UNIVERSITY OF CAMBRIDGE
Faculty of Economics

Cambridge Working Papers in Economics

Cambridge Working Papers in Economics: 2015

THE HARD PROBLEM OF PREDICTION FOR CONFLICT PREVENTION

Hannes Mueller Christopher Rauh

10 March 2020

There is a growing interest in prevention in several policy areas and this provides a strong motivation for an improved integration of machine learning into models of decision making. In this article we propose a framework to tackle conflict prevention. A key problem of conflict forecasting for prevention is that predicting the start of conflict in previously peaceful countries needs to overcome a low baseline risk. To make progress in this hard problem this project combines unsupervised with supervised machine learning. Specifically, the latent Dirichlet allocation (LDA) model is used for feature extraction from 4.1 million newspaper articles and these features are then used in a random forest model to predict conflict. The output of the forecast model is then analyzed in a framework of cost minimization in which excessive intervention costs due to false positives can be traded off against the damages and destruction caused by conflict. News text is able provide a useful forecast for the hard problem even when evaluated in such a cost-benefit framework. The aggregation into topics allows the forecast to rely on subtle signals from news which are positively or negatively related to conflict risk.

The Hard Problem of Prediction for Conflict Prevention

Hannes Mueller and Christopher Rauh*

March 10, 2020

Abstract

There is a growing interest in prevention in several policy areas and this provides a strong motivation for an improved integration of machine learning into models of decision making. In this article we propose a framework to tackle conflict prevention. A key problem of conflict forecasting for prevention is that predicting the start of conflict in previously peaceful countries needs to overcome a low baseline risk. To make progress in this hard problem this project combines unsupervised with supervised machine learning. Specifically, the latent Dirichlet allocation (LDA) model is used for feature extraction from 4.1 million newspaper articles and these features are then used in a random forest model to predict conflict. The output of the forecast model is then analyzed in a framework of cost minimization in which excessive intervention costs due to false positives can be traded off against the damages and destruction caused by conflict. News text is able provide a useful forecast for the hard problem even when evaluated in such a cost-benefit framework. The aggregation into topics allows the forecast to rely on subtle signals from news which are positively or negatively related to conflict risk.

*Hannes Mueller, Institut d'Analisi Economica (CSIC), Barcelona GSE, MOVE and CEPR. Christopher Rauh: University of Cambridge, Trinity College Cambridge. We thank Elena Aguilar, Bruno Conte Leite, Lavinia Piemontese and Alex Angelini for excellent research assistance. We thank the discussants and seminar and conference audiences at the Paris School of Economics, the BdE, University of Geneva, ISA Toronto, INFER conference, the Barcelona GSE, Tokio University, Osaka University, GRIPS, Uppsala University, Quebec Political Economy Conference, University of Montreal, SAEe Barcelona, the German Foreign Office, Geneva University, Warwick University, the Montreal CIREQ workshop on the political economy of development and the Barcelona workshops on conflict prediction. Mueller acknowledges financial support from the Ayudas Fundación BBVA. The authors declare that they have no competing financial interests. All errors are ours.

1 Introduction

Civil wars are a serious humanitarian and economic problem. According to data from the United Nations Refugee Agency (UNHCR) in 2017 on average 44,400 people around the world had been forced from home every day, the large majority by armed conflict. Once started, a small armed conflict can quickly escalate and lead to repeated cycles of violence that have the potential to ruin society for a generation. International organizations like the UN, the World Bank, the IMF and the OECD have therefore all identified fragility as a key factor for long-term development. Most recently, this has led to calls for more resources and institutional reforms aimed at preventing civil wars (United Nations and World Bank 2017 and OECD 2018). Most explicitly this general trend was expressed by the former President of the World Bank Jim Yong Kim (World Bank 2017*b*): *"[...], we need to do more early on to ensure that development programs and policies are focused on successful prevention."*

However, developing optimal policy for such a prevention problem is difficult because it requires the policymaker to take actions in the present directed at preventing an uncertain future outcome. Be it economic crisis, climate change, armed conflict, crime or a flew epidemics; what we require is not only knowledge about the causal impact of a policy given a state of the world but also the best possible prediction of this state of the world (Kleinberg et al. 2015). In other words, we need to combine prediction with policy aimed at causal factors. The question of whether prevention is economically viable is then not only a question of how effective interventions are but how well they can be targeted. If it is impossible to predict where conflict will break out then policies aiming at prevention would need to be implemented very broadly - perhaps making them inefficient. The conflict prevention problem therefore raises two questions: What is the precision that can be reached in a cross-country ranking out-of-sample, i.e. without knowing the future? Is this precision high enough to make prevention economically viable?

This article aims to contribute to the conflict prevention problem in two ways. First, we

provide a forecast directly targeted at the idea of preventing conflict before it breaks out. Our forecast uses data on conflict histories together with a corpus of over 4 million newspaper articles in a combination of unsupervised and supervised machine learning to predict conflict in countries. This methodology helps foresee outbreaks in countries that would otherwise be off the radar. Text has the huge advantage of being available in real time and can therefore be used to implement an early warning system.[1]

Second, we introduce a conceptual framework that could allow policymakers to integrate knowledge on the effectiveness and costs of preventive actions with knowledge on the precision of the forecast. Our framework builds on the standard classification toolbox in machine learning. In this framework instances are ranked and interventions are targeted at the cases with the highest risk. We use this framework to show that a key trade off is between the effectiveness of the policy and the precision of the forecast and that a key margin of adjustment in prevention is the risk cutoff at which policymakers would intervene. Low average policy effectiveness and low average precision both lead to less interventions and a focus on few cases in which the forecast indicates the highest risk.

A crucial problem for prevention is often that it faces a low baseline risk, i.e. heavily imbalanced classes. In conflict prevention the problem is particularly severe as policymakers face the so-called conflict trap which is well-known in the conflict literature (Collier and Sambanis 2002). Countries get stuck in repeated cycles of violence and, as a consequence, conflict history is an extremely powerful predictor of risk. Forecasts which, explicitly or implicitly, rely on conflict history are then not useful for the declared goal of conflict prevention because they cannot predict when countries are at the verge of falling into the conflict trap. If we want to prevent conflict, we need to pay more attention to previously peaceful countries, which means we need to predict cases with a baseline risk of below 1%. We call this the *hard problem* of conflict prediction. The hard problem is typically not explicitly taken into account when evaluating

[1] To illustrate this we will provide such a forecast publicly starting in summer 2020.

forecasting models but is of first-order importance for prevention.[2]

We train and test a prediction model in sequential non-overlapping samples which allows us to evaluate the out-of-sample performance and at the same time mimics the problem that policymakers face. In the evaluation of our forecasts we focus particularly on conflict outbreaks which are hard to predict, i.e. in countries that were previously peaceful. We find that random forest models perform extremely well in this task and provide substantial benefits over other models. The reason is that the tree structure allows the model to adapt to the hard problem by placing indicators of conflict history high up in the tree and using topics at the bottom nodes. We find that topics which are not directly related to violence and negatively associated with risk, like sports, business, justice, economics or trade, receive increasing importance when predicting hard onsets.

There is a long history in prediction in economics, and for macroeconomic variables like inflation or economic growth it has long been a goal of academic research. But it is also becoming more common for other outcomes as well.[3] However, for conflict the economics literature has mostly focused on understanding causal mechanisms.[4] As a consequence, the literature has made huge strides in understanding the causes of conflict.[5] However, these efforts are often not effective for forecasting. The reason is that causal mechanisms which can be identified need not be good predictors of conflict (Ward, Greenhill and Bakke 2010; Mueller and Rauh 2018) and the methodology used to estimate parameters does not reveal whether a model is overfit to the data, i.e. cross-validation is typically not used. However, the policy context for conflict

[2]Our paper is complementary to Bazzi et al. (2019) who focus on prediction within countries with a known, recent history of violence. Our method aims to provide a forecast for cases without a recent history but at the country level.

[3]For an overview over the more classic literature see Timmermann (2006) and Elliott and Timmermann (2008, 2013). For recent efforts see Böhme, Gröger and Stöhr (forthcoming), Giglio, Kelly and Pruitt (2016), Costinot, Donaldson and Smith (2016). For an overview over prediction efforts and methodology see Mullainathan and Spiess (2017) and Athey and Imbens (2019). In other applications, machine-learning predictions are used to measure rather than forecast outcomes, such as poverty (Jean et al. 2016; Blumenstock, Cadamuro and On 2015).

[4]Two exceptions are Celiku and Kraay (2017) and Bazzi et al. (2019).

[5]For an overview of the earlier literature see Blattman and Miguel (2010). For recent contributions in economics see Besley and Persson (2011); Esteban, Mayoral and Ray (2012); Dube and Vargas (2013); Bazzi and Blattman (2014); Burke, Hsiang and Miguel (2015); Michalopoulos and Papaioannou (2016); Berman et al. (2017).

studies is clearly one where forecasts are valued and valuable and in political science this has led to a large, sophisticated literature.[6] Here, we have an approach in mind in which automated forecasts, expert opinion and the growing evidence on optimal policies go into a cost analysis framework to provide a benchmark for budgeting decisions (and allocation of attention) across countries.[7] We present an extremely simple version of such an intervention framework to show that our forecast would be useful even if policy interventions are costly, imperfect and targeted with the precision of the actual forecasting model.

An advantage of approaching the prediction policy problem within a machine learning classification framework is that we can give different cost weights to false positives (high risk that never escalates into conflict) and false negatives (conflict outbreaks that were not anticipated) and derive optimal forecasts based on these cost weights. This is relevant for economics far beyond the field of conflict research. Svensson (2017), for example, argues that the possibility of a financial crisis needs to be taken into account when conducting monetary policy but proposes a simple logit regression to forecast the occurrence of a crisis.

An additional benefit of the forecasts we provide is that they provide both measures of risk and measures of forecast errors. In other areas of economics this has already produced important insights.[8] In our application, the side-product of the forecast is a quarterly conflict risk measure for nearly 200 countries for the period 2000Q1 to 2019Q4. The fact that our model produces useful forecasts for low baseline risks means we are able to provide political risk estimates for all these countries. Apart from guiding better models of decision-making under risk and uncertainty, this output of the forecasting framework could also provide propensity scores for cross-country comparisons or could provide useful data on unforeseen outbreaks of conflict or help study the role of stabilizing factors.

[6] For summaries of the political science literature see Hegre et al. (2017).

[7] In ranking a large number of diverse countries according to their risk is where we see a significant value added of automated prediction over expert opinion.

[8] Take, for example, Blanchard and Leigh (2013), Jurado, Ludvigson and Ng (2015), Rossi and Sekhposyan (2015) and Tanaka et al. (2019).

There is a growing interest in the use of text to generate data, i.e. feature extraction.[9] Baker, Bloom and Davis (2016) and Ahir, Bloom and Furceri (2018) use relative frequencies of predetermined keywords positively related to economic uncertainty in order to provide a measure uncertainty for the US and 143 countries, respectively.[10] For our feature extraction we rely on the full text but reduce the dimensionality through the Latent Dirichlet Allocation (LDA) or topic model (Blei, Ng and Jordan 2003). Topic models provide an extremely useful way to analyze text because they do not rely on strong priors regarding which part of the text will be useful. In addition, the LDA is in itself a reasonable statistical model of writing and we show that it is able to reveal useful latent semantic structure in our newstext corpus.[11] We find both positive and negative relationships with conflict risk in the topics. And, perhaps surprisingly, a lot of the predictive performance comes from some topics reducing their share before conflict breaks out. While the interpretation of this fact is difficult, the LDA provides at least the possibility to understand the factors that provide predictive power.

Our goal is to maximize forecasting performance and so we cross-validate the model to find the optimal depth and number of trees in the random forest model, i.e. we regularize the model conservatively out-of-sample.[12] Our regularization suggests that a model that uses a handful of variables is optimal. The reason lies in the so-called “small n large p" problem we face when forecasting macro events like conflict. The number of cases is limited and so the forecasting problem cannot be simply solved through a sophisticated supervised machine learning model. A way forward in these situations is to use theory to build priors regarding the variables and model to use. An alternative, which we follow here, is to use unsupervised learning for dimensionality reduction. This method has a long tradition in macroeconomics (Stock and Watson 2006) but also outside the social sciences, in particular in application with few positive events to train the

[9] See Gentzkow, Kelly and Taddy (2019) for an overview.

[10] In a similar fashion Baker et al. (2019) capture equity market volatility.

[11] This feature mirrors findings in Rauh (2019) and Larsen and Thorsrud (2019) who forecast economic activity, and Hansen, McMahon and Prat (2017) who study the effect of increased transparency on debate in central banks.

[12] Systematic regularization is not common practice even in research which clearly aims to provide predictions.

model such as medicine (Mwangi, Tian and Soares 2014).

In summary, this project advances on several fronts. First, we explicitly take conflict history into account. This treatment of history allows us to evaluate performance for cases without a violent past. Second, we use a new full-text archive of 4.1 million newspaper articles dense enough to summarize topics at the quarterly level for nearly 200 countries. This, in turn, allows us to combine feature extraction using unsupervised learning with supervised machine learning. We show that, over time, the supervised model slightly improves its forecasting performance. This suggests that the supervised learning is actually benefitting from generalizable, subtle signals contained in the extracted features. Third, we propose integrating the forecast data into a cost function to evaluate optimal policy. Such a cost framework could be used to develop models of decision-making and evaluate the gains from prevention in other applications as well. Finally, we rely on innovations in the estimation of the topic model (Blei and Lafferty 2006) to solve the computational challenges implied by the need to re-estimate the topics from millions of articles for every quarter. This makes our method particularly useful for actual policy applications that rely on timely risk updates using vast amounts of text.

In what follows we first explain the importance of the so-called conflict trap for conflict forecasting. We also show that countries seem to transition in an out of the trap so that treating it as a characteristic of the country which is fixed, as in Mueller and Rauh (2018), is not doing justice to the dynamic nature of the trap. We then present our forecasting model and the way we evaluate our forecasts in a rolling out-of-sample test. In Section 4 we present the results of our prediction exercise and in section Section 5 we integrate this forecast into a cost-minimization problem. In Section 7 we present case studies of our risk measure before we conclude.

2 The Hard Problem of Conflict Prediction

The most encompassing conflict data is provided by the UCDP Georeferenced Event Dataset (Sundberg and Melander 2013; Croicu and Sundberg 2017). We include all battle-related deaths in this dataset and collapse the micro data at the country/quarter level. We focus on the quarterly level to have more cases to train but also provide forecasts a year ahead. The data offers three types of conflict which we all merge together. This implies that we mix terror attacks and more standard, two-sided violence. An important question arises due to the fact that zeros are not coded in the data. We allocate a zero to all country/quarters in which the country was independent and where data from GED is available. The only exception is Syria which is not covered in all years by the GED downloadable data.

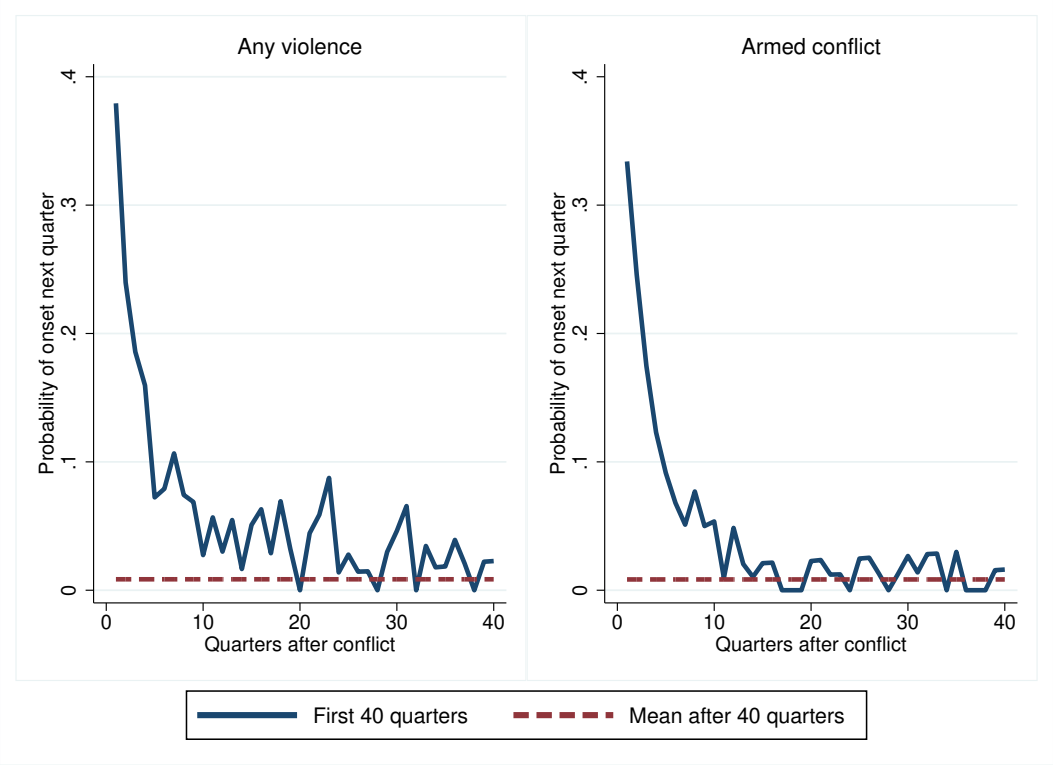
The conflict literature often relies on absolute fatality counts to define conflict. However, these are typically defined at the yearly level and it is not obvious how to translate these definitions to the quarterly level. In addition, onsets of intense violence are relatively easy to predict with ongoing less-intense violence. We therefore use two definitions of conflict. The first takes a quarter with one or more fatalities as conflict (any violence), and the second assumes that conflict is a quarter with at least 50 fatalities (armed conflict). [13] We only consider onset, i.e. only the quarter conflict breaks out. Subsequent quarters of conflict are set to missing. This is important as predicting outbreaks is much harder than predicting conflict. In our data we have 753 onsets of any violence and 453 onsets of armed conflict.

The hard problem can be understood through a simple figure which illustrates the extremely high risk of onset post-conflict. In Figure 1 we plot the likelihood in our sample that a conflict breaks out for the quarters after the end of the previous conflict episode for both our definitions of conflict - any violence and armed conflict. Both figures show that the risk of a renewed onset of conflict is higher than 30 percent right after conflict. Conflict risk falls continuously thereafter but remains substantial in the years following conflict. This is what the conflict literature has

[13] In the Appendix we also look at large scale conflicts with at least 500 battle deaths.

dubbed the conflict trap. Countries get caught in cycles of repeating violence.

Fig. 1: Likelihood of Conflict Relapse



Note: Figure shows the sample likelihood of conflict relapse after violence ended (at 0) conditional on remaining in peace. Any violence is a quarter with any fatalities. Armed conflict is a quarter with more than 50 fatalities.

However, outside the ten year period the baseline risk of conflict is around 1%. In Figure 1 this is illustrated by the red dashed line. In other words, inside the conflict trap onset is relatively likely and easy to forecast using conflict history. Outside the trap onset is very unlikely and hard to forecast. Providing risk estimates for countries that are coming out of conflict therefore provides little added value beyond what most policymakers would already understand intuitively. Good predictions are then particularly hard but also particularly useful outside the conflict trap. The problem of forecasting conflict for cases outside the ten year period is what we call the *hard problem*. We explicitly evaluate the forecast performance of our model for these cases.

Of course, it might be tempting to instead focus on cases that are easier to predict - and indeed this is what the current system of peacekeeping is geared to do. But the dynamics of

the conflict trap imply that avoiding destabilization has huge benefits in the long run. This is because the expected future violence levels in conflict or post-conflict are suprisingly similar but differ dramatically to pre-conflict peace. When a country experiences an outbreak of violence in a hard-to-predict scenario it falls into the trap and therefore its future expected violence changes dramatically. This is much less true for an outbreak of violence post-conflict as countries will tend to switch back and forth between conflict and peace post-conflict. Prevention in hard problem cases will then have considerable dynamic payoffs.

As the quote from the introduction makes clear, prevention is also of interest for the international community. All big international organisations treat fragility and conflict risk as key problems. The need to forecast hard cases follows directly from the need to "do more early on" . Once conflict has broken out, prevention has failed and so, by definition, conflict prevention requires a risk evaluation for hard problem cases, i.e. cases without a recent conflict history.

3 Simulating the Policy Problem

We propose the use of machine learning in two steps to bring large quantities of news text to forecasting conflict and test out-of-sample performance. We first use a dynamic topic model (Blei and Lafferty 2006), which is an unsupervised method for feature extraction. The advantage of this method is that it allows us to reduce the dimensionality of text from counts over several hundred thousand terms to a handful of topics without taking a decision regarding which part of the text is most useful for forecasting conflict.

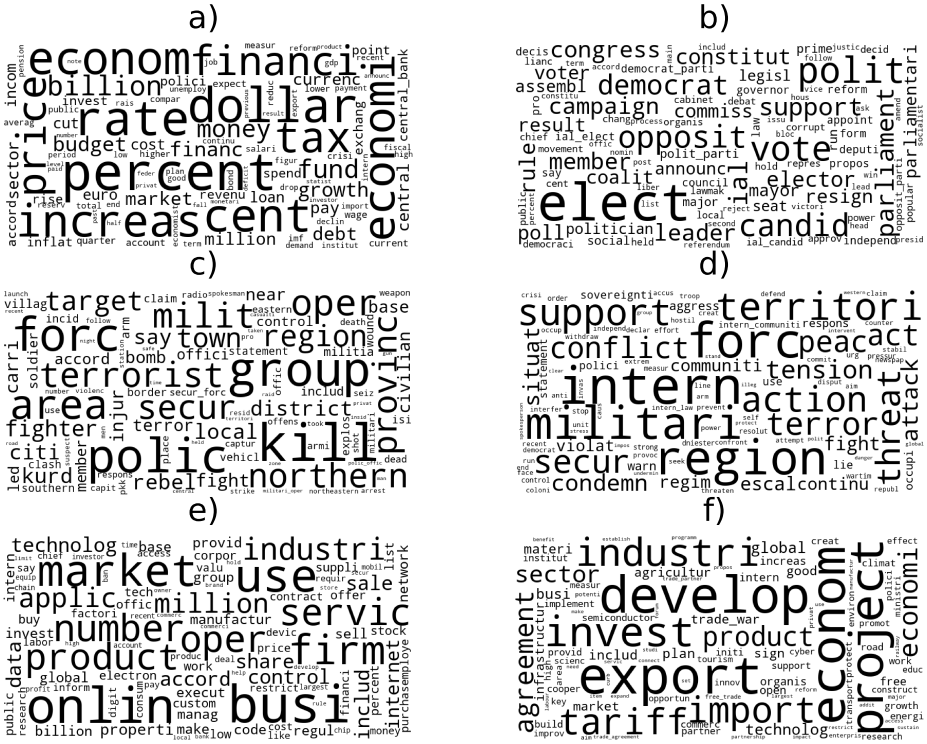
As a basis of our method we use a new unique corpus of over 4 million documents from three newspapers (New York Times, Washington Post and the Economist) and two news aggregators (BBC Monitor and LatinNews). All sources except for LatinNews are downloaded from LexisNexis. A text is downloaded if a country or capital name appears in the title or lead paragraph. The resulting data is described in detail in the Online Appendix. Using the dynamic

topic model we derive topic models with $K = 5$, 10, 15, 30 and 50 topics. The reason we choose relatively few topics is to avoid topics adapting to particularly newsworthy cases of conflict, regions or countries. Topic models between 5 and 30 topics tend to contain topics that can be attributed to generic content like politics or economics, whereas 50 topics are more likely to be specific to certain situations or countries.

Figure 2 shows word clouds for six out of 30 topics estimated on the 2019Q4 sample depicting the most likely terms proportional to their importance in size. The topic model does not label these probability distributions but, as the topic model is based on a reasonable model of writing it uncovers distributions which are easy to interpret or at least distinguish. This is a real strength of using a (statistical) model of writing to extract features from text.

The first topic in Panel a) is what we describe as the economics topic. It features terms such as "econom" ,"dollar" , and "growth" prominently. Panel b) displays a topic which features mostly terms related to politics. Similarly, Panels c) and d) present other topics related to violence and military (infantry), with terms such as "terrorist" and "kill" being keywords, respectively. In Panel e) we see tokens such as "firm" , "manag" , and "busi" (business), while Panel f) exhibits "tariff" , "import" , and "export" (trade).

Fig. 2: Word Clouds of Topics



Note: The word clouds represent the most likely terms of 6 out of 30 topics estimated using all text until 2019Q4. The size of a token is proportional to the importance of the token within the topic. The location conveys no information. Panel a) we consider the Economics topic, b) Politics, c) Terror, d) Violence, e) Business, and f) Trade.

With the estimated topic model we then calculate the share of topics for all countries in each quarter between $1989Q1$ and T. We then use these shares, together with a set of dummies which capture the post-conflict risk, in a random forest model to forecast conflict out of sample. In this step we take the perspective of a policymaker who observes all available text and conflict until period T and has to make a forecast for period $T+1$. We summarize the K topic shares and the log of the word count in the vector, θ_{it}. We then train a forecasting model with all data available in T using the model

$$y_{it+1} = F_T(\mathbf{h}_{it}, \theta_{it}) \tag{1}$$

where y_{it+1} is the *onset* of conflict in quarter $t+1$, $\mathbf{h}_{it}$is a vector of dummies capturing post-conflict dynamics and lower levels of violence. The role of machine learning in this step is to discipline the regularization of the model $F_T(\mathbf{h}_{it}, \theta_{it})$. We fix the hyperparameters in the first sample (1989Q1-2000Q1) through cross-validation. The newest conflicts that break out in the training sample are those that break out in T. Note that this implies that, during training, we only use data for $\mathbf{h}_{it}$ and θ_{it} available until $T-1$. With the resulting model we then produce predicted out of sample values

$$\hat{y}_{iT+1} = F_T(\mathbf{h}_{iT}, \theta_{iT}).$$

which we compare to the true values y_{iT+1}.

We then update our topic model with the news written in the next quarter, add the new information on conflicts, retrain the prediction model, and predict the probabilities of outbreaks in the following quarter. For testing, we thereby produce sequential out-of-sample forecasts, $\hat{y}_{iT+1}$, for $T+1 = 2000Q2, 2000Q3, ..., 2018Q4$. We then compare these forecasts with the actual realizations y_{iT+1}. In this way we get a realistic evaluation of what is possible in terms of forecasting power in actual applications as we never use any data for testing which has been used for training purposes.

To generate $F_T(.)$ we tested predictions from k-nearest neighbor, adaptive boosting, ran-

dom forests, neural network, logit lasso regression, and ensembles of all previously mentioned models. The hyperparameters are chosen by maximizing the AUC through cross-validation within the sample up to $2000Q1$. We found that the random forest model provides the best forecast overall. This is important as it indicates that a method with built-in safeguards against overfitting performs best in our out-of-sample test.

4 Solving the Hard Problem

4.1 The Standard Forecasting View

In order to understand the performance of forecasting results we need to compare the continuous forecast values, $\hat{y}_{iT+1}$ to the actual discrete realizations y_{iT+1}. In the forecasting literature this is done by picking a cutoff c and discretize using the condition

$$\hat{y}_{iT+1} > c.$$

An increase of the cutoff c increases the number of predicted 0s (negatives) which means that there are more false negatives (not predicted outbreaks) and true negatives (peace without warnings). A lower cutoff c increases the number of 1s (positives) which means that there are more false positives (false alarms) and more true positives (correctly predicted outbreaks). We will show that the optimal choice of the cutoff c provides an interesting view on the distribution of $\hat{y}_{iT+1}$. But for now we stick to general ways of displaying the trade-off between false positives and false negatives.

Receiver operator characteristic (ROC) curves are one way to display this trade-off. On the y-axis they report the true positive rate (TPR) as a function of the cutoff c

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

which is the share of all actual positives that are detected by the classifier. More false negatives will lower the TPR so that a TPR of 1 is the best possible value. On the x-axis ROC curves

report the false positive rate (FPR)

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

which is the share of all actual negatives that are detected wrongly by the classifier. More false positives will increase the FPR and the best possible FPR is therefore 0.

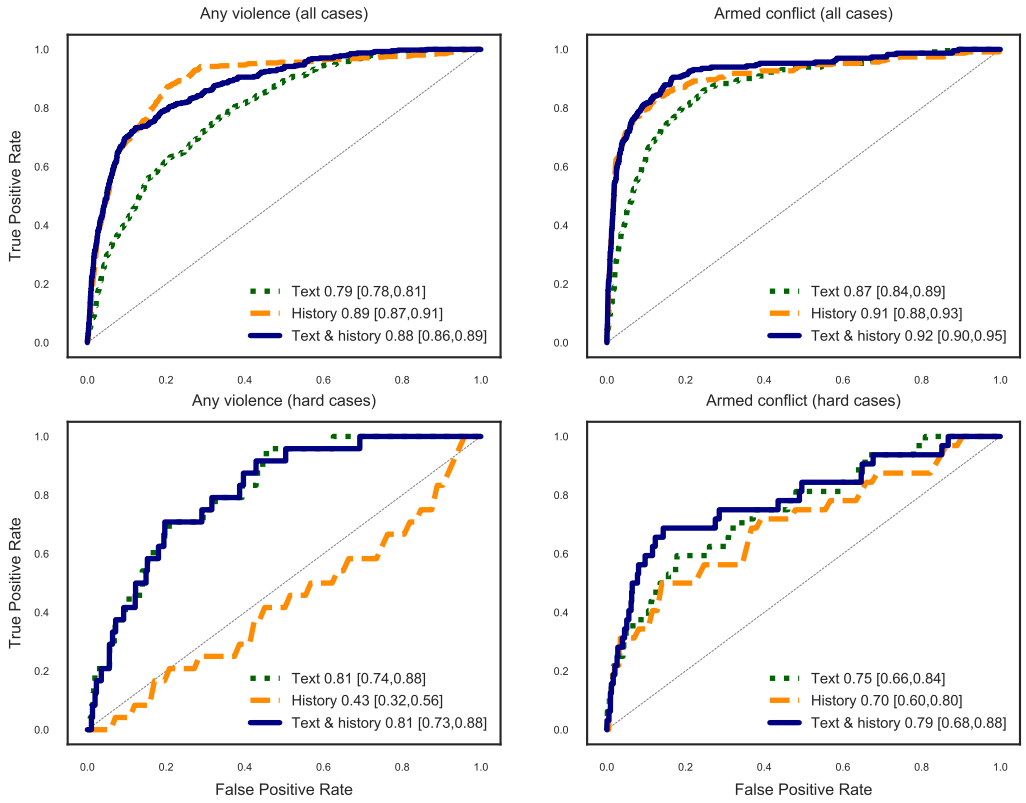
A high cutoff c represents a very conservative forecast which warns only of few onsets, tends to get it right (low FPR) but misses a lot of actual onsets (low TPR). Lowering the cutoff c will typically increase both the FPR and the TPR. If the TPR increases by more than the FPR with falling c the ROC rises above the 45 degree line which indicates a better-than-random forecast. A policymaker who is very afraid of not being able to intervene before most conflicts will choose a low cutoff leading to a high TPR and a high FPR.

The area under the curve (AUC) of the ROC curve is often used to evaluate forecasting models. This makes sense in settings when the relative costs of false negatives and false positives is not known. The ROC, and hence the AUC, expresses a possibility frontier of the forecasting model. In our forecasting model and interpretation of the models we therefore stick to this measure. However, in Section 4.2 we turn to a much more specific interpretation of the policy task.

Figure 3 shows receiver operating characteristics (ROC) curves for the two cutoffs we analyze, i.e. at least 1 (any violence) and 50 (armed conflict) battle deaths, respectively.[14] In each panel, we show the forecasting performance of three forecasting models: (i) A model using just topics and word counts as predictors, which is labeled as *text*, (ii) a model using only information about present or previous violence, which is labeled as *conflict info*, and (iii) a model that draws from both. More specifically, conflict info contains four dummies capturing conflict history: an indicator whether there was conflict (i) last quarter, (ii) 2-4 quarters ago, (iii) 2-5 years ago, or (iv) 6-10 years ago. Moreover, for armed conflict the set of predictors contains a dummy indicating whether any violence is present.

[14] In the Appendix we also present results for a large cut-off of 500 battle deaths.

Fig. 3: ROC Curves of Forecasting Any Violence (left) and Armed Conflict (right)



Note: The prediction method is a random forest. 'Text' contains 30 topics and token counts and 'history' contains 4 dummies indicating the first quarter, quarters 2-4, years 2-5 and years 6-10 after the last conflict and a dummy for the presence of any violence when predicting armed conflict. Top and bottom ROC curves are alternative evaluations of the same forecasting model. Hard cases (bottom) are defined as not having had armed conflict in 10 years. The bottom ROC curves are evaluated only for those cases. The numbers in the legends represent the respective area under curve with bootstrapped 95% confidence intervals in square brackets.

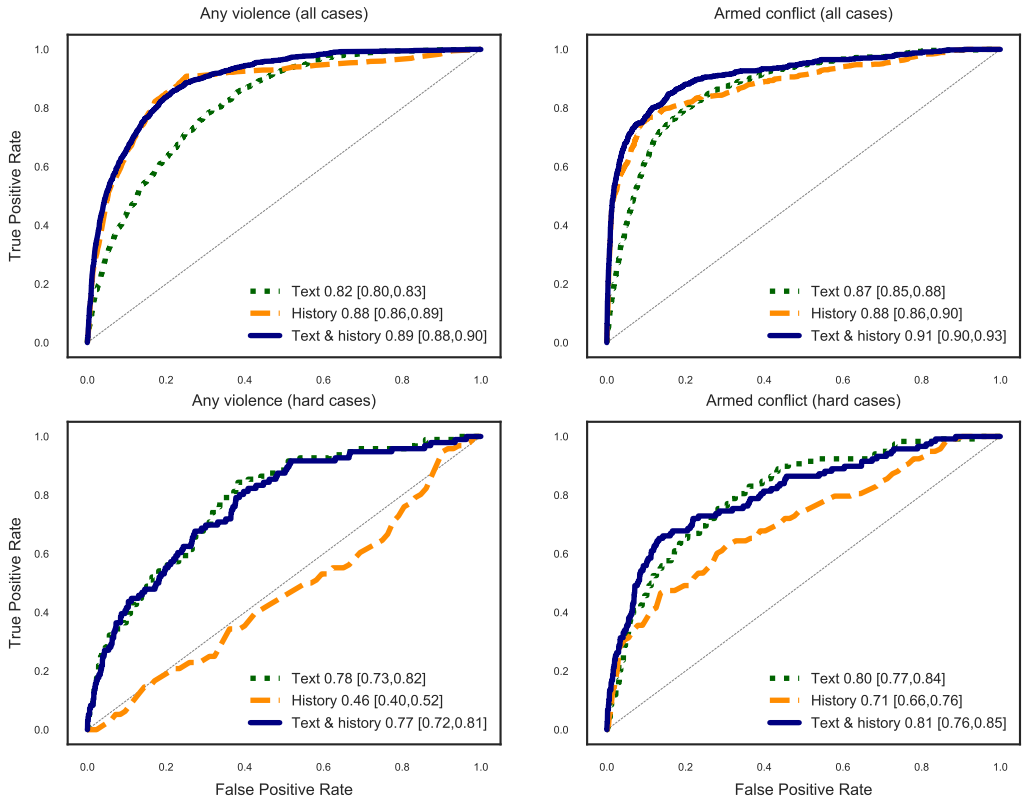
On the top of Figure 3 we see the overall performance of all three sets of predictors when forecasting any violence (left) and armed conflict (right). Text alone (green dotted line) provides some forecasting power and this forecast is comparable to what is common in the literature when predicting at the quarterly level. But it is clear that the information about conflict history (orange dashed line), a simple model of four or five dummies, dominates the text forecast. The combined model reaches an AUC of 0.88 for any violence and an AUC of 0.92 for armed conflict. We generated bootstrapped confidence intervals for the AUC but, even at the

lower bound of these confidence intervals, the AUCs of the combined model are relatively high. Importantly, the AUC could not be improved significantly by adding more variables.

Next we evaluate our forecasting models on the hard problems (bottom). Here we use the same predicted values from our model but only evaluate its performance on the cases without a conflict history. It is important to note that this is information that a policymaker would have and would therefore be able to condition on when evaluating a prediction. Trivially, the conflict information model (orange dashed line) now fails completely to provide a useful forecast, i.e. it is not significantly better than random. Text, however, still provides useful forecasting power and the combined model (blue solid line) now draws its power from the topics. The ability of text features to provide a forecast in cases which experienced no violence for at least a decade is remarkable as these include instabilities like the beginning of terror campaigns, insurgencies or revolutions.

For many applications in prevention a prediction of one year ahead is more desirable. In Figure 4 we therefore provide evaluations of a prediction model that considers an onset if conflict breaks out within any of the four following quarters. The predictive performance of the model remains strong. A forecast of onset up to a year ahead produces an AUC of 0.89 for any violence and 0.91 for armed conflict and topics now adds significant forecasting power even in cases with a conflict history. This is due to the fact that an immediate conflict history is less informative about a renewed outbreak after a year. In other words, the text features are now used by the random forest model to distinguish different post-conflict dynamics.

Fig. 4: ROC Curves For Predictions of Onset Within Next Year



Note: The prediction method is a random forest. 'Text' contains 30 topics and token counts and 'history' contains 4 dummies capturing time passed since the last conflict and dummies for the presence of lower levels of violence. Hard cases are defined as not having had conflict in 10 years. The numbers in the legends represent the respective area under curve with bootstrapped 95% confidence intervals in square brackets.

An important question is how much improvement is possible by adding variables to this framework. To benchmark the performance of text, we compare the predictive power to two sets of predictors from the economics and forecasting literature. The first is a standard set of predictors based on (Goldstone et al. 2010) which includes political institutions dummies based on various dimensions of the Polity IV data, number of neighboring conflicts from UCDP, data on child mortality from the World Bank and the share of population discriminated against using data from the Geographical Research On War, Unified Platform (GROWup) (Girardin et al. 2015). In the second case, we use a dataset of 60 commodity prices which we combine with constant commodity export weights from Bazzi and Blattman (2014) to generate a measure for

commodity income shocks. These are updated on a monthly basis by the World Bank which could provide a good basis for forecasts.[15] Many commodity export weights and variables such as infant mortality are not available for as many countries and years. For the sake of comparability, we only use overlapping predictions for evaluation, i.e. country-quarters in which the availability of data allow predictions for both sets of variables.

Results and a comparison to these models are shown in Figures 5 and 6. We find that both the standard variables and commodity prices contain some forecasting power but that the forecast of all cases from conflict history alone is much better. We also experimented more broadly with additional macro-variables. Other frequently used variables such as executive constraints, ethnic composition, population, GDP growth or GDP levels do not provide a better forecast than conflict history. This is a somewhat surprising finding but it is important to stress that the cross-sectional variation is extremely powerful and finding useful time-varying predictors of conflict is hard. This is in line with findings in Mueller and Rauh (2018) who show that the time-variation in most empirical work on conflict has too low power to predict conflict. Here we go one step further and instead focus on conflict history - just four indicators derived from on one single variable - and again find that it is extremely challenging to improve risk predictions beyond this.[16] Interestingly, in Figure 6 we see that text does only slightly better than commodity shocks in forecasting all cases. However, for hard cases, text alone performs better. A possible explanation is that commodity exporters often experience repeated cycles of violence which is best captured by conflict history.

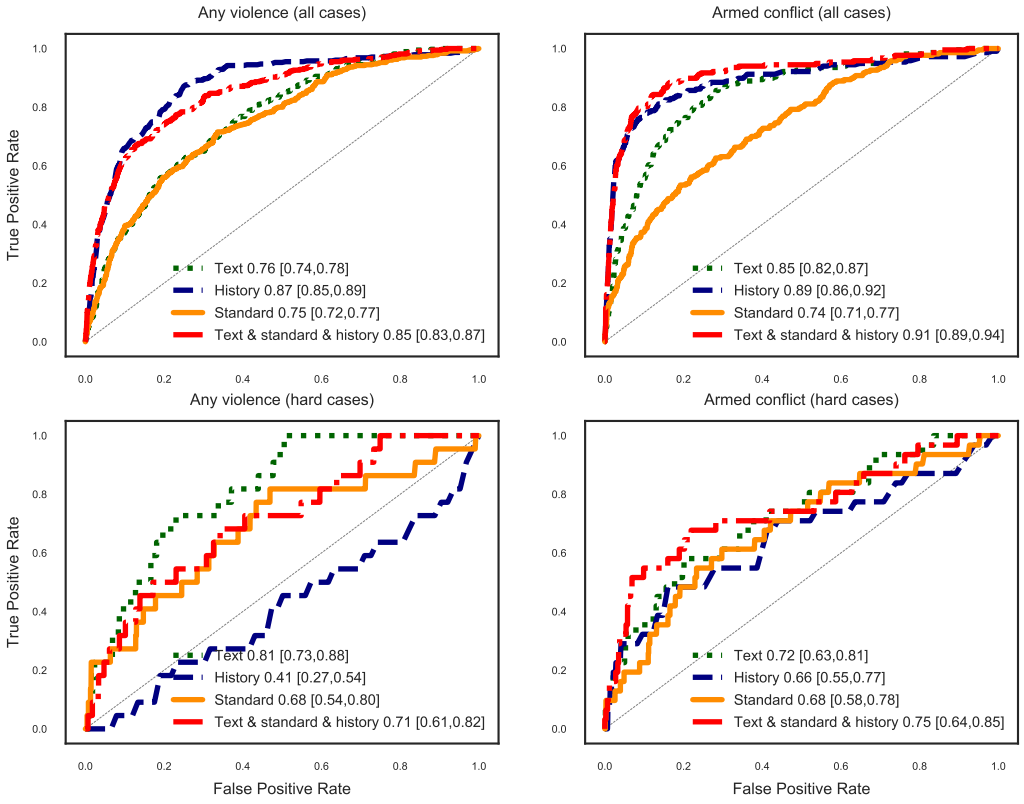
Importantly, a model based on history and text together with standard variables or commodity prices does not consistently perform better than our main model based on history and text alone in Figure 3. This is an important implication as it suggests that adding variables to the history and text model does not add forecasting power. An additional benefit of focusing on

[15] We fix the weights following Ciccone (2018) who finds a strong impact of commodity-price shocks on conflict.

[16] The reason we do not simply use the number of years since conflict as one single variable is because for linear models, such as the logit lasso, this poses a problem. While non-linearities do not pose a problem for the random forest, we want to maintain comparability across prediction methodologies.

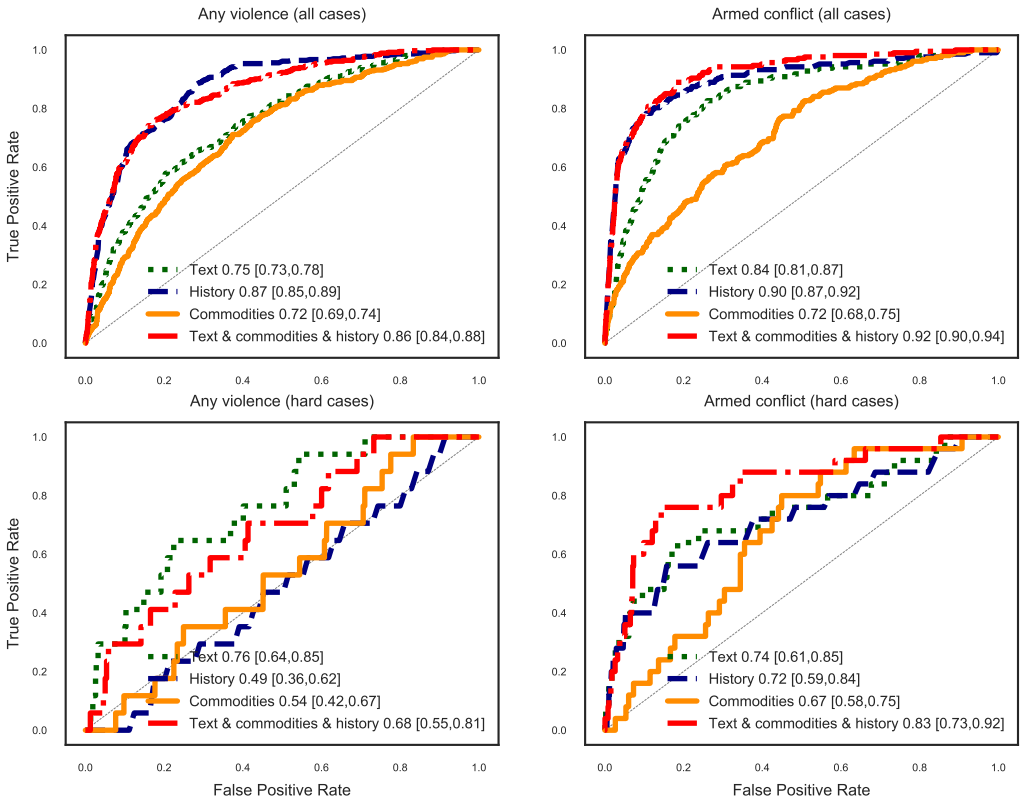
text is its coverage and its availability on the daily basis for all countries going back to 1989. Many standard variables are available only for a subset of countries and with lags up to several years. More problematic is the practice of recoding macro variables which means some of the variables in the standard model use future information. In what follows, we therefore focus on the simple model based on text and history dummies.

Fig. 5: ROC Curves of Forecasting Any Violence (left) and Armed Conflict (right) Compared to Standard Variables



Note: The prediction method is a random forest. 'Text' contains 30 topics and token counts, 'history' contains 4 dummies capturing time passed since the last conflict and dummies for the presence of lower levels of violence, and 'standard' contains infant mortality, political institutions, share of discriminated population, and neighboring conflicts. Hard cases are defined as not having had conflict in 10 years. The numbers in the legends represent the respective area under curve with bootstrapped 95% confidence intervals in square brackets.

Fig. 6: ROC Curves of Forecasting Any Violence (left) and Armed Conflict (right) Compared to Commodity Prices

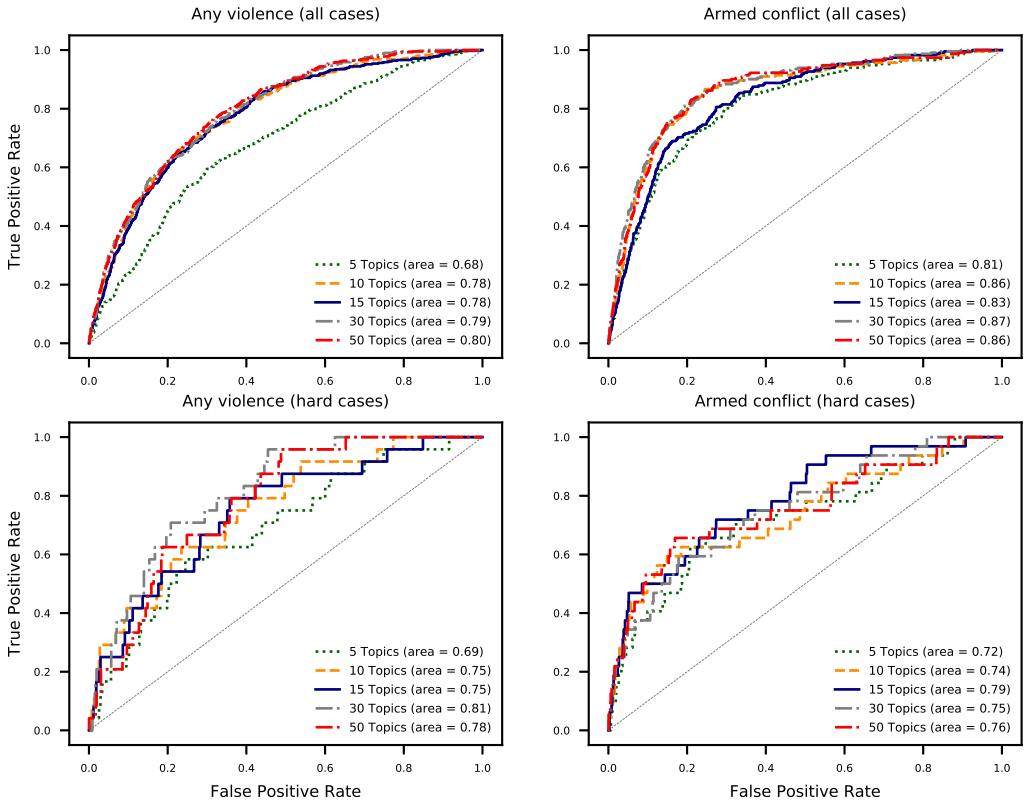


Note: The prediction method is a random forest. 'Text' contains 30 topics and token counts, 'history' contains 4 dummies capturing time passed since the last conflict and dummies for the presence of lower levels of violence, and 'commodities' contains 60 commodity export weights interacted with the quarterly mean price for the commodity. Hard cases are defined as not having had conflict in 10 years. The numbers in the legends represent the respective area under curve with bootstrapped 95% confidence intervals in square brackets.

In Figure 7 we show that the models performance is not specific to 30 topics. The model performs similarly well for 5, 10, 30, and 50 topics. For 5 topics, however, the performance seems slightly worse as the topics are likely to be too general.

A detailed discussion of parameters, additional results, and robustness checks are reported in the Online Appendix. Two findings worth highlighting are: First, more than 40 variables generated from text-based event data do not provide a better forecast than our topics. If anything, topics produce higher AUCs. In addition, we find that when forecasting armed conflict,

Fig. 7: AUC Curves of Forecasting Violence with Random Forest Using Text while Varying Number of Topics



Note: Predictors include specified number of topics and token counts. Hard cases are defined as not having had conflict in 10 years.

our topics and the event data have complementarities in the sense that a model with conflict history plus both sets of variables performs better than a model that relies on only one of the two. These findings are in line with the idea that the event data is more able at capturing a situation which might escalate, whereas the forecast in the topic model relies only in parts on escalation. Second, we also find that the random forest model performs particularly well when compared to other methods of supervised machine learning like logit lasso regressions or neural networks.

4.2 A Key Problem for Policy: Precision

The main problem of ROC curves is that they separate the two dimensions of positives and negatives. The y-axis shows the share of positives that are correctly predicted and the x-axis shows the share of negatives that incorrectly predicted. If there are many more negatives than positives to predict, this will not become visible in a ROC curve. This is a particularly big problem when the samples are as heavily imbalanced, as they are in our sample.

For this purpose we can link the two dimensions through precision which, at cutoff c, is given by

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

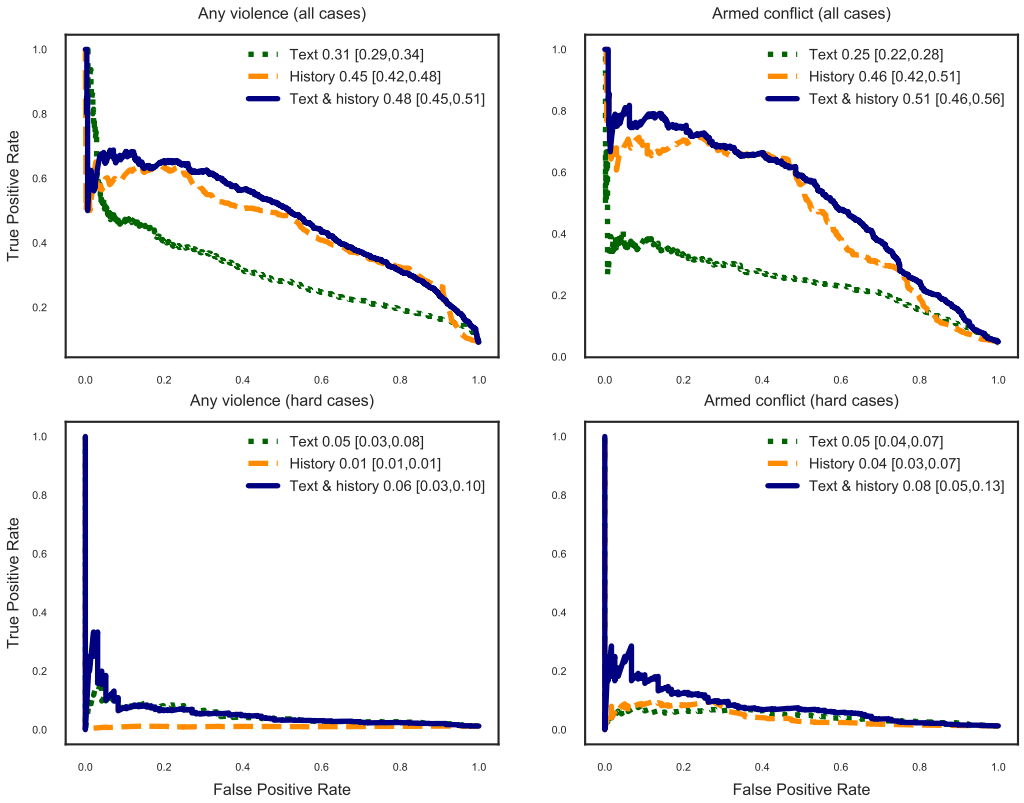
Here the number of FP, which are the negatives the classifier labels as positives, relates the negatives to the dimension of positives represented by TP. Precision falls with more 0s in the sample. As the cutoff falls, precision will converge to the share of 1s in the entire sample, i.e. to around 1% in our hard cases sample.

Figure 8 shows precision-recall curves for the one-year ahead forecasts. In all four graphs the x-axis displays the true positive rate, while the y-axis summarizes the precision, i.e. the share of alarming situations where conflict actually broke out. On the top we show the results for all onsets. Precision overall is very good (around 50-60% for any violence and 60-80% for armed conflict when the TPR is between 10-50%).

On the bottom we show results for hard onsets. As expected, precision deteriorates when predicting hard onsets (around 10 percent when the TPR is between 10-50%). This is due to the extreme imbalance in the hard onset sample. A precision score of over 10% is a significant improvement in this context but this still raises the question of whether precision is high enough to support preventive effort.

Precision shows a policymaker how much prevention effort would follow false alarms, i.e. how much resources are wasted on false positives. A precision of 10% implies that the ef-

Fig. 8: Precision-Recall Curves of Forecasting Violence



Note: The prediction method is a random forest. 'Text' contains 30 topics and token counts and 'history' contains 4 dummies capturing time passed since the last conflict and dummies for the presence of lower levels of violence. Hard cases are defined as not having had conflict in 10 years. Bootstrapped 95% confidence intervals of the average precision are reported in square brackets.

fective cost of preventive action is ten times higher than it would be with a perfect forecast. Precision/recall curves therefore provide some notion of the waste of resources when trying to anticipate the outbreak of violent conflict. But to evaluate whether the forecast is precise enough to justify preventive action we need to develop a model which is directely targeted at the policy problem of prevention.

5 Integrating Forecasting and Prevention

Building on the findings in the standard forecast framework we now derive a way to interpret the model output $\hat{y}_{iT+1}$ for the policy problem we describe in the introduction. This also provides a natural link to the literature on causal drivers of armed conflict because we need to model interventions which would naturally benefit from a better understanding of the drivers of conflict.

Assume that the policymaker wants to minimize expected total costs by choosing a forecast cutoff c, i.e. she is facing the following problem

$$\begin{aligned} \min_c E[Cost_c] &= Cost_{TP} \times E[TP_c] + Cost_{FP} \times E[FP_c] \\ &\quad + Cost_{FN} \times E[FN_c] + Cost_{TN} \times E[TN_c] \end{aligned} \tag{2}$$

where $Cost_J$ with $J \in \{TP, FP, TN, FN\}$ are the different cost weights on the cost function and $E[TP_c]$, for example, is the expected number of true positives at cutoff c. Note, that this cost function is shared by many policy problems like preventing crime or implementing preventive measures against an epidemic at different locations. The framework could even be adapted to a dynamic decision problem in which a policymaker takes repeated actions over time to prevent a financial crisis or economic contractions. We now make additional assumptions to simplify and adapt this general form of the cost function to the policy problem to give a little more interpretative quality to the cost weights.

We assume that the past out-of-sample performance of the forecast model can serve as a benchmark of its future performance, i.e. that $E[TP_c] = TP_c$ in the forecast model. This is only realistic if the existence of the forecast or the presence of the policymaker do not change the performance of the forecast itself. We will return to this point and other caveats in the concluding section.

The objective of the policymaker is to minimize the costs of conflict and interventions. In this context false negatives, FN_c, mean conflicts break out and cause present discounted damage

V_D. The discounted future cost of peace today is V_P which is the cost weight of true negatives TN_c.

Call I the cost per intervention, i.e. per predicted positive. This cost needs to be paid regardless of whether the intervention is successful and therefore generates a cost of false positives, FP_c. In the case of a true positive, TP_c, the policymaker intervenes in a situation in which a conflict would otherwise break out for sure. In other words, the intervention will try to convert a true positive into a peaceful outcome. We assume for simplicity that prevention works with some constant likelihood p. This is where research on causal mechanisms is relevant because it should lead to better policies and an increase in the likelihood p. This highlights the complementarity between research on causal links and the forecasting problem stressed by Kleinberg et al. (2015).

With these assumptions the cost function in equation (2) can then be rewritten:

$$\begin{aligned} min_c E[Cost_c] = (pV_P + (1-p)V_D + I) \times TP_c + (V_P + I) \times FP_c \\ + V_D \times FN_c + V_P \times TN_c. \end{aligned} \tag{3}$$

Lowering the cutoff c will put more weight on the first line of the cost function. As $I > 0$ we therefore need that

$$pV_P + (1-p)V_D + I < V_D$$

or

$$I < p\,(V_D - V_P)$$

for there to be any use in preventive action. This is intuitive as prevention today costs I and leads to an expected benefit of $p\,(V_D - V_P)$.

The forecast framework plays an important but subtle role here. A decrease in the cutoff c will generate more positives and less negatives. This will lower the number of FN_c and TN_c and at the same time increase both TP_c and FP_c. With high precision a relatively large share of cases change from being FN_c to being TP_c which generates a benefit of $p\,(V_D - V_P) - I$. But

with low precision, a larger share of cases change from being TN_c to FP_c which increases costs by I. As observations are ordered by the fitted values, $\hat{y}_{T+1}$, precision will first be high and then fall so that the expected cost minimzation problem is well-behaved.

As an illustration of this integrated prevention framework, we now assume some, relatively conservative. parameters and plot the resulting cost curves. Assume that the outbreak of a conflict causes discounted costs of 100 billion USD and that peace has no costs.[17] Good estimates of prevention costs and effectiveness are much harder to come by - mostly because there is currently no institutional framework for prevention outside countries with a conflict history. Instead, we now have a good understanding of responses to conflict both in terms of humanitarian aid and peacekeeping troops. A key factor is whether military interventions are necessary so that we would expect interventions in hard cases to be much cheaper.[18] We assume that prevention costs for all cases are 1 billion USD per quarter of intervention but only 0.5 billion USD in hard cases. This is a very high amount per quarter and we assume that prevention only works 10% of the time ($p = 0.1$).

We then have that the cost weight on the number of true positives is 91 billion USD and the cost weight on false positives is 1 billion USD. The most costly outcome are false negatives with 100 billion USD and the least costly outcome are true negatives with 0 costs. To highlight the role of understanding causal drivers of conflict we also show the case that interventions are less effective ($p = 0.05$). This will changes the costs of a true positive to 96 billion USD.

We use our year-ahead forecasts and assume that the policymaker treats the forecasts as independent in the sense that several warnings in a row can, but do not have to occur before an onset. Figure 9 illustrates the resulting cost function for armed conflict. On the x-axis of the figure, we show the cutoff c and on the y-axis we plot total costs. We generate the costs using

[17] The World Bank estimated, for example, that from 2011 until the end of 2016, the cumulative losses in GDP from the Syrian civil war were 226 USD billion (World Bank 2017*a*). This is obviously just a small share of the total costs which includes also the costs of the humanitarian response. In the actual application this number could also depend on population size and many other factors.

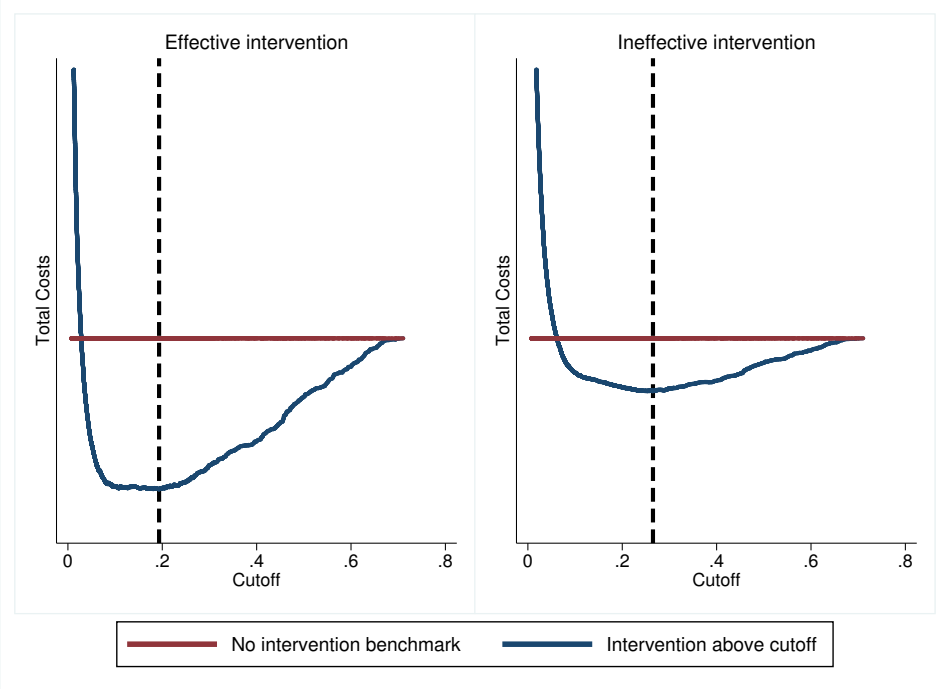
[18] See the discussion in Chalmers (2007).

equation (3) and the out-of-sample forecast performance of the text combined with the conflict history model. In order to benchmark the costs with intervention, we also show the costs without any interventions as a red horizontal line. The cost function (blue line) will always converge towards this benchmark with higher cutoffs as less and less interventions are conducted.

In the left panel of Figure 9 we show the outcome with relatively effective interventions ($p = 0.1$). For low cutoffs there are a lot of interventions and this generates plenty of false positives and costs which are higher than without interventions. With higher cutoffs the number of interventions falls and with it total cost. At a cutoff just under $c = 0.2$ the cost curve takes a minimum. At this point an increase of the cutoff increases costs because precision now is so high that the cases with interventions cover a lot of actual outbreaks so that raising the cutoff creates many costly false negatives. In the right panel, we show the same cost curves under the assumption that interventions are less effective ($p = 0.05$). Note that the total costs without interventions remains the same. However, total costs are now much higher with interventions so that cost savings with prevention are lower. Keep in mind that the ordering of observations by their forecast values produces some convexity in the cost function. As precision needs to be higher, the number of interventions needs to fall and this implies a higher optimal cutoff value which is now above $c = 0.2$.

This trade-off between policy effectiveness and the cutoff is a special feature of the framework we propose here. Policy effectiveness influences the way a policymaker interprets the forecasts coming out of a forecasting model. Worse policy tools make the policymaker more conservative in her forecast. The usefulness of our forecast model therefore critically depends on whether conflicts can be resolved and at which costs. But the reverse is also true; optimal intervention policy will react to the precision of the forecast. To see this, keep in mind that low precision implies that a lot of false positives, FP_c, are produced compared to true positives, TP_c, and this raises the cost of intervening at every cutoff. This means the policymaker needs to raise the optimal cutoff just as she would with a lower p.

Fig. 9: Cost Curves: Armed Conflict (All)



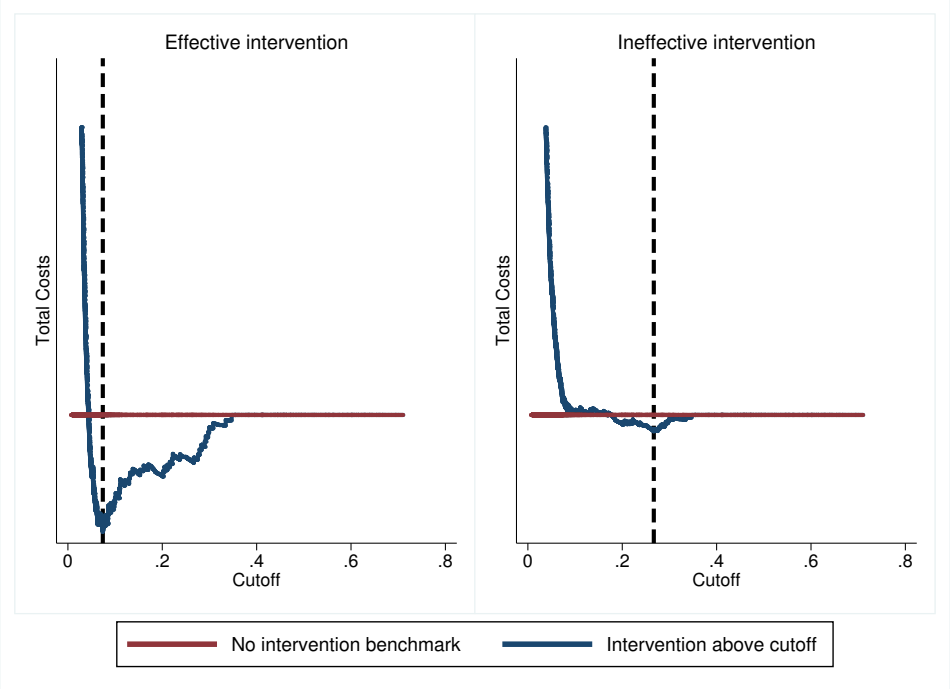
Note: The figure contrasts two scenarios for armed conflict using equation (3) and the out-of-sample forecast information of the text + conflict history model. The left cost curve assumes interventions are relatively effective ($p = 0.1$) and the right figure assumes interventions are relatively ineffective ($p = 0.05$). Intervention costs are 1 billion USD and if an outbreak is prevented this saves damages of 100 billion USD.

A similar image emerges in the case of hard onsets shown in Figure 10. Here we assume that intervention costs only 0.5 billion USD given that prevention efforts should require less resources and have higher benefits.[19] Again, the policymaker will pick a higher cutoff and engage in relatively little prevention with low effectiveness. Note, however, that Figure 10 also suggests that prevention is about to hit a boundary with low effectivenss given that the cost curve with interventions is now barely below the no-intervention benchmark. Still, despite the extremely low precision in hard onsets there are reasonable parameter values for which predicting hard onsets is a worthwhile exercise.

In the Appendix we show how interventions would have been allocated across time if our simple prevention framework had been used in the past. The assumptions we make would

[19] See the discussion in United Nations and World Bank (2017).

Fig. 10: Cost Curves: Armed Conflict (Hard Cases)



Note: The figure contrasts two scenarios for armed conflict using equation (3) and the out-of-sample forecast information of hard problem cases from the text + conflict history model. The left cost curve assumes interventions are relatively effective ($p = 0.1$) and the right figure assumes interventions are relatively ineffective ($p = 0.05$). Intervention costs are 0.5 billion USD and if an outbreak is prevented this saves damages of 100 billion USD.

have led to many interventions post conflict and fewer interventions in hard cases. Still, for the prevention of armed conflict we get a significant level of activity even for hard onsets of around ten interventions each quarter. This means that, despite the incredibly low baseline risk of around 1%, intervention would have been economically feasible under our assumptions. Also, we only reduced intervention costs but gave the hard problem cases no special treatment in terms of prevented costs, V_D, which is probably not realistic. Our exploration here should therefore be regarded more as description of the existing system of late interventions. Such a system will pay much more attention to interventions post-conflict which is exactly what we see in reality.

6 How Machine Learning Solves the Hard Problem

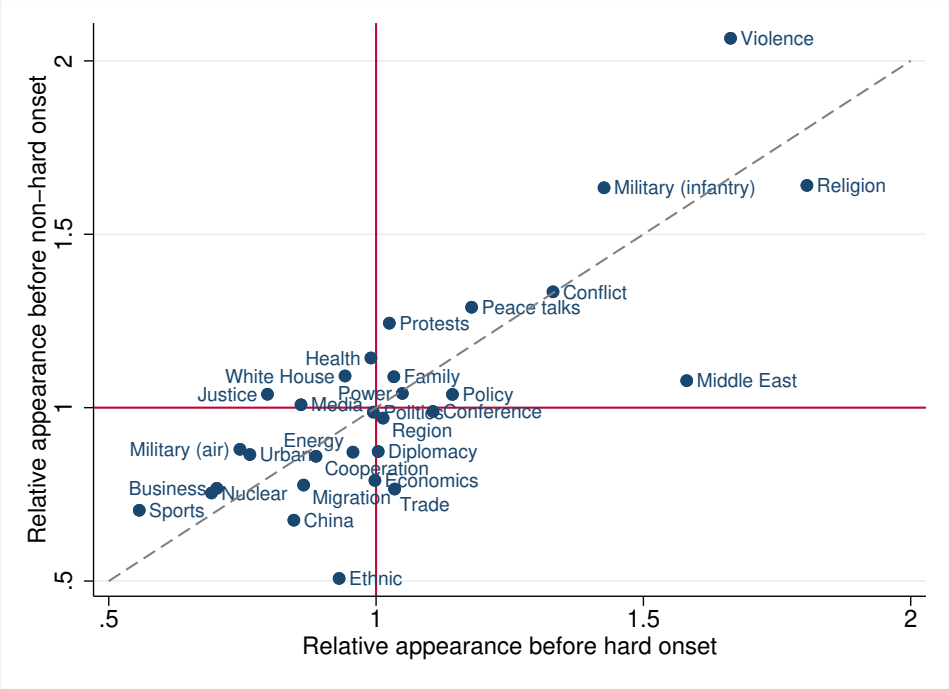
We use a methodology which is standard in other areas, such as inflation forecasting or pattern recognition, and apply it to the prediction of conflict. First, unsupervised learning is used for feature extraction. Then these features (topics) are used to predict conflict. An important advantage of this approach is that the model can rely on positive and negative associations of specific topics with violence. In Figure 11 we show the shares of each of the 30 topics in quarters before an onset in hard cases (x-axis) and in onset cases where the country has a conflict history (y-axis) relative to quarters in which there is no onset in the next period. For instance, the violence topic is 1.5 more likely to appear before a hard onset relative to a peaceful quarter but almost twice as likely to appear before an onset in a country with a conflict history. A striking topic which is not directly, semantically related to conflict is the religion topic which contains terms like islam and hindu and is strongly, positively associated with risk. Other topics like sports, business or trade appear less before onsets. News stories on business, for example, are more than 20% less likely before all onsets. In this way the topics provide signals which the supervised learning is able to exploit.

The random forest relies on these different associations by combining conflict history and the various topics in decision trees. In the top panels of Figure 12 we show the relative total importance of the topics compared to conflict history in our random forest model. For simplicity, we combine the importance of the topics in one bar and only distinguish topics by whether they contain tokens that indicate violence prominently. In this way we separate the signals contained in the five semantically related topics (violence, military etc.) from the other topics.[20]

The gray bars in Figure 12 indicate the relative total importance of topics when predicting conflict generally and the black bars indicate their importance in hard cases. Topics provide more than 50% of total importance of which non-violent topics account for most of the predic-

[20]The list of topics from the full sample and our classification into violence vs. non-violence are presented in the Online Appendix.

Fig. 11: Topic Shares Before the Onset of Any Violence Relative to Peaceful Quarters

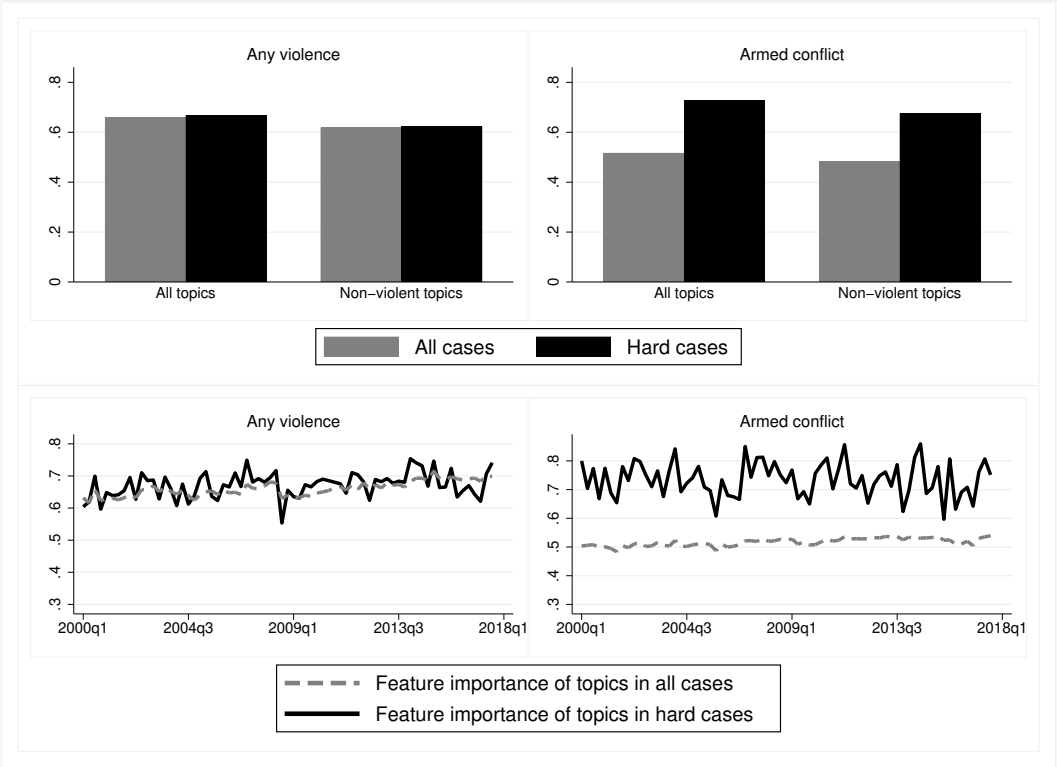


Note: Each dot represents the average appearance of a topic across country/quarters relative to peaceful quarters. The x-axis represents the relative appearance in quarters preceding hard onsets while the y-axis shows the relative appearance in quarters before onsets in countries with a conflict history.

tive power. The rest can be attributed to the token count and the variables capturing conflict history or low levels of violence. Importantly, the predictive power of non-violent topics is high when predicting hard cases and in armed conflict this is particularly visible. In other words, the forecast of hard cases seems to rely to a large degree on parts of the news text which are not direct reports of violence and are, apart from religion, in the lower left corner of Figure 11.

But why is the random forest model using this subtle variation better than other models? A detailed analysis of our forecasting models shows that the decision trees tend to pick conflict history at the top of the tree to divide the sample. For example, the dummy indicating the first quarter post-conflict receives the largest importance score and is more often used in top nodes of the tree. Topics are then introduced in lower branches. In other words, the random forest model is automatically geared towards picking up more subtle risks with topics when conflict history

Fig. 12: Feature Importance of Topics in Random Forest



Note: The feature importance is calculated on sequential out-of-sample predictions of random forests with conflict history and text using a tree depth of 5 and 300 trees for any violence and a depth of 4 and 200 trees for armed conflict.

is absent. In this way the forecasting model works around the importance of the conflict trap by conditioning on conflict history and at the same time uses information contained in the text. An additional, even more subtle aspect of this process is that the model uses topics to capture stabilizations in countries with a conflict history. News on business vary dramatically across countries and time. In Angola they started to appear, for example, when the country started to stabilize and the algorithm can therefore use this topic to drive down the false positive rate.

An important aspect of our method is that the sample available to the model is increasing in time. In the bottom panels of Figure 12 we use this fact to show how the total importance of topics changes over time, i.e. with a growing sample. Again, we see that the importance of topics is higher when predicting hard cases of armed conflict. In addition, importance of the

topics is increasing slightly for any violence. This means that the random forest model relies more and more on the text to separate high from low risk. Using cross validation we confirm that the overall predictive performance of the resulting random forests also tends to increase over time. To some extent this is surprising given the dramatically changing international context and new instabilities in the period 2000 to 2018.

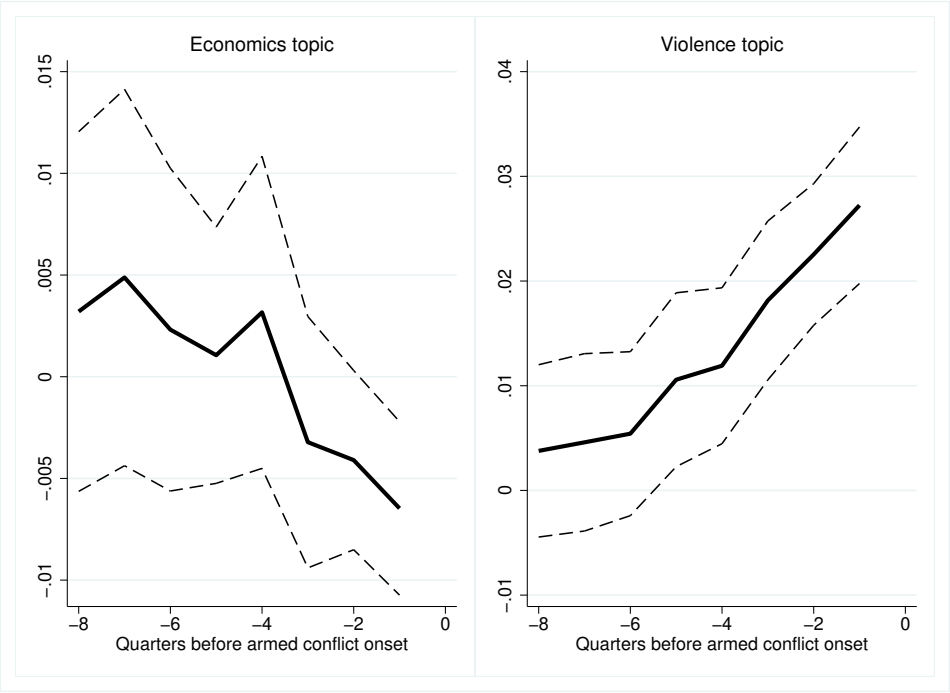
7 Dynamic Predicted Risk and Case Studies

There are several strong positive and negative associations between different topics and conflict onset. Importantly, these relationships are not only varying between countries but also within countries over time so that meaningful dynamic risk profiles result from our forecasts. To illustrate the dynamic properties of two topics, Figure 13 shows the movement of the economics (left) and violence (right) topic shares and the 95% confidence interval before the onset of armed conflict in the full sample. The figure is based on regressions which include country fixed effects so that they are showing the change in topics within countries over time. The increase of reporting on violence and the decline of reporting on economics leading up to the outbreak of violence is very clearly visible.

As a result of such movements, the risk evaluations that come out of our methodology are changing over time. In Figure 14 we show the rolling out-of-sample risk estimates coming out of the text-only model controlling for country fixed effects as onset approaches. The predicted risk is clearly increasing both when forecasting any violence (left) and armed conflict (right).

Given the dynamic movements of estimated risk within countries reflect actual onset risk, it makes sense to treat our risk estimates as data to be analyzed. We illustrate the nature of the risk forecast based on the text model for eight countries in Figures 15 and 16. The red dashed lines report the risk of armed conflict (right y-axis) whereas the blue lines report the risk of any violence (left y-axis). In all cases the risk estimate is a year ahead.

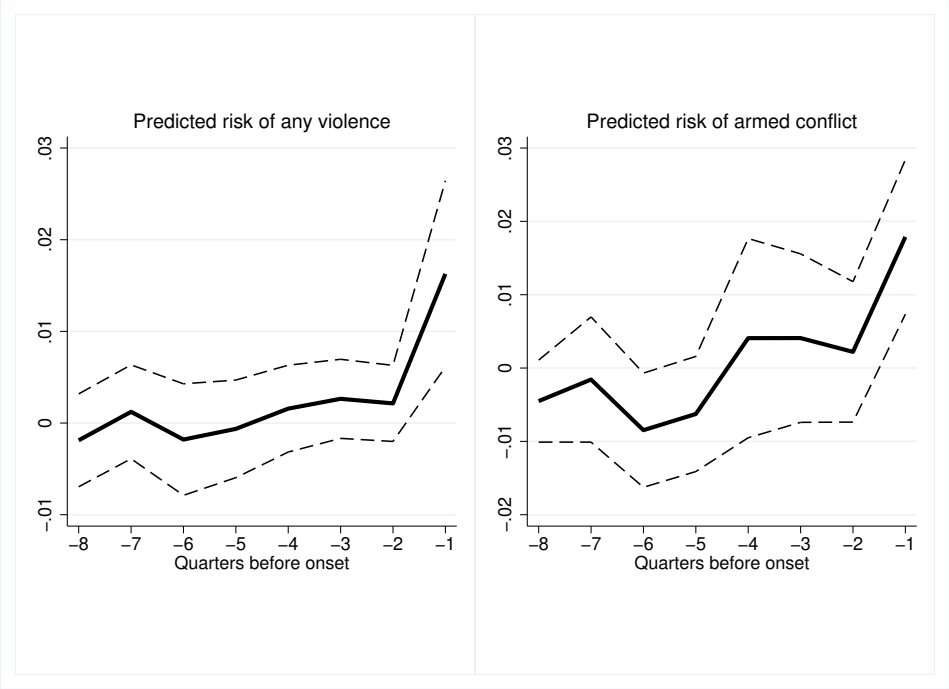
Fig. 13: Share Written on Economics and Military Topic Before Onset of Armed Conflict



Note: The topic share residuals relative to other quarters in the same country are represented by the solid lines. The data is generated through regressions of the topic shares on dummies for the number of quarters before the onset of conflict and country fixed effects. Dashed lines mark 95% confidence intervals.

What is clear from the figures is that risk reacts to violence. For example, the change in risk after the September 11 terror attacks in the United States is visible as a large shock. Similarly, terror attacks in Spain, Germany, and Tunisia all brought changes in predicted conflict risk with them. Also, the figures give a very clear idea of high risk and low risk periods in the respective countries. Germany entered a period of relatively high risk during the refugee crisis in 2015 and Tunisia after 2010. In the United States risk has fallen to relatively low levels in recent years. Spain had a relatively calm period but, given a terror attack, its secessionary movements and the central government's responses, is recently experiencing higher risk again. What is important, however, is that the forecast model produces low risk estimates the moment a situation calms down like in Spain and Germany but does not in Tunisia. In this way the model captures general low and high risk situations.

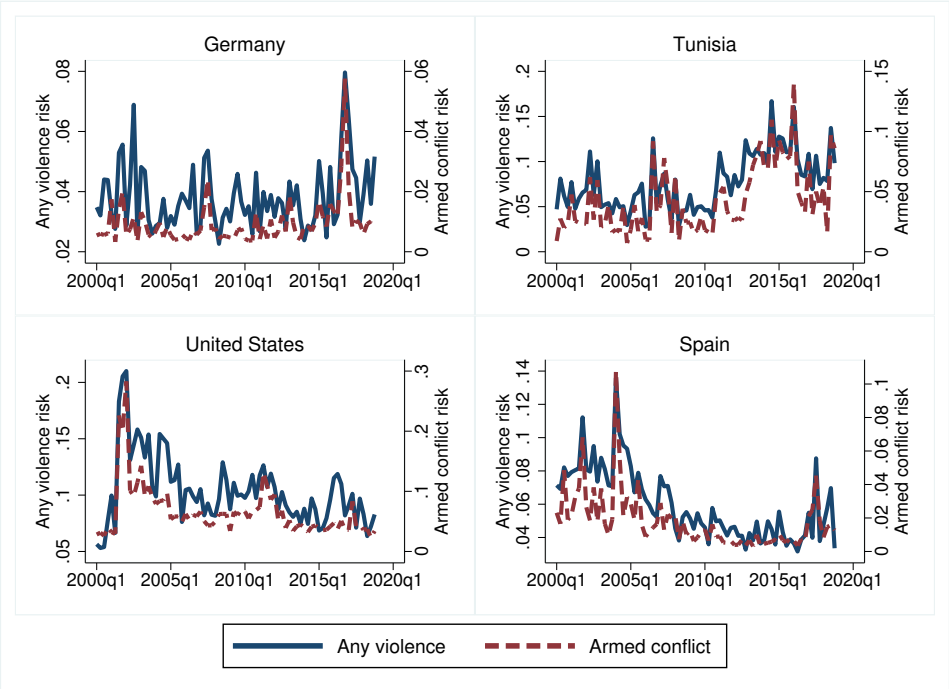
Fig. 14: Predicted Risk Before the Onset of Conflict



Note: Risk residuals relative to other quarters in the same country are represented by the solid lines. The data is generated through regressions of the out-of-sample predicted risk on dummies for the number of quarters before the onset of conflict and country fixed effects. Dashed lines mark 95% confidence intervals.

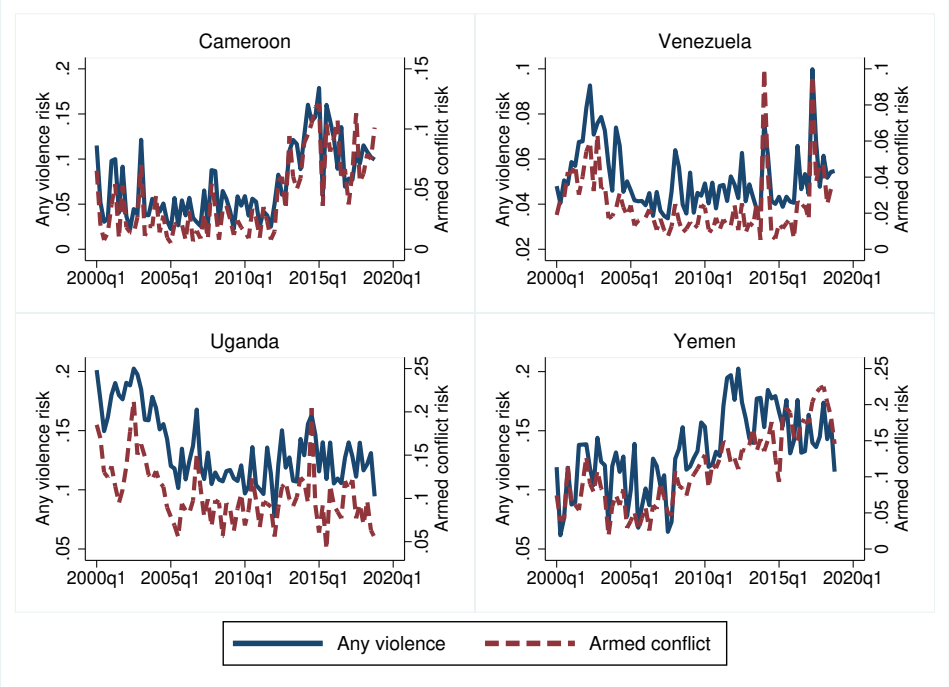
In order to look at cases which have tend to have higher risk levels we select four cases: Cameroon, Venezuela, Uganda and Yemen. The different trajectories of risk are visible with Cameroon, Venezuela and Yemen facing dramatically increasing risk while our risk model predicts stabilization in Uganda after violence stops. It is cases like these which are important parts of the risk model as they inform the model which topics are useful to predict (relative) stabilization. Risk for Venezuela shows a clear uptick in 2017.

Fig. 15: Predicted Risk of Any Violence and Armed Conflict (Case Studies I)



Note: Predictors include 30 topics and token counts.

Fig. 16: Predicted Risk of Any Violence and Armed Conflict (Case Studies II)



Note: Predictors include 30 topics and token counts.

8 Conclusion

The prevention of conflict requires attention to cases with a low baseline risk, i.e. cases in which the country is experiencing a sudden destabilization after long periods of peace. Research can help here by providing forecasting models which are able to pick up subtle changes in risk. We contribute to this agenda by providing a forecasting model which combines unsupervised and supervised machine learning to pick up subtle conflict risks in large amounts of news text. This allows us to forecast cases which would otherwise remain undetected and, at the same time, overcomes the problem of lack of good and timely published data which is a crucial problem in applications.

Our results paint a positive picture of the role of supervised learning in longer time series. The model increases its reliance on text and its performance as the sample size increases. This suggests that the dimensionality reduction with LDA helps to reveal deep, underlying features which are recognized when enough data is available. Yet, dimensionality reduction using unsupervised learning is rarely used in conflict forecasting. Applying unsupervised learning to the large amounts of available event data seems a particularly useful way forward.

Forecasting models like ours also provide objective risk evaluations for countries which never experienced violence. This is not only potentially useful for policymakers but it has the advantage of providing the basis for research on prevention itself. Here is where we see considerable potential for future research. The fact that some conflicts are harder to forecast than others might also yield insights into the role of exogenous factors like economic shocks and endogenous internal political factors.

Our second contribution is the interpretation of the forecast in a policy framework which would allow a policymaker to analyze trade-offs between intervention effectiveness and forecast performance. We show that under reasonable parameter assumptions a policymaker might want to use our forecasts to engage in preventive action. However, the framework also reveals strong

pressures to focus on a reaction to recent violence as it is currently the case the real world. We believe such insights to be of use far beyond the prediction of conflict.

However, there are several caveats in our current framework which are fruitful areas for future research. First, our model is static which means it does not take into account the fact that the policymaker will also intervene in the future which would, in turn, change the costs and expected damages today. In this context our framework should be regarded as a way to evaluate a one-off policy intervention in one quarter instead of a full intervention strategy. Second, we assumed that risks do not react to the forecast or the prospects of intervention. If forecasts are made public they may, however, have an impact on conflict risk by changing beliefs of local or global actors. In this regard our policy regime might face its own Lucas critique or constitute self-fulfilling prophecies. In this case the publication of the forecast would render it wrong.

Appendix

A Data Description

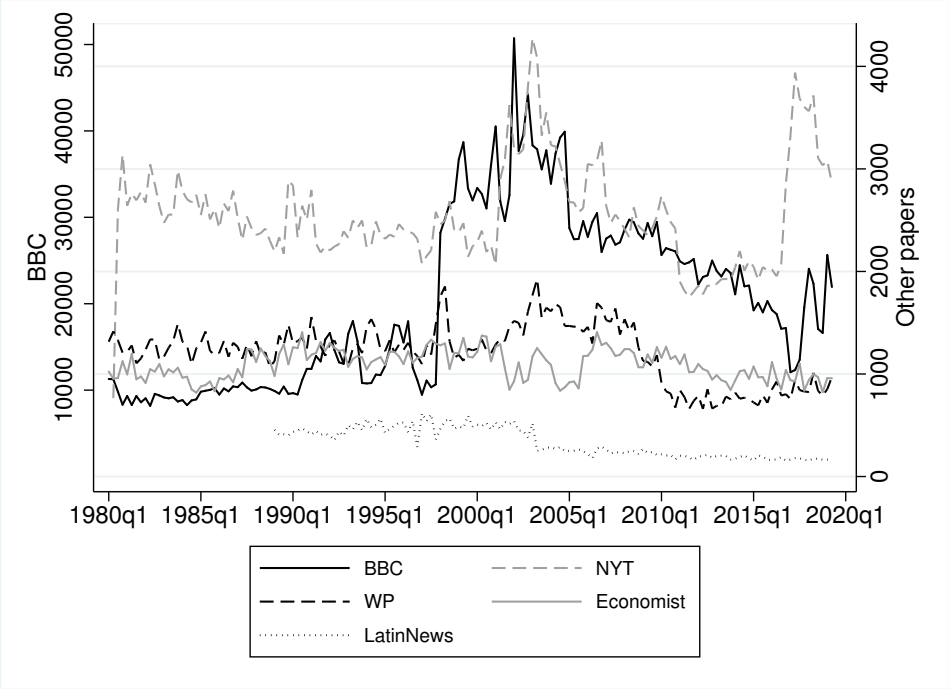
All our text data is downloaded manually from Lexis Nexis and Latin News. Due to copyright issues the raw newspaper articles cannot be shared. Summarized topics and all other data and codes will be made available upon publication. The key factor in choosing our news sources is that they should be english-speaking, offer as much text as possible and long time-series. We therefore chose the New York Times (NYT), the Washington Post (WP), the Economist, Latin News, and the BBC Monitor (BBC). The latter source represents the bulk of our data and tracks broadcasts, press and social media sources in multiple languages from over 150 countries worldwide and produces translations in English. Latin News is a news aggregator and commentator that specializes on countries in Latin America. We chose this source as the BBC Monitor has a geographic focus on Asia and Africa and we wanted to get a better signal

for Latin America. A potential problem with our approach of mixing different sources with different weights of economics and politics is that measurement error increases. However, we see continuous quarterly coverage of smaller countries as crucial.

We download an article if the name of the country or its capital appears in the title of the article. This gives us a panel of articles from all sources for over 190 countries for the period 1989Q1 to 2019Q4. In total we have 4,125,954.1 million articles of which 406,912 articles are from the New York Times, 215,398 from the Washington Post, 40,861 from Latin News, 192,385 from the Economist, and 3,270,398 million articles from the BBC Monitor. This means that the BBC Monitor articles dominate our data. Figure A.1 shows the number of articles we have for each quarter. From this is clear that the number of BBC news available from Lexis Nexis increases around 2000. Part of this increase came from a change in the headlines, which around this period often began with the country name followed by a colon and then a traditional headline. This temporary change does not affect the regional distribution substantially. In any case, the increase in the amount of news is only problematic for our forecasting exercise if the increase or decrease in the number of news somehow affects the share of news written on a specific topic. It would then become impossible to use this data effectively for forecasting as the training of the model would not produce useful forecast in the testing sample.

In Table D.1 we summarize the different sets of predictors we use. The first model is based on our text data. This includes 5, 10, 15, 30 or 50 topic shares and the log of the word count of that quarter. The word count varies between 5 and 1226371 and is log-normally distributed with a mean of 4274 words, while the topics sum to one within each country-quarter. The second model is based on the violence data from GED. It includes dummies for the conflict history and dummies for ongoing low-level violence. Finally, we use the ICEWS event database to generate a quarterly panel between 1995Q1 and 2016Q4. We only use events in which the source and the target of the action were in the same country. We then make a count of all 20 event types on the Conflict and Mediation Event Observations (CAMEO) integer scale and

Fig. A.1: Number of Articles by Source



Note: The y-axis on the left exhibits the quarterly sum of BBC articles, while the y-axis on the right exhibits the quarterly sum of articles from The Economist, New York Times, Washington Post, and Latin News.

another count of all 20 events on the CAMEO scale that involve the government either as target or as source. In addition, we generate a count of all protest events, the average CAMEO code of events involving the government, and the average CAMEO code of all events taking place in the country.

B Discussion of Estimated Topics

In this section we discuss various aspects of the estimated topics. We estimate the topic model repeatedly starting with all text until 2000Q1 and then we update every quarter using a dynamic topic model (Řehůřek and Sojka 2010). We allow the weight variational hyperparameters for each document to be inferred by the algorithm. Before feeding the text to the machine learning algorithm we conduct standard procedures when working with text. We remove overly frequent

Table A.1: Sets of Predictors

Name	Variables
Topics	Estimated topics using dynamic topic model and total number of tokens
Conflict info	Conflict history and low-level current violence indicators
Standard	Infant mortality, political institutions, share of discriminated population, and neighboring conflicts
ICEWS	Count of all event types, events involving government, all protests, overall average CAMEO code, and average CAMEO code of events involving the government

words defined as stopwords. Then we stem and lemmatize the words before also forming two and three word combinations. Next we remove overly frequent tokens, i.e. those appearing in at least half of the articles. Finally, we also remove rare expressions appearing in less than 100 documents.

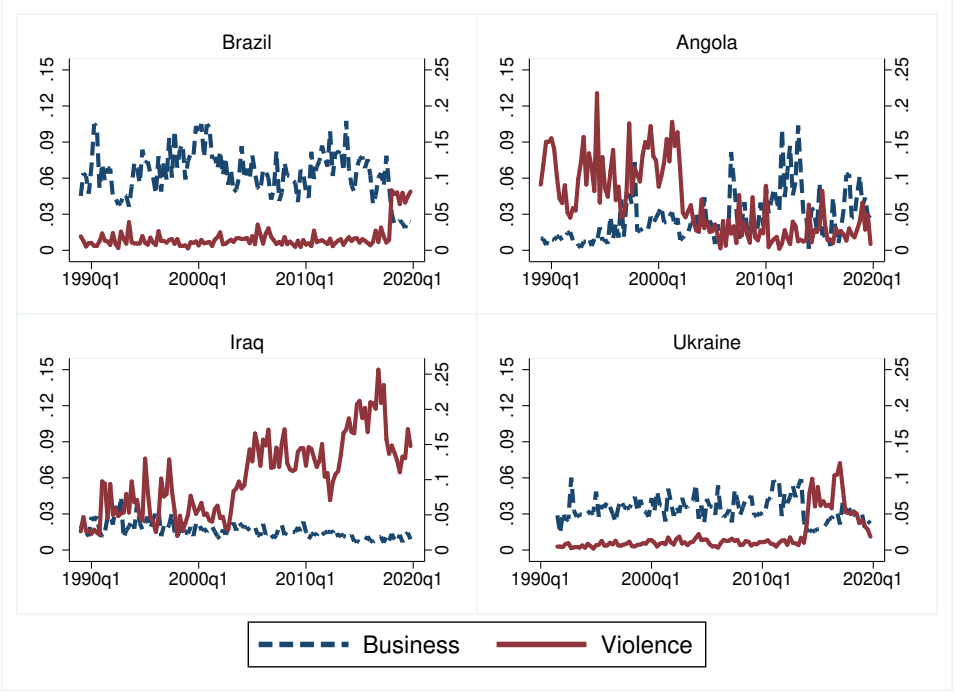
In what follows we will focus on topic models estimated in 2019Q4 as these describe all relevant text we use in our forecasting framework. Table B.1 summarizes the top 10 terms in the $K = 30$ topic model estimated in 2019Q4. There are five topics which are clearly related to conflict. Topic 1, which we label the military topic, contains terms like missile and military and terms indicating military combat. Topic 2, which we label the violence topic, has terms like attack, kill, group, police and force which could capture non-militarized conflict.

The striking feature of the topic model is, however, that it provides an overview over the entire news landscape with topics capturing peace talks, trade and development, politics and business. This allows our forecast to rely on much more the presence or absence of conflict topics.

In Figure B.1 we show the timeline for two topics, business and violence, in our sample period for Brazil, Angola, Iraq and Ukraine. Two features are remarkable. First, as expected, news on business disappear rapidly as news on violence surge. Brazil, a country which was characterized by a lot of news on business in particularly remarkable in this regard. In Angola writing on violence decreases dramatically following the cease-fire in 2002. However, what is

even more remarkable is that business news come in only slowly after a few years and fluctuate significantly. We know that the forecasting model relies to a large extent on variation like this - especially in its forecast of hard onsets. In Iraq the invasion of 2003 is very clearly visible and Iraq is clearly depicted as an extremely violent place in the news. In addition, business is almost never discussed in Iraq. In Ukraine the start of the turmoil in 2014 and outright war later stand out. Of course, such movements in conflict topics are only helpful for forecasting if they anticipate conflict. In Figure 13 in the main text we show this for the economics and violence topic.

Fig. B.1: Violence and Business Topic Shares Across Time and Countries



C Prediction Algorithms

To explore the gains from supervised learning we look at five different algorithms specified in Table C.1 which are trained with the available data using a Python implementation (Pedregosa

et al. 2011). We standardize the data in order to improve the performance of machine learning algorithms such as neural networks.

The five individual supervised prediction algorithms we use are a logistic lasso regression, k-nearest neighbor (kNN), neural network, AddaBoost, and random forest. Providing very brief summaries, the logit lasso estimates the log odds of an event using a linear expression while choosing which variables to include through a penalizing term. kNN is a non-parametric method used for classification in which the algorithm classifies a vector according to similarity. If a vector of predictors looks similar to those with many onsets, then it is more likely to classify a given set of predictors as an onset. Neural networks are a complex web of artificial neurons split into layers which are meant to resemble the functioning of neurons in a brain. Thereby, the technique can capture non-linearities through feedback effects between the multiple layers and because neurons might not fire until reaching a threshold. AdaBoost, which is short for Adaptive Boosting, uses output of other learning algorithms, referred to as 'weak learners', aggregated as a weighted sum. In our case, the weak learner is chosen to be a decision tree of depth one. AdaBoost is adaptive in the sense that weak learners are tweaked in favor of instances misclassified by previous classifiers.

Random forests construct many decision trees at training time and then averages across the predictions of the entire collection of trees, i.e. the forest. This way of modeling risk has the particular appeal that important features like conflict history will be chosen early if available, and the model therefore adapts automatically to the hard problem. We discuss this feature in the main text.

While the final evaluation of our model is carried out strictly out-of-sample, i.e. in the future without using any contemporaneous or future information, the training of the models is performed through cross-validation. More specifically, the method used is k-fold cross-validation, where the training set is split into k smaller sets. For each of the k 'folds' the following procedure is performed: A model is trained using $k-1$ of the folds as training data; using the

remaining data a test is carried out by computing our chosen performance measure, the AUC. The performance measure reported by k-fold cross-validation is then the average of the values computed in the loop. Each individual algorithm also requires the specification of hyperparameters by the user. For each set of predictors, we choose these hyperparameters by doing a grid search using the sample until the year 2000 and then selecting the hyperparameters that generate the highest AUC. Note, that this will understate the performance of the forecasting model slightly if more information leads to a deeper or modified model in later years.

In Table C.2 we present the chosen hyperparameters for the random forest. We see that with text alone, random forests tend to be deeper than when adding conflict history and information about current violence.

Table B.1: Top Ten Keywords of 30 Topic Model Using All Text Until 2019Q4

Nr	Label	Keywords
1*	Military (air)	militari air defenc exercis missil defens aircraft forc launch sea
2*	Violence	attack kill group polic forc area oper terrorist milit provinc
3	Conference	meet visit prime offici held deleg attend leader discus discuss
4	Power	like polit power time say want way make war long
5*	Conflict	war intern forc region militari territori support action threat secur
6	Urban	like time citi work dont way make look place open
7*	Military (infantry)	forc militari armi secur arm command oper troop defenc arm_forc
8	Economics	bank percent rate dollar cent econom increas economi tax price
9	Media	medium video say newspap channel daili pro broadcast languag journalist
10	Religion	islam taliban muslim islam_state group religi leader offici hindu majli
11	Justice	court investig case polic offic law arrest charg crimin justic
12	Business	compani busi onlin use market firm number servic oper product
13	Protests	protest right polic human demonstr human_right moral group activist organis
14	Diplomacy	ministri statement offici issu tokyo diplomat foreign_ministri affair embassi decis
15	Region	region head republ council servic inform centr deputi administr accord
16	Cooperation	cooper relat bilater develop region visit issu secur tie meet
17	White House	offici administr hous white secur washington depart white_hous intellig week
18	Energy	websit oil gas power energi plant compani electr product project
19*	Nuclear	nuclear missil sanction test weapon launch bjp rang ballist ballist_missil
20	Sports	team game los player play leagu sport cup fernandez second
21	Peace talks	talk peac agreement negoti meet deal council agre process issu
22	Family	woman famili child student univers school life work young old
23	Ethnic	nato ethnic church languag tamil congress_parti member lianc minor support
24	China	beij parti communist mainland offici communist_parti central committe xinjiang leader
25	Politics	parti elect polit vote opposit candid democrat ial support parliament
26	Policy	issu time need way want make problem import situat polit
27	Migration	border migrant island refuge airport cross port travel sea ship
28	Health	citi health area water local region provinc medic district resid
29	Middle East	arab moham base uae lebanes princ gaza gulf king salman
30	Trade	trade export develop econom project invest industri tariff import product

Note: The labels are arbitrary and have no influence on the prediction model.
The topics marked by '*' are considered violence topics.

Table C.1: Models

Technique	Brief description
Logit	Linear estimation of log-odds
K-nearest neighbor	Classifies a vector according to similarity
Neural network	Artificial neurons split into layers including feedback effects
AdaBoost	Weighted sum of other learning algorithm ('weak learner')
Random forest	Average over many decision trees
Stacking	Ensemble using logit based on five predictions

Table C.2: Hyperparameters

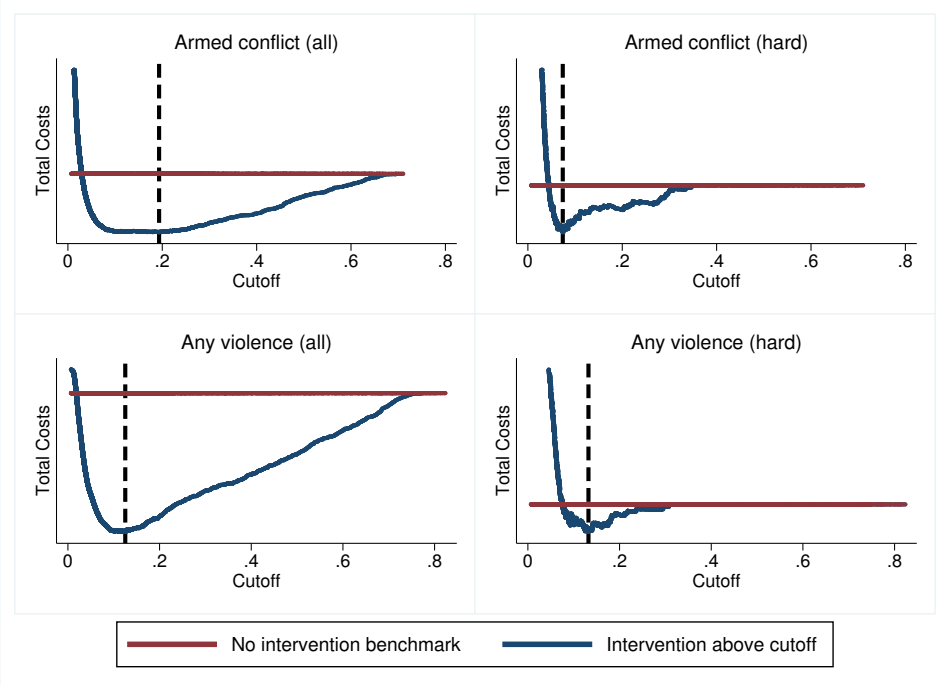
Predictors	Random forest	
	Depth	Trees
Any violence		
Text	6	350
Conflict info	4	10
Text & conflict info	8	225
Armed conflict		
Text	7	75
Conflict info	4	50
Text & conflict info	8	350
Civil war		
Text	7	150
Conflict info	1	125
Text & conflict info	2	375

Note: Hyperparameters chosen through cross-validation within the sample before year 2000.

D Using the Optimal Interventions Model

To illustrate the cost function approach we show what would have happened if these cost parameters had been used to intervene in the past. For this we first assume high effectiveness ($p = 0.1$) and calculate the optimal cutoff for all cases and hard cases in armed conflict and any violence. The optimal cutoffs and resulting cost functions are displayed in Figure D.1. In the left panel, we show the cost functions for all cases and on the right we show the costs for hard onsets only. We show the cost functions for armed conflict on the top and cost functions for any violence at the bottom. In all four cases there is a clear case for prevention. Even in the hard onset cases optimal interventions appear to reduce total costs.

Fig. D.1: Cost Curves: Model Summary

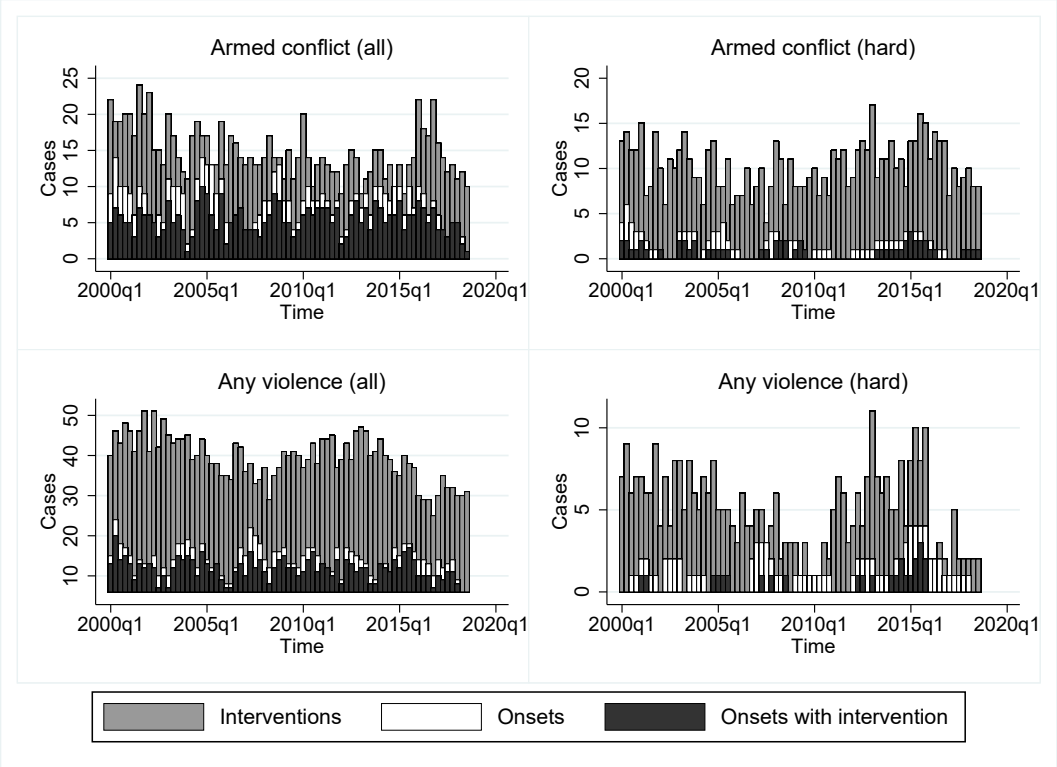


Note: The figure contrasts cost curves for armed conflict and any violence using equation (3) and the out-of-sample forecast information of all cases and for hard problem cases from the text + conflict history model. The cost curves assume that interventions are relatively effective ($p = 0.1$). Intervention costs are 1 billion USD for all cases and 0.5 billion USD for hard cases. If an outbreak is prevented this saves damages of 100 billion USD.

In Figure D.2 we then show the hypothetical distribution of onsets and interventions over

time if a policymaker had used the advice coming out of the model with the four optimal cutoffs derived from Figure D.1. The grey bars in the figure indicate false positives, i.e. interventions without an onset. The white bars indicate false negatives and the black bars true positives. Precision here can be grasped by comparing the black bars to the grey bars. The true positive rate is given by comparing the white bars to the black bars.

Fig. D.2: Simulating Timing and Frequencies of Interventions Using our Model



Note: The predictions underlying the figure are based on a model using 30 topics and token counts as well as 4 dummies capturing time passed since the last conflict and a dummy for the presence of lower levels of violence. Hard cases are defined as not having had conflict in 10 years. The cutoff for interventions is chosen to minimize costs as displayed in Figure D.1. The grey bars indicate the number of false positives (interventions without onsets). The white bars indicate false negatives (onsets without interventions) and the black bars true positives (onsets with interventions).

The different policy prescriptions for all and hard onsets becomes immediately clear from Figure D.2. To minimize costs, the policymaker would intervene before most onsets of conflict if she applied the model to all onsets. For any violence this would lead to over 40 interventions

in some quarters, i.e. an extremely high level of activity. This high number of interventions implies that only a few outbreaks of violence are not covered by an intervention. In this way, the policymaker would intervene in about 10 situations which would otherwise have escalated for sure within a year. Interestingly, the number of interventions is trending downwards which is in line with the fact that there tend to be less outbreaks in later years. Remember, that these are effectively all interventions in post-conflict situations which is in line with the idea that there is a strong incentive to intervene in a situation in which conflict risks are obvious.

For hard cases interventions would be much less common and the share of onsets that would be covered by intervention falls dramatically. This effect is particularly strong for any violence onsets with less than ten interventions in most quarters. Around 2010, for example, no interventions would be conducted. This is a result of the worse forecastability of hard onsets and of any violence in particular. This means that the policymaker is too conservative to intervene, i.e. there is no situation that looks threatening enough to risk a false positive.

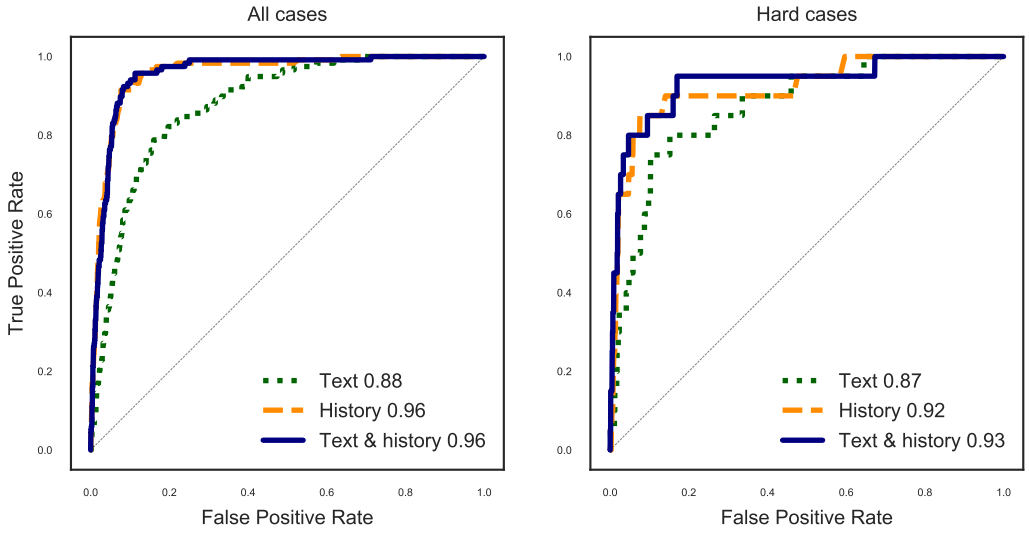
However, for armed conflict we do get a significant level of activity for hard onsets of around ten interventions each quarter. This means that, despite the incredibly low baseline risk of around 1%, intervention becomes feasible. Also, we only reduced intervention costs but gave the hard problem cases no special treatment in terms of prevented costs, V_D, which is not realistic. Figures D.1 and D.2 should therefore be regarded more as description of the existing system of late interventions rather than the description of a system which tries to prevent that countries fall into the conflict trap. Such a system will pay much more attention to interventions post-conflict which is exactly what we see in reality.

E Additional Results

In the following section we show additional results, including for a greater cut-off of 500 battle deaths. In Figure D.3 we see that for the very large cutoff in terms of battle deaths, the onset of

conflict becomes relatively easy to predict. The harder it is to predict conflict, the more topics add to the forecasting power. In particular, when forecasting the hard cases of any violence, the text-only model provides a relatively good forecast given the difficulty of predicting these events. When predicting civil war, the presence of any violence or armed conflict are powerful predictors, even in the hard cases, which is why it is difficult to augment the prediction of further escalation even with text. However, one should note that text alone also achieves high levels of accuracy for all and the hard cases.

Fig. D.3: ROC Curves of Forecasting Civil War

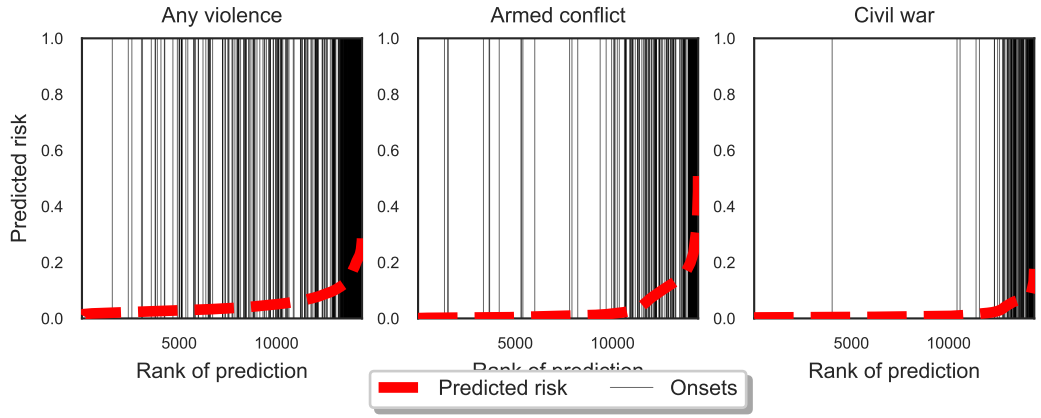


Note: The random forest has a tree depth of 2 and 125 trees. 'Text' contains 30 topics and token counts and 'history' contains 4 dummies capturing time passed since the last conflict and a dummy each for the presence of any violence and armed conflict. Hard cases are defined as not having had civil war in 10 years.

In Figure D.4 we show separation plots for each of the outcomes for predictions using topics and conflict information. The figures order predictions by their rank on the x-axis and plot the predicted level of risk using the red dashed line on the y-axis. The black vertical lines indicate actual onsets. For all outcomes, onsets tend to be bunched on the right side of the panel where the predicted probabilities are highest. But separation plots have the additional advantage of providing an idea of where the model fails to predict conflict. The 5000 lowest risk observations

contain only 19 onsets without clear common features.

Fig. D.4: Separation Plot of Forecasting Violence using Text and Conflict Information

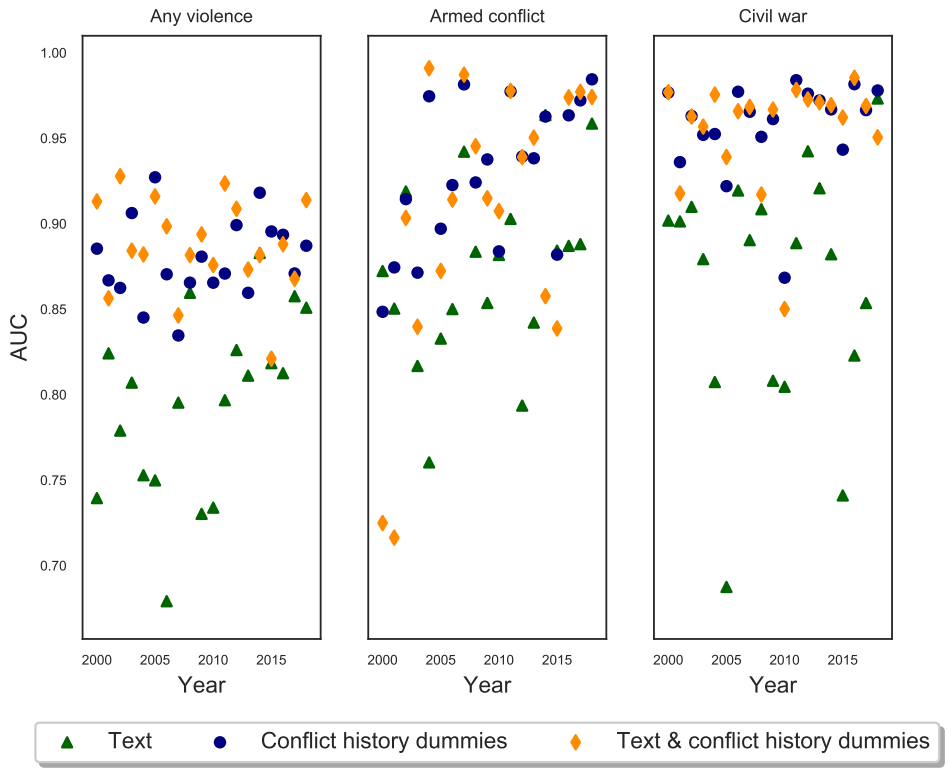


Note: 'Text' contains 30 topics and token counts and 'history' contains 4 dummies capturing time passed since the last conflict and dummies for the presence of lower levels of violence.

In Figure D.5 we show the AUC for ROC curves computed for every single year for each of the three outcomes. We see that the AUCs stay constant or seem to increase slightly over time especially when using text only. We attribute this positive trend to the increase in the training sample over time.

A problem in Figure D.5 is that we have only relatively few onsets, which means that the general trend in model performance is hard to evaluate due to high volatility. In Figure D.6 we therefore show the results of a cross-validation exercise in which we fix the number of trees in the forest but run a gridsearch over the optimal tree depth and record the maximal AUC of this cross validation. The results again suggest a clear upward trend in the cross-validated AUC which is in line with the out-of-sample AUC in Figure D.5. Interestingly, the cross validation AUC is significantly below the true out-of-sample AUC. This is most likely because the folds in the cross validation do not take the panel structure into account and train models on very different parts of the data. Our out-of-sample always uses the most recent past data to predict one quarter or year ahead which is less challenging. The optimal tree depth fluctuates from quarter to quarter but there are no broader trends in the optimal depth.

Fig. D.5: AUC by Year of Forecasting Violence

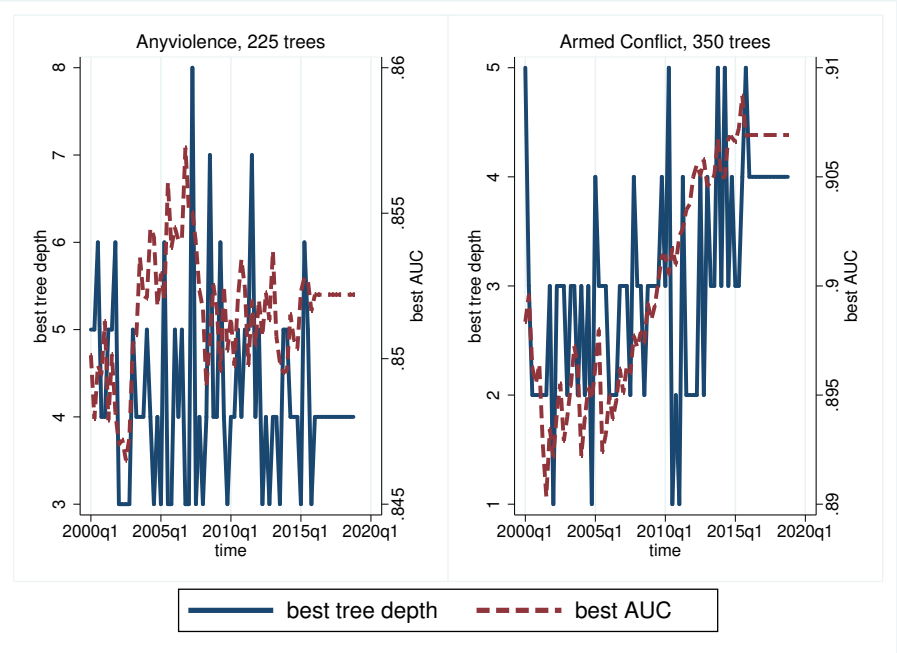


Note: Figures show the overall forecast performance by year.

Overall, Figures D.5 and D.6 provide some evidence against the idea that no generalized forecasting model can be developed as changes in the international context prevent generalization. In both figures forecast performance with text tends to improve with increasing sample size despite a dramatically changing international context and a completely new set of violence onsets.

In Figure D.7 we compare the performance of the topics to events from the Integrated Conflict Early Warning System (ICEWS) database. The ICEWS model we build relies on over 40 event counts. We use 20 counts of all CAMEO event categories that have their target on the territory of the country. We also take the 20 counts involving the government. In addition, we add the average CAMEO scale number of all events. Here, again, we find that topics combined with conflict history dummies perform at least as good the event data combined with conflict history

Fig. D.6: Cross Validated AUC Over Time



Note: Figures show the cross validated AUC as a black solid line and the optimal tree depth as a dashed line. The cross validation sample always runs from 1989Q1 to the time given on the x-axis.

dummies for all cutoffs evaluated. Adding events to topics with conflict info only provides an improvement when forecasting armed conflict. This is interesting because it suggests that ICEWS events provide a good way to capture a situation which might escalate but that the risk of any violence onset is too diffuse to be identified with supervised learning. The unsupervised learning approach we choose instead is more useful here.

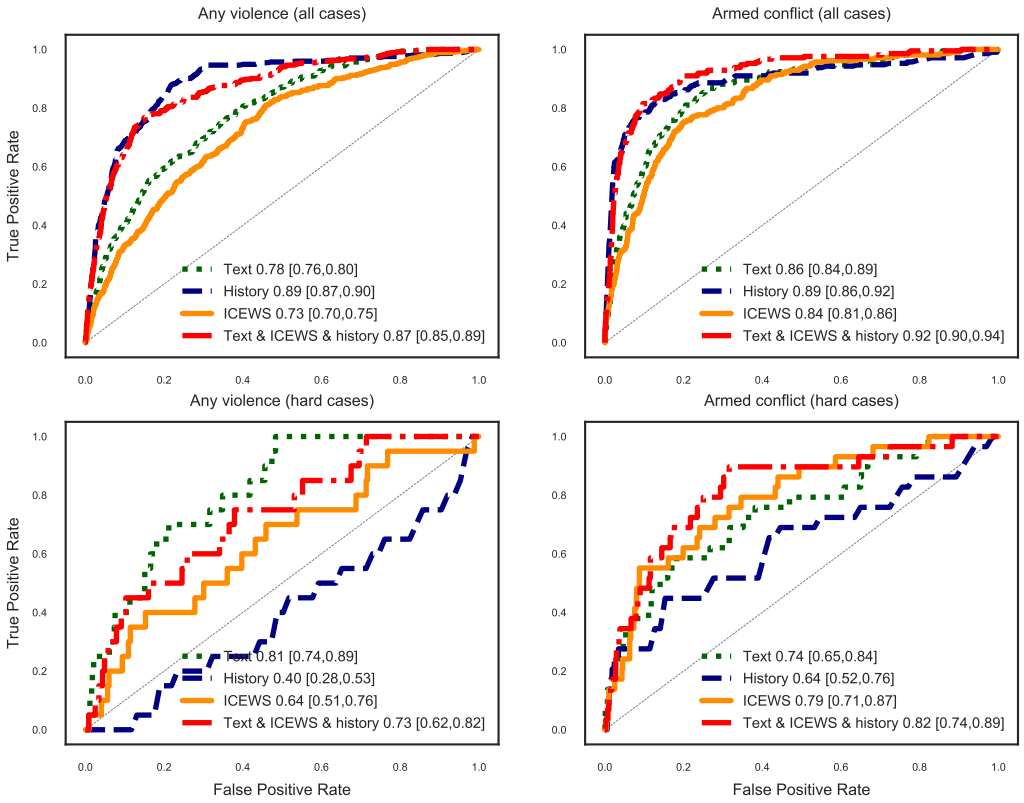
We show the performance of each of these individual prediction models using text only (Figure D.8) and both text and conflict information (Figure D.9). Across most dimensions it seems that the random forest is the algorithm performing best. But it stands out particularly when predicting the hard cases of any violence with conflict history and text. Here the random forest reaches an AUC of 0.81 whereas the logit lasso only reaches and AUC of 0.68. This is consistent with the idea that the random forest receives an advantage because the model is able to use the information contained in the text conditional on conflict history. This is less important when predicting armed conflict as violence escalation is much more important there.

Table D.1: Sets of Predictors

Name	Variables
Topics	Estimated topics using dynamic topic model and total number of tokens
Conflict info	Conflict history and low-level current violence indicators
Standard	Infant mortality, political institutions, share of discriminated population, and neighboring conflicts
ICEWS	Count of all event types, events involving government, all protests, overall average CAMEO code, and average CAMEO code of events involving the government

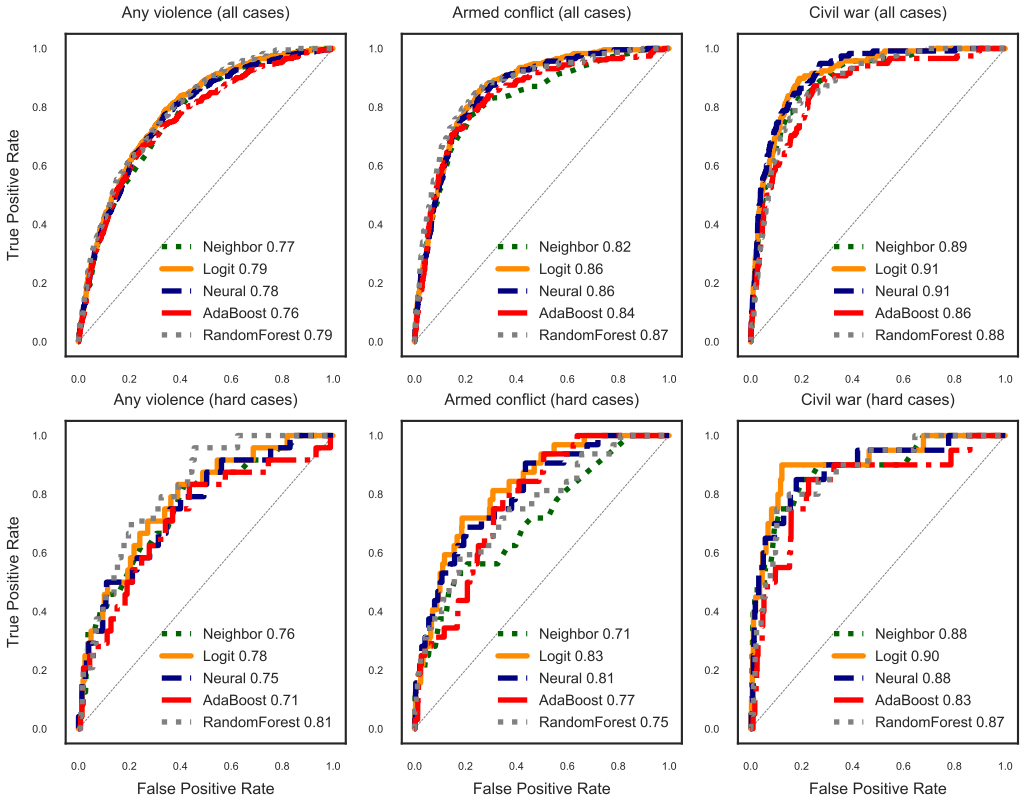
All in all, the Figures paint a consistent picture: Conflict history and present violence are very good predictors of the outbreak of violence. Nonetheless, text summarized by topics adds useful information to predict topics, in particular in countries without current violence or a conflict history.

Fig. D.7: AUC Curves of Forecasting Violence Using Text Compared to ICEWS



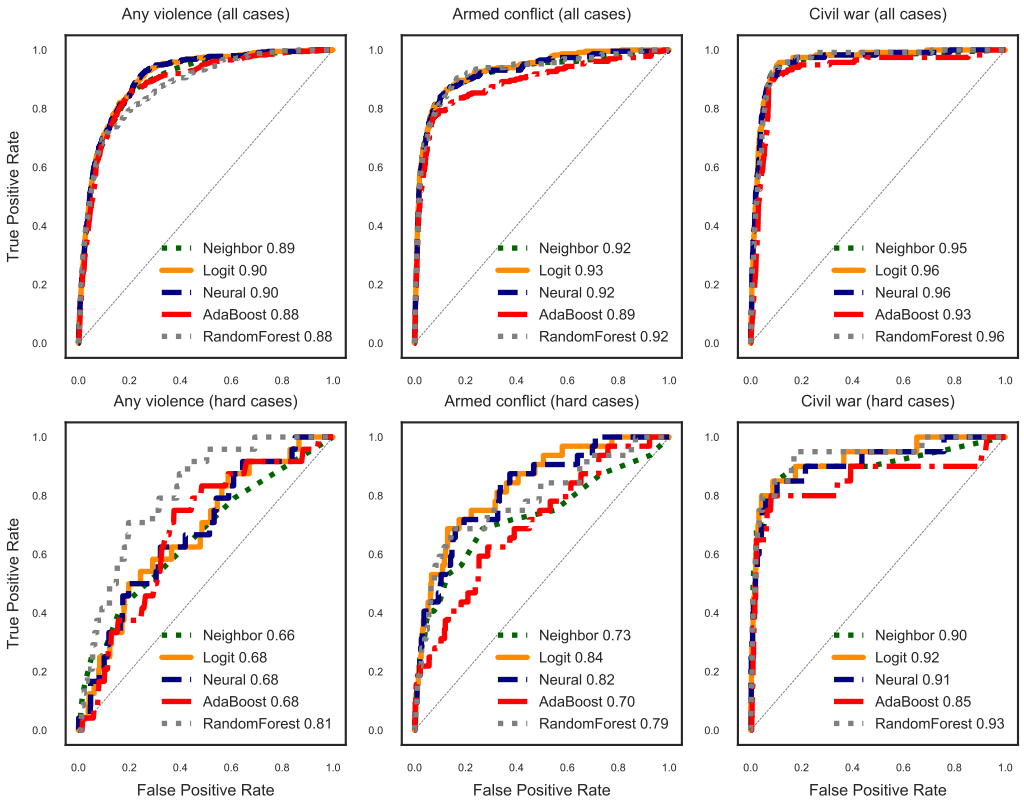
Note: 'Text' contains 30 topics and token counts, 'history' contains 4 dummies capturing time passed since the last conflict and dummies for the presence of lower levels of violence, and 'ICEWS' contains a count of all event types, events involving government, all protests, overall average CAMEO code, and average CAMEO code of events involving the government. Hard cases are defined as not having had conflict in 10 years.

Fig. D.8: ROC Curves of Forecasting Violence with Individual Prediction Models Using Only Text



Note: 'Text' contains 30 topics and token counts. Hard cases are defined as not having had conflict in 10 years.

Fig. D.9: ROC Curves of Forecasting Violence with Individual Prediction Models Using Text and Conflict History Information



Note: 'Text' contains 30 topics and token counts and 'history' contains 4 dummies capturing time passed since the last conflict and dummies for the presence of lower levels of violence. Hard cases are defined as not having had conflict in 10 years.

References

Ahir, Hites, Nicholas Bloom, and Davide Furceri. 2018. "The World Uncertainty Index." *Available at SSRN 3275033*.

Athey, Susan, and Guido W Imbens. 2019. "Machine learning methods that economists should know about." *Annual Review of Economics*, 11: 685–725.

Baker, Scott R, Nicholas Bloom, and Steven J Davis. 2016. "Measuring economic policy uncertainty." *The Quarterly Journal of Economics*, 131(4): 1593–1636.

Baker, Scott R, Nicholas Bloom, Steven J Davis, and Kyle J Kost. 2019. "Policy News and Stock Market Volatility." National Bureau of Economic Research.

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

Bazzi, Samuel, Robert A Blair, Christopher Blattman, Oeindrila Dube, Matthew Gudgeon, and Richard Merton Peck. 2019. "The Promise and Pitfalls of Conflict Prediction: Evidence from Colombia and Indonesia." National Bureau of Economic Research.

Berman, Nicolas, Mathieu Couttenier, Dominic Rohner, and Mathias Thoenig. 2017. "This mine is mine! How minerals fuel conflicts in Africa." *American Economic Review*, 107(6): 1564–1610.

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

Blanchard, Olivier J, and Daniel Leigh. 2013. "Growth forecast errors and fiscal multipliers." *American Economic Review*, 103(3): 117–20.

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

Blumenstock, Joshua, Gabriel Cadamuro, and Robert On. 2015. "Predicting poverty and wealth from mobile phone metadata." *Science*, 350(6264): 1073–1076.

Böhme, Marcus, André Gröger, and Tobias Stöhr. forthcoming. "Searching for a Better Life: Predicting International Migration with Online Search Keywords." *Journal of Development Economics*.

Burke, Marshall, Solomon M Hsiang, and Edward Miguel. 2015. "Climate and conflict." *Annual Reviews of Economics*, 7: 577–617.

Celiku, Bledi, and Aart Kraay. 2017. "Predicting conflict." *The World Bank*.

Chalmers, Malcolm. 2007. "Spending to save? The cost-effectiveness of conflict prevention." *Defence and Peace Economics*, 18(1): 1–23.

Ciccone, Antonio. 2018. "International commodity prices and civil war outbreak: new evidence for Sub-Saharan Africa and beyond."

Collier, Paul, and Nicholas Sambanis. 2002. "Understanding civil war: a new agenda." *Journal of Conflict Resolution*, 46(1): 3–12.

Costinot, Arnaud, Dave Donaldson, and Cory Smith. 2016. "Evolving comparative advantage and the impact of climate change in agricultural markets: Evidence from 1.7 million fields around the world." *Journal of Political Economy*, 124(1): 205–248.

Croicu, Mihai, and Ralph Sundberg. 2017. "UCDP GED Codebook version 18.1." *https://ucdp.uu.se/downloads/*.

Dube, Oeindrila, and Juan F Vargas. 2013. "Commodity price shocks and civil conflict: Evidence from Colombia." *The Review of Economic Studies*, 80(4): 1384–1421.

Elliott, Graham, and Allan Timmermann. 2008. "Economic forecasting." *Journal of Economic Literature*, 46(1): 3–56.

Elliott, Graham, and Allan Timmermann. 2013. *Handbook of economic forecasting*. Elsevier.

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

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy. 2019. "Text as data." *Journal of Economic Literature*, 57(3): 535–74.

Giglio, Stefano, Bryan Kelly, and Seth Pruitt. 2016. "Systemic risk and the macroeconomy: An empirical evaluation." *Journal of Financial Economics*, 119(3): 457–471.

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

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

Hansen, Stephen, Michael McMahon, and Andrea Prat. 2017. "Transparency and deliberation within the FOMC: a computational linguistics approach." *The Quarterly Journal of Economics*, 133(2): 801–870.

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

Jean, Neal, Marshall Burke, Michael Xie, W Matthew Davis, David B Lobell, and Stefano Ermon. 2016. "Combining satellite imagery and machine learning to predict poverty." *Science*, 353(6301): 790–794.

Jurado, Kyle, Sydney C Ludvigson, and Serena Ng. 2015. "Measuring uncertainty." *American Economic Review*, 105(3): 1177–1216.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. "Prediction policy problems." *American Economic Review*, 105(5): 491–95.

Larsen, Vegard H, and Leif A Thorsrud. 2019. "The value of news for economic developments." *Journal of Econometrics*, 210(1): 203–218.

Michalopoulos, Stelios, and Elias Papaioannou. 2016. "The long-run effects of the scramble for Africa." *American Economic Review*, 106(7): 1802–48.

Mueller, Hannes, and Christopher Rauh. 2018. "Reading Between the Lines: Prediction of Political Violence Using Newspaper Text." *American Political Science Review*, 112(2): 358–375.

Mullainathan, Sendhil, and Jann Spiess. 2017. "Machine learning: an applied econometric approach." *Journal of Economic Perspectives*, 31(2): 87–106.

Mwangi, Benson, Tian Siva Tian, and Jair C Soares. 2014. "A review of feature reduction techniques in neuroimaging." *Neuroinformatics*, 12(2): 229–244.

OECD. 2018. "States of Fragility 2018." OECD Publishing, Paris.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. "Scikit-learn: Machine learning in Python." *Journal of Machine Learning Research*, 12(Oct): 2825–2830.

Rauh, Christopher. 2019. "Measuring uncertainty at the regional level using newspaper text."

Řehůřek, Radim, and Petr Sojka. 2010. "Software Framework for Topic Modelling with Large Corpora." 45–50. Valletta, Malta:ELRA. `http://is.muni.cz/publication/884893/en`.

Rossi, Barbara, and Tatevik Sekhposyan. 2015. "Macroeconomic uncertainty indices based on nowcast and forecast error distributions." *American Economic Review*, 105(5): 650–55.

Stock, James H, and Mark W Watson. 2006. "Forecasting with many predictors." *Handbook of economic forecasting*, 1: 515–554.

Sundberg, Ralph, and Erik Melander. 2013. "Introducing the UCDP georeferenced event dataset." *Journal of Peace Research*, 50(4): 523–532.

Svensson, Lars EO. 2017. "Cost-benefit analysis of leaning against the wind." *Journal of Monetary Economics*, 90: 193–213.

Tanaka, Mari, Nicholas Bloom, Joel M David, and Maiko Koga. 2019. "Firm Performance and Macro Forecast Accuracy." *Journal of Monetary Economics*.

Timmermann, Allan. 2006. "Forecast combinations." *Handbook of economic forecasting*, 1: 135–196.

United Nations and World Bank. 2017. "Pathways for Peace: Inclusive Approaches to Preventing Violent Conflict-Main Messages and Emerging Policy Directions." *World Bank, Washington*.

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.

World Bank. 2017*a*. "The Toll of War : The Economic and Social Consequences of the Conflict in Syria." World Bank, Washington, DC.

World Bank. 2017*b*. "United Nations and World Bank leaders call for stronger international efforts to prevent violent conflict." Press release.