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

Hannes Mueller Christopher Rauh

September 29, 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): Tenured Scientist at IAE (CSIC), Barcelona GSE (address: Institut d'Analisi Economica, CSIC Campus UAB, 08193 Bellaterra, Spain; email: h.mueller.uni@gmail.com); Rauh: Assistant Professor at University of Montreal (address: Département de sciences économiques, Université de Montréal, C.P.6128 succ. Centre-Ville, Montréal H3C 3J7, Canada; 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.

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 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

[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.

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, then we present a way to evaluate the ability of a model to forecast the timing of conflict before we present our methodology of aggregating news text into topics followed byt the main results. Next we demonstrate the close link between predicting the timing of

conflict and predicting new, otherwise unforeseen conflicts. Finally, we explore why the topic model provides such useful forecasts of the timing before we conclude.

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 (e.g., Fearon and Laitin 2003; Collier and Hoeffler 2004; Collier et al. 2009; Gleditsch and Ruggeri 2010; 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

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

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 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

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

[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.

of the Bank of England. We contribute to this literature by applying the topic model to newspaper text which spans large cross-country panels.

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.

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 section

[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.

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.

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.[7] 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.

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

[7] 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 $T-1$.

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 (Beck 2015). 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.[8]

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 Online Appendix C.[9]

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 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).[10] 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.[11]

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

[8] 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.

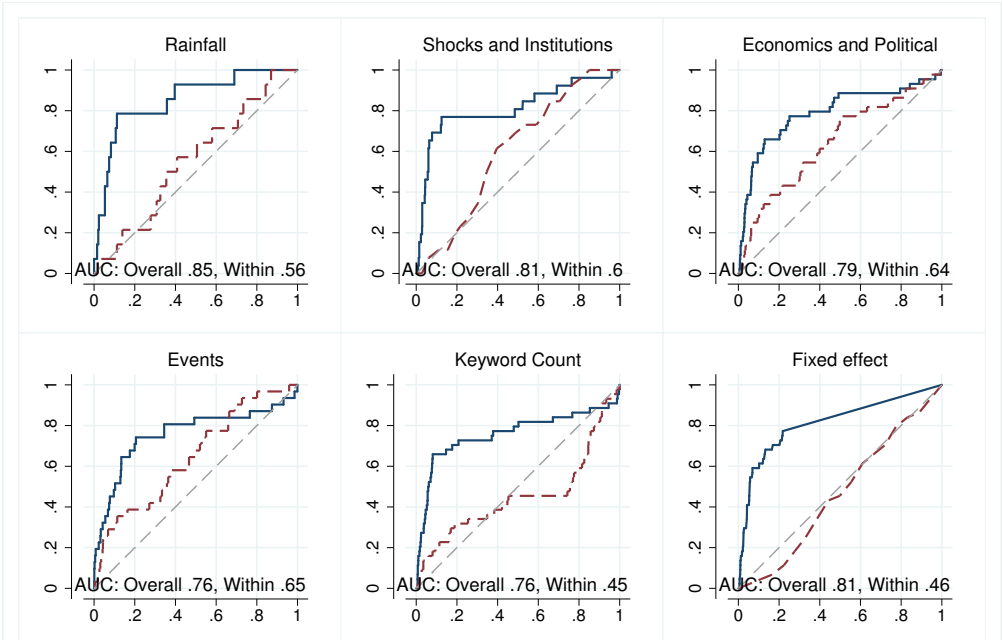
[9] The five models we recreate are published in Miguel and Satyanath (2011), Besley and Persson (2011*b*), Goldstone et al. (2010), Ward et al. (2013) and Chadefaux (2014). Wherever available, we use the respective replication datasets.

[10] 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.

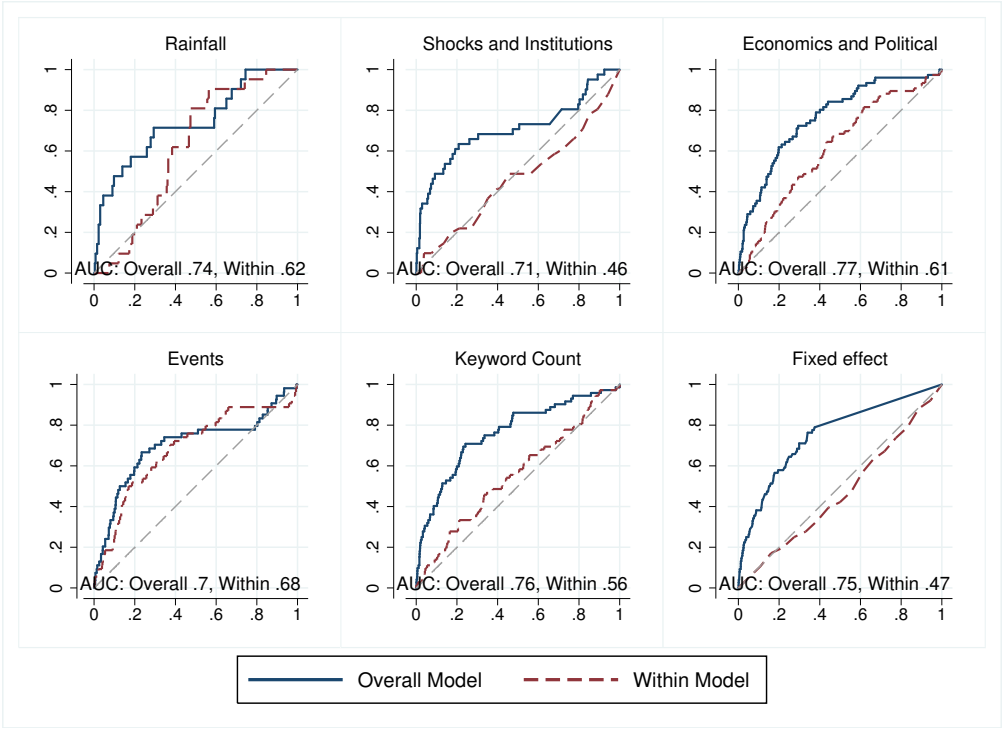
[11] All recent casualties in Afghanistan are, for example, coded as stemming from an internationalized internal conflict. We ran extensive robustness checks regarding our definition. For a detailed discussion see the Online Appendix.

Figure 1: ROC Curves for Onset (X-Axis: FPR, Y-Axis: TPR) [bw print/col online]

(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 Appendix Table C.1. The within model is the overall model net of country fixed effects as presented in equation (3).

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, 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. However, in many applications this is not satisfactory and leads to forecasting mistakes if peaceful countries destabilize or violent countries stabilize.

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.

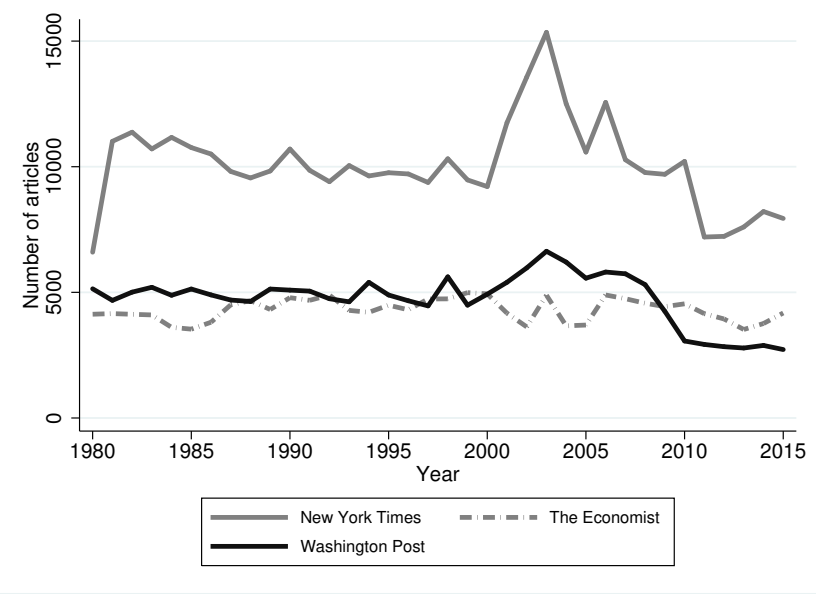
News Text

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.[12] 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. 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 [bw print/col online]



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.[13] 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).[14] 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

[12] 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.

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

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

"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.

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 , is composed of 15 topics and hyperparameters $\alpha = 3.33$ and $\beta = 0.01$.[15]

[15] 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.

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.[16]

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.

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.[17] In column (1) we show the

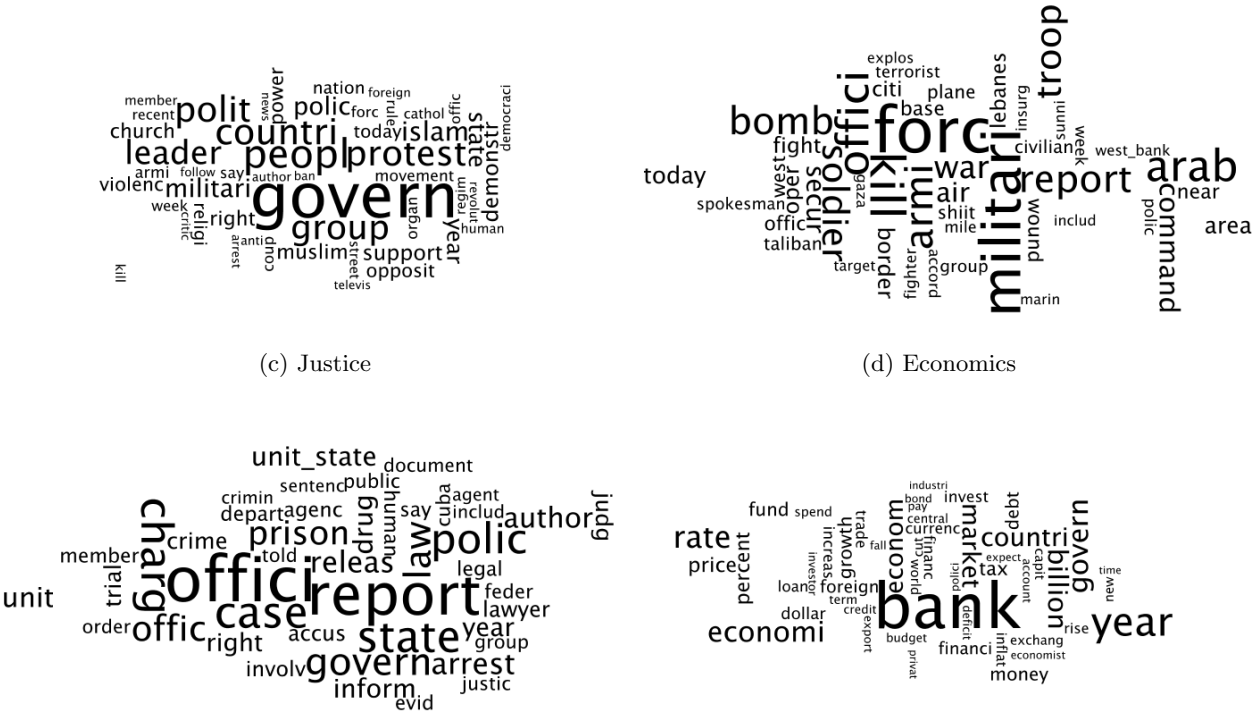
[16]The restriction is actually the conflict data. We have text written until 2015 and used it to predict 2016/2017. Published working paper versions contain these forecasts.

[17]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

Figure 3: Word Clouds of Topics [bw print and online]

(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.

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 context and surge in asymmetric armed conflicts.[18] 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.

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

similar probability distributions that appear both in 1995 and 2015.

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

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.

then

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

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.[19] 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

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

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).

PREDICTING CONFLICT WITH NEWSPAPER TOPICS

In this section, we combine the forecast evaluation from Section with the topic model from section . 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

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

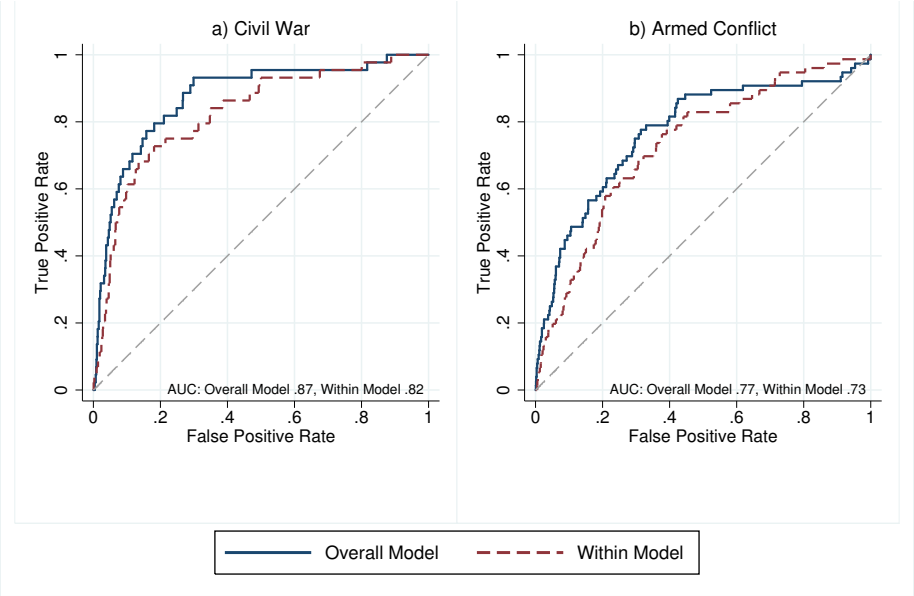
Using a set of varying cutoffs c and our estimates for $\hat{y}_{iT+1}^{within}$ and $\hat{y}_{iT+1}^{overall}$, we then calculate the TPR_c and FPR_c for each cutoff c, which we present in standard ROC curves as explained in Section .

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 performs worse. Again, the within variation seems to be an

Figure 4: ROC Curves for Onset (Topics Model) [bw print/col online]



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).

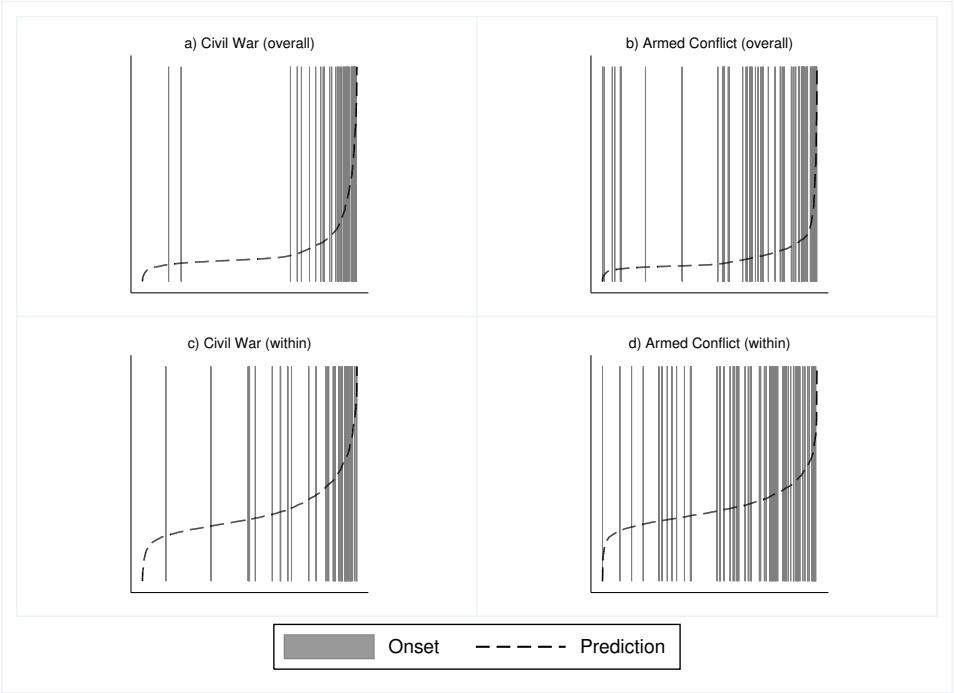
important driver of the ability to forecast conflict. The AUC only drops from 0.87 and 0.77 in the overall model and to 0.82 and 0.73 in the within model for civil war and armed conflict, respectively. This is an important difference to the other models evaluated previously, which suffer a drop of around 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.

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

Figure 5: Separation Plots for Onset (Topics Model) [bw print and online]



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 predictions are ordered from low to high likelihood.

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.

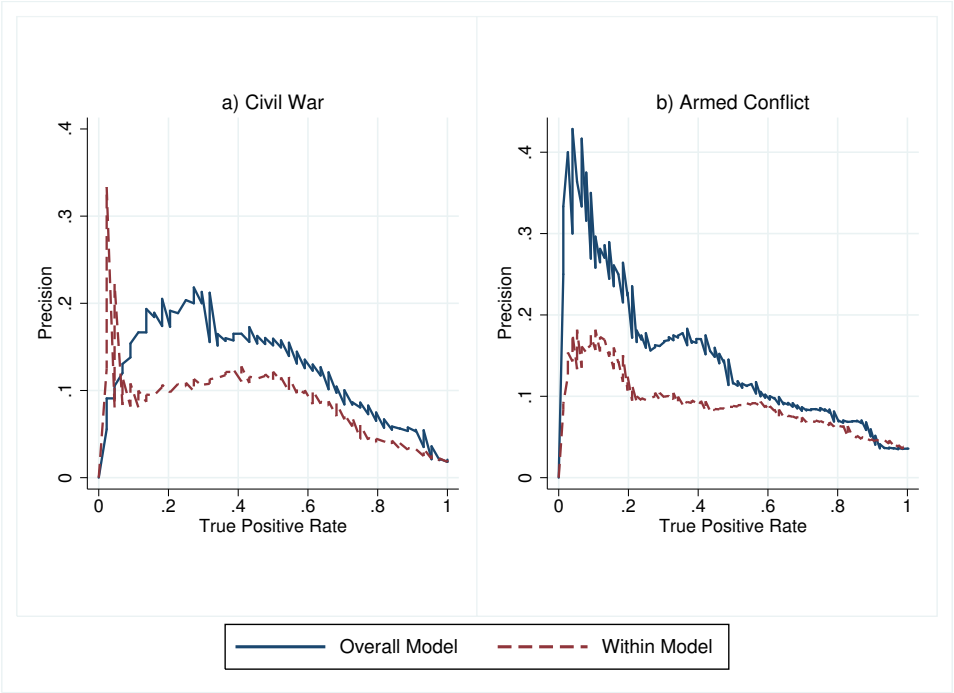
We run several robustness checks regarding these basic results. Details and corresponding Figures are all reported in Appendix E. First, in Figure E.1 we show that prediction results when using topics for conflict incidence, unsurprisingly, perform even better than for the prediction of onset.

Second, we use more and less topics.[20] 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

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

Figure 6: Precision Recall Curves (Topics Model) [bw print/col online]



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).

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 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 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.

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.[21]

[21] 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.

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 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.

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.[22] 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

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

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 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.

READING BETWEEN THE LINES

We now explore why topics provide such useful forecasting power on the time dimension.[23] 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.[24] 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.[25]

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.

[23] 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.

[24] 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).

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

Table 3: Lasso Model

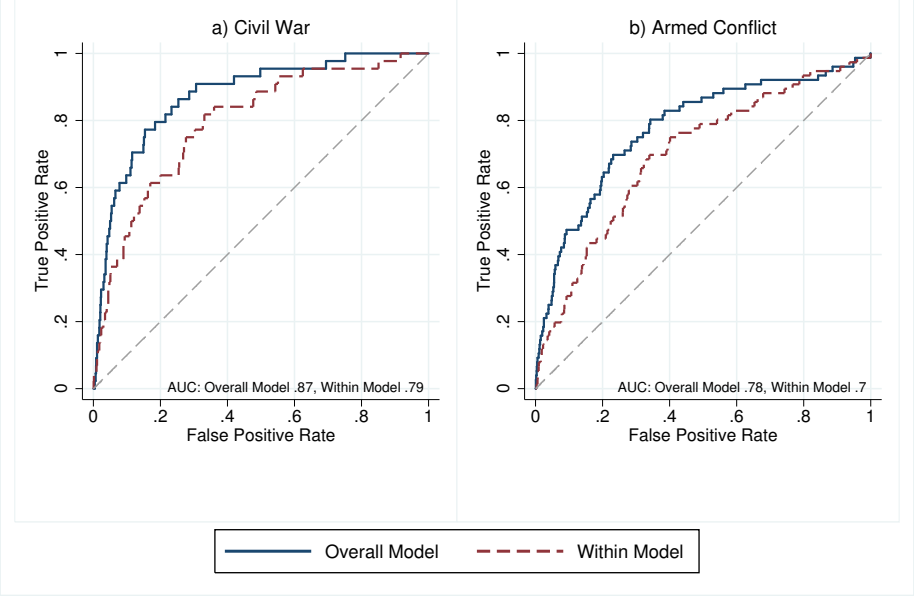
Selectivity level	mild	regular	very	mild	regular	very
	civil war onset next year			armed conflict onset next year		
	(1)	(2)	(3)	(4)	(5)	(6)
Topic shares						
conflict1	0.0366	0.0564		0.306**	0.259**	0.275***
	(0.0685)	(0.0599)		(0.121)	(0.103)	(0.0999)
conflict2	0.256**	0.300***	0.281***	0.304**		
	(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.

The first message from Table 3 is that conflict topics are chosen by the LASSO despite the fact 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.[26] 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.[27]

Figure 7: ROC Curves for Onset (Only Non-Conflict Topics) [bw print/col online]

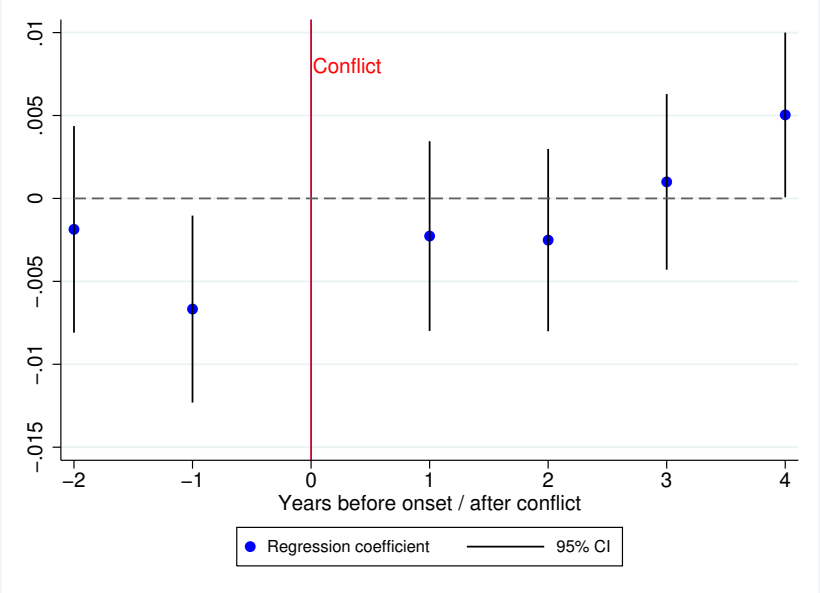


Notes: Predictions result from a panel estimated as in equation (2). The topic model contains the 12 non-conflict topics as θ_{it} derived using LDA with $\alpha = 3.33$ and $\beta = 0.01$ and are aggregated at the country-year level. We excluded the 3 conflict topics and re-normalized so that the 12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

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

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.[28] 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.[29]

Figure 8: The *Justice* Topic Before and After the Outbreak of Conflict [bw print/col online]



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.

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 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.

[28]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.

[29]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.

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

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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

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