	5.633
total conflict words^2	5,097	2971693	9005532	0	3.28E+08	$\geq$4 armed conf. neighbours	2,556	0.034	0.181	0	1
poility 2 score	5,097	1.424	7.135	-10	10	share of pop. discriminated	2,556	0.033	0.101	0	0.848
years since last conf.	5,097	26.090	21.951	0	67	civil war	2,556	0.057	0.232	0	1
years since last conf.^{2}	5,097	1162.429	1305.943	0	4489	armed conflict	2,556	0.180	0.385	0	1
years since last conf.^{3}	5,097	57806.980	79030.080	0	300763		2,556	0.180	0.385	0	1
civil war	5,097	0.084	0.277	0	1						
armed conflict	5,097	0.193	0.395	0	1						

Notes: All variables are intersections of the variables and the conflict variables "armed conflict" and "civil war". The rainfall data includes only African countries. For variable descriptions see Section C. Topic shares and number of articles are from the year 2013. The most prominent words of the 15 topics in 2013 are exhibited in Table I.1.

D Formal Discussion of LDA and the Gibbs Sampler

The LDA generates a stream of observable words, $w_{m,n}$, partitioned into documents, which are vectors of words, w_m, i.e. the order of words does not matter.[35] The model assumes that for each of these documents, a vector of topic proportions, η_m, is drawn from a Dirichlet distribution $Dir(\alpha)$. From this, topic-specific words are emitted. That is, for each word, a topic indicator $z_{m,n}$ is sampled according to the document-specific mixture proportion, and then the corresponding topic-specific term distribution, $\varphi_{z_{m,n}}$, is used to draw a word. The topics φ_k are sampled from a Dirichlet distribution $Dir(\beta)$ once for the entire corpus. The key in estimating this model is that only the w_m are actually observed. Everything else needs to be backed out. Typically, the elements of the vectors α and β are assumed to be the same for all documents and topics, respectively. The LDA model can therefore be described by three parameters α, β and the number of topics K.

For statistical inference we use a Gibbs sampling technique, which is a Markov chain Monte Carlo method. At the very heart of the algorithm is the likelihood that a word i in a document m is attributed to topic k in a step of the chain. Denote the term of word i as t and assume the total number of terms is V. The likelihood of attributing i to k is then

$$p\left(z_i = k \mid z_{-i}, w\right) \propto \frac{n_{k,-i}^{(t)} + \beta}{\sum_{t=1}^{V} n_{k,-i}^{(t)} + \beta} \cdot \frac{n_{m,-i}^{(k)} + \alpha}{\left[\sum_{k=1}^{K} n_m^{(k)} + \alpha\right] - 1}$$

where $n_{k,-i}^{(t)}$ is the frequency by which word i was attributed to the same topic generally and $n_{m,-i}^{(k)}$ is the frequency by which all other words in the same document m are attributed to the topic. The equation above highlights the role played by co-occurrence. The algorithm forms topics around terms t that appear together in many documents. The probability $\left(n_{m,-i}^{(k)} + \alpha\right)\left(\left[\sum_{k=1}^{K} n_m^{(k)} + \alpha\right] - 1\right)^{-1}$ ensures that if a lot of terms in a text are attributed to the same topic then it is more likely that token i in the same text will also be attributed to the same topic. High values of α imply that each article is likely to consist of a mix of many topics. Analogously, a high value of β favours a topic to contain a mixture of most words, whereas low values allow topics to consist of a limited number of prominent words.

[35] The following description is based on Heinrich (2009).

We let the chain run for 1000 iterations.[36]

E Robustness of Main Findings

In this section we discuss the robustness of our main findings.

E.1 Conflict Incidence

In principle, both the prediction of onset and incidence should be of interest. Predicting onset is a lot more demanding as we have to estimate the parameters of the model from a reduced sample of years. Incidence is of interest as it produces an overall measure of start, continuation and end of conflict. However, since most conflicts contain a large number of consecutive conflict years, findings here are driven to a large degree by conflict continuation and not onset or end of conflict.

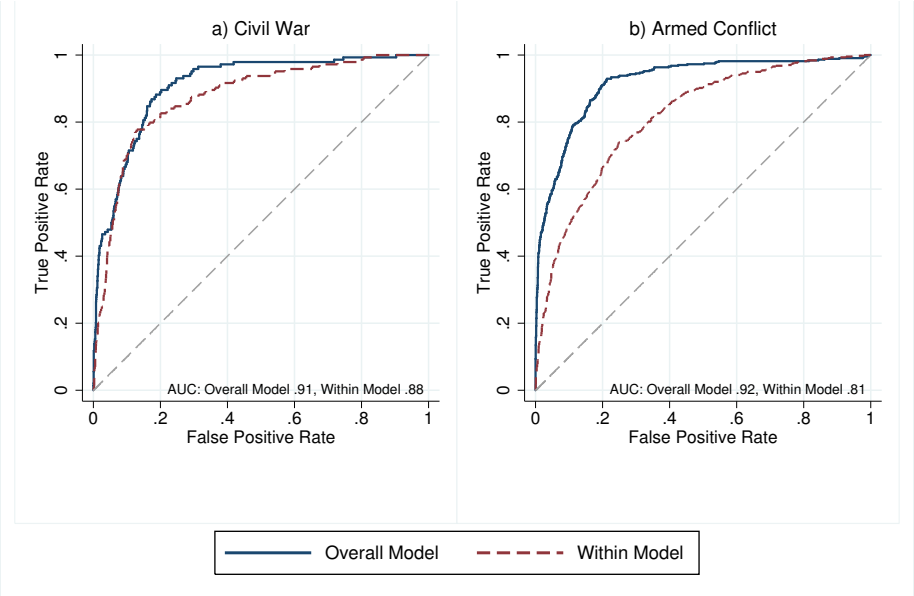
Our topic model performs extremely well when predicting conflict incidence, as can be seen in Figure E.1. The overall model can predict 90 percent of both civil wars and armed conflicts correctly at a FPR of only 20 percent. At a false positive rate of 50 percent the true positive rate is close to 1. A large share of this forecasting power comes from the within variation. When predicting civil wars the within variation reaches a TPR rate of 80 percent for a FPR of 20 percent. For armed conflict the within variation of the news model performs less well but still generates a TPR of above 60 percent for a FPR of 20 percent. The better performance of the incidence models is driven, to a large degree, by the fact that conflict follows conflict. Nonetheless, we believe that the fact that topics can pick this up is useful.

E.2 More or Less Topics

One important question is whether the forecasting performance is sensitive to the number of topics we have chosen. Note that for each topic model we estimate, we need to estimate the model for multiple final years T. As a consequence the estimation of a topic model with 15 topics takes about ten days. For more topics the time increases considerably. While we have mostly stuck to standard

[36]The C++ Gibbs Sampler we use is provided by Phan and Nguyen (2007). We use the default values for burn-in and thinning. For a detailed and user-friendly description of the usage of LDA for topic modelling, we refer to Heinrich (2009).

Figure E.1: ROC Curves for Incidence (Topics Model)



Notes: Predictions result from a panel estimated as in equation (2). The topic model contains 15 topics as θ_{it} derived using LDA with $\alpha = 3.33$ and $\beta = 0.01$ and are aggregated at the country-year level. The within model is the overall model net of country fixed effects as presented in equation (3).

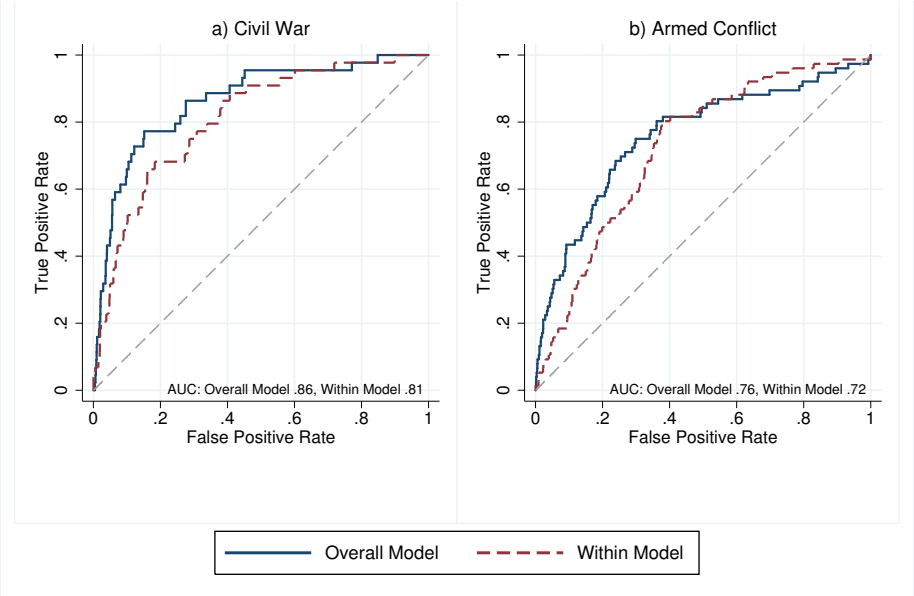
assumptions, our experiments with different values for α and β did not reveal any systematic effect on predictions. The number of topics, however, does affect performance.

Figures E.2, E.3, and Figure E.4 show the performance with 5, 50, and 70 topics, respectively. When we move to less topics the performance only suffers slightly, whereas the drop in predictive power is more pronounced with substantially more topics due to overfitting.

E.3 Combining Topics With Other Information

An important question is whether our news model can add to an existing standard model. This is the approach typically taken by the political science literature, which augments existing models with news data. As it is the natural benchmark for our analysis, we use a model which includes four political regime dummies from polity IV, infant mortality, the share of the population that is discriminated against, and a dummy that captures whether more than three neighboring countries had an armed conflict. To this we add the 15 topics. In Figure E.5 we show that, building on this model (based on Goldstone et al. 2010) the topic shares add forecasting power. For civil war the AUC of the overall model increases from 0.79 to 0.84 and from 0.77 to 0.78 for armed conflict due

Figure E.2: ROC Curves for Onset (Five Topics)



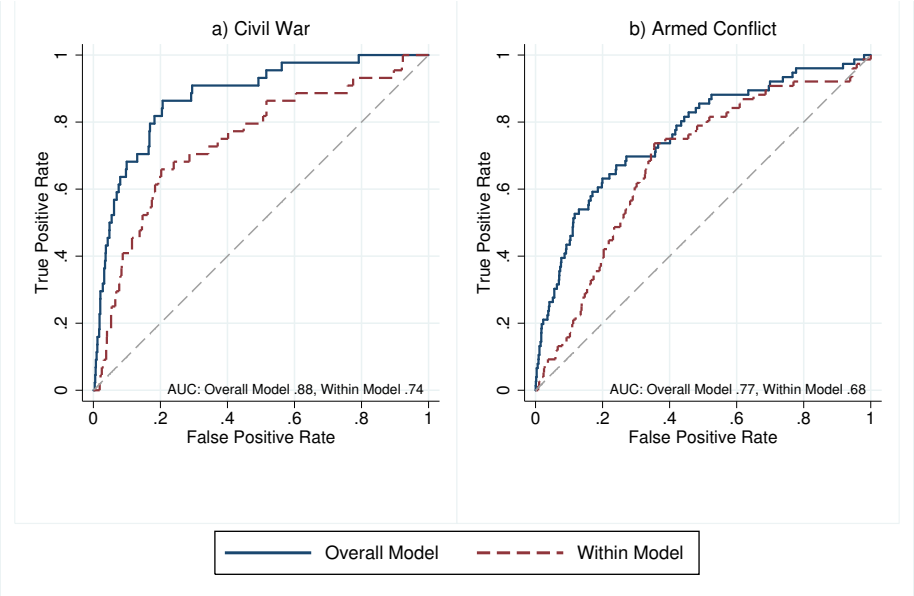
Notes: Predictions result from a panel estimated as in equation (2). The topic model contains five topics as θ_{it} derived using LDA with $\alpha = 10$ and $\beta = 0.01$ and are aggregated at the country-year level. The within model is the overall model net of country fixed effects as presented in equation (3).

to the inclusion of our topics. The greatest gains are for the within model, with an increase of AUC from 0.64 to 0.78 and 0.61 to 0.69, respectively. However, note that the augmented model performs worse than the topic model alone across all dimensions, except for the AUC of the overall model for armed conflict, which stays the same.

Given that civil war often follows weaker forms of violence, we address the question whether topics are capable of improving upon a prediction that uses armed conflict to predict civil war onset. In panel (a) of Figure E.6 we present ROC curves for a prediction model including country fixed effects and an armed conflict dummy. Therefore, the model basically is meant to predict conflict escalation. We see that both the overall and the within model have strong predictive power. However, in panel (b) we see that when we add our topics, the AUC increases, in particular for the within model. Therefore, topics do not merely provide information about lower levels of violence, but add substantial information to the prediction of civil war (such as the disappearance of "peaceful" topics emphasized in Section 7).

In Figure E.7 we report the precision, i.e. the fraction of onset warnings that actually turn out to be onsets, for the within model using the armed conflict dummy versus combining the armed

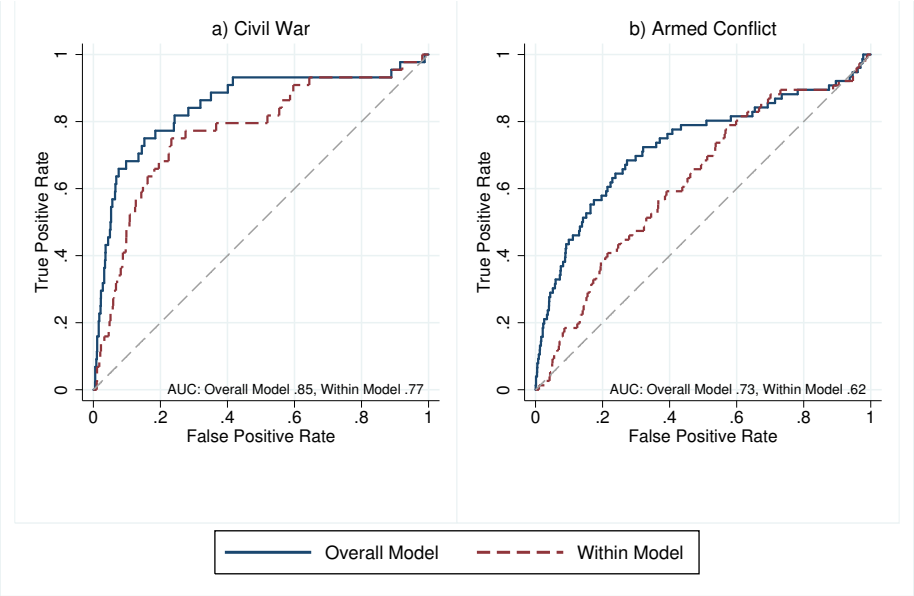
Figure E.3: ROC Curves for Onset (50 Topics)



Notes: Predictions result from a panel estimated as in equation (2). The topic model contains 50 topics as θ_{it} derived using LDA with $\alpha = 1$ and $\beta = 0.01$ and are aggregated at the country-year level. The within model is the overall model net of country fixed effects as presented in equation (3).

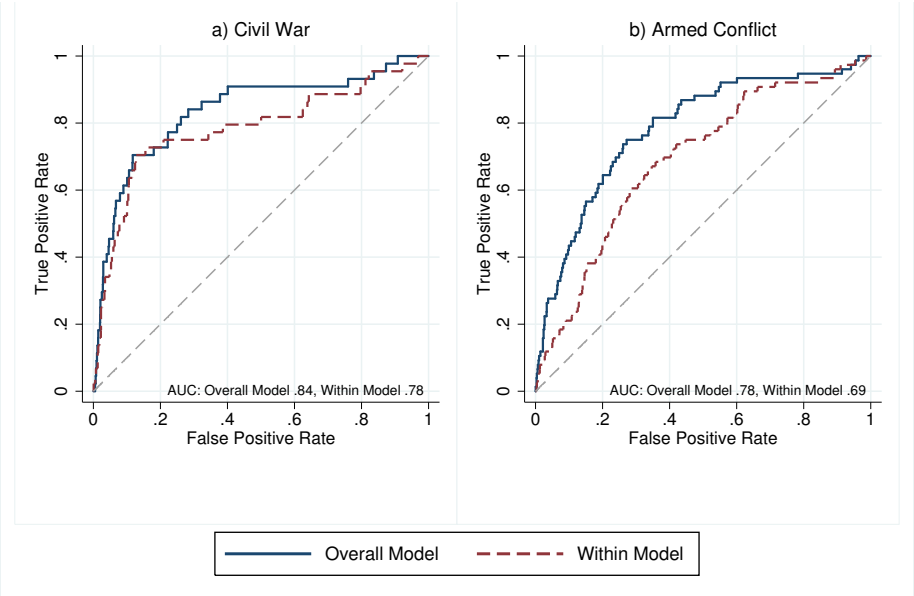
conflict dummy with our topics. We see that the within model using the armed conflict dummy achieves a precision of about 10 percent, while combined with our topics the precision doubles for most true positive rates.

Figure E.4: ROC Curves for Onset (70 Topics)



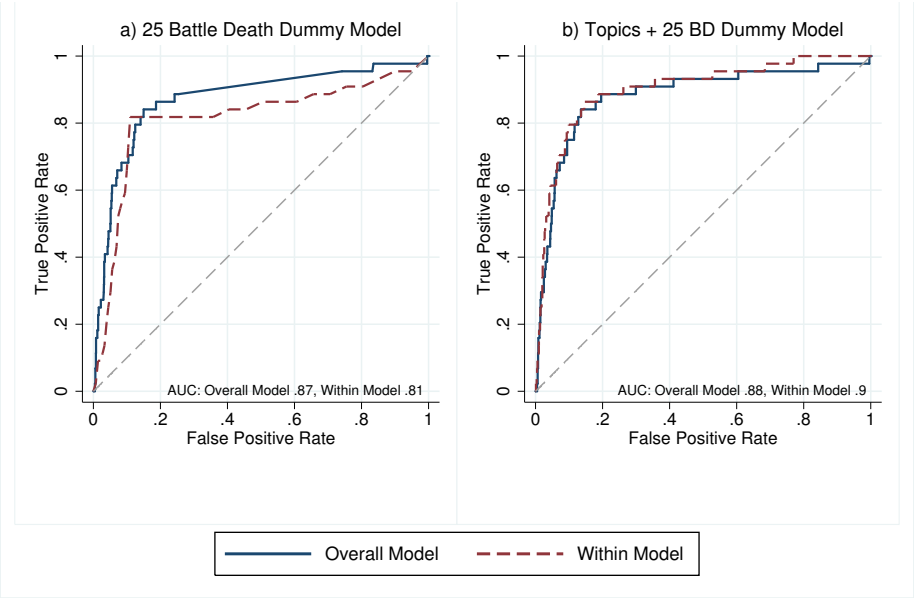
Notes: Predictions result from a panel estimated as in equation (2). The topic model contains 70 topics as θ_{it} derived using LDA with $\alpha = 0.67$ and $\beta = 0.01$ and are aggregated at the country-year level. The within model is the overall model net of country fixed effects as presented in equation (3).

Figure E.5: ROC Curves for Onset (Economic and Political Model Plus Topics)



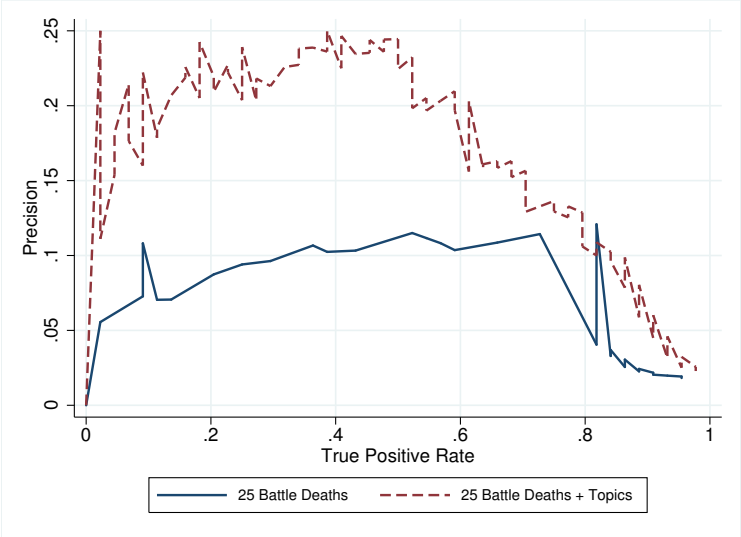
Notes: The Economic and Political is based on Goldstone et al. (2010) and includes four political regime dummies from polity IV, infant mortality, the share of the population that is discriminated against, and a dummy that captures whether more than three neighboring countries had an armed conflict to which we add topics based on 15 topics computed using LDA with $\alpha = 3.33$ and $\beta = 0.01$. The within model is the overall model net of country fixed effects as presented in equation (3).

Figure E.6: ROC Curves for Onset of Civil War (Topics Model Combined with Armed Conflict Dummy)



Notes: Predictions result from a panel estimated as in equation (2). The topic model contains 15 topics as θ_{it} derived using LDA with $\alpha = 3.33$ and $\beta = 0.01$ and are aggregated at the country-year level. The within model is the overall model net of country fixed effects as presented in equation (3).

Figure E.7: Precision Recall Curves Using Only Within Model (Topics Model Combined with Armed Conflict Dummy)



Notes: Predictions result from a panel estimated as in equation (2). The topic model contains 15 topics as θ_{it} derived using LDA with $\alpha = 3.33$ and $\beta = 0.01$ and are aggregated at the country-year level. Both lines are from the within model, which is the overall model net of country fixed effects as presented in equation (3).

E.4 Other Conflict Definitions and Other Outcomes

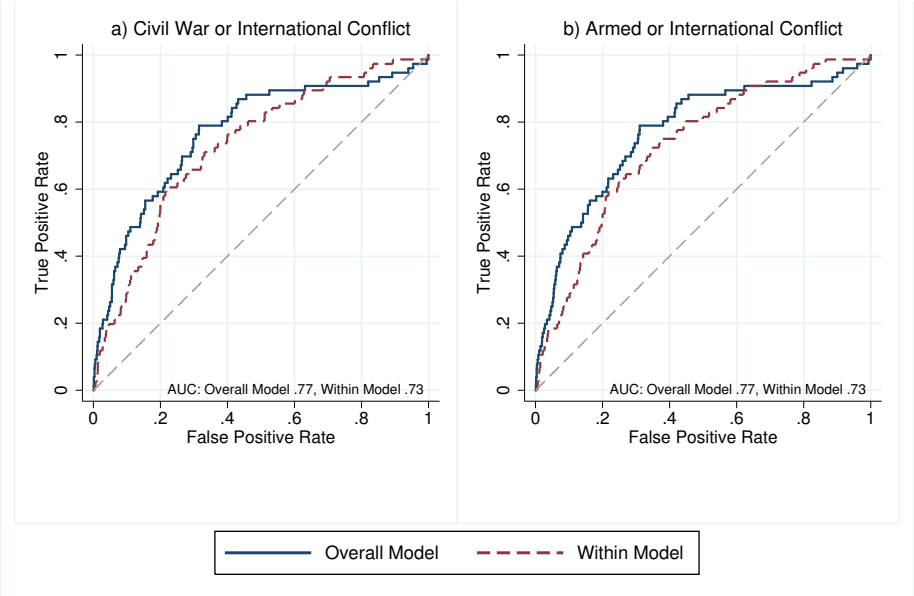
We have also experimented with the type of conflict we predict. To do this we have changed our conflict definition to include all types of conflict (including external wars). Results remain as can be seen in Figure E.8. We have also used only battle-related deaths occurring in internal wars and have used only the best estimate of those. Our within results are getting slightly stronger under this more restrictive definition of internal conflict. In addition, we have experimented with a different dataset on political violence used by Besley and Persson (2011*b*). Here, violence includes purges from the dataset of Banks (2005) and data on armed conflict from the Armed Conflict Database. Our model is able to forecast both incidence and onset of political violence in this data (panel (a) in Figure E.9). In particular, the within and overall model perform similarly. However, the predictive power when forecasting onset is reduced considerably. This likely reflects the fact that only relatively few onsets occur in the shorter period of time (1996-2005). We also follow Mueller (2016) and define conflict as an armed conflict that exceeded an intensity of 0.08 battle-related deaths per 1000 inhabitants. The idea here is that the importance of the event at the country level should follow a per-capita logic. A conflict with 25 casualties in India, for example, might not be as newsworthy for national news agencies as if the same event would take place in Venezuela. Again our topic model exhibits high predictive power using the within and overall variation as can be seen in panel (b) of Figure E.9.

In Figure E.10 we use the upper bound estimated for battle deaths by PRIO to define armed conflict and civil war. Compared to using the "best" estimate provided by PRIO as in Figure 4, there are no notable changes to the AUC. This finding provides further evidence that our methodology is not sensitive to conflict definitions or any specific type of conflict.

Moreover, we test our predictive power concerning refugees, an outcome closely related to violence and reported by the UNHCR. We use this data to construct the total number of refugees who have left their country of origin. In light of the discussion of news biases this data has the advantage that it is collected using registers, surveys, registration processes and censuses.

The number of refugees is almost uniformly distributed from 1 to several million, which makes the choice of the right cutoff difficult. We, therefore, take an agnostic approach and define two cutoffs so that we get ten percent and five percent of country-years with a number of refugees above

Figure E.8: ROC Curves for Onset Including International Conflicts (Topics Model)



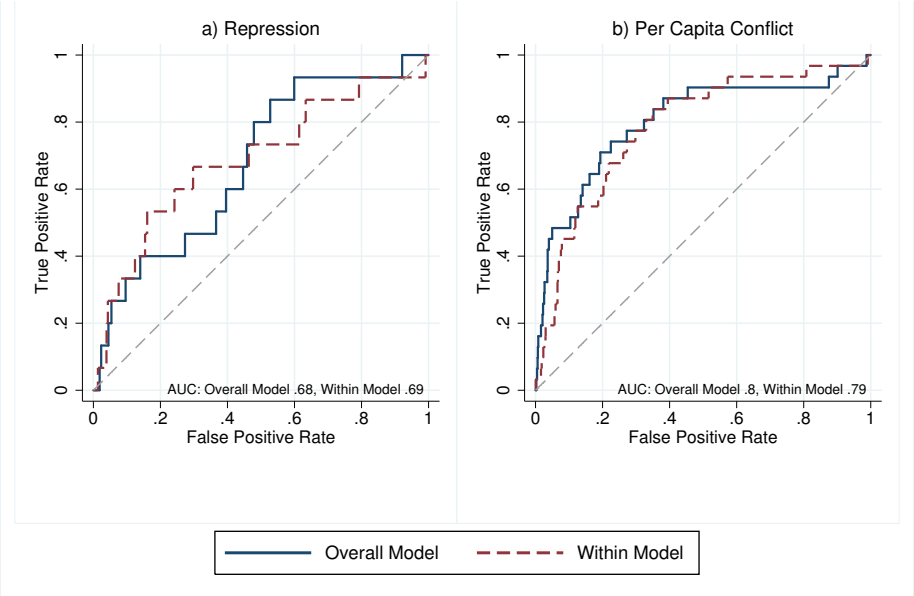
Notes: Predictions result from a panel estimated as in equation (2). The topic model contains 15 topics as θ_{it} derived using LDA with $\alpha = 3.33$ and $\beta = 0.01$ and are aggregated at the country-year level. The within model is the overall model net of country fixed effects as presented in equation (3).

the threshold. This gives us cutoffs of 30,000 and 130,000 refugees. The resulting dummy variables have frequencies comparable to armed conflict and civil war.

We then use our topic model to test whether we can predict whether a large number of refugees will leave the country in the next year. In panel (a) of Figure E.11, we show that the onset of more than 130,000 refugees can be predicted somewhat with our model. In panel (b) we predict the onset of 30,000 refugees and results are very similar. What is striking here is that the overall and within models perform very similarly, with the within model sometimes exceeding the relatively weak predictive power of the overall model. This is important as refugee numbers are often reported by local aid agencies and not by news agencies. News are therefore able to forecast events not mainly collected by news sources (as most of the violence is).

We also use our model to forecast conflict onset one or two years before it happens. The results, displayed in Figure E.12 are qualitatively similar to our main results. However, an interesting change is that while the overall model performs almost as before, the within model performs slightly worse. This underlines the difference in the logic between these two models. Predicting the onset of conflict is harder two years before it occurs if one wants to predict the timing of it.

Figure E.9: ROC Curves for Onset of Repression and Conflict in Per Capita Terms (Topics Model)



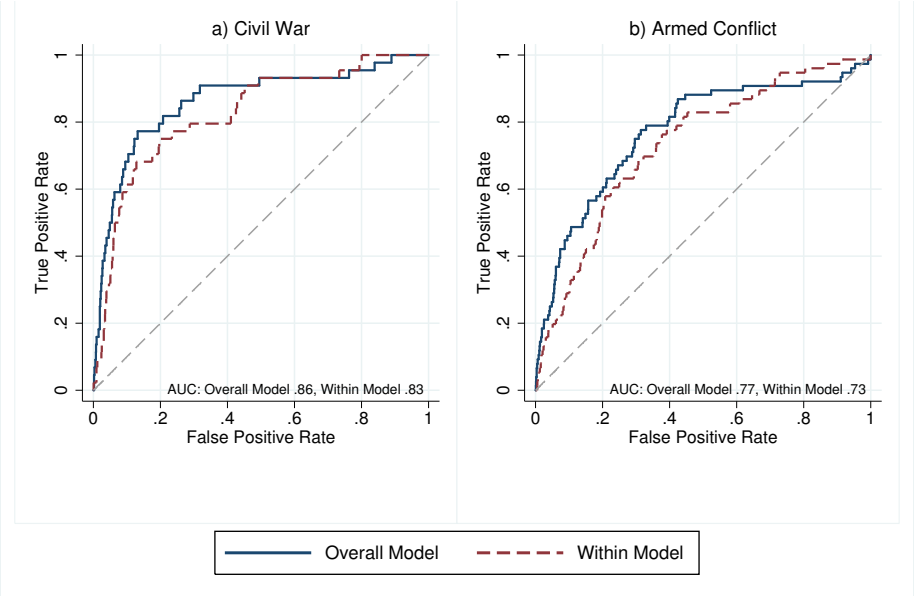
Notes: Predictions result from a panel estimated as in equation (2). The topic model contains 15 topics as θ_{it} derived using LDA with $\alpha = 3.33$ and $\beta = 0.01$ and are aggregated at the country-year level. The within model is the overall model net of country fixed effects as presented in equation (3).

E.5 Robustness Across Space and Time

One remaining question is whether topics are capable of predicting conflict consistently across pace and time or whether the predictions are a very local phenomenon. In Figure E.13 we present evidence that topics are consistently capable of predicting conflict across regions. Here we present the results when comparing ranked predictions within Africa and Asia, and see that the model maintains predictive power across regions. Breaking the ROC down by further subregions does not affect this finding.

In Figure E.14 we exhibit the performance of the model for every single year. The predictive power does not seem to exhibit a time trend or any particularly weak prolonged interval, i.e. the model seems to be valid across time. This finding is strengthened by Figure E.15, for which we rank predictions within each year before generating the aggregate ROC curves. Again, we see that the AUC is high for the overall and within model for both civil war and armed conflict.

Figure E.10: ROC Curves for Onset With High PRIO Estimates (Topics Model)



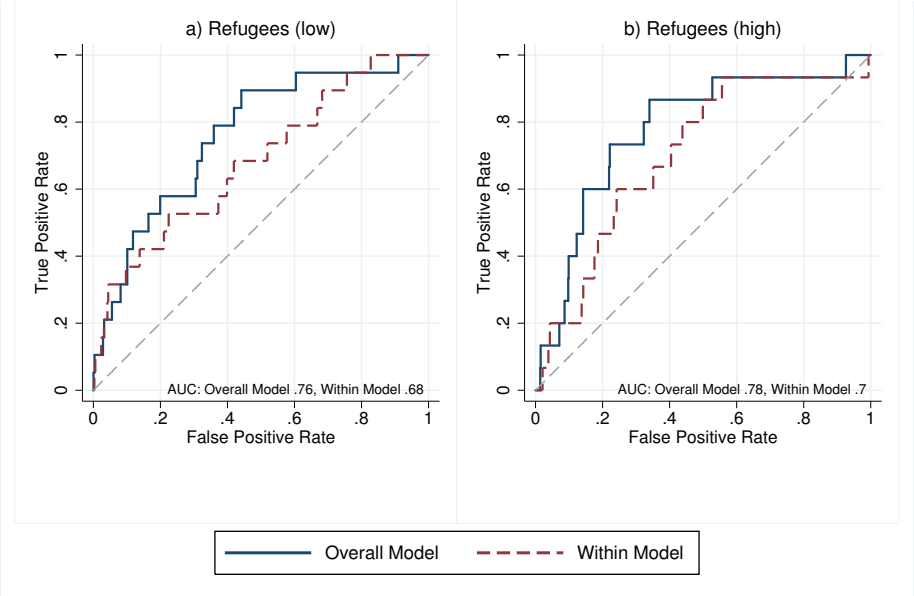
Notes: Predictions result from a panel estimated as in equation (2). The topic model contains 15 topics as θ_{it} derived using LDA with $\alpha = 3.33$ and $\beta = 0.01$ and are aggregated at the country-year level. The within model is the overall model net of country fixed effects as presented in equation (3).

E.6 Comparison to Counts and Events

As explained in the text it is interesting to look at word counts and events without the additional standard variables and compare them to topics. As most standard variables do not provide useful within variation this is the much better comparison. There are two views here. If topics only provide noise through their depth and width then simple keywords and conflict events should be more useful for forecasting. If there is actual information contained in depth and width then topics should help.

In Figure E.16 we first show ROC curves for civil wars and in Figure E.17 we show armed conflicts. To make the comparison easier we always report the ROC curves for the event and keyword count models as blue lines and the ROC curve of the topic model as red dashed lines. We always show the overall performance in panel (a) and the within performance in panel (b). From this it is clear that events and keywords are better predictors without the standard variables included in the original model and that topics are still able to forecast the timing better.

Figure E.11: ROC Curves for Refugee Flows



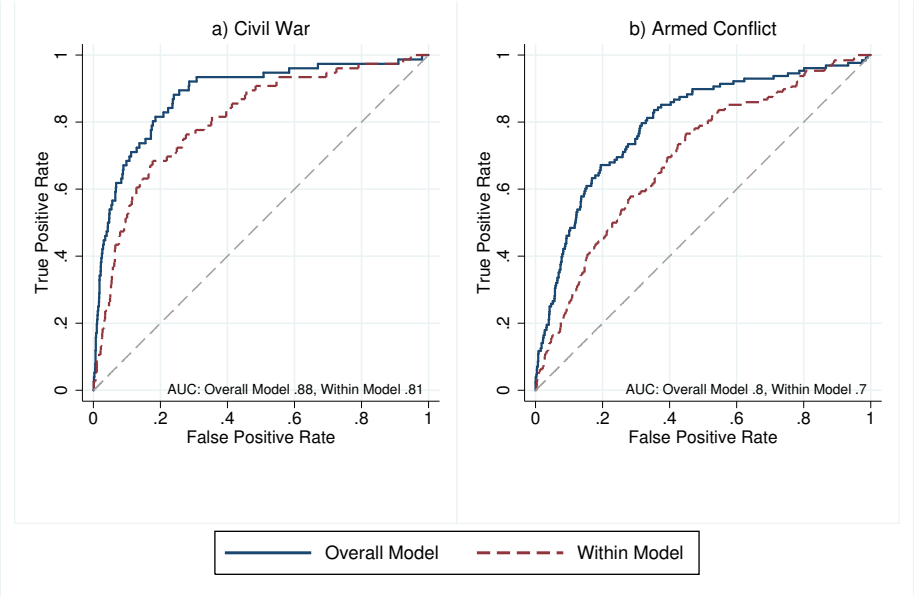
Notes: Predictions result from a panel estimated as in equation (2). The topic model contains 15 topics as θ_{it} derived using LDA with $\alpha = 3.33$ and $\beta = 0.01$ and are aggregated at the country-year level. The within model is the overall model net of country fixed effects as presented in equation (3).

E.7 Robustness Using Logit Without Country Fixed Effects

In Figure E.18 we present the results of a logit model without country fixed effects; a model which has been popular in the literature due to its prediction bound between 0-1, which therefore is interpretable as a probability. In each panel we compare the results for a model from the literature to the performance of the topic model only for years and countries in which we have predictions of both models. We add the keyword count model excluding the variables concerning how many years have passed since the last conflict, as well as the event model with events only (see Appendix C for more information on the included variables). Both of these news-based models rely on additional information in the cross-section, which captures the between variation well, as indicated by the performance drop when removing these variables and the good performance of the model using only the variables capturing years since conflict.

This exercise highlights that the predictive capacity of the information captured by our topic shares is not limited to using a linear probability model with country fixed effects, i.e. topics can predict conflict whether using a logit or a fixed effect framework. Again topics outperform all other models. However, we emphasize that using a logit model comes at the cost of not being able to

Figure E.12: ROC Curves for Onset Two Years Before Conflict (Topics Model)



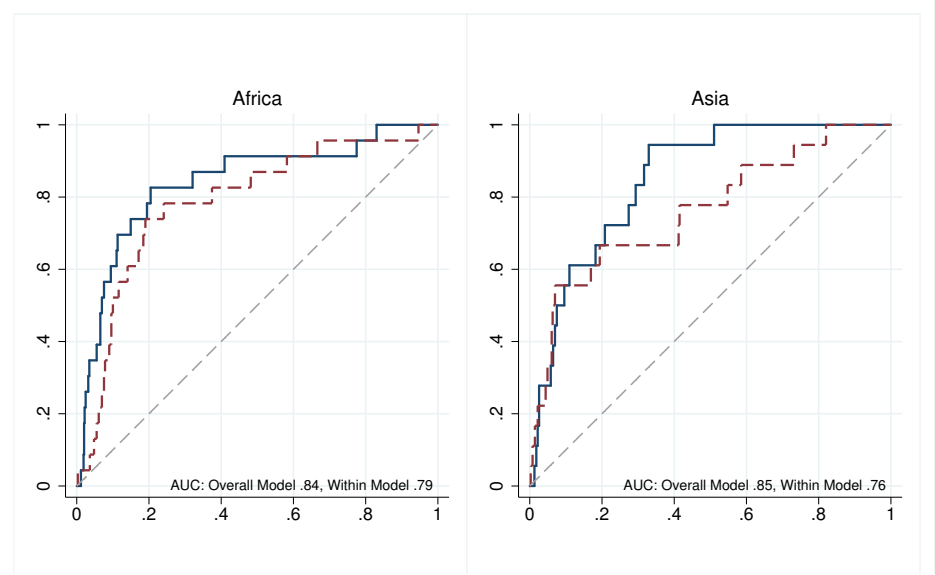
Notes: Predictions result from a panel estimated as in equation (2). The topic model contains 15 topics as θ_{it} derived using LDA with $\alpha = 3.33$ and $\beta = 0.01$ and are aggregated at the country-year level. The within model is the overall model net of country fixed effects as presented in equation (3).

separate the between from the within variation. Using a conditional logit model, which is similar to incorporating country fixed effects, is also not a remedy. In order to predict using a conditional logit model, one needs to make arbitrary assumptions about the "fixed effect" of a country that either never or always has been experiencing conflict. These arbitrary assumptions will determine whether countries without conflict will ever be considered at risk.

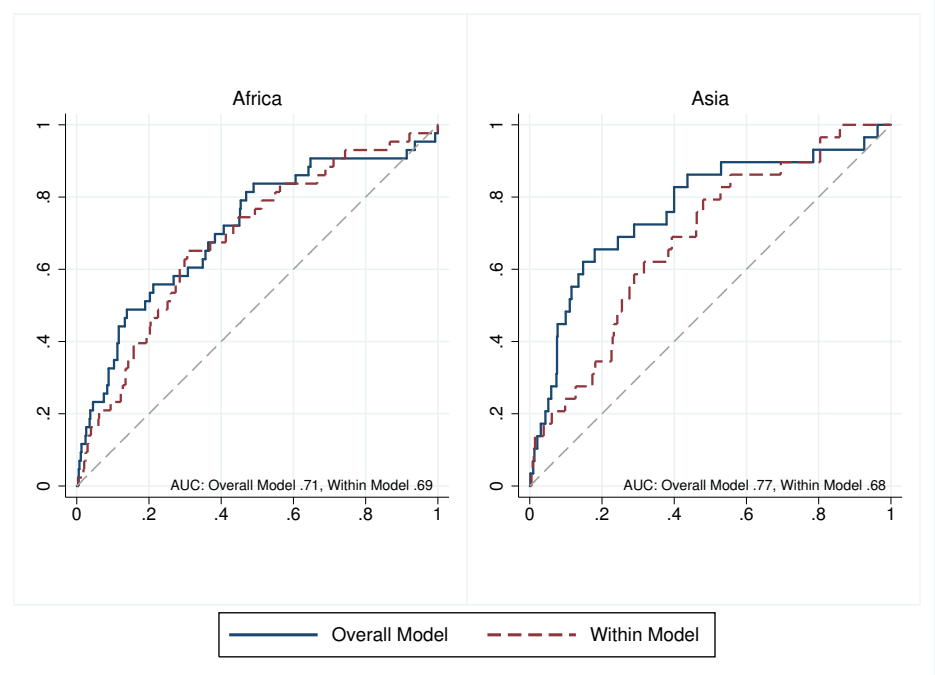
By comparing the ROC curves in E.18 to those produced with a linear probability model with country fixed effects, which are exhibited in Figures I.2 and I.3, we can tell that the logit model seems to produce a slightly higher AUC. The highest AUC our topic provides for civil war is 0.92, compared to 0.87 produced using the linear probability model. For armed conflict, the AUC combining topics with the variables capturing years since last conflict achieve an AUC of 0.84, which is higher than the 0.77 achieved with topics in the linear probability model. However, it is now again unclear whether the forecast relies on useful within variation or simply exploits the between variation.

Figure E.13: ROC Curves for Onset by Region (X-Axis: FPR, Y-Axis: TPR)

(a) Civil War



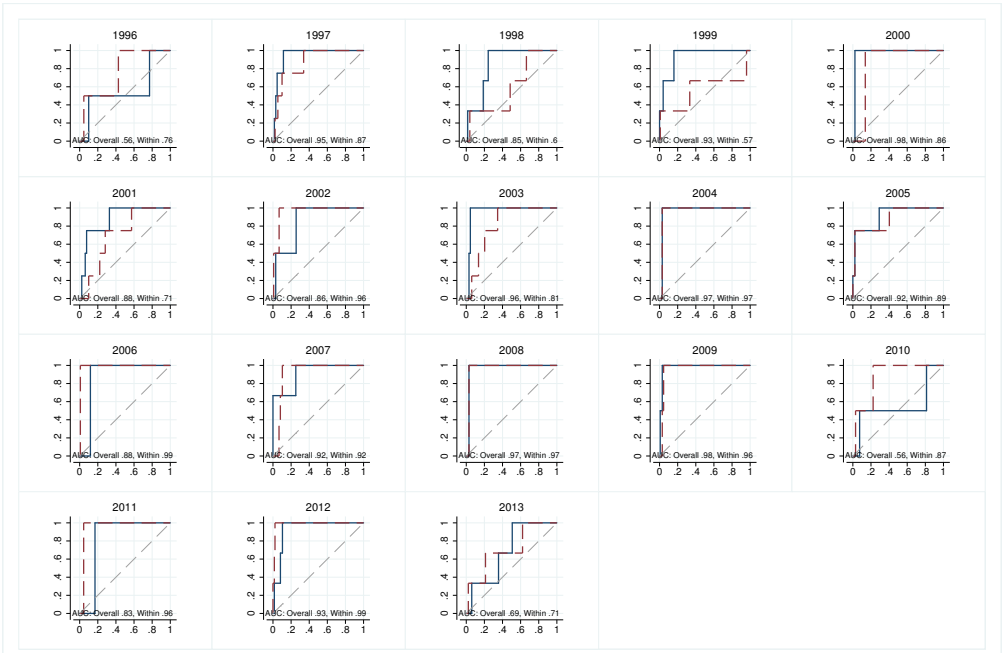
(b) Armed Conflict



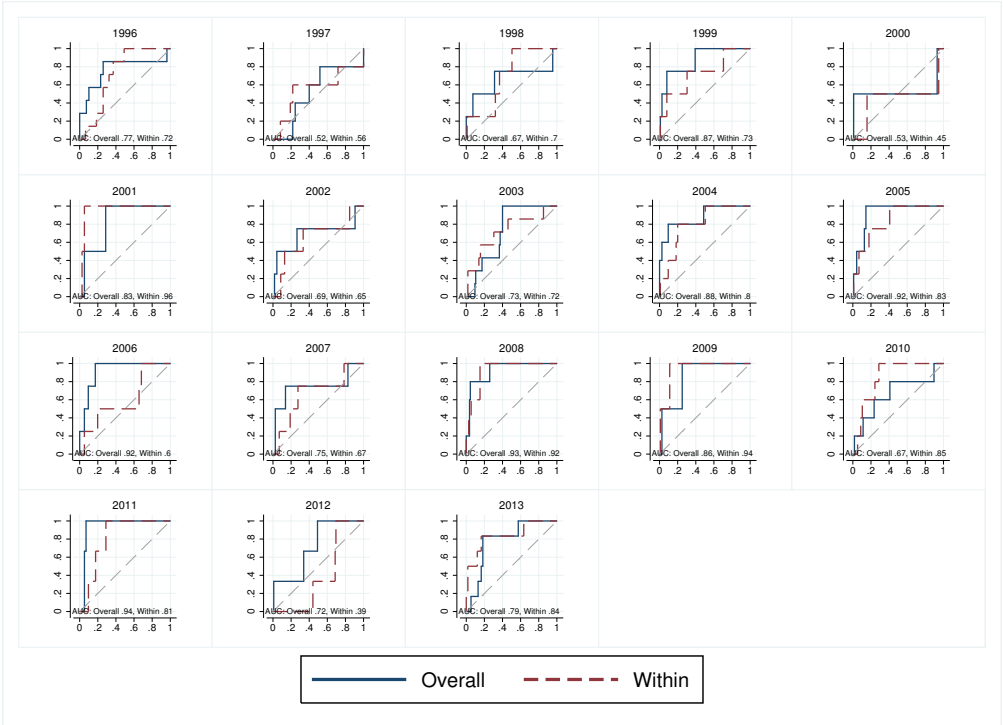
Notes: Predictions result from a panel estimated as in equation (2). The topic model contains 15 topics as θ_{it} derived using LDA with $\alpha = 3.33$ and $\beta = 0.01$, which are aggregated at the country-year level. The within model is the overall model net of country fixed effects as presented in equation (3).

Figure E.14: ROC Curves for Onset by Year (X-Axis: FPR, Y-Axis: TPR)

(a) Civil War

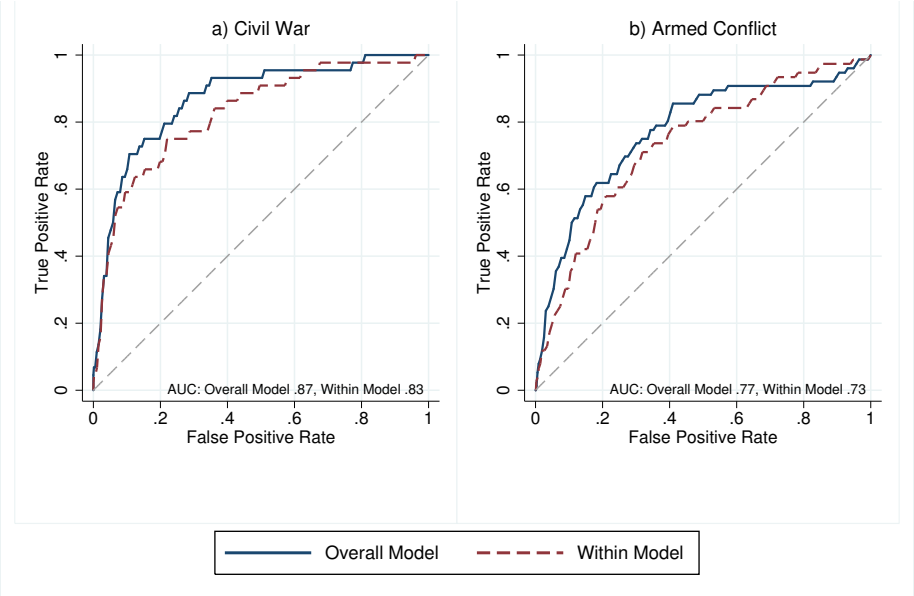


(b) Armed Conflict



Notes: Predictions result from a panel estimated as in equation (2). The topic model contains 15 topics as θ_{it} derived using LDA with $\alpha = 3.33$ and $\beta = 0.01$, which are aggregated at the country-year level. The within model is the overall model net of country fixed effects as presented in equation (3).

Figure E.15: ROC Curves for Onset Ranked Within Year (Topics Model)



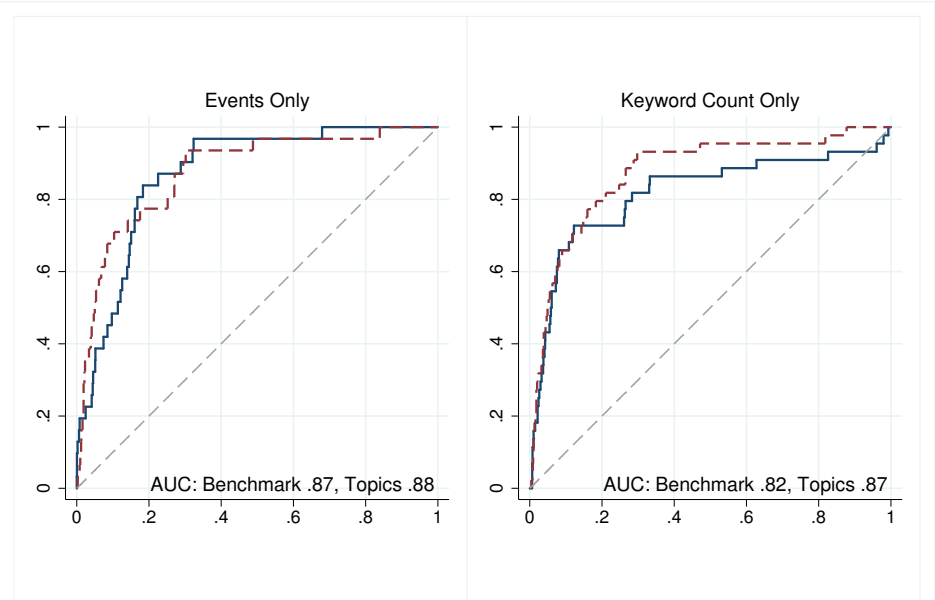
Notes: Predictions result from a panel estimated as in equation (2). The topic model contains 15 topics as θ_{it} derived using LDA with $\alpha = 3.33$ and $\beta = 0.01$ and are aggregated at the country-year level. The within model is the overall model net of country fixed effects as presented in equation (3).

E.8 Using Neural Networks With Topics to Predict Onset

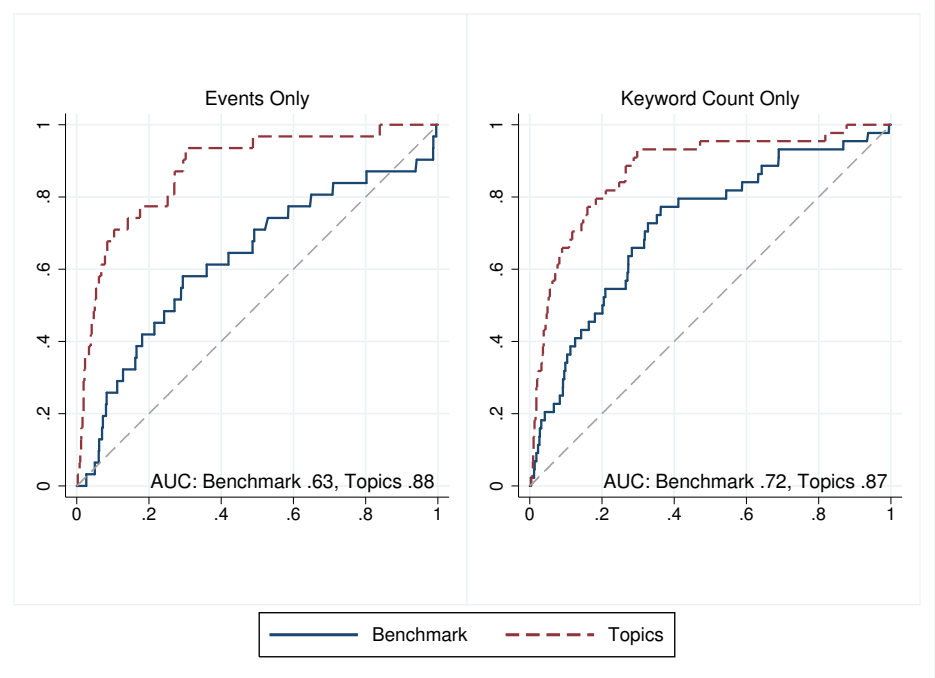
In order to bridge the gap to the existing machine learning and forecasting literature we also use a neural networks technique to forecast conflict with topics. We split the sample with all information until year T into 70% trainings and 30% validation and trained the neural network with 10 hidden layers. As the following step we predict conflict onset in $T + 1$ out-of-sample to generate the ROC curves in Figure E.19. The red dashed line reports a ROC curve of using our topics within a neural network setting. The gains over our within model in Figure 4 are modest. These moderate gains are in line with findings by Goldstone et al. (2010) who find little improvement when switching to neural networks from simpler regression models. In addition, there is no obvious way to investigate whether the forecast uses the within or the between variation to do the forecast. The blue line adds country dummies and here the forecast for armed conflicts becomes much worse. We think this is because of an "overfitting" to the data due to the large number of dummies.

Figure E.16: ROC Curves for Onset of Civil War Comparing Only for Overlapping Predictions of News Models (X-Axis: FPR, Y-Axis: TPR)

(a) Overall



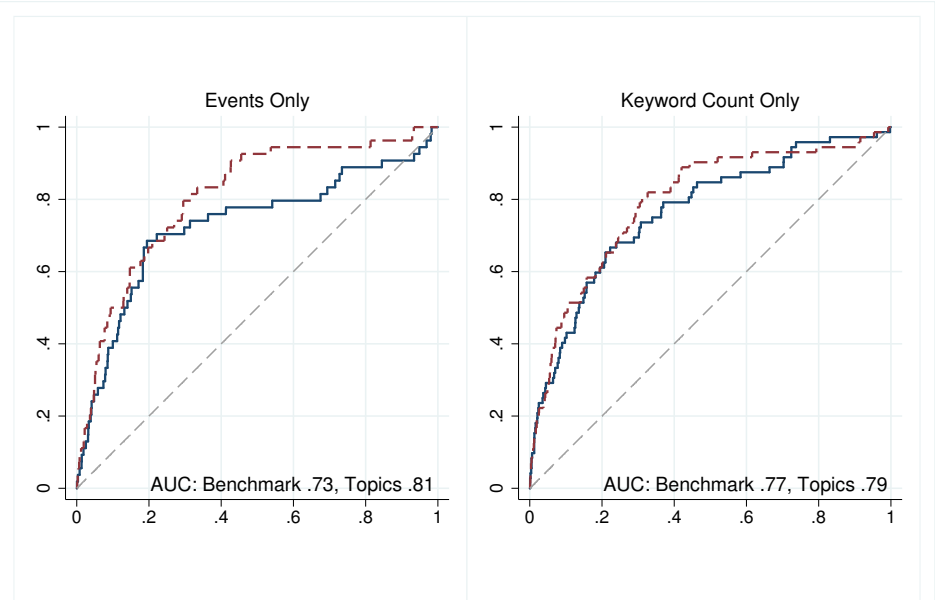
(b) Within



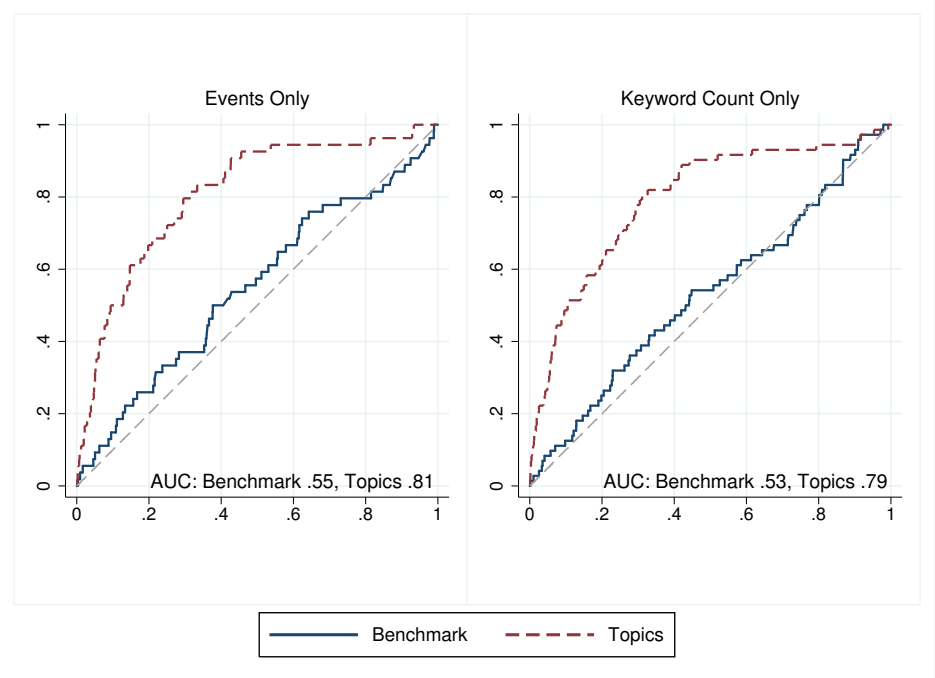
Notes: Predictions result from a panel estimated as in equation (2). The variables included for each model as $\mathbf{x}_{it}$ are specified in Section 3.3 and Appendix Table C.1. The within model is the overall model net of country fixed effects. Here we only include predictions for country-years where we have predictions for both the comparison and the topic model.

Figure E.17: ROC Curves for Onset of Armed Conflict Comparing Only for Overlapping Predictions of News Models (X-Axis: FPR, Y-Axis: TPR)

(a) Overall



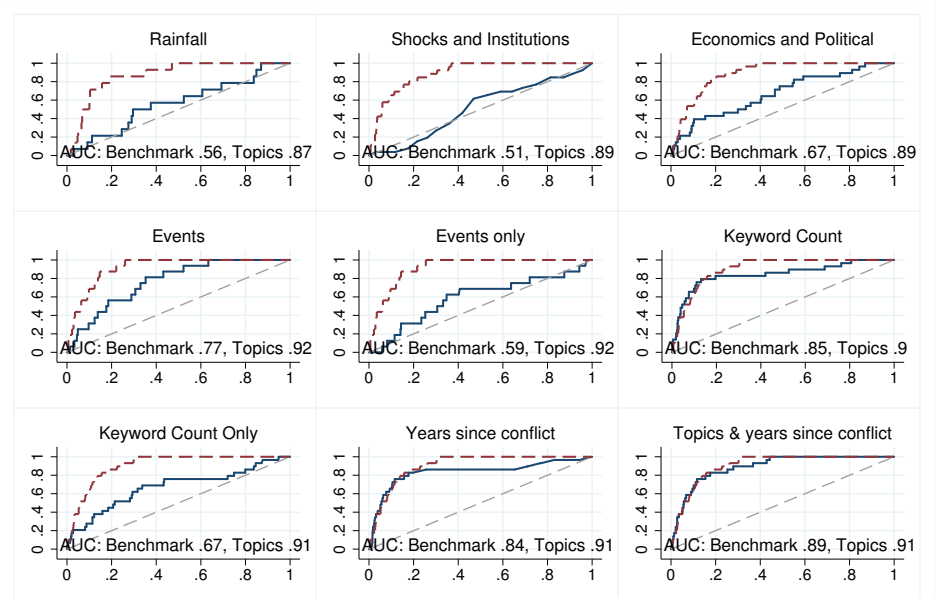
(b) Within



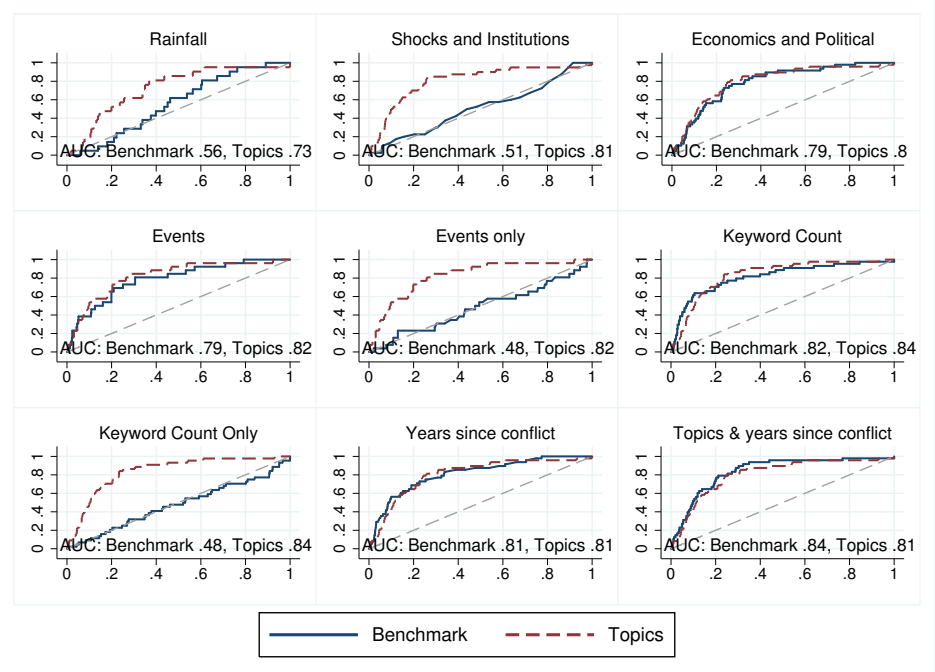
Notes: Predictions result from a panel estimated as in equation (2). The variables included for each model as $\mathbf{x}_{it}$ are specified in Section 3.3 and Appendix Table C.1. The within model is the overall model net of country fixed effects. Here we only include predictions for country-years where we have predictions for both the comparison and the topic model.

Figure E.18: ROC Curves for Onset Comparing Only Overlapping Predictions Using Logit (X-Axis: FPR, Y-Axis: TPR)

(a) Civil War

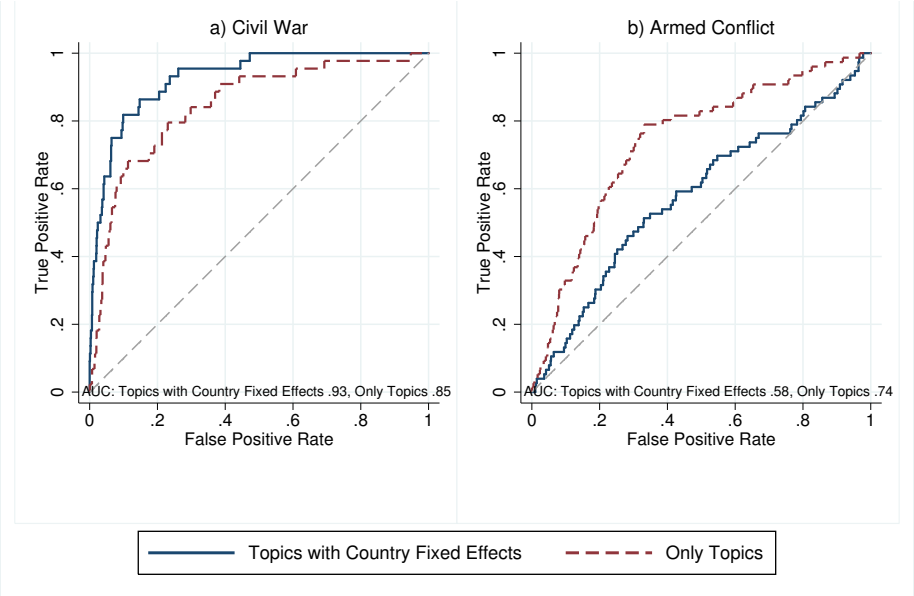


(b) Armed Conflict



Notes: For a description of included variables for each model, we refer to Section 3.3 and E.7 or Appendix Table C.1. In contrast to all other estimations presented in the paper, here we use logit instead of a linear probability model. Here we only include predictions for country-years where we have predictions for both the comparison and the topic model.

Figure E.19: ROC Curves for Onset (Neural Networks)



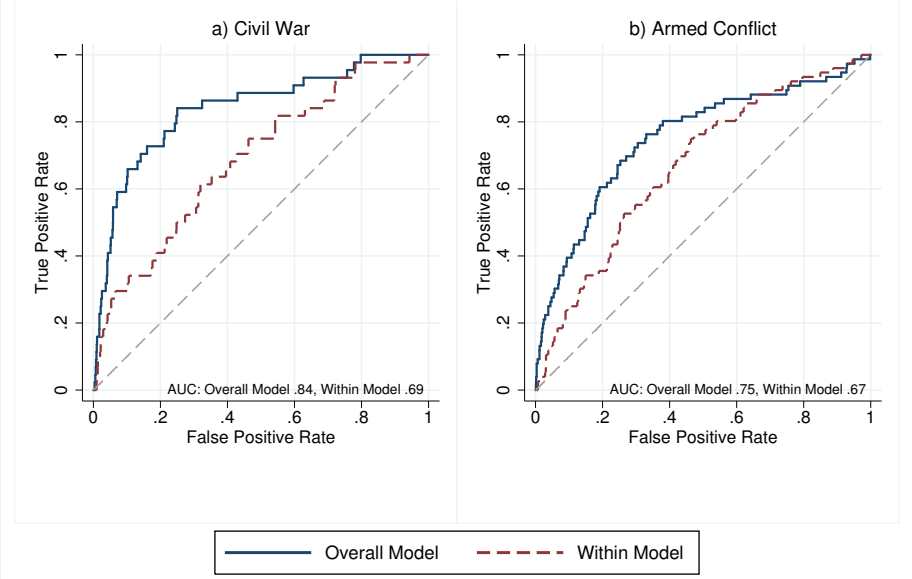
Notes: Predictions result from a neural network with a sample-split of 70% trainings and 30% validation sample and 10 hidden layers. The topics contain 15 topics as θ_{it} derived using LDA with $\alpha = 3.33$ and $\beta = 0.01$ and are aggregated at the country-year level. The country fixed effects models contain a dummy for each country.

E.9 Using Only Conflict Topics to Predict Conflict

In order to evaluate the contribution of conflict topics we run an additional robustness check in which we include only conflict topics (up to 3) in all years.[37] The result of this attempt are in Figure E.20. The predictive power for both armed conflict and civil war is now lower indicating that stabilizing topics play an important role.

[37] We re-normalize the conflict topic shares so they add up to one. Results here are robust to not doing this.

Figure E.20: ROC Curves for Onset (Only Conflict Topics)



Notes: Predictions result from a panel estimated as in equation (2). The topic model contains the three conflict topics as θ_{it} derived using LDA with $\alpha = 3.33$ and $\beta = 0.01$ and are aggregated at the country-year level. We only include the three conflict topics and renormalized so that the three topic shares sum to 1. The within model is the overall model net of country fixed effects as presented in equation (3).

F Bias in the Overall Model Against New Onsets

In this section we show that there can be a bias against new conflicts when using country fixed effects in a forecasting model. We show this first theoretically and then report on a simulation.

Assume that our entire topic model could be captured by a simple dummy $x_{it} \in \{0, 1\}$ which we call "conflict news". High conflict news ($x_{it} = 1$) are followed by conflict onset with a (small) probability of $\eta \in (0, 1)$. This is in addition to a baseline probability of conflict onset ε. Imagine that each country has its own individual probability of generating conflict news $p_i \in [0, 1]$. Given the relationship between conflict and conflict news this implies a country-specific propensity for conflict onset of $\varepsilon + p_i * \eta$.

Assume a situation in which we have T years within sample and are trying to forecast the year $T + 1$. In the fixed effects regression

$$y_{it+1} = \alpha + \beta_i + x_{it}\beta^{FE} + \varepsilon_{it} \tag{5}$$

the estimated β^{FE} will capture the relationship between news and onset a year later within the sample. This relationship is stochastic because η is a probability. In other words, there will be many years in which there are conflict news which are not followed by conflict. Depending on the realisations of news and conflict this stochastic relationship generates different biases in the overall sample compared to the within sample.

Imagine two countries, one with a very low probability of generating conflict news and one with a higher probability of generating conflict news. Assume further that both countries did not experience conflict up until T. However, due to constant reporting on conflict in the country with a lot of news (and higher risk) the estimated fixed effect $\hat{\beta}_i$ will be lower, i.e. more negative, than in the low risk country. The more news are generated without a conflict onset, the lower will be the estimated fixed effect.

In the overall model this has the effect of biasing the playing field against the risky country. Conditional on a conflict-free history we have that

$$E\left[\hat{y}^{overall}_{iT+1} \mid x_{iT}\right]_{low\ risk} > E\left[\hat{y}^{overall}_{iT+1} \mid x_{iT}\right]_{high\ risk}.$$

However, in the within model we have

$$E\left[\hat{y}_{iT+1}^{within} \mid x_{iT}\right]_{low\ risk} = E\left[\hat{y}_{iT+1}^{within} \mid x_{iT}\right]_{high\ risk}.$$

This matters because our evaluation of the forecast relies on the relative ranking between countries in a given year. In the overall model the high-risk country will be ranked lower than the low-risk country for the same true risk. Naturally, the reverse is true for histories with conflict realisations. The overall fitted value will attribute a higher risk in these cases.

How important is this problem in practice? In order to get an idea of the size of the problem we have run 100 iterations of the above model with random realisations of conflict. In each iteration we pick values of p_i and η to match the likelihood of a conflict onset at the country level i in the actual data. We first assume a value for η and then calculate

$$p_i = \bar{y}_i/\eta$$

for each country.[38] For every year we then draw a news shock and if $x_{it} = 1$ we use the likelihood η to generate random onsets in the next year. In addition to this we add some completely random conflict onsets with a small probability of ε.[39]

We then implement our forecasting methodology on the 100 different simulated samples and calculate for each sample the average ranking in country-years followed by conflict which are generated by the two models $\hat{y}_{it+1}^{overall}$ and $\hat{y}_{it+1}^{within}$.

Finally, we separate onsets which take place in country-years without previous onsets and those that followed other onsets. The results of this exercise are striking. The overall model almost completely fails to spot first onsets. The average ranking of first onsets in the model is around 100 out of 146 compared to under 70 in the within model. This changes for repeated onsets where the overall model becomes relatively better and produces an average rank of 34 compared to 39 in the within model.

Note, that here we assumed that the true model is one where there is useful within-country

[38] As far as we can tell the values assumed for η do not change our results. In any case, we want to use a low probability to generate a lot of noise in the model. We report results for $\eta = 0.1$.

[39] For simplicity we treat every conflict onset as an onset even if it is right after an onset in the previous year. Since conflict in the simulated data is iid this is not a problem.

variation, i.e. that $x_{it} = 1$ is a useful predictor. This is what makes the within model powerful.

G Conflict and Non-Conflict Topics

In the robustness section we have shown that even when we exclude conflict topics in our forecasting exercise, our model maintains a good share of its forecasting ability. In this section we demonstrate that the separation of conflict and non-conflict topics means that conflict topics will capture all potential information directly linked to fighting events. We are therefore left with variation which is unlikely to be linked to fighting. We will do this in two steps. First, we illustrate how topics help to summarize text through probability distributions across tokens. The nature of those tokens differs dramatically across topics so that different parts of the text are captured by conflict topics and non-conflict topics. Second, we show that, consistent with this, the share of non-conflict topics falls before conflict onset and falls further (with one exception which we discuss) once conflict starts.

G.1 How Topics Deal with Text

First, we turn towards how the topic model categorizes text into different topics. Take the example of a randomly chosen piece of text published in the NYT on March 29th 1991 on Libya presented in Figure G.1a. After throwing out stop-words and lemmatizing, the piece of text looks as displayed in Figure G.1b.[40] These words (and their two- and three-word combinations) are fed into the Gibbs sampler to generate the topics.

We then use the topics, i.e. the probability distributions over tokens, to identify what topic a text is written on. We now illustrate this step by displaying the probability distribution of two topics in the text. We start with our "tourism" topic which is negatively correlated with conflict. The darker a token in Figure G.1c, the more likely it is to be drawn when the journalist writes about tourism. The most likely tourism tokens in the text are "new", "american" and "year" which are rather generic tokens that are likely in many topics. It is therefore no wonder that our classification method attributed only 4 percent of the text to tourism (i.e. less than if it were random).

However, the picture changes dramatically when we overlay the probability distribution of one

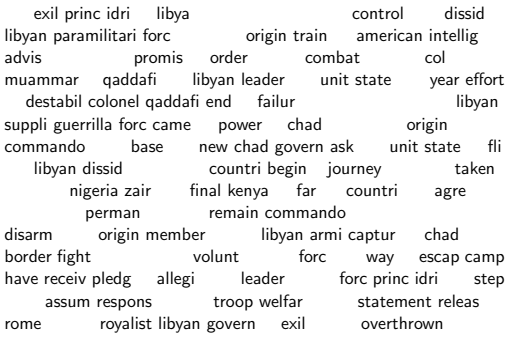
[40] We additionally try to remove names of people, identified by a library of names and the usage of titles, such as "Mr" or "Mrs". We use the Natural Language Toolkit dictionary of names for males "names.words('male.txt')" and females "names.words('female.txt')".

Figure G.1: Transforming Raw Text Into Topics

(a) Raw text

The exiled Prince Idris of Libya has said he will take control of a dissident Libyan paramilitary force that was originally trained by American intelligence advisers, and he has promised to order it into combat against Col. Muammar el-Qaddafi, the Libyan leader. The United States' two-year effort to destabilize Colonel Qaddafi ended in failure in December, when a Libyan-supplied guerrilla force came to power in Chad, where the original 600 commandos were based. The new Chad Government asked the United States to fly the Libyan dissidents out of the country, beginning a journey that has taken them to Nigeria, Zaire and finally Kenya. So far, no country has agreed to take them permanently. The 400 remaining commandos, who have been disarmed, were originally members of the Libyan Army captured by Chad in border fighting in 1988. They volunteered for the force as a way of escaping P.O.W. camps. "Having received pledges of allegiance from leaders of the force, Prince Idris has stepped in to assume responsibility for the troops' welfare," said a statement released in Rome by the royalist Libyan government in exile. It was overthrown in 1969.

(b) Without stop words and lemmatized



(c) Tourism topic intensity (4%)

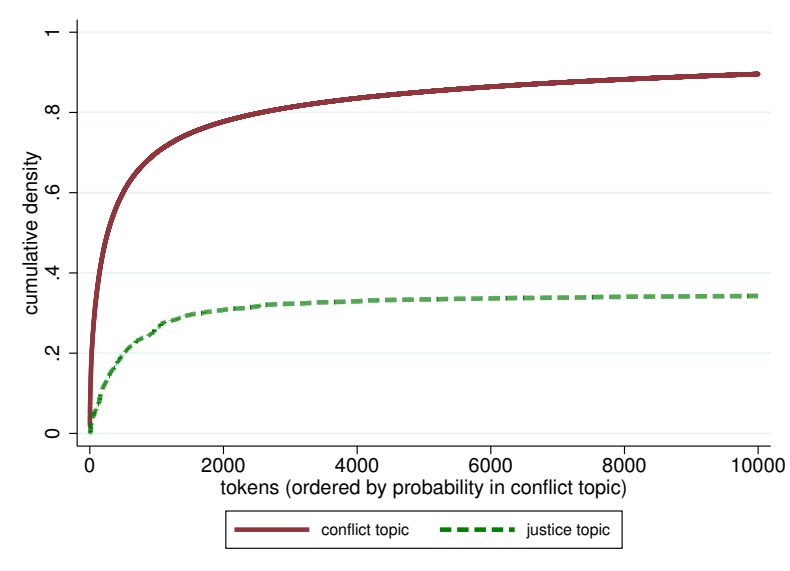
(d) Conflict topic intensity (27%)

Notes: The text in panel (a) was published in the NYT on March 29th 1991. Panel (b) shows the text after removing stop words and lemmatizing. Panel c) shows darker shades for *tourism* words and panel d) for *conflict* words.

of the conflict topics in the Figure G.1d. Now tokens like "forc", "armi", "troop" or "combat" become darker. In general, many more tokens are displayed in darker shade now. This means that the text is much more likely to be generated by a journalist writing on conflict. Not surprisingly, the text has a higher share conflict (27 percent) .

This is a general feature of the topic model we estimate. Conflict topics are identified by a different set of tokens than non-conflict topics. This means that topics use very long lists of terms with particular weights but the lists differ significantly across topics. For an illustration we plot the use of terms in two different topics in Figure G.2. The figure shows the cumulative probability distribution of a conflict topic with respect to the 10,000 most likely phrases in the topic as a red line. The *conflict* topic needs more than 2,000 top terms to reach 80 percent of its probability mass and has not reached 90 percent at the 10,000th term. Other topics consist of different lists of terms. To make this clear, we plot the cumulative distribution of the *justice* topic (as a green dashed line) over the same terms. As the distribution of probabilities in this topic differs markedly the cumulative distribution of the *justice* topic does not even reach 40 percent of its probability mass in the 10,000 terms most likely in the *conflict* topic, i.e. the *justice* topic is composed by a different set of terms among the 800k terms we use.

Figure G.2: Cumulative Overlap Between Tokens in *Conflict* and *Justice* Topic



Notes: This figure displays the number of tokens (x-axis) ordered by probability (from high to low) within the *conflict* topic. The y-axis displays the cumulative density of the *conflict* topic (red solid line) and the *justice* topic (green dashed line) covered.

The rapid increase in both cumulative densities towards the left of the graph is due to common, generic tokens such as "american" in the example above. The top terms which appear in both the justice topic and the conflict topic are listed in the left column of Table G.1. The top shared terms are are words like *report*, *american* or *use*, none of which suggest any fighting or even preparation for armed struggle. For comparison we also display the joint words with the other conflict topic in the right column. The list contains words like *attack*, *kill* and *bomb* which are directly linked violence.

Table G.1: Joint tokens between topics

Joint tokens of *conflict2* and *justice*	Joint tokens of *conflict2* and *conflict1*
offici	attack
report	kill
offic	offici
american	report
unit	bomb
govern	secur
oper	offic
say	citi
accord	govern
use	today

Note: The table displays overlapping words amongst the 50 most prominent words in the *conflict2* and *justice* topic (left column) and the *conflict1* and *conflict2* topic (right column).

This feature of the topic model allows us to exclude the variation most related to conflict by excluding the conflict topics. Another way to see this is to look at the token lists that our non-conflict topics rely on. We report these in Table I.1. Nothing here indicates that these topics are used to describe fighting events.

G.2 Non-Conflict Topics and Conflict

Next we show how writing on non-conflict topics changes before and in conflict. To do so we run fixed-effects regressions of topic shares on the left hand-side. On the right-hand-side we put a dummy for conflict years and, importantly, a dummy that indicates a year before the onset of conflict. If non-conflict topics would capture violence indirectly then we would expect there to be a significant and positive relationship between their topic share and indicators of conflict.

In Panel A of Table G.2 we report the relationship between armed conflict and our topic shares. We find no positive significant relationship between the year before onset and topic shares. In addition, most coefficients are negative and one, on *justice*, is significant. In Panel B we report the relationship with civil war. Here the picture is even more pronounced now with three topics falling before the outbreak of civil war.

In addition, we find that conflict years, both armed conflict and civil war, are strongly and negatively related to the share written on non-conflict topics. The only exception is the topic *int.relations2* which increases significantly with both armed conflict and civil war. Note, however, that the share of this topic only increases significantly after conflict has started. This means it does not predict conflict. From Appendix Table I.1 we can see that the topic contains country names and tokens like "soviet" or "european" but no token that would indicate fighting. We can therefore potentially interpret the positive correlation with fighting as diplomatic actions taken in the face of internal violence.

These negative relationships stand in sharp contrast to the positive association between conflict topics and conflict as shown in Table G.3. All three conflict topics in the 2013 topic model increase before conflict starts and get an additional boost in conflict years. This further highlights how effective topics are in capturing the realities in the respective country. Given that conflict topics contain tokens which would be used to describe fighting and the positive relationship with the conflict year dummy, we cannot rule out that some fighting takes place before PRIO/UCDP code an armed conflict. However, this would not explain the findings of Section 7.

In summary, we have demonstrated that topics are an effective way to exclude tokens which might relate to fighting events. We have shown that this is because the large number of tokens used to define conflict topics are different from the ones used to define non-conflict topics. Finally, if anything, non-conflict topics fall before conflict onset so that a direct relationship between non-conflict topics and fighting events can be ruled out.

Table G.2: Relating non-conflict topics to conflict onset and conflict years

Panel A: Topic share reaction to armed conflict onset and armed conflict years Dependent variables: Non-conflict topic shares												
Topic	(industry)	(civ.life1)	(asia)	(sports)	(justice)	(tourism)	(politics)	(business)	(econ.)	(int.rel2)	(int.rel1)	(civ.life2)
One year before	-0.002	-0.005	0.002	-0.005	-0.010***	-0.003	0.003	-0.006	-0.005	0.001	-0.002	-0.003
	(0.002)	(0.003)	(0.002)	(0.004)	(0.003)	(0.003)	(0.003)	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)
Conflict year	-0.007***	-0.014***	-0.000	-0.010***	-0.007***	-0.012***	-0.007***	-0.017***	-0.015***	0.011***	-0.007***	-0.010***
	(0.001)	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Country FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.006	0.012	0.000	0.004	0.004	0.010	0.002	0.013	0.011	0.008	0.002	0.006
Observations	5367	5367	5367	5367	5367	5367	5367	5367	5367	5367	5367	5367

Panel B: Topic share reaction to civil war onset and civil war years Dependent variables: Non-conflict topic shares												
Topic	(industry)	(civ.life1)	(asia)	(sports)	(justice)	(tourism)	(politics)	(business)	(econ.)	(int.rel2)	(int.rel1)	(civ.life2)
One year before	-0.003	-0.007*	-0.003	-0.006	-0.009**	-0.005	-0.005	-0.010**	-0.009**	-0.001	-0.001	-0.006
	(0.003)	(0.004)	(0.002)	(0.005)	(0.004)	(0.004)	(0.004)	(0.005)	(0.004)	(0.004)	(0.004)	(0.004)
Conflict year	-0.007***	-0.020***	0.001	-0.010***	-0.013***	-0.014***	-0.011***	-0.013***	-0.012***	0.013***	-0.002	-0.014***
	(0.002)	(0.002)	(0.001)	(0.003)	(0.002)	(0.002)	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)
Country FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.004	0.016	0.001	0.003	0.007	0.009	0.004	0.005	0.005	0.008	0.000	0.008
Observations	5367	5367	5367	5367	5367	5367	5367	5367	5367	5367	5367	5367

Notes: The results stem form an OLS regression including a constant with a single topic share specified in the heading as dependent variable. *One year before* is a dummy taking the value one one year before conflict and zero otherwise. *Conflict year* is a dummy taking the value one during conflict and zero otherwise.

Table G.3: Relating conflict topics to conflict

Dependent variable: Conflict topic shares						
	Armed conflict			Civil war		
	(conflict1)	(conflict2)	(conflict3)	(conflict1)	(conflict2)	(conflict3)
One year before	0.016***	0.012***	0.006**	0.010**	0.035***	0.019***
	(0.003)	(0.004)	(0.003)	(0.004)	(0.006)	(0.004)
Conflict year	0.013***	0.053***	0.029***	0.006***	0.060***	0.035***
	(0.002)	(0.003)	(0.002)	(0.002)	(0.003)	(0.002)
Country FE	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.010	0.075	0.056	0.002	0.067	0.056
Observations	5367	5367	5367	5367	5367	5367

Notes: The results stem form an OLS regression including a constant with a single topic share specified in the heading as dependent variable. *One year before* is a dummy taking the value one one year before conflict and zero otherwise. *Conflict year* is a dummy taking the value one during conflict and zero otherwise.

H Case Study

In Table H.1 we report the 5 countries with the highest risk of conflict onset in the following year according to the within and overall model. We report this list for the years 2010, 2005 and 2000. Estimates for 2010, for example, are out-of-sample predictions for onset in 2011. We chose 2010 as it precedes particularly many onsets in 2011. We then chose 2005 and 2000 to provide equally spaced alternative years. As can be seen in Figure E.14 the AUCs in the selected years are slightly higher than average for civil wars but around the average for armed conflict. The table also reports onsets which were not among the first five countries with their respective rank in that year.

Table H.1: Case study

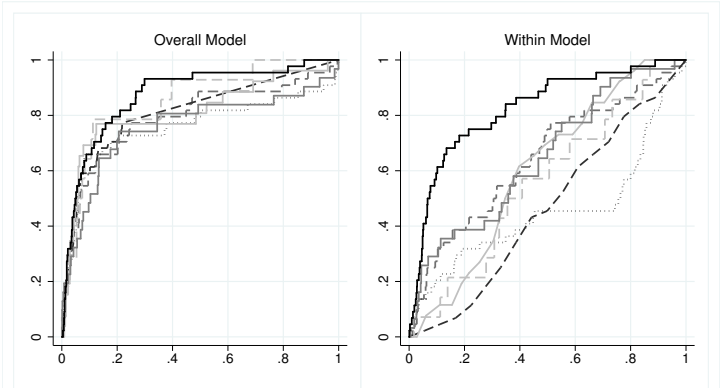
Panel A: Civil war			
	Overall model		Within model
Rank	Country	Rank	Country
Year 2000			
1	Tajikistan	1	Eritrea
2	Ethiopia	2	Lebanon
3	Uganda	3	Albania
4*	Rwanda	4	Sierra Leone
5	Sierra Leone	5	Ethiopia
		19*	Rwanda
Year 2005			
1*	Chad	1*	Sudan
2	Uganda	2	Congo
3	Angola	3	Cote d'Ivoire
4	Philippines	4	Burundi
5*	Sri Lanka	5*	Sri Lanka
6*	Sudan	6*	Israel
43*	Israel	58*	Chad
Year 2010			
1	Chad	1	Central African Republic
2	Sri Lanka	2	Mauritania
3	Uganda	3	Guinea-Bissau
4	Colombia	4	Uzbekistan
5	Philippines	5*	Yemen
11*	Yemen	32*	Libya
111*	Libya		
Panel B: Armed conflict			
	Overall model		Within model
Rank	Country	Rank	Country
Year 2000			
1	Cambodia	1	Eritrea
2*	Somalia	2	Togo
3	Niger	3	Albania
4	Yemen	4	Lebanon
5	Congo	5	Congo
109*	Central African Republic	19*	Somalia
		111*	Central African Republic
Year 2005			
1	Angola	1	Congo
2*	Somalia	2*	Central African Republic
3	Senegal	3	Cote d'Ivoire
4	Tajikistan	4	Togo
5	Congo	5	Lebanon
7*	Pakistan	10*	Somalia
18*	Central African Republic	24*	Pakistan
21*	DR Congo	52*	DR Congo
Year 2010			
1	Angola	1	Lebanon
2	Sri Lanka	2	Mali
3*	Senegal	3	Kyrgyzstan
4	Indonesia	4	Sri Lanka
5	Niger	5	Guinea-Bissau
15*	Nigeria	7*	Syria
30*	Cote d'Ivoire	12*	Cote d'Ivoire
51*	Syria	15*	Nigeria
108*	Libya	32*	Senegal
		38*	Libya

Notes: Asterisk (*) indicates an onset in the following year. Panel A contains civil war, whereas panel B armed conflict. The top five countries in terms of risk in the following year are displayed for predictions in $T = 2000, 2005, 2010$ according to the topic model. The left columns are based on the overall model, whereas the right columns are based on the within model. The table also displays countries with an onset that were not amongst the five highest risk predictions. For these countries the rank is listed as

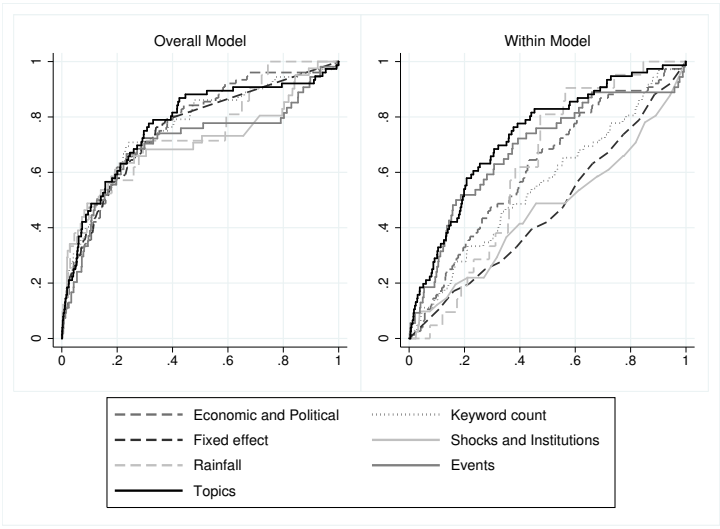
I Additional Figures and Tables

Figure I.1: ROC Curves for Onset (X-Axis: FPR, Y-Axis: TPR)

(a) Civil War



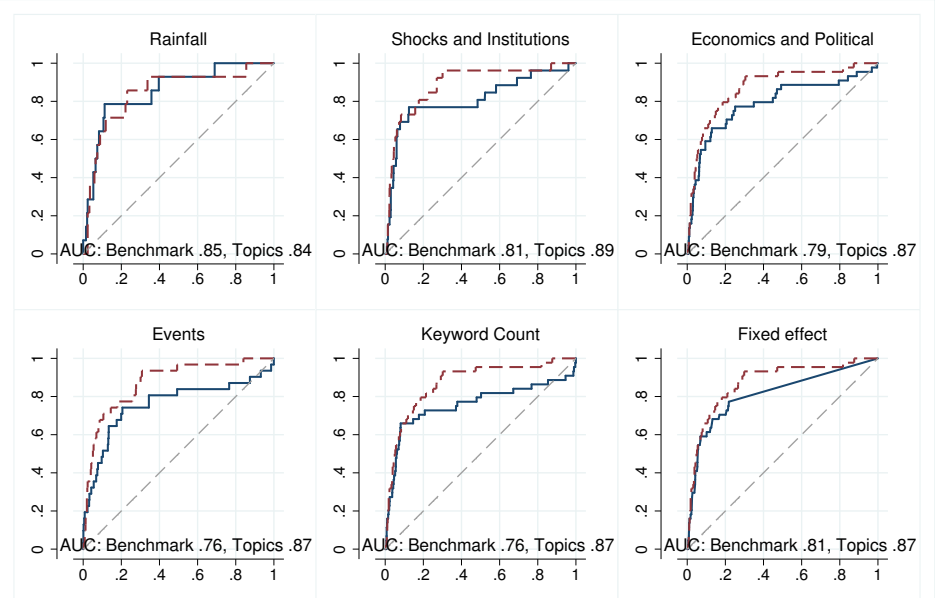
(b) Armed Conflict



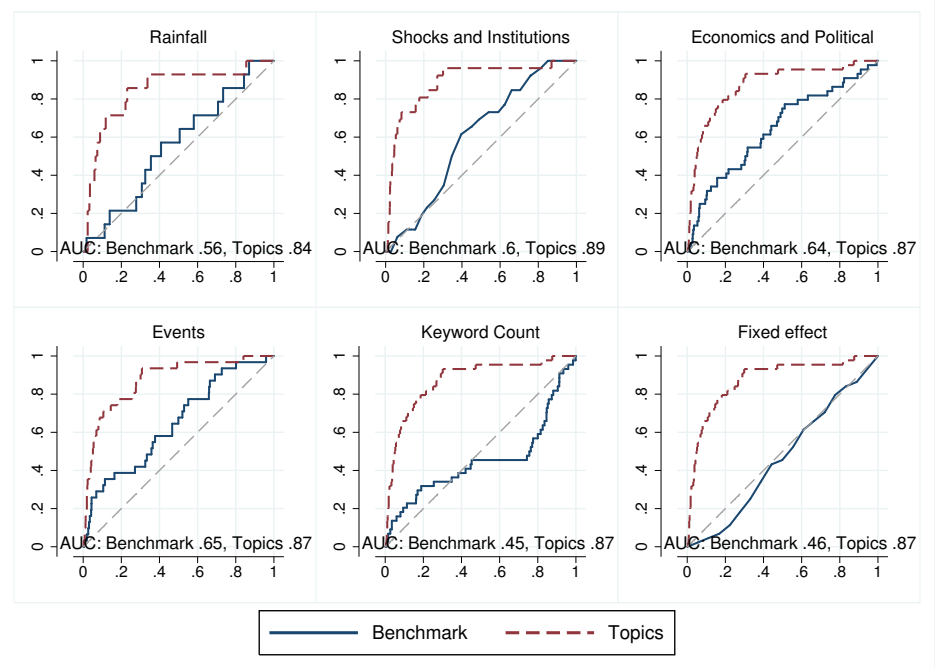
Notes: Predictions result from a panel estimated as in equation (2). The variables included for each model as $\mathbf{x}_{it}$ are specified in Section 3.3 and Appendix Table C.1. The within model is the overall model net of country fixed effects.

Figure I.2: ROC Curves for Onset of Civil War Comparing Only for Overlapping Predictions (X-Axis: FPR, Y-Axis: TPR)

(a) Overall



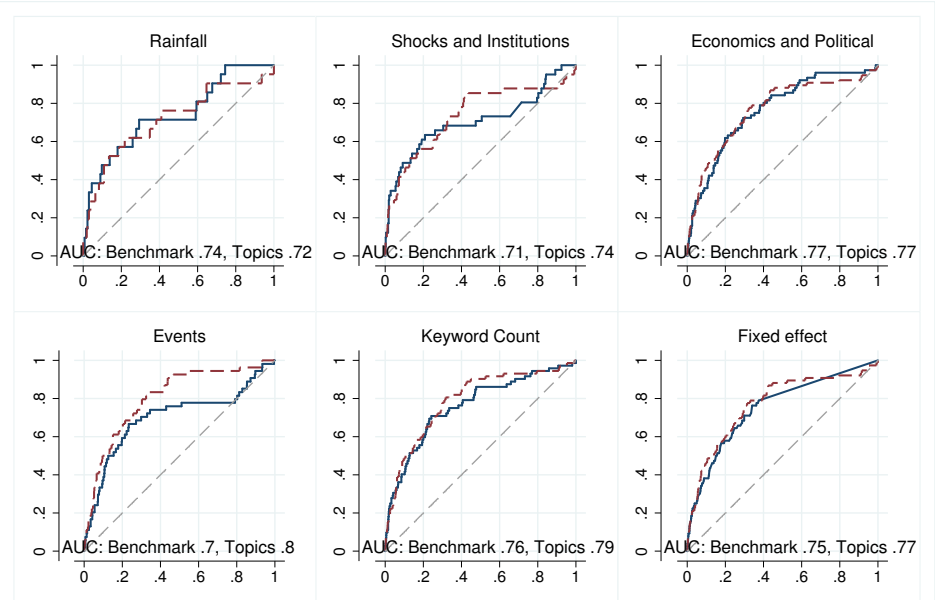
(b) Within



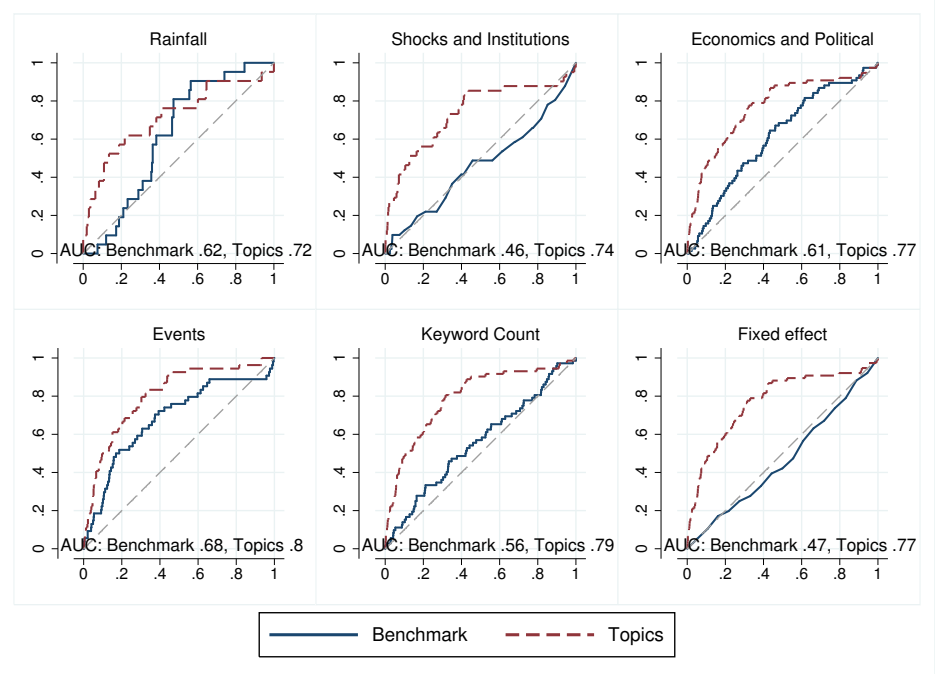
Notes: Predictions result from a panel estimated as in equation (2). The variables included for each model as $\mathbf{x}_{it}$ are specified in Section 3.3 and Appendix Table C.1. The within model is the overall model net of country fixed effects. Here we only include predictions for country-years where we have predictions for both the comparison and the topic model.

Figure I.3: ROC Curves for Onset of Armed Conflict Comparing Only for Overlapping Predictions (X-Axis: FPR, Y-Axis: TPR)

(a) Overall

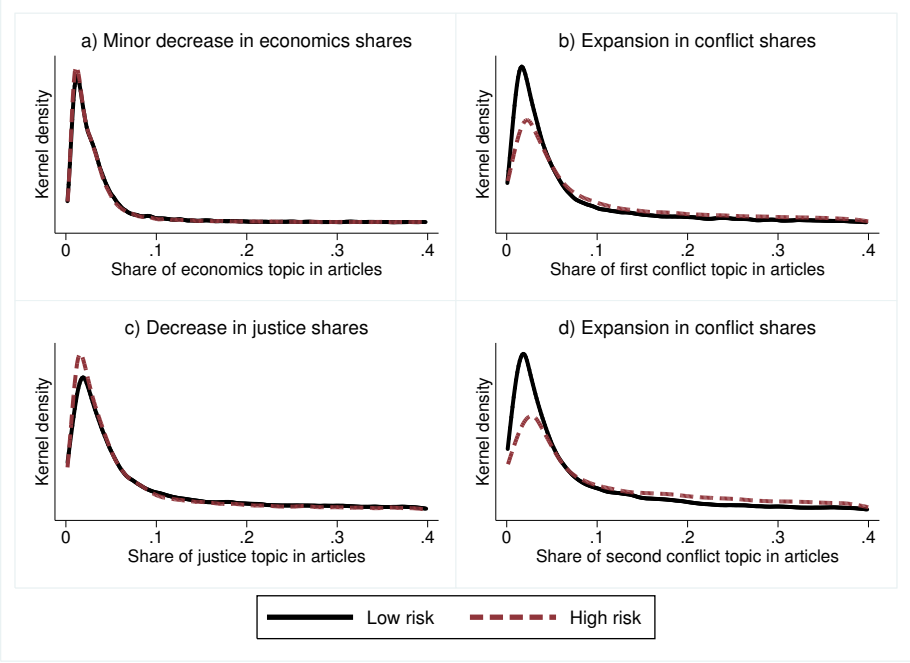


(b) Within



Notes: Predictions result from a panel estimated as in equation (2). The variables included for each model as $\mathbf{x}_{it}$ are specified in Section 3.3 and Appendix Table C.1. The within model is the overall model net of country fixed effects. Here we only include predictions for country-years where we have predictions for both the comparison and the topic model.

Figure I.4: Topic Shares of Economics, Justice, and Conflict in the Universe of Articles when Risk Is High vs Low



Notes: Shares represent average topic shares of all articles (not aggregated at the country-year level). High risk is defined as a predicted fitted value of onset above 0.05.

Table I.1: Topic content 2013

Topic title	15 most prominent words
Industry	compani, busi, market, firm, year, japanes, new, industri, share, american, sale, million, billion, manag, make
Civic life1	peopl, year, work, famili, say, child, woman, school, home, live, life, like, univers, old, citi
Asia	china, chines, south, korea, offici, year, taiwan, north, vietnam, hong, kong, hong_kong, beij, foreign, govern
Sports	team, game, play, second, year, point, time, world, won, player, win, score, run, final, minut
Justice	offici, report, court, state, case, charg, polic, investig, govern, law, offic, prison, arrest, releas, author
Tourism	citi, like, hotel, street, room, restaur, travel, place, good, food, hour, time, open, new, hous
Politics	parti, elect, polit, govern, vote, democrat, prime, new, power, leader, parliament, opposit, support, year, campaign
Conflict1	govern, peopl, countri, protest, polit, group, leader, islam, militari, polic, state, year, demonstr, support, muslim
Business	oil, year, countri, trade, world, state, import, export, produc, develop, price, govern, product, plant, new
Economics	bank, year, rate, govern, economi, billion, countri, market, econom, price, percent, tax, growth, fund, money
Inter. relations1	state, unit, unit_state, american, offici, administr, washington, nation, nuclear, meet, militari, bush, weapon, secur, talk
Inter. relations2	soviet, european, union, german, germani, europ, west, countri, east, britain, western, new, british, soviet_union, foreign
Conflict3	govern, nation, war, forc, rebel, unit, african, refuge, unit_nation, countri, south, peac, peopl, serb, guerrilla
Civic life2	work, new, like, book, world, time, film, art, life, cultur, year, music, american, centuri, war
Conflict2	forc, attack, militari, kill, offici, arab, armi, american, report, bomb, troop, soldier, war, command, air

Notes: Topics are listed in same order as in Table C.1. The order of the words within a topic reflects the prominence of the words within each topic. Topic titles are chosen by the authors ex-post based on most prominent words within a topic. Comparability between topics across time is established by counting the number of words that coincide, where words receive a weight inversely proportional to their frequency across all topics within a given year.

Table I.2: Area under Curve of ROC

	Area under Curve		
Model	Complete	Within	Δ
Civil War			
Economic and Political	0.79	0.64	-19%
Keyword Count	0.76	0.46	-39%
Fixed Effect	0.81	0.46	-43%
Shocks and Institutions	0.81	0.60	-26%
Rainfall	0.85	0.56	-34%
Events	0.76	0.65	-14%
		Average	-29%
Topics	0.87	0.82	-6%
Armed Conflict			
Economic and Political	0.77	0.61	-21%
Keyword Count	0.76	0.56	-26%
Fixed Effect	0.75	0.47	-37%
Shocks and Institutions	0.71	0.46	-35%
Rainfall	0.74	0.62	-16%
Events	0.70	0.68	-3%
		Average	-23%
Topics	0.77	0.73	-5%

Notes: The area under the curve (AUC) is computed using the curves exhibited in Figures 1 and 4. Predictions result from a panel estimated as in equation (2). The variables included for each model as $\mathbf{x}_{it}$ are specified in Section 3.3 and Appendix Table C.1. The within model is the overall model net of country fixed effects.