UNIVERSITY OF CAMBRIDGE
Faculty of Economics

Institute for New Economic Thinking

Cambridge-INET Institute

Cambridge-INET Working Paper Series No: 2016/05

Cambridge Working Paper Economics: 1630

READING BETWEEN THE LINES: PREDICTION OF POLITICAL VIOLENCE USING NEWSPAPER TEXT

Hannes Mueller (Barcelona GSE)
Christopher Rauh (University of Cambridge)

This article provides a new methodology to predict conflict by using newspaper text. Through machine learning, vast quantities of newspaper text are reduced to interpretable topic shares. We use changes in topic shares to predict conflict one and two years before it occurs. In our predictions we distinguish between predicting the likelihood of conflict across countries and the timing of conflict within each country. Most factors identified by the literature, though performing well at predicting the location of conflict, add little to the prediction of timing. We show that news topics indeed can predict the timing of conflict onset. We also use the estimated topic shares to document how reporting changes before conflict breaks out.

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

Hannes Mueller   Christopher Rauh

May 3, 2016

Abstract

This article provides a new methodology to predict conflict by using newspaper text. Through machine learning, vast quantities of newspaper text are reduced to interpretable topic shares. We use changes in topic shares to predict conflict one and two years before it occurs. In our predictions we distinguish between predicting the likelihood of conflict across countries and the timing of conflict within each country. Most factors identified by the literature, though performing well at predicting the location of conflict, add little to the prediction of timing. We show that news topics indeed can predict the timing of conflict onset. We also use the estimated topic shares to document how reporting changes before conflict breaks out.

Keywords: Conflict, Forecasting, Machine Learning, Panel Data, Topic Models, Latent Dirichlet Allocation.

[*] Mueller: IAE (CSIC), Barcelona GSE (email: h.mueller.uni@gmail.com); Rauh: University of Cambridge, INET Institute (email: cr542@cam.ac.uk). We thank Hector Galindo, Stephen Hansen, Daniel Ohayon, Anand Shrivastava, Jack Willis, and the participants of the ENCoRe Barcelona meeting, the internal Political Economy Workshop Cambridge, and the EPCS Freiburg 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 from 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. It is easier to predict whether a country is at risk *in general* rather than *when* a country is particularly at risk.

An additional problem in forecasting the timing of armed conflict is that it is rare and at the same time relatively concentrated in some countries. Imagine that at the end of the second world war one had attributed a prediction of conflict to all years in some countries and the prediction of no conflict to all years of the remaining countries. This forecast would have been able to reach a striking level of accuracy of over 90 percent. However, such a model would be completely time invariant and, hence, tell us nothing about the timing of conflict.

This is problematic because it implies that the variation between countries can dominate the analysis unless the between and within variation are separated explicitly. Empirical models that are overall quite accurate can therefore be nonetheless 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.[2] Yet, policymakers and academics alike might be interested in a measure of conflict risk that is meaningful on the time dimension and performs well out-of-sample.

As a solution to this problem we propose data generated from news sources.[3] News content has strong country-specific elements and is available in real time - both of these elements give it a comparative advantage in predicting the time-dimension of conflict. At the same time, journalists are not merely neutral observers. They curate facts and connect them to build an overall impression of a situation. In this regard their work is similar to the work of country experts working for governments. Summarizing this expert opinion in a quantifiable measure could therefore add forecasting power which goes beyond the available data and even other methods of summarizing text.

To this end, we propose a new, fully 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. We show that it is possible to summarize news text in a

[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.

[2] Ward, Greenhill and Bakke (2010) demonstrate that focusing on statistically significant relationships does not necessarily contribute to the prediction of conflict. However, they do not emphasize that many of these significant relationships suffer from a lack of time variation.

[3] We are not the first to propose news sources to forecast conflict. See, for example, Chadefaux (2014) and Ward et al. (2013), which we discuss in detail below.

meaningful way in a topic model. More surprisingly, perhaps, the topics which summarize the text are able to forecast conflict across countries and decades. Moreover, we demonstrate that writing on specific topics consistently increases, while other topics disappear before conflict. The result is a model which can even forecast the onset of armed conflict, an event of much lower intensity, a year before it occurs. 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 from LexisNexis and collect words and series of words, referred to as tokens, in one vector for each article. Newspaper articles offer several advantages for the analysis of conflict. 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. Also, newspapers provide a high degree of consistency regarding the density of news per year that can be reported. This makes token counts at least somewhat comparable across years and even decades. We downloaded all articles on 185 countries from the New York Times, the Washington Post, and the Economist for all available years since 1980. 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 prior on which part of the text is important when predicting conflict - we can let the data speak.

As the final step, we use 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 to predict conflict in the following year. We show that the timing of conflict can be predicted out-of-sample with changes in topic shares. The work conducted by journalists can be used in this way to provide a new quantitative measure of country risks.

We use our estimated models to simulate the problem of a policy maker who attributes costs to false positives and false negatives when forecasting conflict. This view reveals that a model, which is geared towards predicting the timing of conflict, produces a very different perspective on the forecasting problem. The main advantage of using time-varying measures of risk is that one has a higher chance of anticipating conflicts which break out in formerly fairly peaceful countries, like Egypt or Libya. Another, perhaps less obvious, advantage of using time-varying risk measures is that previously violent countries which stabilize do not generate false positives.

We argue that there are two dimensions to the ability of topic models to predict the onset of armed conflict. First, since topics are a collection of words that co-occur, the model is valid across countries and time. Insurgencies, for example, will trigger certain keywords that are shared across all countries and time even if the specific context differs. Second, the model uses positive and negative associations with conflict to predict onset. We find, for example, that news which describe judicial procedures systematically decrease when conflict risk increases. We also use our estimates of the content of articles to show that our predictions are driven by large changes in the content of some articles rather than subtle changes in many articles.

We proceed as follows. We first discuss related literature in Section 2. In Section 3 we argue that standard linear fixed effects regressions can be used to distinguish between location and timing in forecasting. In Section 4 we present our methodology of aggregating news text into topics and the forecasting method. Section 5 presents the main results, while in Section 6 we discuss policy implications. In Section 7 we investigate the content of topics and its relation to conflict prediction. 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 like 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) or natural resources (Brückner and Ciccone 2010, Bazzi and Blattman 2014). This literature is more concerned with identification and less with forecasting power. Another part of the literature has looked at using a mix of political and economics variables 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).

Most recently, attention has shifted towards forecasting armed conflict with variables used by this literature. Rost, Schneider and Kleibl (2009) 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 like political institutions or infant mortality. Hegre et al. (2013) forecast conflict for the period 2010-2050 using a combination of variables like population, infant mortality and education.[4] Ward et al. (2013) use a combination of event data and more standard variables to make monthly

[4] The intriguing point in such a long-term forecast is that the timing of conflict is less important so that the standard framework seems quite adequate.

predictions up to six months ahead. Their model has a striking degree of accuracy in predicting civil war onset and performs well out-of-sample. They also discuss the purpose of forecasting and argue for out-of-sample prediction as a possible gold standard of model development in the field of conflict studies. Chadefaux (2014) 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.

We add to this forecasting literature in three ways. First, we provide a novel way to summarize news in few dimensions, which we see as complementary to the existing event data and word counts. The summaries we generate do not use any prior assumption on the words that could predict conflict, i.e. our predictions use almost the entire newspaper text written on each country. Secondly, an important conceptual contribution of this project is that we build a forecasting models which allows us to focus on within-country variation. In other words, we explicitly focus on predicting the timing of violence out-of-sample. This is an important distinction to most existing studies.[5] Thirdly, we attempt to predict conflict with news data up to two years before conflict occurs. The longer time period compared to other studies which use news provides a longer period to react to early warning.

Another part of the literature has tried to predict conflict locations within ongoing conflicts. Here the problem of predicting timing is alleviated as it is already clear that the baseline-conflict risk is fairly high. Research can therefore focus on distinguishing the determinants of risk in the cross-section. Blair, Blattman and Hartman (2014) use 56 risk factors to predict locations of conflict within Liberia. They find that especially ethnic diversity and polarization, two slow-moving variables, consistently predict the location of violence over time. Similarly, Schutte (2014) predicts location of conflict using structural factors like population or the distance to the capital. An interesting exception are studies which predict the timing of local violence using violence in neighbouring geographic units. Weidmann and Ward (2010), for example, show that a reliable predictor of violence in a given period is violence in near regions in the previous period.

The topic model has been used by Quinn et al. (2010) to categorize over 100,000 speeches in the US congress. They estimate a topic model of 42 topics to show that topics can be used to analyze democratic agenda dynamics over a long time period. Topics have also been used by Hansen, McMahon and Prat (2014) to quantify discussions in the central bank committee of the Bank of England. The approach has the big advantage of requiring no human input except for the choice of two distribution parameters. We contribute to this literature by applying the topic model to newspaper text and by using the estimated

[5] Ward et al. (2013) use country as a group variable in their logit framework. This is similar to country fixed effects but it means that countries, which are always peaceful or always in conflict in the sample period, get dropped from the sample. This is particularly important as sudden outbreaks of civil war in a previously peaceful country cannot be predicted.

topic shares in cross-country panels. We show that topics form around useful categories such as “conflict”, “economics”, “sports”, “tourism” and “justice” and, perhaps most surprisingly, that re-estimation of the topic model across years preserves these categories to a large extent. However, within our method, the approach we use is the simplest possible. It would, for example, be possible to use a structural topic model as suggested by Roberts et al. (2013), in which covariates are embedded in the topic generation. The appeal of the method we choose here is that it is fully automatic except for the choice of number of topics and two additional parameters. We are not imposing a prior regarding the structure of the text.

News sources have previously been used to generate data on expectations and perceptions. Kuziemko and Werker (2006) use the frequency of the United Nations and the Security Council being mentioned in the New York Times as a proxy for its political importance. Ramey (2011) shows that increases in military spending can be predicted through news reports several quarters before they occur. Brückner and Pappa (2015) show that news on the Olympic Games, not the Games themselves, drive investment in countries that host them. Baker, Bloom and Davis (2015) show that word combinations such as “economic uncertainty congress” can be used to measure political uncertainty. The approach chosen in these studies is possible because there is a clear prior regarding which news reports should capture perceptions. An exception in this literature is Gentzkow and Shapiro (2010), who develop a measure of political bias of newspaper outlets in the US. To do this, they generate a list of expressions that indicate Republican or Democratic slant. They generate this list by looking at what expressions distinguish political speeches by Republicans and Democrats. Our methodology follows this basic idea but instead tries to understand how news text changes in the years before political instability compared to other years.

An important issue of concern, is that news data has been criticised as a source of data. For example, Woolley (2000) discusses several issues related to using newspaper data. He criticises the use of counts in environments with widely varying news volumes and recommends deflating counts. Weidmann (2016) shows forcefully that media coverage can lead to a bias in reporting. He demonstrates that this is a problem if conflict data is used to study the effect of media access. Indeed, as in most other studies of conflict, our left-hand-side variable is based on an event count which is partly informed by news agencies. We address this concern by looking at topic shares on the right-hand-side, which means we deflate the right-hand-side news data. Our predictors rely on content rather than quantity of reporting. In addition, we add country fixed effects, which controls for reporting biases at the country level. We show that changes in content can predict changes in reported counts 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 UNHCR.

3 Forecasting the Timing of Conflict

In this section, we show that the main difficulty of forecasting conflict lies in forecasting the timing. We show this in two steps. First, we argue that separating the variation between countries from the variation within a country is essential to understanding the timing of conflict. We then use a linear fixed effects model to study three empirical models of conflict suggested in the literature. This reveals huge differences in their ability to forecast the between and within variation in conflict onset. By far most explanatory power in most models comes from the fixed effects. Predicting the timing of the onset of armed conflict is particularly difficult.

3.1 The Time Dimension of Conflict

Our aim is to train our model to forecast by comparing observations in country i and year t that were followed by conflict within one or two years (treatment) to observations which did not experience conflict later (control). One way to do this is by regressing a dummy y_{it} that indicates one year before conflict on a set of country characteristics $\vec{x}_{it}$.[6] The most standard way to do this is the logit model, which has the formal representation

$$\Pr(y_{it} = 1) = F(\alpha + \vec{x}_{it}\vec{\beta}) \tag{1}$$

where $\Pr(y_{it} = 1)$ is the probability of observing conflict within the next year and F is the cumulative logistic distribution. Less common is the linear probability model (LPM) which takes the form

$$y_{it} = \alpha + \vec{x}_{it}\vec{\beta} + \varepsilon_{it}. \tag{2}$$

There are several problems associated with this model and a broad academic debate has brought pro- and counter-arguments for its adoption.[7] We nonetheless choose the linear model for two reasons. First, theoretically it is much easier to discuss the difference between within-country variation and between-country variation. Empirically, due to the fact that we are forecasting, we are not interested in the coefficients $\vec{\beta}$ and do not mind that the fitted values, $\hat{y}_{it}$, are not bounded between 0 and 1. Second, we are particularly interested in using fixed effects and produce forecasts for countries that within-sample never experienced an armed conflict or civil war.

It is easy to show that estimating the model in equation (2) encompasses learning about pre-cursors to conflict in two ways: from the differences between countries and from what happens across time within a country. Formally, we can write the estimated coefficients $\vec{\beta}$ in terms of the two sums of squares which capture the two sources of variation

$$\vec{\beta} = \left[\vec{S}_{xx}^{total}\right]^{-1}\left(\vec{S}_{xy}^{within} + \vec{S}_{xy}^{between}\right) \tag{3}$$

[6] Throughout we will indicate vectors through arrows. The only exception are means of vectors

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

where $\vec{S}_{xx}^{total}$ is the total sum of squares of the model $\vec{x}_{it}$.[8] It will be useful to inspect the nominator of equation (3) separately. The first term in the nominator, the within sum of squares, describes deviation of $\vec{x}_{it}$ and y_{it} from the country means $\bar{x}_i$ and $\bar{y}_i$ through

$$\vec{S}_{xy}^{within} = \sum_i \sum_t \left(\vec{x}_{it} - \bar{x}_i\right)\left(y_{it} - \bar{y}_i\right), \tag{4}$$

which captures the association of $\vec{x}_{it}$ and conflict y_{it} across time within a country, i.e. the *within variation.*

The second term in the nominator describes the deviation of the country means, $\bar{x}_i$ and $\bar{y}_i$, from the overall means $\bar{x}$ and $\bar{y}$ through

$$\vec{S}_{xy}^{between} = \sum_i T\left(\bar{x}_i - \bar{x}\right)\left(\bar{y}_i - \bar{y}\right) \tag{5}$$

which captures the association of the average values of $\bar{x}_i$ and average conflict $\bar{y}_i$ between countries, i.e. the *between variation.* Variation in $\vec{x}_{it}$ affects both the within and between sum of squares in equations (4) and (5). A change in a given year first immediately affects $\vec{x}_{it}$ in (4). If $\vec{x}_{it}$ changes permanently in a country this implies a change in $\bar{x}_i$.

Equations (4) and (5) illustrate why the vector $\vec{x}_{it}$ can become a predictor of conflict. The first option is that $\vec{x}_{it}$ deviates from its mean $\bar{x}_i$ exactly when conflict becomes imminent, i.e. when $y_{it} = 1$. An economic crisis might, for example, trigger conflict which would imply that a deviation of $\vec{x}_{it}$ from $\bar{x}_i$ coincides with a change of y_{it} from 0 to 1. The second option is that countries whose average characteristics $\bar{x}_i$ deviate from the overall average $\bar{x}$ are more (or less) likely to enter conflict than the average country. Political institutions, which differ vastly between countries but change rarely, have often been found to predict conflict in the cross-section.

Learning from $\vec{S}_{xy}^{between}$ exclusively means that the timing of instability cannot be predicted. The between variation allows us to rank countries according to their likelihood of destabilization but these rankings will not change from one year to the next. This also means that countries which have been free of violence in the past are likely to be "off the radar". It is therefore questionable whether a model that relies on large values of $\vec{S}_{xy}^{between}$ alone can be used effectively to forecast conflict in the long run.

In many applications a policymaker might be interested in whether a particular country is more likely to enter a crisis than the year before. This is especially important if sudden developments change established dynamics in a country. Learning about the timing of instability is only possible from the within variation, $\vec{S}_{xy}^{within}$, in equation (4). From the equation we can see that countries which remain in peace or only experience episodes with $y_{it} = 1$ throughout the sample cannot contribute to learning about timing as $y_{it} = \bar{y}_i$ in all years. This might be a subtle point but it turns out to be of crucial importance in the actual application because it prevents the use of the fixed effects logit framework to predict conflict in previously peaceful countries.

[8] For a more detailed discussion see Greene (2003).

3.2 Out-of-Sample Forecasting

The previous discussion made clear that regressions like in equation (2) mix the between and within variation to estimate the parameters $\vec{\beta}$. A straightforward way to separate the within from the between variation is the fixed effects model

$$y_{it} = \beta_i + \vec{x}_{it}\vec{\beta}^{FE} + \varepsilon_{it}, \tag{6}$$

where β_i is a set of country fixed effects. The estimate of $\vec{\beta}^{FE}$ relies entirely on the within variation in equation (4).[9] This prevents us from making any conjectures regarding the transition to conflict from countries which are always in conflict or always in peace. Instead, the learning process about what constitutes a risk to stability is gathered exclusively from countries that have destabilized or stabilized. In what follows, we adopt this linear model because it allows us to produce forecasts for countries which have only $y_{it} = 1$ or $y_{it} = 0$. If we want to estimate the probability of future conflict in a country that has been stable, this is an essential property.

To illustrate our forecasting method, we forecast the timing of conflict out-of-sample using three different conflict models, $\vec{x}_{it}$, suggested by the literature. For each model we proceed in five steps:

1) We use all the data from all countries until year T as our estimation sample. We estimate a fixed-effects regression as in equation (6) where the vector $\vec{x}_{it}$ is given by the respective empirical models described above and where y_{it} is a dummy that takes a value of one if conflict occurs one year later. The exhibited results are always for the cases of the onset of conflict but hold for incidence as well. We define as onset, the outbreak of a conflict following at least one peaceful year, so a 1 preceded by at least one 0, whereas incidence is any conflict year. When onset is our dependent variable, the following conflict years are coded as missing.

2) We then use the respective estimates for $\hat{\beta}_i$, $\hat{\beta}^{FE}$ to produce the fitted values

$$\begin{aligned} \hat{y}_{it}^{overall} &= \hat{\beta}_i + \vec{x}_{it}\hat{\beta}^{FE} \\ \hat{y}_{it}^{within} &= \vec{x}_{it}\hat{\beta}^{FE} \end{aligned}$$

for the same set of countries but year $t = T+1$. The within fitted values, $\hat{y}_{it}^{within}$, capture the risk of conflict compared to the country's average propensity, i.e. without taking into account whether the country was low or high risk in the past. If we use the estimated $\hat{y}_{it}^{overall}$, we use both the within and between variation contained in the model.

3) We predict the values of y_{it} in $T + 1$. In particular, we forecast conflict if the fitted value in $T + 1$ is higher than a cutoff c.[10] Conversely, if the fitted value is below the cutoff threshold c, we assume no conflict is going to take

[9] To see this note that $\vec{\beta}^{FE} = \left[\vec{S}_{xx}^{within}\right]^{-1}\left[\vec{S}_{xy}^{within}\right]$.

[10] In other words, we use a model estimated with data until T, applied to data available in $T + 1$ to predict onset in year $T + 2$.

place a year later. By comparing our estimates for $\hat{y}_{it}^{within}$ and $\hat{y}_{it}^{overall}$ to a set of varying cutoffs c, we generate the total number of true positives (TP_c), the number of true negatives (TN_c), the number of false positives (FP_c) and false negatives (FN_c).

4) We calculate the true positive rate (TPR)

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

and the false positive rate (FPR)

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

for each of the cutoffs c. The TPR is the share of all actual conflicts which is correctly identified. A TPR of 0.4 implies that 40 percent of all conflicts are correctly anticipated. The FPR captures the share of stable years, which are falsely thought of as preceding conflict. A FPR of 0.4 implies that 40 percent of all years without a conflict in the following year are falsely thought of as preceding conflict.

5) This procedure is repeated for all years in the respective sample. We then use the total TPR_c and FPR_c at different cutoffs to produce receiver operating characteristic (ROC) curves. These depict the trade-off between TPR_c and FPR_c and therefore capture the power in a simple and nonetheless meaningful way. Note, that these summaries minimize the problems brought about by our linear probability model. The transformation of fitted values to TPR_c and FPR_c allows all possible values to be easily converted into the same trade-off. In this way, the two dimensions of TPR_c and FPR_c provide a common space in which to interpret the fitted values $\hat{y}_{it}^{within}$ and $\hat{y}_{it}^{overall}$, despite the fact that $\hat{y}_{it}^{within}$ is centered around zero.

We separately test three different models x_{it} using different samples, which we discuss in detail in the Appendix. First, we use a model of rainfall shocks on the African continent.[11] Secondly, we use foreign aid shocks and income shocks interacted with the country's institutional environment to predict future onset.[12] Thirdly, we use a combination of endogenous economic and political variables to forecast conflict.[13] In all three models, we predict armed conflict and civil war onset as measured by Uppsala Conflict Data Program (UCDP) (and PRIO) battle related deaths.[14] 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

[11] We use the replication data provided by Miguel and Satyanath (2011).

[12] We use data and variable definitions from Besley and Persson (2011*b*).

[13] The model we use is based on Goldstone et al. (2010). This 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 neighbouring countries had an armed conflict.

[14] See Sambanis (2004) for a discussion of the conflict data.

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.[15]

The results for all three models are summarized in ROC curves in Figure 1. These depict the true positive rate (TPR) on the y-axis and the false positive rate (FPR) on the x-axis. Optimally, one would want a TPR of 1 at a FPR of 0. This would mean that all conflicts are predicted without raising any false alarms. The 45 degree line in ROC curves is the benchmark that would be reached by random forecasts.

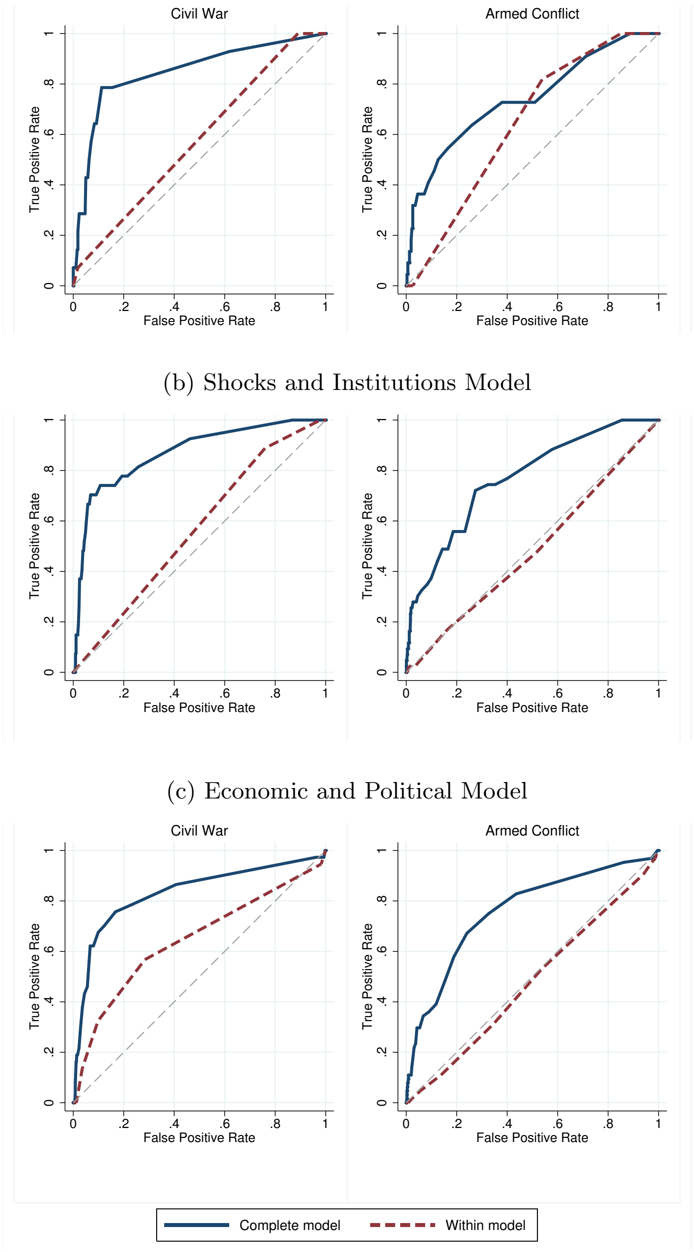
We first discuss the rainfall model which uses two precipitation growth rates to forecast conflict. The two ROC curves in the left and right panel of Figure 1a represent the overall model's ability to forecast (blue solid line) and the within model's ability (red dashed line). When we predict onset with the overall model, the results look promising. Both armed conflict and civil war onset are predicted reasonably well with TPR far above the FPR, i.e. the 45 degree line. However, the within variation contained in the model, captured by the red dashed line, contributes relatively little to the ability to forecast. Summarizing the predictive capacity in terms of the area under the curve (AUC), we see that for the left panel the AUC drops from 0.85 to 0.58, and for the right panel from 0.74 to 0.64. The second model combines proxies of external economic shocks together with political institutions in the country. Again, the blue solid lines in the panels of Figure 1b represent the forecasting capabilities of the overall model, $\hat{y}_{it}^{overall}$ while the red lines represent the forecasting power of $\hat{y}_{it}^{within}$. The within model now provides even less ability to forecast onset. We record a drop in the AUC from 0.87 to 0.57 on the left, and from 0.77 to 0.48 on the right. The third model presented in Figure 1c was developed to forecast instability. And, indeed, we now see that the within variation provides some ability to forecast civil wars out-of-sample. Nonetheless, the general impression is maintained. There is a considerable difference in the forecast capabilities between the overall and the within model, indicated by the fall in the AUC from 0.83 to 0.64 for civil war, and from 0.76 to 0.48 for armed conflict. The difference is particularly pronounced for armed conflict onset, which typically precedes civil war.[16]

By separating within and overall variation, we have established that most of the forecasting power in the overall model comes from country fixed effects with the AUC dropping by an average of 30 percent for both civil war and armed conflict. In other words, conflict onset is predicted to a large extent by the fact that it had previously taken place in a country. With some exceptions, existing economic and political variables have a hard time predicting the timing of conflict onset.

[15] 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.

[16] We have checked that this result is not due to bad luck concerning the way in which we combined variables. We have used a Lasso technique to select variables from a broad set of economic and political variables and the variables selected in this way are strikingly close to the model depicted in Figure 1c.

Figure 1: ROC Curves for Onset



(a) Rainfall Model

(b) Shocks and Institutions Model

(c) Economic and Political Model

Notes: The rainfall model is a model of rainfall shocks on the African continent using the replication data provided by Miguel and Satyanath (2011). The shocks and institution model is based on foreign aid shocks and income shocks interacted with the country's institutional environment using data and variable definitions from Besley and Persson (2011*b*). The economic and political model is based on Goldstone et al. (2010), 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 neighbouring countries had an armed conflict. The within model is the complete model net of country fixed effects.

4 A Topic Model of Newspaper Text

In light of the difficulty of predicting the timing of conflict, our hope is that news reports will generate time variation that goes beyond the economic and political variables discussed in the previous section. However, to do this, we first need to generate meaningful summaries of the news text written by journalists. In this section, we first discuss the news reports we rely on, and then discuss how we summarize them with the help of a topic model. Finally, we 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.[17] 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 100 articles are written on a country in a given year. However, the extent of coverage varies drastically with the type of country so that we observe between 1 and more than 5500 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.[18] However, total news articles in our dataset by newspaper are fairly constant across time as shown in Figure 2, i.e. there is not a large increase in available news which is typical for analysis that use all available sources of news.[19] This is quite intuitive. Total space for articles is fairly constant and so attention seems to shift towards countries which seem newsworthy to journalists and editors. 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. This is important as it facilitates forecasting across countries.

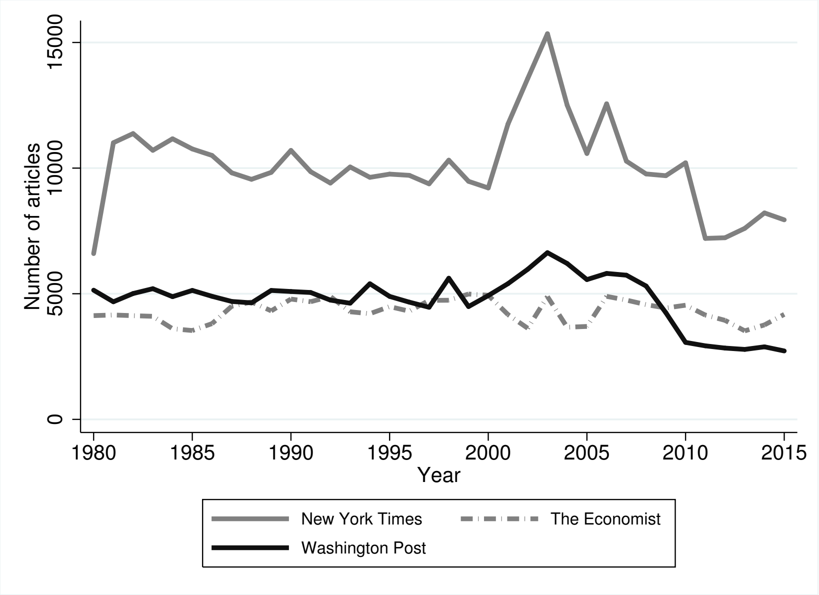
In order to improve the performance of our machine learning algorithm, we process the raw texts of articles of all three newspapers according to standard

[17] In the case of the Economist we also search in the leading paragraph as the title rarely contains a country or capital name.

[18] The findings come from simple OLS regressions. The findings on conflict are robust to the introduction of country fixed effects. Regression results are available from the authors on request.

[19] The NYT generally seems to shift in an out of discussions of foreign affairs more than the other two publications. Also, the period after the financial crisis saw a decline in news from the NYT and the WP which could be explained by a stronger focus on internal affairs. Around this time US troops were pulled out of Afghanistan and Iraq.

Figure 2: Number of articles by news source over time



text mining procedures. First, we remove a library of common words, which in text mining are referred to as stop words, such as "to" or "that".[20] Second, we lemmatize and then stem words using the Snowball algorithm, which is an updated version of the algorithm from Porter (1980).[21] Lemmatizing groups variant 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. Finally, since for our project we intend to capture general rather than content specific to a single country, we remove country names and names of people, identified by a library of names and the usage of titles, such as "Mr" or "Mrs".[22] 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. Dropping rare tokens, in particular, means that we drop a lot of tokens from the list without losing a lot of text. Even after this procedure we 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

[20] See `http://norm.al/2009/04/14/list-of-english-stop-words/` for the list of stop words.

[21] The Python package for lemmatizing is available at `http://www.nltk.org/_modules/nltk/stem/wordnet.html` and for stemming at `http://snowball.tartarus.org`.

[22] We use the Natural Language Toolkit dictionary of names for males "names.words('male.txt')" and females "names.words('female.txt')".

typically reduced dimensionality by focusing on particular words.

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 an 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" and 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 "win" might be more of a sports-topic word and will, therefore, indicate 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. However, topics themselves are also backed out. 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.

In more technical terms, LDA generates a stream of observable words $w_{m,n}$, partitioned into documents, which are vectors of words $\vec{w}_m$, i.e. the order of words does not matter.[23] The model assumes that for each of these documents, a vector of topic proportions, $\vec{\eta}_m$, is drawn from a Dirichlet distribution $Dir(\vec{\alpha})$. 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, $\vec{\varphi}_{z_{m,n}}$, is used to draw a word. The topics $\vec{\varphi}_k$ are sampled from a Dirichlet distribution $Dir(\vec{\beta})$ once for the entire corpus. The key in estimating this model is that only the $\vec{w}_m$ are actually observed. Everything else needs to be backed out. Typically, the elements of the vectors $\vec{\alpha}$ and $\vec{\beta}$ 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. This is proportional to

$$p\left(z_i = k \mid z_{-i}, w\right) \propto \frac{n_{k,-i}^{(t)} + \beta}{{}_{t=1}^{V} n_{k,-i}^{(t)} + \beta} \cdot \frac{n_{m,-i}^{(k)} + \alpha}{\left[{}_{k=1}^{K} n_{m}^{(k)} + \alpha\right] - 1} \tag{7}$$

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

where $n^{(t)}_{k,-i}$ is the frequency by which the token of word i was attributed to the same topic generally and $n^{(k)}_{m,-i}$ is the frequency by which all other words in the same document m are attributed to the topic. This highlights the role played by co-occurrence. The algorithm forms topics around tokens that appear together in many documents. The term $\left(n^{(k)}_{m,-i}+\alpha\right)\left(\left[{}^{K}_{k=1}n^{(k)}_{m}+\alpha\right]-1\right)^{-1}$ ensures that if a lot of tokens in a text are attributed to the same topic then it is more likely that 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.

We let the chain run for 1000 iterations.[24] 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 $\alpha = 3.1$ and $\beta = 0.01$. Concerning α and β we follow the literature; but we estimated models for 5, 10 and 30 topics. All of these yield very similar results.

4.3 Topic Estimation Results

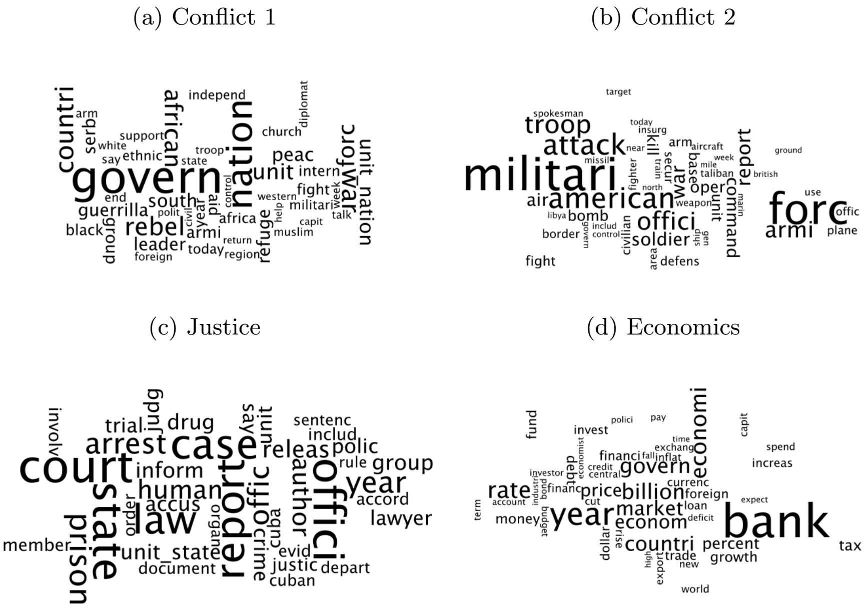
In order to be able to use the estimated topics in out-of-sample forecasting we need to estimate the model for each year. We start forecasting in 1995 so that the first topic model we estimate uses all articles between 1975 and 1995. We estimate one model for each consecutive year, where the last model uses all text from 1975 up to 2015.

The K distributions over terms, $\vec{\varphi}_1, ..., \vec{\varphi}_K$, are called topics. In our application to the news content the estimated topics seem natural and intuitive, i.e. when we look at the most common words in each topic it is fairly easy to come up with a title for the topic. For example, in all years topics appear which we can classify as conflict, sports, tourism, the economy and politics.

In Figure 3, we present four topics from the 2015 estimation as word clouds of the top 50 words 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 quite clearly suggest (potential) violence. Words like force and military indicate as much. 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 hundreds of thousands of tokens. This is important as the full list of tokens associated with the topics in Figure 3, for example, could include factors that trigger or at least anticipate conflict. Figure 3c seems to summarize processes in the justice system.

[24] The C++ Gibbs Sampler we use is provided by Phan and Nguyen (2007) and is available at `http://gibbslda.sourceforge.net`. 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 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 $\alpha = 3.1$ and $\beta = 0.01$. The size of a word represents its probability within a given topic.

The topic in Figure 3d describes economics. In our discussion of which topics predict conflict, we will treat this topic as the omitted category.

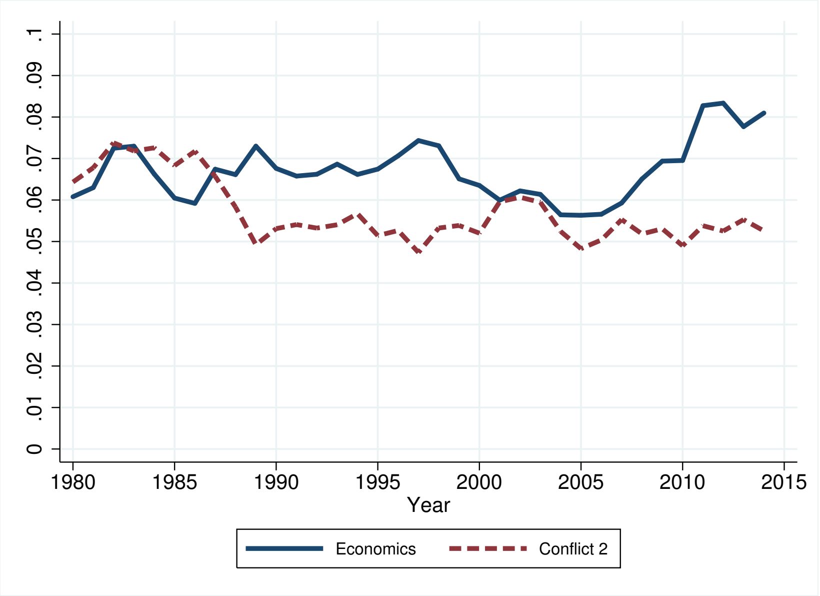
After estimating the topic model, we are in possession of a dataset containing the composition of each article m in terms of the K topics, $\vec{\eta}_m$. The question remains of how to aggregate the shares in each article to receive a topic distribution in a country-year. We use a simple method that takes 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

$$\vec{\theta}_{it} = ({}_{m\in M_{it}}\vec{\eta}_m N_m + \alpha)/\,({}_{m\in M_{it}} N_m + K\alpha) \qquad (8)$$

where ${}_{m\in M_{it}} N_m$ 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. If a lot of words are written, the posterior topic distribution can deviate strongly from the prior.

In order to use our estimates when forecasting out-of-sample, we need to estimate a full panel of topic shares $\vec{\theta}_{it}$ for each year T. Figure 4 plots the average share of the economics and one of the conflict topics between 1980 and 2015 as seen through the lens of topics estimated in 2015. The average share is fairly stable across time. Even when the global financial crisis breaks out in 2009, the economics topic gains only about 2 percentage points of all words written in the three newspapers on average. This does not mean, however, that topic shares do not change dramatically at the country level. We return to the content of topics in Section 7 after presenting the main results.

Figure 4: Average Topic Shares of Economics and Conflict and Over Time



Notes: Average topic shares are computed using aggregated country/year topic shares.

5 Predicting Conflict with Newspaper Topics

In this section, we use the estimated topic shares in linear fixed effects regression to forecast conflict out-of-sample. Our estimation method follows the one described in Section 3.2 with only slight changes. In each year T between 1995 and 2013, we estimate a topic model using text written between year 1975 and year T. We obtain a vector of 15 topic shares $\vec{\theta}_{it}$ in country i at time t, which we calculate as in equation (8). Then we use these shares as our explanatory variables $\vec{x}_{it}$ in our estimation equation (6).

As before, we calculate two sets of fitted values

$$\hat{y}_{it}^{overall} = \hat{\beta}_i + \vec{\theta}_{it}\hat{\beta}^{topics}$$

and

$$\hat{y}_{it}^{within} = \vec{\theta}_{it}\hat{\beta}^{topics}$$

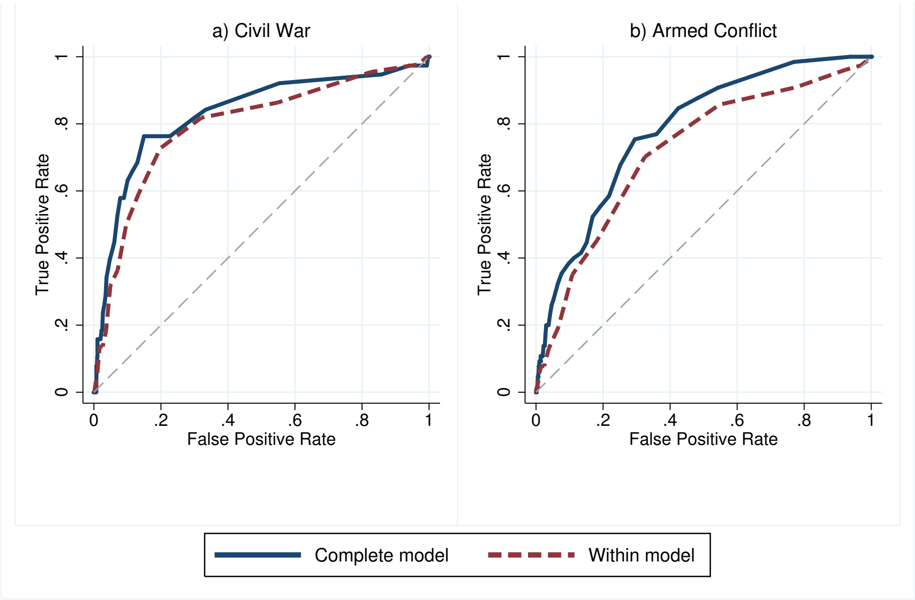
for the same set of countries and year $t = T + 1$. Using a set of varying cutoffs c and our estimates for $\hat{y}_{it}^{within}$ and $\hat{y}_{it}^{overall}$, we calculate the true positive rate TPR_c and the false positive rate FPR_c for each of the cutoffs c. The results can be presented in standard ROC curves.

5.1 Main Results

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 and end.

Our main results is shown in the two graphs in Figure 5, which show receiver operating characteristic (ROC) curves for the outbreak of both civil war and armed conflict.[25] The blue lines in all four panels show the forecasting performance using the fitted values from the overall news model $\hat{y}_{it}^{overall}$ while the red lines provide the ROC curve of the within model $\hat{y}_{it}^{within}$. As before, one would want a true positive rate of 1 at a false positive rate of 0. This would mean that all conflicts are predicted without raising any false alarms. The 45 degree line in ROC curves is the benchmark that would be reached by random forecasts.

Figure 5: ROC Curves for Onset (Topic Model)



Notes: The topic model is based 15 topics computed using LDA with $\alpha = 3.1$ and $\beta = 0.01$, which are aggregated at the country/year level. The within model is the complete model net of country fixed effects.

Figure 5 shows that news topics fair well at predicting onset of both civil war and armed conflict. When predicting civil war onset, the news model generates

[25] The results hold not only for the onset of conflict but also incidence as well as can be seen in the Appendix in Figure C.1.

a TPR of about 70 percent for a FPR of 20 percent. Furthermore, the predictive power of the within model is very close to the predictive power of the full model. This is quite a striking finding given the difficulty of forecasting the timing of such rare events. The model of predicting the onset of armed conflict performs worse but the same basic pattern as with civil war onset is maintained.[26] Again, the within variation seems to be the main driver of the ability to forecast conflict. The AUC only drops from 0.83 (0.80) in the complete model to 0.78 (0.72) in the within model. This is an important difference to variables which capture the economic or political fundamentals presented in Section 3.2, which suffer an average drop of 0.3.

Our topic model performs extremely well when predicting conflict incidence. The overall model can predict 90 percent of both civil wars and armed conflicts correctly at a false positive rate (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 worse but still generates a TPR of 60 percent for a FPR of 20 percent. The better performance of these 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.

Topics provide an automated summary of text. This is a particular advantage for forecasting if parts of the text become useful which one would not have payed attention to otherwise. To check this we contrast our model with a model based on a set of word counts as suggested by Chadefaux (2014).[27] These counts are useful for forecasting under the prior that conflict words will anticipate conflict. Figure 6 shows ROC curves for the word-count model in blue and the ROC curves for our topic model as dashed red lines. When predicting the timing of civil war onset, word counts provide quite a lot of forecast capacity. At a false positive rate of 20 percent they reach a true positive rate of over 50 percent. However, the topic model still adds about 20 percentage points to the true positive rate - based on exactly the same text. When predicting armed conflict onset the difference between the two ways of summarizing the text becomes even more striking. Words counts now do not predict onset. This is interesting because it indicates that what provides our topic model with its forecasting power when predicting onset is not the rise in tokens directly related to conflict. We return to this insight in section 7. The topic model also performs better when predicting incidence but the gap between topics and word counts is much smaller.

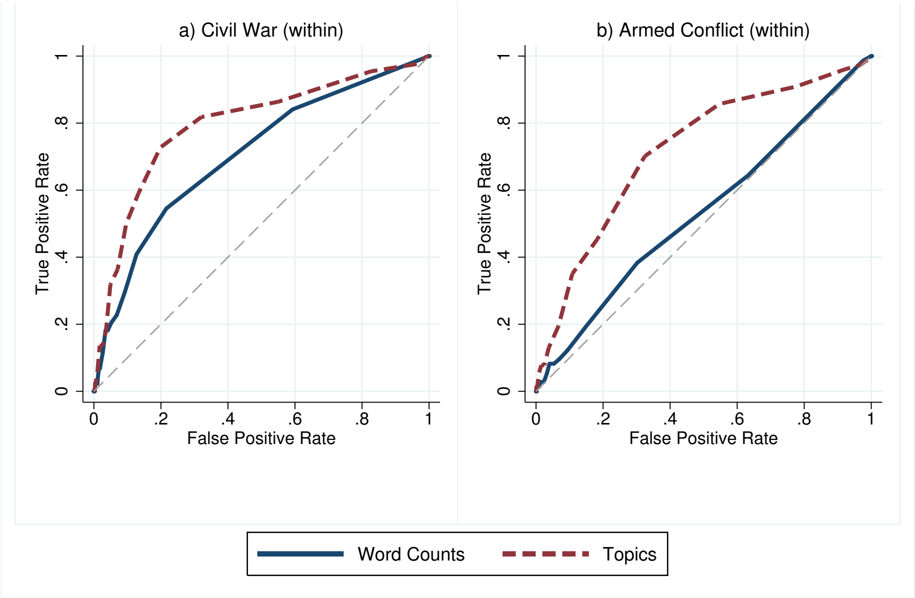
We have run several robustness checks regarding these basic results which we report in Appendix B. First, we also used more and less topics.[28] As shown

[26] The onset of armed conflict is particularly difficult to predict as it is preceded by years without organized violence.

[27] We follow his methodology closely and count the words in our news data for every country/year. We then generate a model based on the four word-count variables also used by Chadefaux (2014).

[28] We adjust α accordingly to maintain the ratio $\alpha = 50/K$. We have resisted the temptation

Figure 6: ROC Curves for Onset (Topic Model vs Word Counts)



Notes: Word counts are computed by counting the frequency of the list of conflict words defined by Chadefaux (2014) per country/year. The topic model is based 15 topics computed using LDA with $\alpha = 3.1$ and $\beta = 0.01$, which are aggregated at the country/year level. The within model is the complete model net of country fixed effects.

in the Appendix in Figure B.1, results remain largely unchanged with five topics but forecasting power drops slightly when predicting armed conflict onset. With 10 or 30 topics, the results are very similar to 15 topics. It seems reasonable to assume that if we would increase topics further, the model would fit very will within-sample but worse out-of-sample. We made this experience when trying to fit more standard variables. The model tends to overfit to specific situations, which then do not generalize out-of-sample.

Secondly, we show that, building on a standard model, the topic shares add forecasting power. This indicates that it is not simply reporting on basic economic and political facts that helps us forecast.

Thirdly, we add an indicator for contemporaneous conflict to the incidence model to see whether the news model provides forecasting power beyond the simple logic that conflict follows conflict. News reports add forecasting power beyond an already very high benchmark. This confirms the idea that the topic model has some value.

Fourthly, we discuss using different definitions of conflict in the appendix. Again, we find similar results. We have also analysed how well our model

of testing different topic models close to 15 topics as this would put the idea behind running out-of-sample tests on its head.

performs when forecasting conflict two years before onset. The within model performs only slightly worse. Interestingly, the overall model performs very similarly which confirms the idea that the between variation dominates the overall model. If the forecast is time invariant, it does not matter whether conflict breaks out one or two years later.

Fifth, we contrast our forecasting model with the event data generated by the Integrated Conflict Early Warning System (ICEWS). The comparison is particularly interesting as event data provides an alternative summary of news reports. We find that topics are, generally, able to forecast better when forecasting onset and that there is no clear dominance when forecasting incidence.

Finally, we use our topic model to predict refugee movements. To do so, we use data on refugees from the UNHCR and try to predict the onset of a large number of refugees. We use two different cutoffs, 30,000 persons and 130,000, which is similarly common to armed conflict and civil war. Again we find that 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 reported by the news.

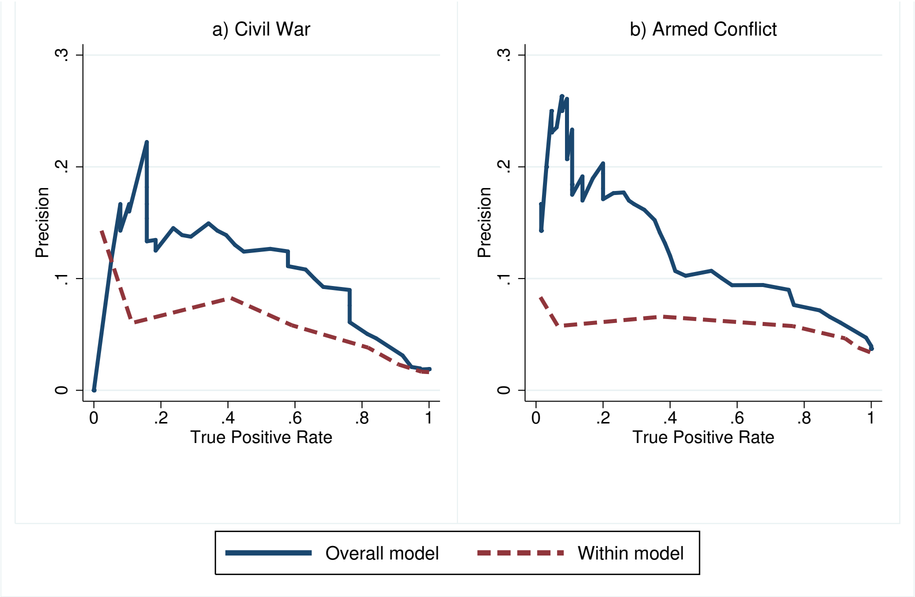
The main takeaway from all these tests is that the timing of conflict can be predicted using automated summaries of news reports. The topic models produce a relatively high true positive rate for relatively low rates of false positives. However, for rare events such as civil war and armed conflict, false positives are a problem even at low rates, as the majority of country/years are non-conflict years. A way to capture this problem is *precision*,

$$P_c = \frac{TP_c}{FP_c + TP_c},$$

which gives the number of years in which conflict was predicted correctly divided by all conflict predictions. Hence, the difference between P_c and the TPR_c is that true positives are not set into a relation to all years of conflict ($FN_c + TP_c$) but to all years in which conflict was predicted ($FP_c + TP_c$). In Figure 7 we compare precision to the TPR_c for our main model with 15 topics. Note that the new curves can take values of P_c of 0 and 1 for $TPR_c = 0$, depending on whether the first positives are true or false positives. As the TPR_c goes towards 1, precision will converge towards the ratio between conflict and non-conflict observations.

In the Appendix in Figure C.2 we can see that precision starts out very high when predicting incidence. At a true positive rate of 80 percent, precision is still above 20 percent when predicting civil wars and above 50 percent when predicting armed conflict. This means that the large majority of armed conflicts are correctly anticipated at the cost of raising false alarms in only half the cases. The within variation contained in the model is, again, an important part of the overall variation. Precision is generally much lower when predicting onset. This reflects the fact that onset is harder to predict and is much rarer. Still, the perception that the time variation is useful in the news model is maintained.

Figure 7: Precision Recall Curves for Onset



Notes: The curves are based on the results of the topic model computed using LDA with $\alpha = 3.1$ and $\beta = 0.01$, which are aggregated at the country/year level. The within model is the complete model net of country fixed effects.

6 How the Within View Changes Early Warning

In this section we imagine the results from the previous section would be used by an agent who is interested in forecasting conflict events correctly. This could be, for example, inhabitants and firms inside the country itself, foreign investors, a ministry for foreign policy in a different country or a risk insurer. Implicitly all of these actors would be interested in setting up a cutoff, c, which would trigger some sort of response or closer scrutiny. It turns out that this way of thinking about this problem of forecasting delivers a simple way to think about how risk perceptions change with the adoption of the within and between model.

As pointed out by Kennedy (2015), the relevant information for a decision problem of how to set up a cutoff, c, in such a scenario combines the results in the previous section with the agent's attribution of weights to the four outcomes TP_c, FP_c, TN_c and FN_c. Let us denote the costs for these as k_{TP}, k_{FP}, k_{TN}, and k_{FN}, respectively. The policymaker will then minimize total costs

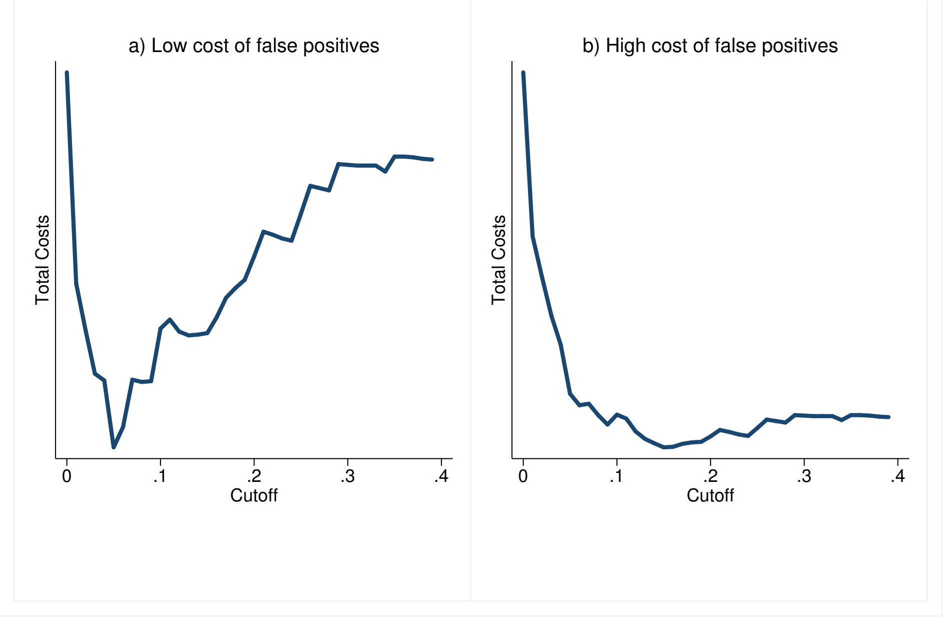
$$\min_{c} costs\,(c) = TP_c \times k_{TP} + FP_c \times k_{FP} + TN_c \times k_{TN} + FN_c \times k_{FN}. \quad (9)$$

Potentially this problem could be fed with country-specific cost parameters. One could, for example, give allies, trade partners, or countries with larger populations more weight. For now, we only aim to illustrate the change in perspective that the within variation would give on this choice problem.

First, the policymaker would need to choose a cutoff implied by her model. This will be a function of the cost parameters assumed. To simplify the discussion, assume that the costs of correct predictions are $k_{TP} = k_{TN} = 0$. Typically one would attribute a high cost to k_{FN} because a negative surprise might be particularly costly for the policymaker, $k_{FN} > k_{FP}$. We then look at two scenarios regarding the remaining parameters. First, assume that false positives are relatively cheap, for example that they cost only five percent of a false negative, $k_{FP} = 0.05 * k_{FN}$. Second, assume that the costs of false positives are relatively high, ten percent of the costs of false negatives $k_{FP} = 0.1 * k_{FN}$.

We use these parameter values to evaluate the total costs generated by different cutoffs, c. For every country/year we first generate a dummy that takes a value of 1 if the condition $\hat{y}_{it}^{overall} \geq c$ is satisfied. We then compare that dummy to the variable y_{it} to generate the set of variables TP_c, FP_c, TN_c and FN_c. Finally, we calculate $costs\,(c)$ from equation (9).

Figure 8: Cost Curves for Armed Conflict Incidence



Notes: Cost curves are computed using $k_{FP} = 0.05k_{FN}$ (low) and $k_{FP} = 0.1k_{FN}$ (high).

In panel a) of Figure 8, we show the results under the assumption that false positives carry a low costs. The cost function takes a U-shape with a minimum at around $c = 0.05$. At this value the model generates a large TPR of around 75 percent and a FPR of 29 percent.[29] For lower values of c the number of false

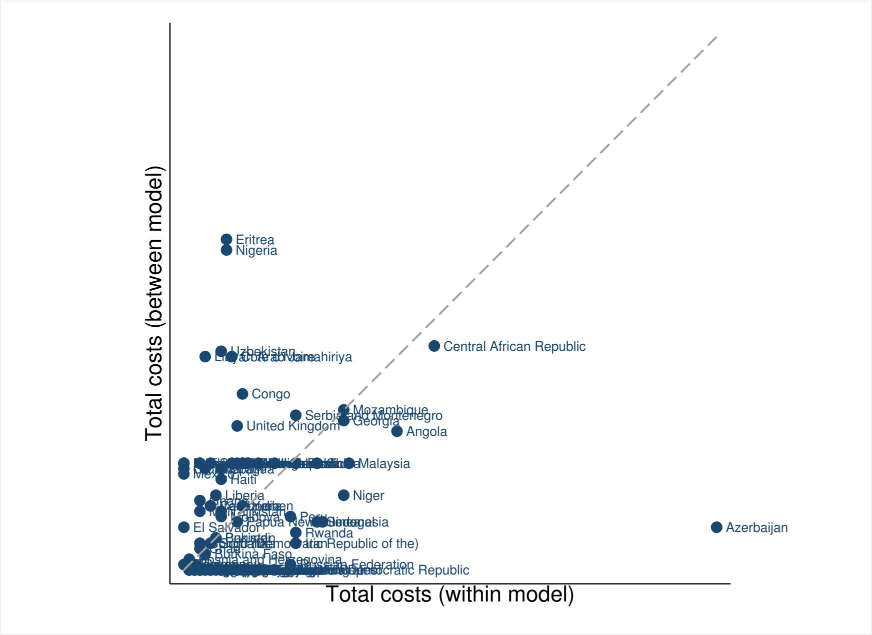
[29] It is easy to verify that this point is indeed on the blue line in Figure 5, panel b). The line in the TPR and FPR space flattens considerably at this point which means our model

positives would increase too much compared to the gain in true positives. For higher values of c we would get more false negatives and higher costs.

The picture changes significantly in panel b) of Figure 8 where we assume a higher relative cost of false positives. Now the minimum cost is reached just below $c = 0.15$. The higher cutoff reflects the fact that false positives have become more costly. At this cutoff the model generates a TPR of 35 percent and a FPR or 8 percent which, again, is a point on the blue line in Figure 5, panel b). Clearly, we have now less false positives per false negatives.

An important advantage of assuming a cost structure is that we can simplify the four dimensions of forecasting models into one single dimension. This allows us to compare our within and between model with a simple cost measure. In Figure 9 we compare the total minimal cost by country generated in the within and between model if we assume a low cost of false positives.[30] Under these assumptions the between model produces more false negatives and the within model produces more false positives. In other words, the within model anticipated more conflicts but generates more false positives while it is trying to predict the timing of conflict. Strikingly, the within model would have predicted the most recent onsets in Egypt and Libya while the between model would not have.

Figure 9: Cost Comparison of Within and Overall Model



Notes: The costs are computed using $c = 0.044$ and $k_{FP} = 0.05k_{FN}$.

produces a lot of additional false positives for few true positives at this point.

[30] In the between model we use $c = 0.033$ and in the within model we use $c = 0.04$ as this minimizes costs.

In addition, there is are large differences with regard to which countries produce high costs in the two models. Azerbaijan, for example, generates very high costs in the within model as several onsets of armed conflict took place in a row, two of which remained undetected by the model. The between model generates larger costs in, for example, Eritrea, Uzbekistan, Libya and Nigeria. This is evidence that, at the very least, the within model can add a new perspective on the problem of forecasting even when compared to the overall model. However, such a model should only be used if it generates meaningful variation, as does our proposed topic model.

7 How News Topics Forecast Timing

We have shown that topics generated from news reports can be used to forecast the timing of conflict, i.e. the within variation. We have also shown that this can lend a new perspective on how country risk is perceived. We now turn towards analyzing the exact news content which is responsible for our ability to forecast.

The topic summaries of our articles can provide useful clues as to what predicts conflict. As already mentioned in Section 4.3 several topics appear consistently and with similar word lists across years. We can therefore test each topics' relation to conflict in several samples.[31] In order to test which topics predict conflict, we first categorize word lists in each of the years in our test samples with T running from 1995 to 2013. For each sample lasting until year T, we then use these codings in fixed effects panel regressions and record the estimated coefficients on the variables $\vec{\theta}_k$ and their standard errors. As before, the dependent variable is a dummy which indicates one year before conflict onset. The topic of economics, for example, appears consistently across all years. In order to have a consistent benchmark from one year to the next, we make this topic our omitted category. This means all estimated coefficients should be interpreted relative to this topic.

Results for four of our topics are summarized in Figure 10. Each dot represents the coefficient of the respective topic share from a single fixed effect panel regression. The coefficients are ordered by magnitude with the x-axis exhibiting the relative rank of the coefficient. The blue dots represent coefficients which are significantly different from 0, i.e. different from the effect of the economics topic benchmark at the 10 percent level, whereas red dots represent insignificant ones. The thin vertical black lines illustrate the 90 percent confidence interval.[32]

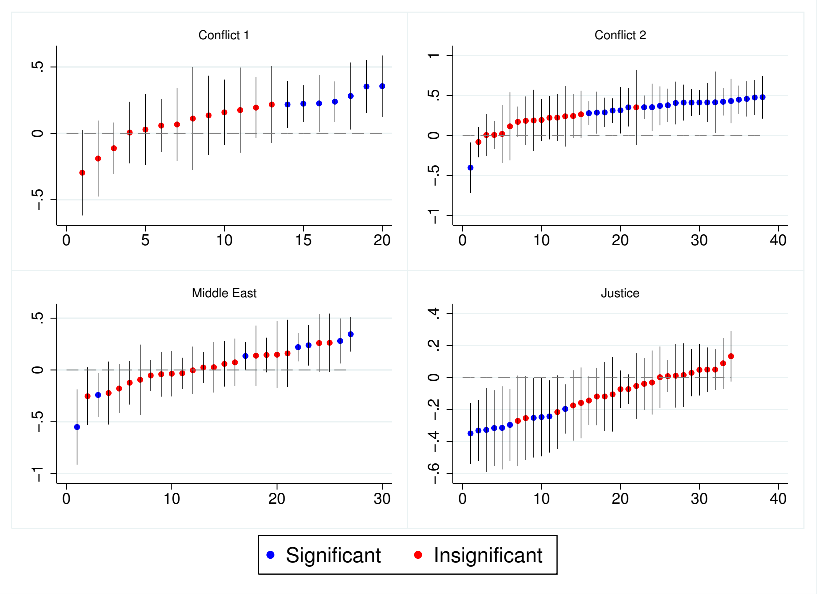
Topics which contain conflict words increase relative to the economics topic before conflict breaks out. In addition, words that describe judicial procedures seem to decrease before conflict while words describing the Middle East increase. For the remaining topics not shown here we find that they seem to behave similar to economics, i.e. the estimated coefficients we get are centered around

[31] Note that here we are looking at relations of topics to one year before conflict *in-sample*, rather than *out-of-sample* as in Section 4.3.

[32] Some Figures show less coefficients because not all topics appear in all years.

zero and mostly insignificant. The coefficient of the topic describing industry or tourism, for example, are sometimes positive, sometimes negative, but generally not significantly different from economics. Note, however, this can still mean that news on industry, economics and tourism all decrease relative to the conflict topic one year before conflict.

Figure 10: Coefficients of Topic Shares for Onset



Notes: The coefficients and confidence intervals for each of the four to of 15 topics are obtained from a fixed effect panel regression with a dummy for one year before the onset of conflict as dependent variable. We use both armed conflict and civil war as dependent variable for all available samples from 1995 to 2015. The coefficients are ordered by magnitude.

Next, we can look into the cases in which conflict was predicted by our model through high fitted values and look inside the distribution of topic shares for all articles written during these years. In other words, we now look at the article shares $\vec{\eta}_m$ and compare years before conflict onset to other peaceful years. As a measure of our prediction we take the cutoff of $c = 0.05$ derived in the previous section. With this cutoff, we predicted 35 conflict onsets in 25 countries correctly.[33] In what follows we focus on these cases and compare high risk to low risk years within the same countries.

We want to know how writing changes before conflict. To understand this we first look at the full distribution of topic shares in the articles. Figure

[33] The countries are Bangladesh, Central African Republic, Chad, Congo, DR Congo, Cote d'Ivoire, Egypt, Eritrea, Georgia, Guinea-Bissau, Iraq, Lebanon, Lesotho, Liberia, Libyan Arab Jamahiriya, Mali, Mauritania, Niger, Nigeria, Pakistan, Senegal, Somalia, Sri Lanka, United Kingdom, Uzbekistan, Yemen.

11 shows the kernel densities for all articles written during peacetime in our conflict countries. On the y-axis we display the kernel density and on the x-axis we display the respective topic share. In the top panels we compare the distribution of the shares for $\vec{\eta}_{m,conflict}$ and $\vec{\eta}_{m,economics}$ in the years with high and low risk.[34] A value of $\vec{\eta}_{m,economics} = 0.2$ in panel a) means, for example, that one-fifth of the tokens in these articles were written on economics. If more is written on economics during conflict, we would observe that more articles contain a higher share of economics during conflict, which would change the red dashed line.

From panel a) in Figure 11 it is clear that there are no obvious observable changes in the shares written on economics. The distributions of $\vec{\eta}_{m,economics}$ displayed in panel a) is very similar in high and low risk years. The peak in each of the kernel densities is close to the prior of 1/15, and the mass of articles in the tails is almost identical in the two distributions. We only observe a slight concentration of articles with low economics shares close to the prior. There are, however, more visible changes in the distribution of $\vec{\eta}_{m,conflict}$. Now the density around the prior drops visibly one year before conflict onset. At the same time, many more articles contain higher conflict shares.

In the bottom of Figure 11, we look at two additional topic dimensions - justice and the second conflict dimension. Again, there are significant changes. News that contain tokens related to justice decrease more visibly, albeit not drastically. The density around the prior increases slightly and less articles with high justice shares are written. In addition, we can detect a visible shift on the second conflict dimension. The density of articles with low conflict shares decreases and we get visibly more articles with shares above 0.10.

What this illustrates is that the shifts in topic shares $\vec{\theta}_{it}$ are not driven by shifts in which all journalists suddenly report much more on conflict or less on economics. Instead, some articles seem to change character while most maintain their topic mix.[35] This reveals how the journalistic work conducted in these countries adds information. It is about a completely new kind of news story making headlines and not about changing nuances within already pre-existing news stories.

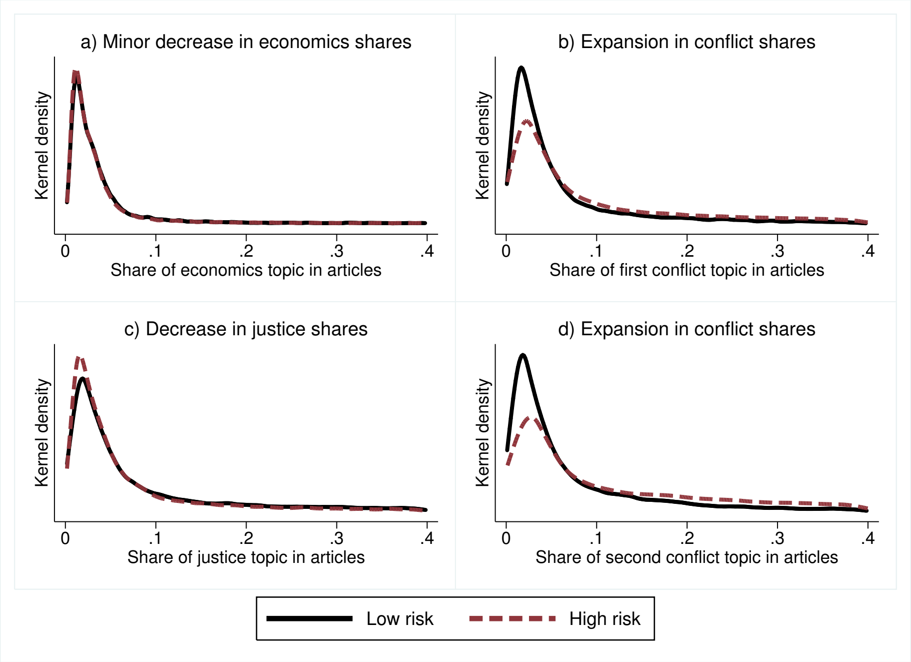
As an additional illustration of how the justice topic helps predicts conflict, we show how it evolves before and after conflict by regressing the justice topic share on dummies for the number of years before the onset of conflict and the number of years after conflict has ended.[36] As can be seen in Figure 12, the justice topic decreases significantly one year prior to conflict and increases after conflict has ended. Strikingly, there seems to be a significant boom of news stories after conflict. The fact that stories on trials and justice disappear before conflict could indicate that judicial institutions are less active before

[34] To be as consistent as possible, we always use the share estimates as estimated from the topic estimation in the last year, 2013.

[35] The change of only some articles to high conflict shares is consistent with the findings in 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.

[36] In this regression we control for country fixed effects and the remaining topic shares.

Figure 11: Topic Shares of Economics, Justice, and Conflict in the Universe of Articles when Risk Is High vs Low

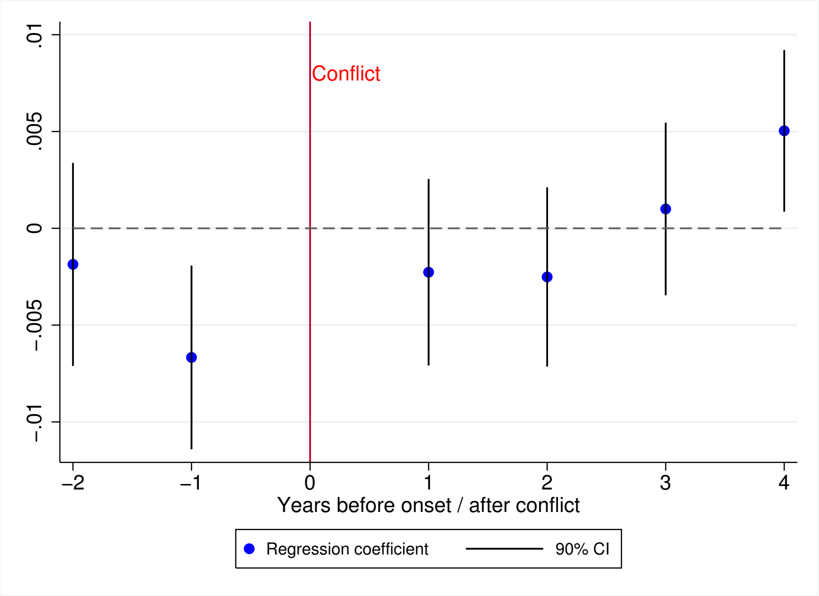


Notes: Shares represent average topic shares of all articles (not aggregated at the country/year level). High risk is defined as a predicted probability of onset above five percent.

violence breaks out. A second margin through which justice relates to the outbreak of conflict is that once conflict has ended, articles about war trials and police prosecution begin to surface, which seem to indicate the sustainable end of a conflict and the (re)establishment of a functioning state. The fact that economics, for example, does not react as strongly could indicate that judicial institutions and processes have a particularly strong effect on post-conflict stability. In any case, the pattern we see in the data is consistent both with the idea that post conflict justice is a crucial factor for stability or that legal checks and balances play an important role in keeping existing tensions from becoming violent.

An important takeaway from Figures 10, 11, and 12 is that our forecasting model can rely on more than one dimension of the news content when forecasting. Writing on justice, for example, decreases clearly before conflict. In this context, it is important to stress that the coefficients we display in Figure 10 are coming from regressions which control for the conflict topic shares. In other words, justice adds predictive power beyond conflict and relative to economics. We show in Figure C.3 in the Appendix that for cases in which we predict conflict, we also find visible changes in the distribution of articles on industry

Figure 12: 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.

and tourism.[37] Our model is, therefore, able to spot a relatively diverse set of risks to internal peace. In addition, increases in the non-conflict dimensions can indicate a stabilization. This is important because forecasting correctly requires that the model exhibits falling risk at times in which countries stabilize.[38]

8 Conclusions

In this article we present a new method of aggregating news text in a meaningful way. Topic models have the ability to diminish the dimensionality of text from counts of more than one million expressions to, for example, 15 topics. We have argued that, aggregated this way, news text can be used to predict the timing of conflict.

Our findings highlight that models need to be tested for whether their within variation is meaningful. If not, policymakers might rely on meaningless changes

[37] See Figure C.3 in the Appendix.

[38] For example, the risk of a civil war onset fell by over 9 percentage points in Angola when the most intense violence ended in 2001. And this was despite the fact that armed conflict flared up several times during this time. An important correlate of this fall in risk was an expansion in news stories about industry together with a decrease in the conflict topic.

of risk across time and this has the potential to lead to large errors. We have shown, for example, that the within variation of a standard model has surprisingly little power when forecasting the timing of armed conflict. This is a finding that should be taken into account when interpreting existing studies that do not distinguish within from between variation.

Ultimately, forecasters might face a trade-off between prediction with maximum accuracy overall or using a less accurate model that generates useful variation across time. At the very least, using a model with useful variation across time should provide a useful addition for forecasters. Having useful variation in conflict risk across time might be of value on its own right. There are many cases for which our model reports sudden increases in conflict risk which were not followed by conflict. There are two options regarding this variation. First, it might be due to reporting in the newspapers we use. For example, the invasion of Iraq generated a lot of "conflict" news which might have spilled over to other countries. Second, risk might have actually increased but policies prevented destabilization. Our within measure of conflict risk for Egypt, for example, shows several increases and decreases before the most recent outbreak of violence.

Topic models could provide a useful alley for research in political events more generally.[39] We have used the most simple, off-the-shelf, version of the various algorithms available and have used the same text collection for estimating the topic model and calculating topic shares. One could instead train a topic model on specific subsets of texts or on a separate set of texts and then use this to spot the generated topics in the main body of newspaper articles. Applications include training a topic model on academic articles or specialized country reports and then using the generated topics to figure out which set of topics lead to better forecasts. 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).

[39] In addition, our method could be used for nowcasting as there is a striking degree of accuracy in which conflict periods and even intensity is captured by our model.

Appendix

A Standard Models

We use three models to forecast conflict. The first 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*). The third model we use is from Goldstone et al. (2010). We follow their generation of the political institutions dummy closely. We also add data on infant mortality from the world bank and the share of population which is discriminated from the GROWup dataset. The latter 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 this lowers the ability of this model to forecast. Finally, we add a count of adjacent countries with an ongoing armed conflict.

We merge this 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 ensure comparability across models we only include countries with more than 1 million inhabitants. Summary statistics for all variables in our sample are in Table 1.

We also tried looked at the within-sample performance for all three models. The results were fairly inconsistent in the model following Besley and Persson (2011*b*), broadly consistent in the model following Goldstone et al. (2010) and very consistent in the case of Miguel and Satyanath (2011). However, these differences can, perhaps, largely be explained by the different methodology we use. Besley and Persson (2011*b*) try to explain incidence with contemporaneous data, Goldstone et al. (2010) use a non-linear model, produce random samples from their sample to reduce the number of zeros and add sub-continent fixed effects instead of country fixed effects. Only Miguel and Satyanath (2011) use a very similar framework. In light to the out-of-sample performance of these three models, it is worth noting the contrast between the rainfall model and the model inspired by Goldstone et al. (2010). In the latter case, we only find one coefficient which is significantly different from zero. Yet, this model performs

Table 1: Summary statistics

	Obs.	Mean	SD	Min	Max		Obs.	Mean	SD	Min	Max
Main sample						***Pillars of prosperity sample***					
Topic 1 share	5538	0.081	0.048	0.005	0.529	Natural disaster	5988	0.306	0.461	0	1
Topic 2 share	5538	0.068	0.039	0.004	0.403	Natural dis. * good institutions	5988	0.054	0.226	0	1
Topic 3 share	5538	0.075	0.073	0.006	0.694	Security council member	5988	0.067	0.250	0	1
Topic 4 share	5538	0.095	0.092	0.003	0.623	Sec. c. m. * good inst.	5988	0.014	0.118	0	1
Topic 5 share	5538	0.054	0.040	0.004	0.542	Sec. c. m. * cold war	5988	0.045	0.207	0	1
Topic 6 share	5538	0.069	0.055	0.003	0.685	Sec. c. m. * c. war * good inst.	5988	0.009	0.095	0	1
Topic 7 share	5538	0.078	0.057	0.007	0.594	Armed conflict	5988	0.171	0.377	0	1
Topic 8 share	5538	0.084	0.057	0.006	0.513	Civil war	5988	0.084	0.278	0	1
Topic 9 share	5538	0.060	0.042	0.004	0.480						
Topic 10 share	5538	0.055	0.058	0.004	0.624	***Early warning sample***					
Topic 11 share	5538	0.042	0.050	0.001	0.501	Partial autocracy	5298	0.199	0.399	0	1
Topic 12 share	5538	0.071	0.048	0.005	0.382	Part. democ. w/ factionalism	5298	0.179	0.383	0	1
Topic 13 share	5538	0.063	0.051	0.005	0.686	Part. democ. w/o factionalism	5298	0.116	0.320	0	1
Topic 14 share	5538	0.052	0.059	0.003	0.519	Full democracy	5298	0.206	0.405	0	1
Topic 15 share	5538	0.056	0.047	0.002	0.507	Infant mortality	5298	54.766	45.174	2.2	215
Articles	5538	121.324	243.798	1	5542	Share of discriminated pop.	5298	0.023	0.151	0	1
Armed conflict	5538	0.187	0.390	0	1	4+ armed conf. adjacent	5298	0.055	0.152	0	0.98
Civil war	5538	0.081	0.272	0	1	Armed conflict	5298	0.190	0.393	0	1
						Civil war	5298	0.084	0.277	0	1
Rainfall sample											
Precipitation growth	965	0.020	0.218	-0.609	1.677						
Precip. growth $t-1$	965	0.021	0.215	-0.550	1.677						
Armed conflict	965	0.254	0.435	0	1						
Civil war	965	0.122	0.328	0	1						

Notes: All samples are intersections of the variables and the conflict variables ”armed conflict” and ”civil war”. The rainfall sample includes only African countries. For variable description see the Appendix text. Topic shares and number of articles are from the year 2015.

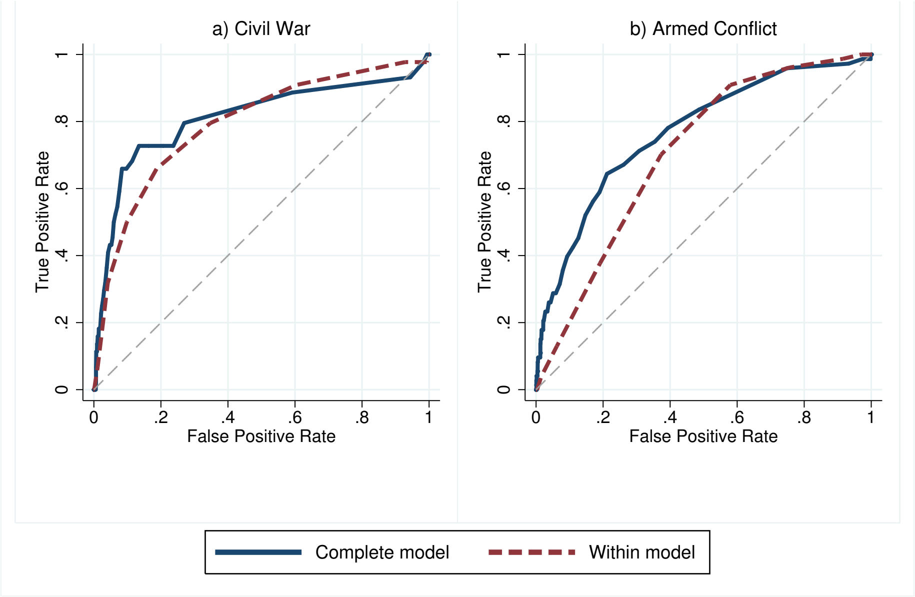
extremely well when predicting the timing of civil war onset out-of-sample. The rainfall model looks like the more robust model but has only little power when forecasting onset out-of-sample.

B Robustness of Main Findings

In this section we discuss the robustness of our main findings. The, perhaps, most important question is whether the parameters of the topic model we have chosen has repercussions for its performance in forecasting. Note that for each topic model we estimate we need to estimate the model every year so that 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 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.

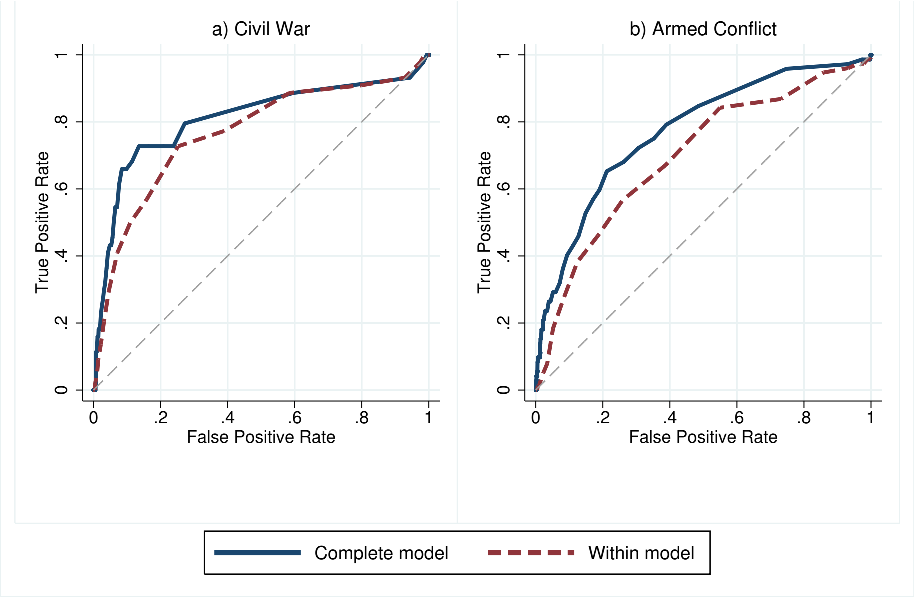
When we move to less topics the performance only suffers slightly. Appendix Figure B.1 shows the performance of a topic model with 5 topics. We have also run robustness checks with 10 topics and 30 topics which both yielded very similar results to 15 topics. Appendix Figure B.2 shows our main results regarding onset for our model with 30 topics.

Figure B.1: ROC Curves for Onset (Five Topics)



Notes: The topic model is based five topics computed using LDA with $\alpha = 10$ and $\beta = 0.01$, which are aggregated at the country/year level. The within model is the complete model net of country fixed effects.

Figure B.2: ROC Curves for Onset (Thirty Topics)

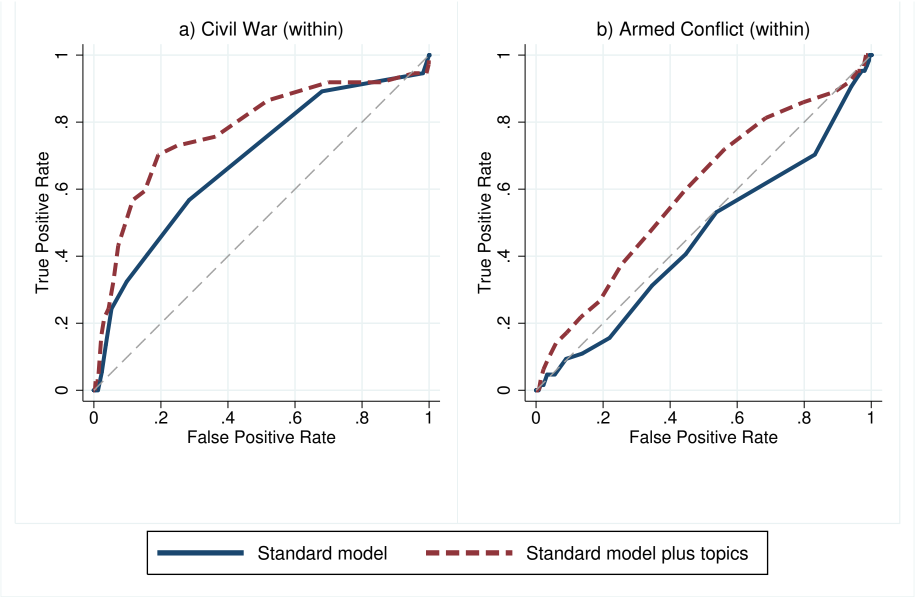


Notes: The topic model is based on thirty topics computed using LDA with $\alpha = 1$ and $\beta = 0.01$, which are aggregated at the country/year level. The within model is the complete model net of country fixed effects.

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 neighbouring countries had an armed conflict. To this we add the 15 topics. In Figures B.3 and B.4 we show that, building on a standard model, the topic shares add forecasting power. For civil war the AUC of the within model increases from 0.68 to 0.78 and for armed conflict from 0.45 to 0.59 due to the inclusion of our topics. In the case of armed conflict onset, the model is now relatively weak, i.e. the joint model performs worse than the news shares alone. This probably reflects the problem of overfitting, which we mentioned previously.

A similar question is whether our news model of incidence can add anything beyond the simple forecasting model of "conflict follows conflict". To test this we add our news shares to a model in which the dummy y_{it} is regressed on a dummy which captures contemporaneous conflict, y_{it-1}. The result is shown in Figure B.5. From this it becomes clear that the lag is a powerful forecast for incidence and that our news model nonetheless adds some forecasting power. We have also experimented with variables that count the years since the last

Figure B.3: ROC Curves for Incidence (Standard Model Plus Topics)



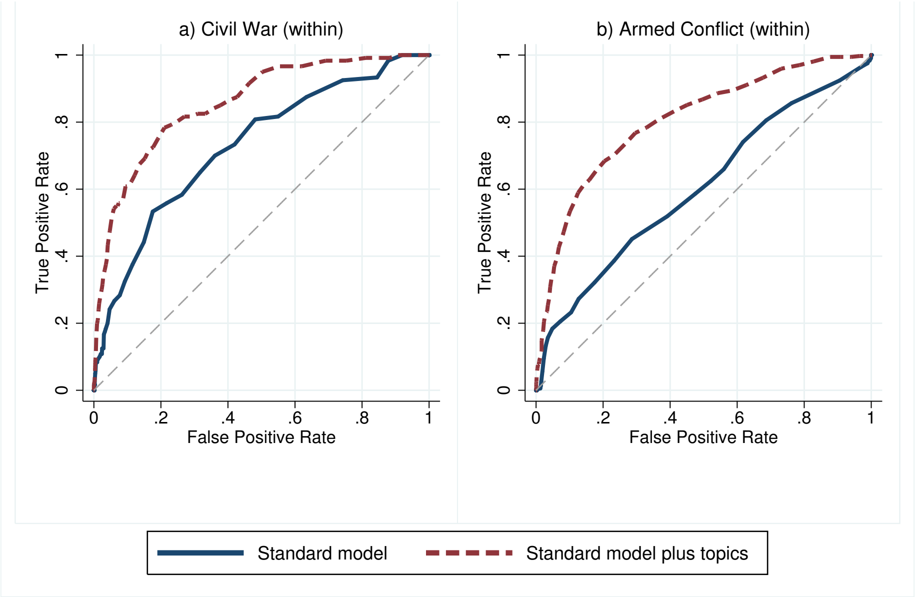
Notes: The standard model includes four political regime dummies from polity IV, infant mortality, the share of the population that is discriminated against, and a dummy that captures whether more than three neighbouring countries had an armed conflict to which we add topics based on 15 topics computed using LDA with $\alpha = 3.1$ and $\beta = 0.01$. The within model is the complete model net of country fixed effects.

conflict in the prediction of onset, and found that onset is not easy to predict even by polynomials of this variable, which means our model provides a lot more additional power in this case.

We also used our model to forecast conflict onset one or two years before it happens. The results, displayed in Figure B.6 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 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. But the between risk is very similarly so that the overall model performs almost equally well.

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) and results remain as can be seen in Figure B.7. We have also used only battle-related deaths occurring in internal wards and have used only the best estimate of those. Our within results are getting slightly stronger under this much more restrictive definition of internal conflict. We also follow Mueller (2016) and define conflict as an armed conflict that exceeded an intensity of 0.08

Figure B.4: ROC Curves for Incidence (Standard Model Plus Topics)

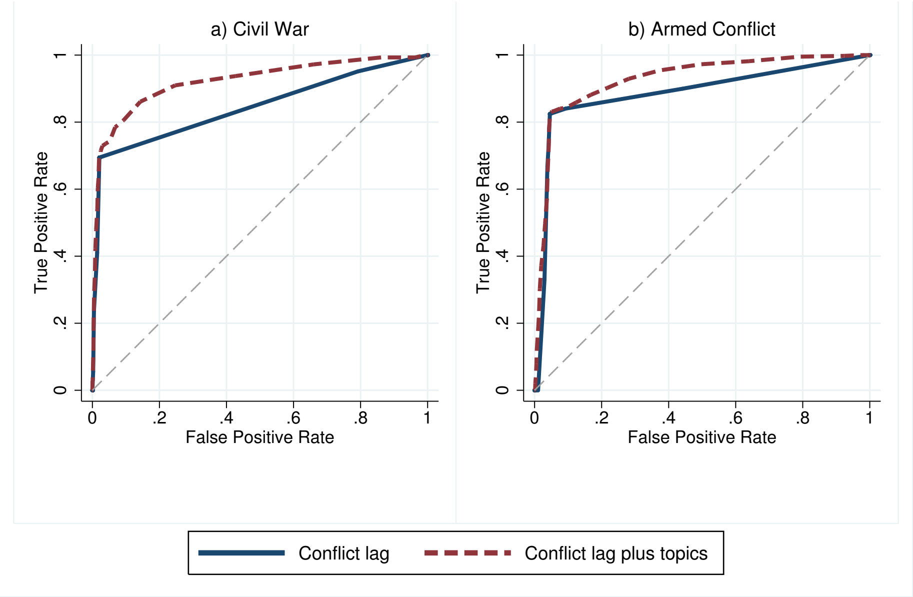


Notes: The standard model includes four political regime dummies from polity IV, infant mortality, the share of the population that is discriminated against, and a dummy that captures whether more than three neighbouring countries had an armed conflict to which we add topics based on 15 topics computed using LDA with $\alpha = 3.1$ and $\beta = 0.01$. The within model is the complete model net of country fixed effects.

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 within and overall as can be seen in B.8. 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 (Figure B.9). In particular, the within and overall model perform very similarly. However, the forecasting 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).

In the main text we compare our model to the Integrated Conflict Early Warning System (ICEWS) as described in Ward et al. (2013), which uses event data generated from news. We see our work as complementary to this work and Chadefaux (2014) for several reasons. First, our method allows us to use all news content to predict civil war without imposing any priors. The use of negative

Figure B.5: ROC Curves for Incidence (Lagged and Topics)



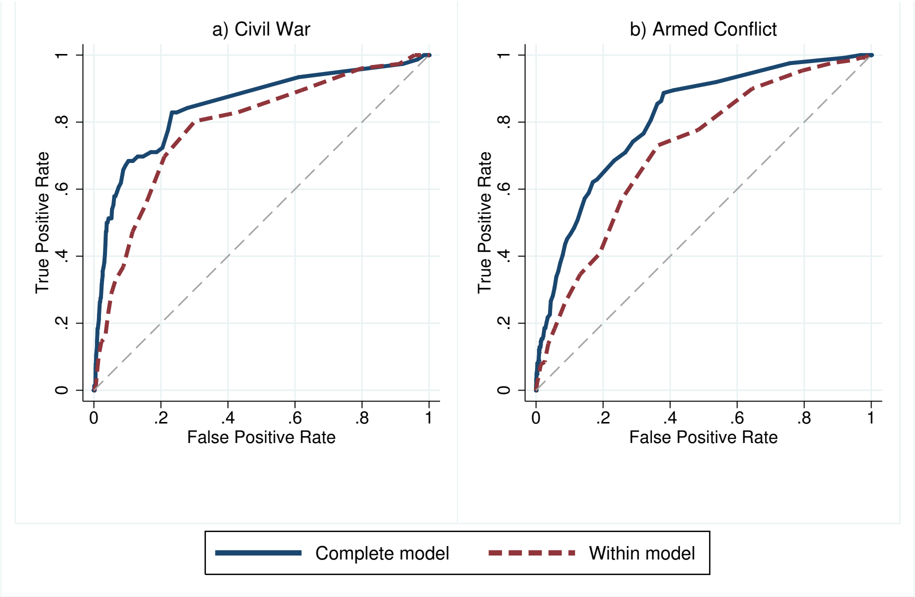
Notes: The conflict lag model includes a lag of conflict to which we add the topic model based on 15 topics computed using LDA with $\alpha = 3.1$ and $\beta = 0.01$.

correlations to elements in the news, for example justice, is a direct result of this. However, this means we need to rely on sources that provide the entire text of articles. Second, the fact that we try to forecast the onset of rare events one and even two years before they happen implies that we need to rely on news sources that are consistently available for decades. We, therefore, relied on three newspapers which gives a little more than 700,000 articles. For comparison, the ICEWS uses more than 30 million news stories, whereas Chadefaux (2014) searches keywords in over 60 million pages of news text.

As we did not have a strong prior regarding which model from Ward et al. (2013) to use, we tried several ones. In the end, the best performance was reached by a model which includes all the events described in the paper but no other variables. In Figure B.10 we compare the forecasting power of the resulting events model and topic model when forecasting onset. The list of event counts, blue solid line, can predict civil war onset fairly well and even performs slightly better than the topic model for small false negative rates. Our news model performs better for most values and in particular when predicting armed conflict onset. When predicting incidence there is no clear dominance of either the topic or the events model, shown in Figure B.10 and Figure B.11.

Moreover, we test our predictive power concerning refugees, an outcome which is 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

Figure B.6: ROC Curves for Onset (Two Years Before Conflict)



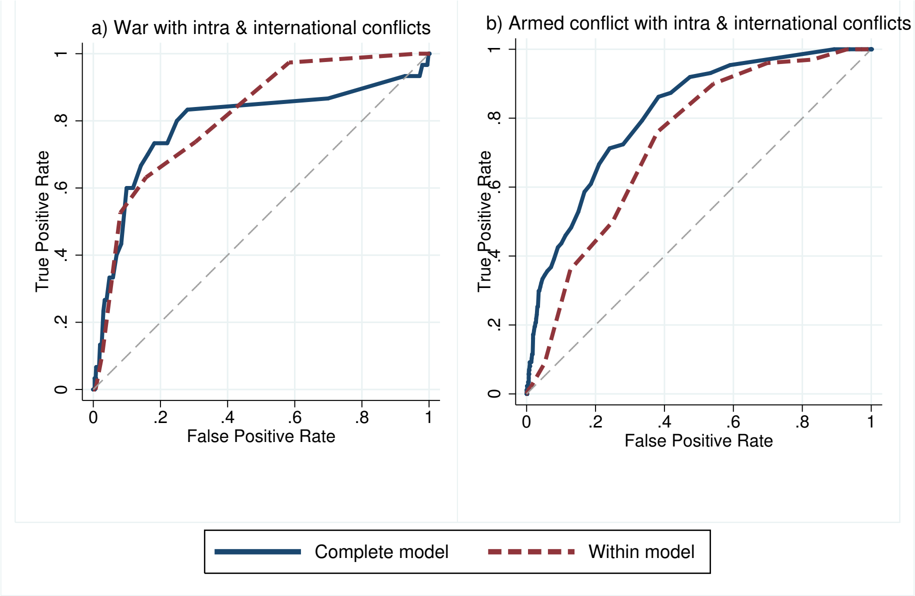
Notes: The curves are based on the results of the topic model computed using LDA with $\alpha = 3.1$ and $\beta = 0.01$, which are aggregated at the country/year level. The within model is the complete model net of country fixed effects.

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 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.

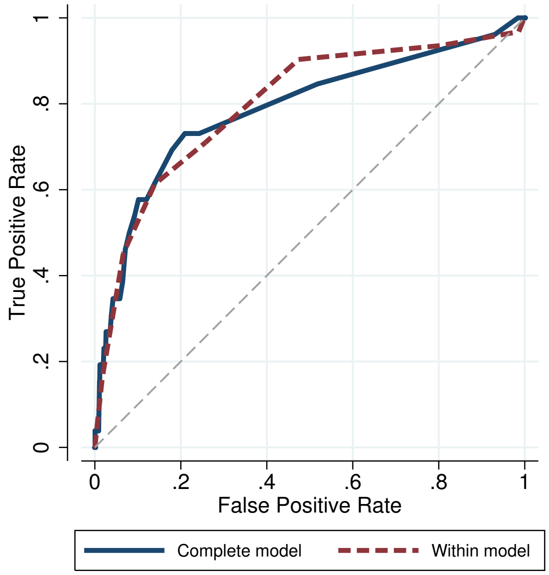
We then use our topic model to test whether we can predict whether a large number of refugees will leave the country in the next year. In panel a) of Figure B.12, we show that the onset of more than 130,000 refugees can be predicted somewhat with our model. In panel b) we predict the onset of 30,000 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 complete 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, which are not mainly collected by news sources (as most of the violence is).

Figure B.7: ROC Curves for Onset (Other Conflicts)



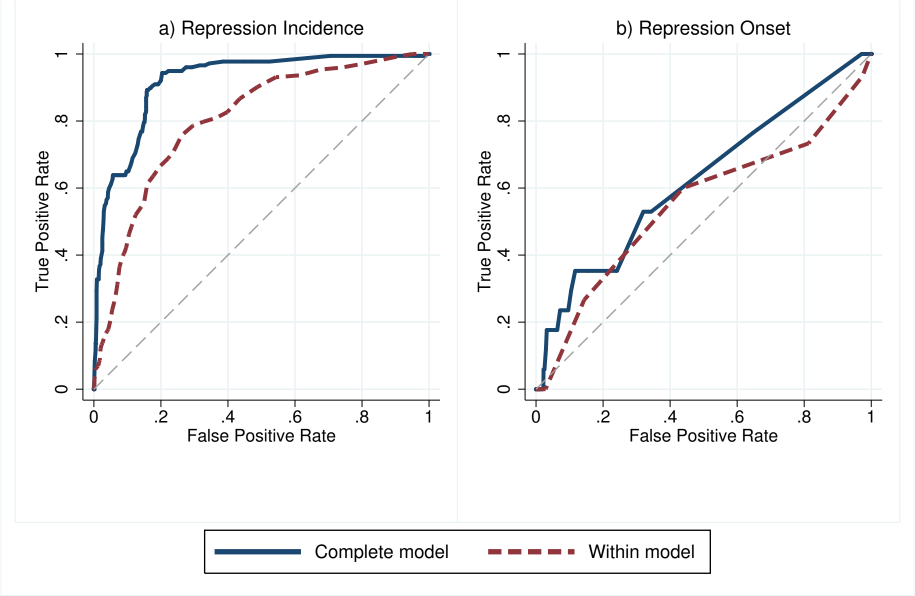
Notes: The curves are based on the results of the topic model computed using LDA with $\alpha = 3.1$ and $\beta = 0.01$, which are aggregated at the country/year level. The within model is the complete model net of country fixed effects.

Figure B.8: ROC Curves for Onset (Conflict in Per Capita Terms)



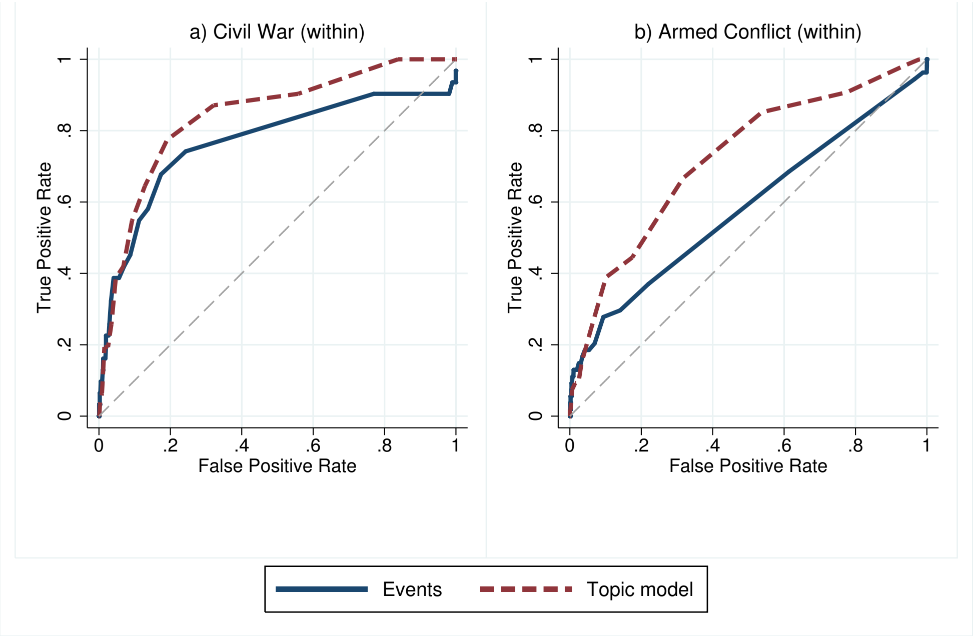
Notes: The curves are based on the results of the topic model computed using LDA with $\alpha = 3.1$ and $\beta = 0.01$, which are aggregated at the country/year level. The within model is the complete model net of country fixed effects.

Figure B.9: ROC Curves for Repression



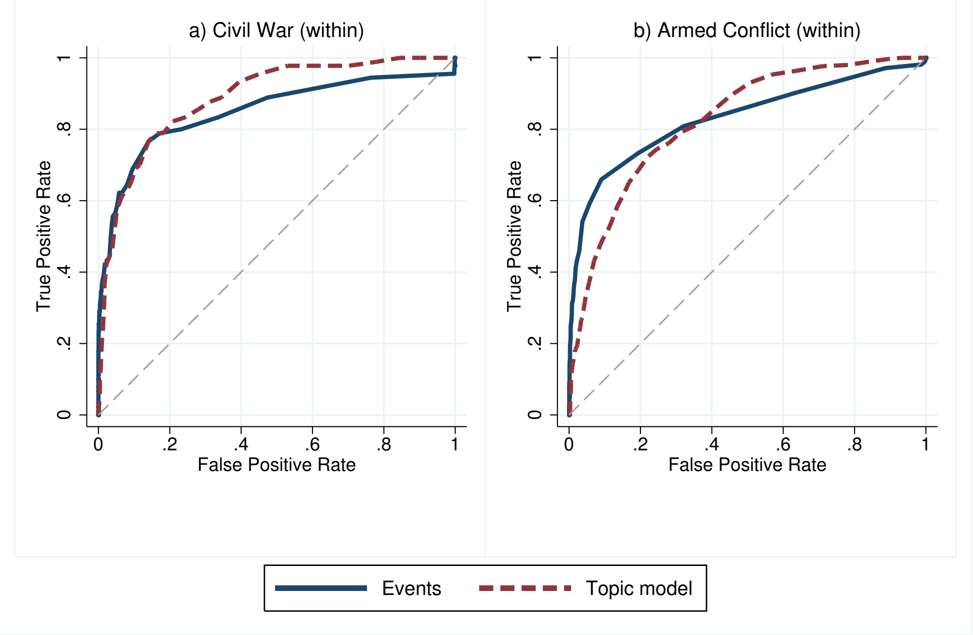
Notes: The curves are based on the results of the topic model computed using LDA with $\alpha = 3.1$ and $\beta = 0.01$, which are aggregated at the country/year level. The within model is the complete model net of country fixed effects.

Figure B.10: ROC Curves for Onset Focusing on Within Variation (Events vs Topic Model)



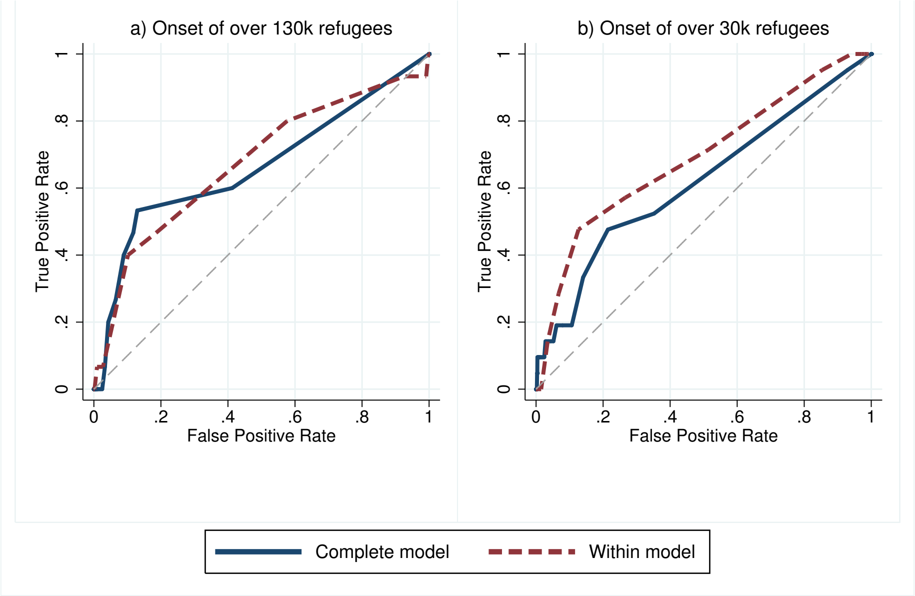
Notes: The event model is based on ICEWS, as described in Ward et al. (2013), which uses event data generated from news. The topic model is based on 15 topics computed using LDA with $\alpha = 3.1$ and $\beta = 0.01$. The within model is the complete model net of country fixed effects.

Figure B.11: ROC Curves for Incidence Focusing on Within Variation (Events vs Topic Model)



Notes: The event model is based on ICEWS, as described in Ward et al. (2013), which uses event data generated from news. The topic model is based on 15 topics computed using LDA with $\alpha = 3.1$ and $\beta = 0.01$. The within model is the complete model net of country fixed effects.

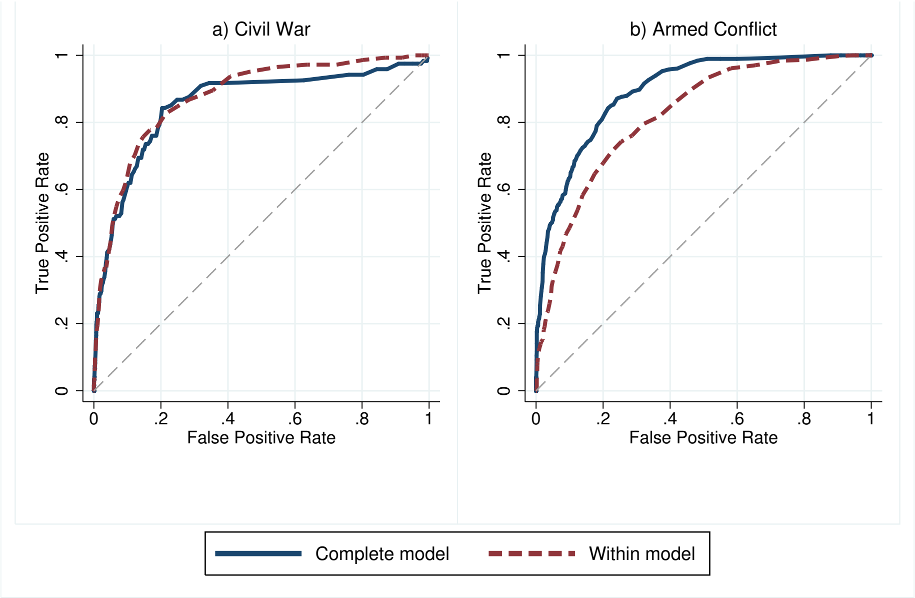
Figure B.12: ROC Curves for Refugee Flows



Notes: The curves are based on the results of the topic model computed using LDA with $\alpha = 3.1$ and $\beta = 0.01$, which are aggregated at the country/year level. The within model is the complete model net of country fixed effects.

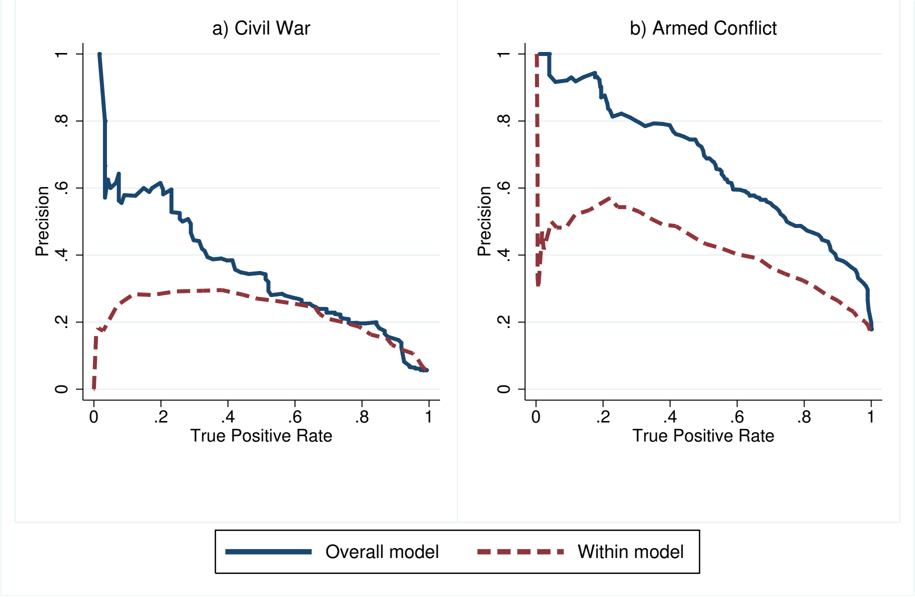
C Additional Figures

Figure C.1: ROC Curves for Incidence



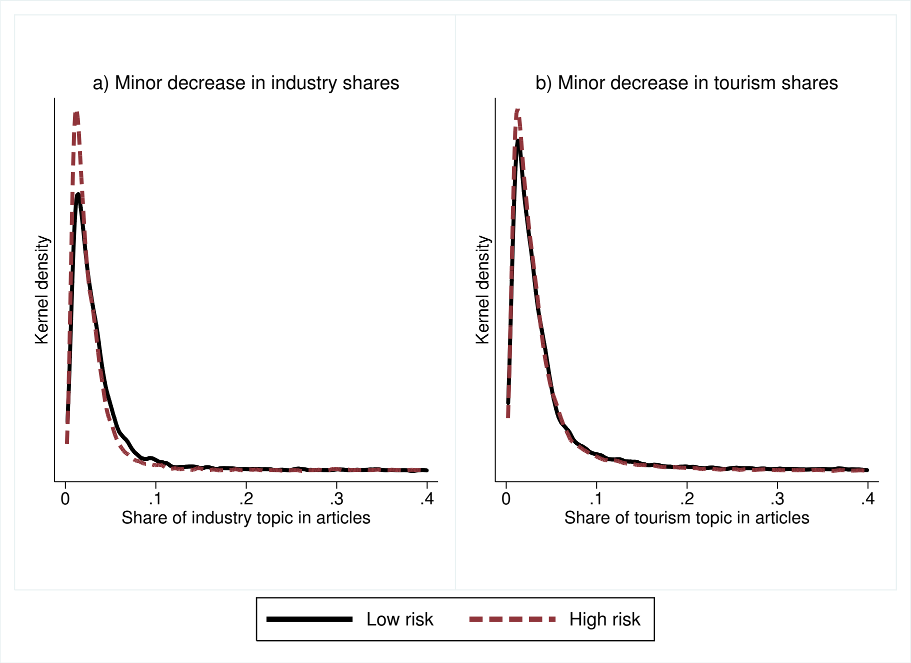
Notes: The topic model is based 15 topics computed using LDA with $\alpha = 3.1$ and $\beta = 0.01$, which are aggregated at the country/year level. The within model is the complete model net of country fixed effects.

Figure C.2: Precision Recall Curves for Incidence



Notes: The curves are based on the results of the topic model computed using LDA with $\alpha = 3.1$ and $\beta = 0.01$, which are aggregated at the country/year level. The within model is the complete model net of country fixed effects.

Figure C.3: Topic Shares of Industry and Tourism 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 probability of onset above five percent.

References

Baker, Scott R, Nicholas Bloom, and Steven J Davis. 2015. “Measuring economic policy uncertainty.” National Bureau of Economic Research.

Banks, Arthur. 2005. “Cross-National Time-Series Data Archive. Jerusalem: 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?” 22–25.

Besley, Timothy, and Torsten Persson. 2011*a*. “The Logic of Political Violence.” *Quarterly Journal of Economics*, 126(3).

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

Blair, Robert A, Christopher Blattman, and Alexandra Hartman. 2014. “Predicting Local Violence.” *Available at SSRN 2497153.*

Blattman, Christopher, and Edward Miguel. 2010. “Civil war.” *Journal of Economic Literature*, 3–57.

Blei, David M, and John D Lafferty. 2006. “Dynamic topic models.” 113–120, ACM.

Blei, David M, Andrew Y Ng, and Michael I Jordan. 2003. “Latent dirichlet allocation.” *the Journal of machine Learning research*, 3: 993–1022.

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.

Brückner, Markus, and Evi Pappa. 2015. “News shocks in the data: Olympic Games and their macroeconomic effects.” *Journal of Money, Credit and Banking*, 47(7): 1339–1367.

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

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

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*, 66–95.

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

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

Gentzkow, Matthew, and Jesse M Shapiro. 2010. "What drives media slant? Evidence from US daily newspapers." *Econometrica*, 78(1): 35–71.

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

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.

Greene, William H. 2003. *Econometric analysis.* Pearson Education India.

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.

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

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.

Kuziemko, Ilyana, and Eric Werker. 2006. "How much is a seat on the Security Council worth? Foreign aid and bribery at the United Nations." *Journal of Political Economy*, 114(5): 905–930.

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.

Mueller, Hannes. 2016. "Growth and Violence: Argument for a Per Capita Measure of Civil War." *Economica*, forthcoming.

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

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.

Ramey, Valerie A. 2011. "Can government purchases stimulate the economy?" *Journal of Economic Literature*, 49(3): 673–685.

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.

Schutte, Sebastian. 2014. "Regions at Risk Predicting Conflict Zones in African Insurgencies." Mimeo.

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.

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

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