UNIVERSITY OF CAMBRIDGE
Faculty of Economics

Institute for New Economic Thinking

CAMBRIDGE WORKING PAPERS IN ECONOMICS
CAMBRIDGE-INET WORKING PAPERS

The Hard Problem of Prediction for Conflict Prevention

Hannes Mueller
Institut d'Analisi Economica

Christopher Rauh
University of Cambridge

Abstract

In this article we propose a framework to tackle conflict prevention, an issue which has received interest in several policy areas. A key challenge of conflict forecasting for prevention is that outbreaks of conflict in previously peaceful countries are rare events and therefore hard to predict. To make progress in this hard problem, this project summarizes more than four million newspaper articles using a topic model. The topics are then fed into a random forest to predict conflict risk, which is then integrated into a simple static framework in which a decision maker decides on the optimal number of interventions to minimize the total cost of conflict and intervention. According to the stylized model, cost savings compared to not intervening pre-conflict are over US$1 trillion even with relatively ineffective interventions, and US$13 trillion with effective interventions.

Reference Details

2103 Cambridge Working Papers in Economics
2021/02 Cambridge-INET Working Paper Series

Published 6 January 2021
Revised 22 February 2022

JEL Codes F21, C53, C55

Websites www.econ.cam.ac.uk/cwpe
www.inet.econ.cam.ac.uk/working-papers

THE HARD PROBLEM OF PREDICTION FOR CONFLICT PREVENTION

Hannes Mueller
Institut d'Analisi Economica (CSIC), BSE, MOVE and CEPR

Christopher Rauh
University of Cambridge, Trinity College Cambridge, and CEPR

Abstract
In this article we propose a framework to tackle conflict prevention, an issue which has received interest in several policy areas. A key challenge of conflict forecasting for prevention is that outbreaks of conflict in previously peaceful countries are rare events and therefore hard to predict. To make progress in this *hard problem*, this project summarizes more than four million newspaper articles using a topic model. The topics are then fed into a random forest to predict conflict risk, which is then integrated into a simple static framework in which a decision maker decides on the optimal number of interventions to minimize the total cost of conflict and intervention. According to the stylized model, cost savings compared to not intervening pre-conflict are over US$1 trillion even with relatively ineffective interventions, and US$13 trillion with effective interventions. (JEL: F21, C53, C55)

Acknowledgments: 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 SE, 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, the Barcelona workshops on conflict prediction, and Heidelberg University. Mueller acknowledges financial support from the Banco de España, the Spanish Ministry of Economy and Competitiveness, through the Severo Ochoa Programme for Centres of Excellence in R&D (CEX2019-000915-S) and the Ayudas Fundación BBVA. The authors declare that they have no competing financial interests. All errors are ours.

E-mail: hannes.mueller@iae.csic.es (Mueller); cr542@cam.ac.uk (Rauh)

1. Introduction

Civil wars are a serious humanitarian and economic problem. According to data from the United Nations Refugee Agency (UNHCR) by mid-2021 over 84 million people around the world have been forced to flee their homes, 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 2017b): *"[...], we need to do more early on to ensure that development programs and policies are focused on successful prevention."*

However, developing an 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 an epidemic; 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). 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. In other words, we need to combine an understanding of causal factors with good predictions of what will happen. If it is impossible to predict where conflict will break out then costly 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 at the monthly level in over 190 countries. Text has the huge advantage of being updated daily and can therefore be used to implement an early warning system.[1]

Second, we introduce a conceptual framework that allows for the integration of 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. A key margin of adjustment in prevention then 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 the forecaster is the low baseline risk, i.e. there are many more zeros, peaceful months, than ones, onsets of conflict. This problem, referred to as imbalanced classes, makes it hard to train an algorithm to predict since there are relatively few situations to learn from. In conflict prevention the problem is

1. To illustrate this, we provide a monthly updated forecast publicly since early 2021 on the webpage www.conflictforecast.org.

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. However, if we want to prevent countries from falling into the conflict-trap 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, outbreaks of violence in countries without a recent history of violence. We find that conflict history is an excellent predictor in countries with recent violence, but text is able to forecast the hard problem cases. Other forecasting studies have encountered the same problem. In their detailed study of Colombia and Indonesia, Bazzi et al. (2019) conclude, for example, that *"Even with such unusually rich data, however, the models poorly predict new outbreaks or escalations of violence."* Whereas the conflict trap is well-known in the literature this has not led to a separate study of hard problem cases and so we do not know whether existing empirical findings are determined by the majority of cases in which conflict simply followed a previous conflict. This is exactly where we break ground by explicitly studying the role played by conflict history in 190 countries and by exploiting news text to go beyond it.

The forecasting procedure is a pipeline composed of two parts. First, we summarize the vast amounts of newspaper articles dating back to 1980 into 30 topics using the Latent Dirichlet Allocation (LDA). We then collapse the resulting topics at the country/month level, with the idea being that this step provides us with the distribution of news summarized into a few interpretable topics about each country across time. Second, we feed these topics together with a small number of variables capturing past violence into a random forest prediction model.[2] We train and test the 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.

As a methodology, the random forest has two advantages. First, a random forest is a collection of many individual decision trees. Therefore, it provides a prediction by averaging across many predictors, a prediction method referred to as an ensemble, which has the advantage that it tends to average out idiosyncratic errors. Moreover, from each decision tree to the next, the training sample is bootstrapped which avoids overfitting to particular situation or noise, a common pitfall when predicting conflict. Second, we show 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 exploiting news text increasingly at the bottom nodes. We find, for example, that topics which are not directly related to violence and negatively associated with risk, such as trade, international relations, business or economics, are increasingly featured in lower branches of the tree.

To evaluate our forecast we provide an intervention framework which is explicitly based on the idea of a policymaker who minimizes total costs of conflict. The idea is that based on predictions and their accuracy, a policymaker can decide to intervene, which is costly and may or may not succeed at avoiding the outbreak of a conflict. As opposed to Kleinberg et al. (2015), we approach the prediction policy problem within a machine learning classification framework which allows us to 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 far beyond the field of conflict research and in applications outside economics where the policymaker

2. Other prediction methods, such as logit lasso or adaptive boosting, also perform well when combining history, i.e. information about past violence, and text. We refer to Appendix E for an extensive comparison.

faces runaway dynamics as in financial crisis, climate change, crime, or epidemics.[3] The integration into a forecast framework allows us to study the trade-offs between forecast precision, the cost and effectiveness of interventions, and the cost of failing to prevent. If conflict is regarded as more costly, for example, then more situations should be classified as positives and there should be more interventions. As a result, the forecast will become less conservative. If forecasting a conflict is difficult or interventions are ineffective then the optimal forecast framework should produce more negatives so that interventions become less common.

There is a long tradition 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.[4] However, for conflict the economics literature has mostly focused on understanding causal mechanisms.[5] As a consequence, the literature has made huge strides in understanding the causes of conflict.[6] However, these efforts are often not effective at forecasting as variables identified to be causal do not have to be, and typically are not, good predictors of conflict.[7] We illustrate this here by showing that variation in commodity prices of 50+ commodities and their export weights do not provide a good forecast for conflict outbreaks despite the fact that they are central to recent causal work on the drivers of conflict.

Most recently, the conflict literature is moving towards formulating policy recommendations for the prevention of conflict and contemplating the effectiveness of potential interventions.[8] It is in this context that the prediction part of the prevention problem becomes a crucial ingredient, and our main contribution lies in providing a framework that is maximizing predictive power. However, we also contribute by providing a first, simple framework that allows both the treatment effect size and forecast precision to be traded off against each other. We illustrate this by showing how a fall in the effectiveness of interventions would affect the optimal use of the forecasts.

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.[9] In our application, the side-product of the forecast is a monthly conflict risk measure for nearly 200 countries for the period 01/2005 to 08/2020.[10] 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, the predictions also inform us about risk levels of countries that did not

3. See Svensson (2017) for similar arguments in a more standard macro framework.

4. For an overview over the more classic literature see Timmermann (2006) and Elliott and Timmermann (2008, 2013). For recent efforts see Böhme et al. (forthcoming), Giglio et al. (2016), Costinot et al. (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 et al. 2015).

5. Two exceptions are Celiku and Kraay (2017) and Bazzi et al. (2019). For summaries of the political science literature see Hegre et al. (2017).

6. For an overview of the earlier literature see Blattman and Miguel (2010). For recent contributions in economics see Besley and Persson (2011); Esteban et al. (2012); Dube and Vargas (2013); Bazzi and Blattman (2014); Berman and Couttenier (2015); Burke et al. (2015); Michalopoulos and Papaioannou (2016); Berman et al. (2017, 2021).

7. See, for example, Ward et al. (2010); Mueller and Rauh (2018).

8. See Blattman and Annan (2015); Hörner et al. (2015); Blattman et al. (2017); Meirowitz et al. (2019); Rohner and Thoenig (2020).

9. Take, for example, Blanchard and Leigh (2013), Jurado et al. (2015), Rossi and Sekhposyan (2015) and Tanaka et al. (2019).

10. In Appendix E we show that performance stays very similar when predicting conflict outbreaks across larger time horizons, within 3 and 12 months, or larger outbreaks of at last 50 fatalities.

see outbreaks of violence, and could therefore provide propensity scores for cross-country comparisons and help study the role of stabilizing factors.

There is a growing interest in the use of text to generate data, i.e. feature extraction.[11] Baker et al. (2016) and Ahir et al. (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.[12] For our feature extraction we rely on the full text but reduce the dimensionality through the LDA topic model (Blei et al. 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 news-text corpus.[13] We find both positive and negative relationships with conflict risk in the topics. And, perhaps surprisingly, predictive power is also derived from topics reducing their share before conflict breaks out.

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.[14] Regularization is necessary due to 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 with unstructured text because of the vast dimensionality due to the quantity of words which would lead to overfitting and computational problems. One way forward in a situation requiring dimensionality production is to use theory to build priors regarding the variables and models to use. An alternative, which we follow, 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 applications with few positive events to train the model (for an example in medical research see Mwangi et al. 2014). Our results suggest that unsupervised learning can give context to even the most dramatic changes in news reporting and make the forecast more robust in this way.

This project advances on several fronts beyond other and our previous work in Mueller and Rauh (2018). We introduce a new full-text archive of more than 4 million newspaper articles dense enough to summarize topics at the monthly level for 190 countries. In dealing with this amount of heterogenous data, 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 month in our test sample. This 12-fold increase in sample size allows us to introduce a step of non-linear supervised machine learning (random forest) at the country/month level. The use of a random forest model instead of a linear regression model provides huge advantages as it allows us to capture conflict history through a dynamic non-linear model instead of fixed effects.[15] This explicit model of conflict history as the key risk factor means we can develop and evaluate our forecast model conditional on the conflict history. The combination of a framework of optimal policymaking and monthly updates makes our method particularly useful

11. See Gentzkow et al. (2019) for an overview.

12. In a similar fashion Baker et al. (2019) capture equity market volatility.

13. This feature mirrors findings in Rauh (2019) and Larsen and Thorsrud (2019) who forecast economic activity, and Hansen et al. (2017) who study the effect of increased transparency on debate in central banks.

14. Systematic regularization is still not common practice even in research which clearly aims to provide predictions.

15. While Bazzi et al. (2019) also experiment with different supervised machine learning algorithms to predict conflict, their rich data are only available for two countries, and many predictors are only observed triennially and the data availability ends in 2014.

for actual policy applications that rely on timely risk updates using vast amounts of text. To the best of our knowledge, we are also the first to propose a framework to evaluate optimal intervention policies based on such forecasting data.

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 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. Finally, in Section 5 we integrate this forecast into a cost-minimization problem and 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/month level. We focus on the monthly level to have more cases to train but it is straightforward to provide forecasts at any larger time-horizon, for example, a year ahead.

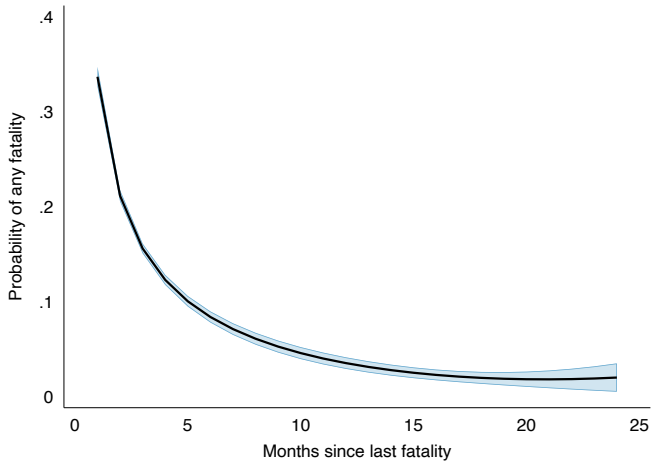
The data offers three types of conflict fatalities which we all add into a single number. This implies that we mix terror attacks and more standard, two-sided violence. An important question arises due to the fact that zeros are not coded in the data. We allocate a zero to all country/months in which the country was independent and where data from GED is available.

We use the simplest possible definition of conflict - the presence or absence of any fatality related to armed conflict.[16] Importantly, we only consider onset, i.e. only the month before conflict breaks out. Subsequent months of ongoing conflict are set to missing. This is important as predicting outbreaks is much harder than predicting conflict incidence. In our data we have 2,112 onsets of conflict according to this definition.

The hard problem can be understood through a simple figure which illustrates the extremely high risk of onset post-conflict and the low-baseline risk otherwise. In Figure 1 we plot the likelihood in our sample that a conflict breaks out for the months after the end of the previous conflict episode. This shows that the risk of a renewed onset of conflict is above 30% right after conflict. Conflict risk falls continuously thereafter but remains substantial in the years following conflict. Almost 90% of all conflict onsets happen within two years of the previous conflict. This is what the conflict literature has dubbed the *conflict trap* or *war trap*. Countries get caught in cycles of repeating violence. Beyond the first ten years after conflict, the risk of conflict is very low. After two years, the baseline risk of conflict outbreaks is 0.5% and outside the ten year period the baseline risk of conflict onsets is only around 0.2%. However, most countries are at peace for long periods, and many outbreaks, therefore, take place long after conflict subsided, even if the baseline risk is extremely small.

16. In the Appendix E we also look at performance when predicting an armed conflict with at least 50 fatalities per month. Here we add the current number of fatalities as a predictor as low levels of violence are a strong indicator of an eminent outbreak.

FIGURE 1. Likelihood of conflict relapse post conflict



Notes: Figure shows the likelihood of conflict relapse after violence ended (at 0) conditional on remaining in peace approximated using a fractional polynomial. Any violence is a month with any fatalities. The shaded area is the 95% confidence interval.

In summary, 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. In our analysis of forecasts we will take an extreme view and try to forecast conflict for cases outside the ten year period. We call this 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 will have huge benefits in the long run. When a country experiences an outbreak of violence in a hard-to-predict scenario it is in danger of falling into a conflict trap. Prevention efforts in hard problem cases, which manage to stabilize countries, will therefore have considerable dynamic payoffs. Of course, these efforts need to be based on an understanding of the causal mechanisms behind conflict to actually change the trajectory of a country.[17]

We take an extreme stance here by focusing on the tail risks in countries that did not experience a conflict in the previous ten years. If these cases can be predicted sufficiently well to prevent them, it will be even easier to predict conflict in cases with a more recent conflict history.

17. As the quote from the introduction makes clear, prevention is also of interest for the international community. All big international organizations 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. Forecast and Test Methodology

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.

3.1. Text Data

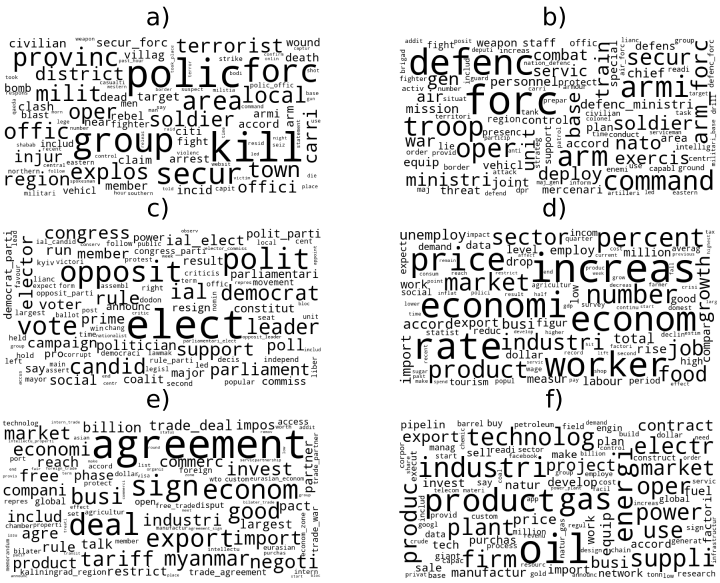
As a basis of our method we use a new unique corpus of over 4 million documents from three newspapers (New York Times, Washington Post and the Economist) and two news aggregators (BBC Monitor and LatinNews). All sources except for LatinNews are downloaded from LexisNexis. A text is downloaded if a country or capital name appears in the title or lead paragraph. We apply standard text mining steps to the text by removing punctuation, stop words and lemmatizing. In addition to single words, we also consider common combinations of two and three words. The resulting data is described in detail in Appendix A. Using the dynamic topic model we derive topic models with $K = 7$, 15 and 30 topics.[18] The reason we choose relatively few topics is to avoid topics adapting to particularly newsworthy cases of conflict, regions or countries. The topic models we estimate tend to contain topics that can be attributed to generic content like politics or economics. A detailed discussion of topics and case studies are presented in Appendix B.

Figure 2 shows word clouds for six out of 30 topics estimated on the 2020m8 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 terror topic. It features terms such as " kill" ,"terrorist" , and "explos" prominently. Panel b) displays a topic which features mostly terms related to the military such as "troop", "armi", and "deploy". Panel c) features expressions related to politics such as "elect", "opposit", and "vote", while d) appears to reflect economic activity, with terms such as " economi", "worker" and "price" being keywords. In Panel e) we see tokens such as "agreement" , "export" , and "tariff" (trade), while Panel f) exhibits "industri" , "product" , and "market" (industry).

18. We need to re-estimate the topic model every month so that it takes around two weeks to estimate all vintages of the 30 topics model.

FIGURE 2. Word clouds of topics



Notes: The word clouds represent the most likely terms of 6 out of 30 topics estimated using all text until $2020m8$. 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 Terror topic, b) Military, c) Politics, d) Economics , e) Trade, and f) Industry.

One potential problem with relying on features generated from news text is that we inherit the news bias of the text. A particularly worrying bias is that autocracies might restrict reporting and, in this way, our forecast would fail for these political regimes. We discuss this possibility in detail in Appendix Section E.1. Our take-away from this is that there is no obvious failure of our model when predicting hard onsets.

3.2. Prediction Method

With the estimated topic model we then calculate the share of topics for all countries in each month between $1989m1$ and T. We then use these topic shares and a set of variables 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 topic model for country i in month t in a vector, $\boldsymbol{\theta}_{it}$ and conflict history in a vector $\mathbf{h}_{it}$. We then train a forecasting model with all data available in T using the model

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

where y_{it+1} is the *onset* of conflict in month $t+1$, $\mathbf{h}_{it}$is a vector of three variables capturing post-conflict dynamics. Specifically, we use the number of months since the last month with any fatality, with more than 50 fatalities and with more than 500 fatalities. The vector $\boldsymbol{\theta}_{it}$ includes the K topic shares, a set of K topic stock variables which capture a discounting stock of previous news topics, and the log of the word count written on country i in month t.

Note, that our model uses only relatively few variables to build the baseline model. Our benchmark, the *history model*, $\mathbf{h}_{it}$ consists of just three variables. The

text model, $\boldsymbol{\theta}_{it}$, adds 61 variables which still leads to a relatively small model of 64 variables when we combine both. We do not follow the typical machine learning approach of allowing the model to do the regularization for the benchmark model as the small data size will immediately lead to lower performance through overfitting.

The role of machine learning in this step is to build a highly non-linear model $F_T(\mathbf{h}_{it}, \boldsymbol{\theta}_{it})$ and at the same time to discipline the regularization of this model. Importantly, the random forest model we use is able to extract a non-linear risk pattern as in Figure 1 from the simple counts used in $\mathbf{h}_{it}$. We fix the hyperparameters in the first sample (1989m1-2005m1) 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 $\boldsymbol{\theta}_{it}$ available until $T-1$.

With the resulting model we then produce predicted out of sample values

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

which we compare to the true values y_{iT+1}. We then update our topic model with the news written in the next month, add the new information on conflicts, retrain the prediction model, and predict the probabilities of outbreaks in the following month. For testing, we thereby produce sequential out-of-sample forecasts, $\hat{y}_{iT+1}$, for $T+1 = 2005m2, 2005m3, ..., 2020m8$. 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, random 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 $2005m1$. 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. Forecast Results

4.1. Prediction of Conflict Outbreaks in Hard Cases

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 literature forecasting conflict this is done by picking a cutoff c and discretizing using the condition

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

An increase in 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 with explicit cost-weights on these different outcomes can provide interesting insights into the distribution of $\hat{y}_{iT+1}$. But for now we stick to standard ways of displaying the trade-off between false positives and false negatives.

Receiver operating 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.

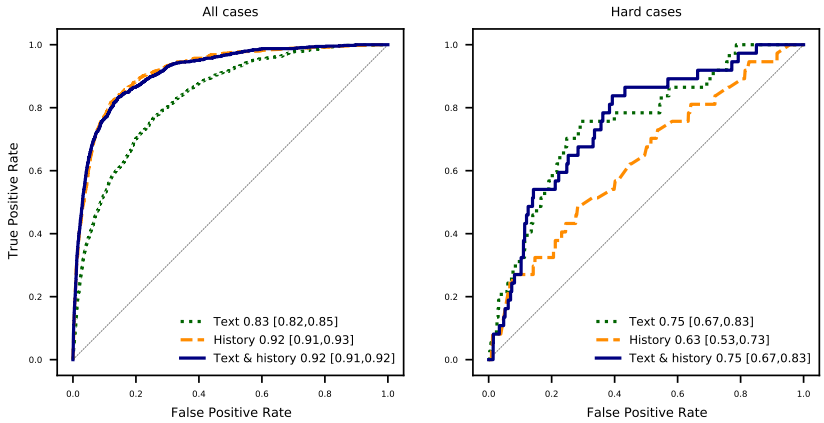
A high cutoff c represents a very conservative forecast which warns only of few onsets, tends to get these cases right (low FPR) but misses a lot of actual onsets (low TPR). Lowering the cutoff c will typically increase both the FPR and the TPR. If the TPR increases by more than the FPR with falling c the ROC rises above the 45 degree line which indicates a better-than-random forecast.

The ROC curve expresses the possibility frontier of the forecasting model. The area under the curve (AUC) of the ROC curve is therefore often used to evaluate forecasting models. This makes sense in settings when the relative costs of false negatives and false positives is not known. We will stick to the AUC as our performance measure for now. However, the ROC curve is not sufficient to solve the prevention policy problem. A policymaker who is very afraid of not being able to intervene before conflicts, for example, will put a high cost to false negatives which is not reflected in the ROC curve. In Section 5 we therefore turn to a much more specific interpretation of the policy task.

Figure 3 shows ROC curves for conflict for the entire pool of out-of-sample predictions. 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 previous violence, which is labeled as *history*, and (iii) a model that draws from both *text & history*.

The *text* model includes the number of tokens as well as 30 topic shares and stocks. Stocks are computed as the discounted sum of topic shares of previous months, with an (arbitrarily) chosen discount factor of 0.8. The idea is that not only immediate events might be of importance, but also the tendencies a country has been experiencing over the past months. Our *history* model consists of three counts of the time that has passed since the last conflict month, the last conflict month with at least 50 fatalities and the last conflict month with at least 500 fatalities. The random forest uses these variables repeatedly to generate a non-linear approximation of the risk as it is shown in Figure 1.

FIGURE 3. ROC curves of forecasting any violence for all onsets (left) and hard onsets (right)



Notes: The prediction method is a random forest. 'Text' contains 30 topic shares and stocks as well as token counts and 'history' contains three variables: the time that has passed since the last conflict month, the last conflict month with at least 50 fatalities, and the last conflict month with at least 500 fatalities. Left and right panels show alternative evaluations of the same forecasting model. The left ROC curves show all cases. Hard cases are shown to the right. Hard cases are defined as not suffering fatalities in ten years. The numbers in the legends represent the respective area under curve (AUC) with bootstrapped 95% confidence intervals in square brackets.

On the left of Figure 3 we see the overall performance of all three sets of predictors when forecasting conflict. *Text* alone (dotted line) provides some forecasting power and this forecast is comparable to what is common in the literature when predicting at the monthly level - an AUC of 0.83. But it is clear that the information on conflict *history* (dashed line), a non-linear model based on only three variables, dominates the text forecast. The model using *text & history* also reaches an AUC of 0.92 which is the same as from *history* alone. We generated bootstrapped confidence intervals for the AUC but, even at the lower bound of these confidence intervals, the AUCs of the combined and *history* models are relatively high.

On the right panel of Figure 3 we evaluate our forecasting models on the hard problem cases, i.e. when a country has not experienced a battle death for at least ten years. 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. The conflict information model (dashed line) is now much less able to provide a useful forecast, i.e. it is only slightly better than random. Text, however, still provides useful forecasting power and the AUC of 0.75 of the combined model (solid line) now comes almost entirely from the text. 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.

Using ROC curves and the AUC as a measure of fit has disadvantages in unbalanced samples like ours. However, the TPR/FPR-space provides a useful standard in which we integrate the decision making problem in Section 5. We will show that ROC curves can be understood as a possibility frontier which captures the ability of the forecast model to reduce the impact of armed conflict. These possibility frontiers can then be directly related to isocost curves, shown in Figure 8, which capture the costs generated by the imbalance in the sample.

The combined model, *text & history*, is made a lot more robust by the newspaper text. As the baseline risk from a recent conflict fades, the role of text becomes more

and more important. In Figure 4 we show how the AUCs of our three models behave with fading information from past conflict. Performance of all models decreases as we evaluate on cases with less and less recent conflict. However, the performance of the history model drops most dramatically as the memory of conflict fades whereas the performance of the *text* model falls more slowly and stabilizes for cases without a recent conflict history. Strikingly, for onsets that occur more than 60 months after the previous conflict, the text-only model provides a better forecast than conflict history. Accordingly, the combined model (solid line) is able to forecast hard problem cases.

FIGURE 4. AUC scores with fading conflict history

Notes: Figure shows the AUC for onsets that happened in the aftermath of a previous conflict episode. Onsets that occur within a window of months displayed on the x-axis are excluded in the evaluation. At a value of 20, for example, all outbreaks are excluded that occur within 20 months of the previous conflict episode. The prediction method is a random forest. 'Text' contains 30 topic shares and stocks as well as token counts and 'history' contains three variables: the time that has passed since the last conflict month, the last conflict month with at least 50 fatalities, and the last conflict month with at least 500 fatalities. Left and right panels show alternative evaluations of the same forecasting model. All three lines show evaluations of the same forecasting models in changing samples.

In Appendix E we show additional results and discuss extensively the performance when looking at different evaluation metrics, types of violence, prediction horizon, and algorithms. When looking at the predictive performance in terms the Matthews Correlation Coefficient (MCC), we find similar patterns, as in *text & history* provide the most predictive power across all cases, but when looking at hard cases we need *text* in order to maintain some predictive power. Our approach proves to be robust to splitting the types of fatalities into non-state, one-sided, and state-based violence. The forecast model is also able to predict violence one-quarter and one-year ahead with very similar predictive performance. The year-ahead *text* model, in particular, performs extremely well with an AUC of 0.91 across all cases and 0.73 in hard cases which implies the model can provide useful early warning even for outbreaks that are occur further into the future. We also show that model performance is not specific to 30 topics. The model performs similarly well for 7, 15, and 30 topics. Finally, we explore the role of different forecast algorithms and show that random forests perform best. In the next section, we turn towards the role

of the random forests in explaining the ability of our methodology to pick up signals for hard problem cases.

4.2. Expanding the Model to Improve the Forecast

An important question is how much improvement is possible by adding more variables to our framework. We have experimented extensively but have not managed to improve the model in a robust way. We illustrate this using two sets of predictors from the economics and forecasting literature. Some representative results are shown in Appendix Figures E.9, E.10, and E.11.

Following up on a large literature that tries to understand commodity as a driver of conflict we use a dataset of 50+ commodity prices which we combine with commodity export weights from Bazzi and Blattman (2014) to generate measures for commodity income shocks. The price data is updated on a monthly basis by the World Bank and could therefore provide a good basis for forecasts.[19] Moreover, we experimented with generating more aggregate revenue features. We tried using a commodity index by extracting the first factor of all the commodity prices, aggregating all weights and prices to one index and to several sub-indexes to capture shocks in minerals, agricultural products, and energy revenue. Adding these variables to our model does not improve the forecast of our main model and there is little indication that commodity shocks can provide useful forecasts by themselves. When predicting all onsets, the commodity price models fluctuate around an AUC of just 0.6, and when predicting hard onsets the AUC is around 0.5, i.e. there is no evidence that commodity price features can help us predict new outbreaks in countries without a conflict history.

This is surprising given the huge attention paid to variation in commodity prices in the conflict literature. There are many potential reasons for this failure. First, the complete failure in hard cases suggests that commodity shocks are mostly relevant for reinforcing existing conflicts rather than making them break-out in the first place. Second, resource revenue shocks might lead to a re-distribution of violence within countries rather than increasing or decreasing it at the macro level. Finally, significance tests have a very different goal than predictive exercises and are typically conducted without out-of-sample checks. This means that models based on causally identified tests of hypothesis are not necessarily good at providing models for prediction (Ward et al. 2010; Lo et al. 2015). However, our results raise the question whether the discussion of causes of conflict should not, at least partially, take explanatory power and out-of-sample performance into account.

We also compare our approach to a standard set of predictors based on Goldstone et al. (2010) which includes political institution dummies based on various dimensions of the Polity IV data, number of neighboring conflicts from UCDP, and data on child mortality from the World Bank. Many standard variables such as infant mortality are not available for as many countries and years. For the sake of comparability, we therefore only use overlapping predictions for evaluation, i.e. country-months in which the availability of data allow predictions for both sets of variables. The results are more encouraging here but it remains impossible to improve the forecast of the full model of *text & history*. In addition, the delay in publishing many of the macro variables used in this model would imply that it cannot rely on up-to-date data when forecasting the future in applications.[20] Adding

19. We fix the weights following Ciccone (2018) who finds a much stronger impact of commodity-price shocks on conflict using fixed weights.

20. It is even more problematic that many macro variables get re-coded over time which means some of the variables in the standard model use future information.

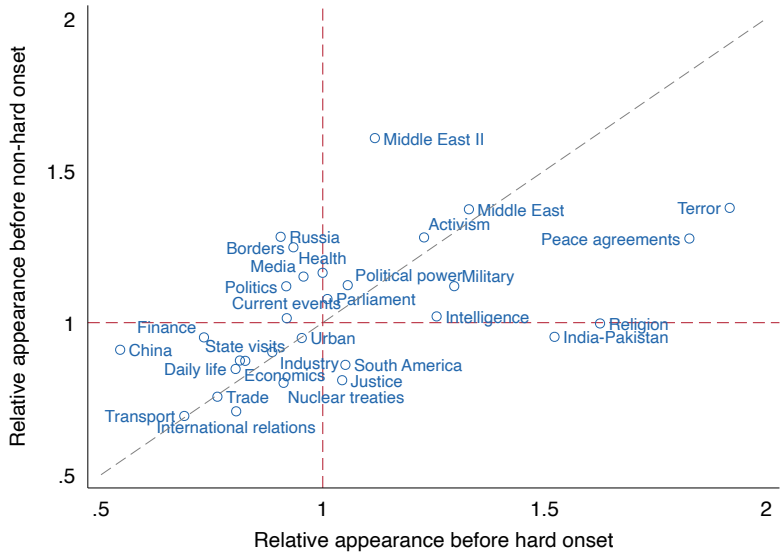
variables to the *text & history* would therefore come at the price of reducing the sample and speed of updates.

Relying on text to forecast armed conflict has its limits as reporting could be affected by media freedom or the interest and power of the involved conflict parties. We show in Appendix Table E.1 that reporting does not follow a clear autocracy/non-autocracy pattern and in Appendix Table E.2 that we do not perform worse in countries with less reporting. However, we do perform worse in autocracies when compared to democracies. This suggests that the informational content of reporting could be higher in non-autocracies but the model always retains some of its forecasting power.

4.3. How Machine Learning Solves the Hard Problem

What explains the good forecast performance of our methodology? A key advantage is that topics allow us to exploit the entire variation in the news text. An important advantage of this is that we can rely on positive and negative associations of topics and risk. In Figure 5 we show the average share of each of the 30 topics in months before an onset relative to months in which there is no onset. On the x-axis we show the relative topic share before hard cases and on the y-axis we show the relative share for all cases. For instance, the terror topic is much more likely to appear before hard and all onsets relative to a peaceful month. On the x-axis we see that newspapers write almost twice as much on terror than in other months before the outbreak of a hard onset. The religion topic appears before hard cases but is not strongly associated with all onsets. Other topics related to the economy, trade, transport and international relations appear less before onsets.[21] In this way, the topics provide signals which the supervised learning is able to exploit.

FIGURE 5. Topic shares before the onset of any violence relative to peaceful months



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

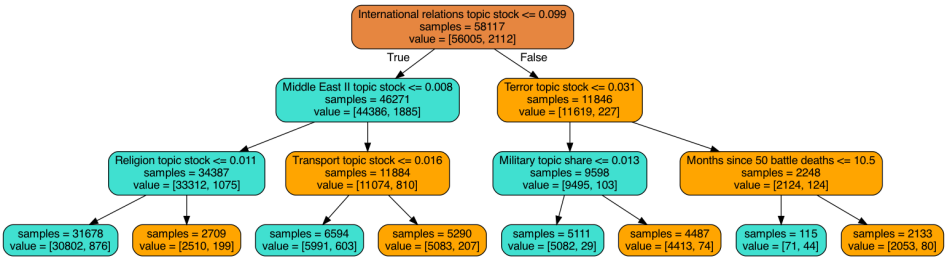
21. These changes are statistically significant and many hold even when we control for country fixed effects.

The random forest builds a forecast model from these text features and the conflict history variables. How does the random forest manage to exploit the variation shown in Figure 5 while at the same time allowing conflict history to play a central role? A random forest is a bootstrap aggregating (bagging) model which aggregates predicted decisions of many decision trees. A decision tree is a structure where at each node a decision is made given a cutoff value of a predictor. A random forest in the case of a binary variable then averages the share of 1s in the final nodes across all trees, and therefore predicts a probability, $\hat{y}_{iT+1}$. The randomness comes about by bootstrapping the sample from one tree to the next which avoids overfitting to the data.

See, for example, a decision tree in Figure 6 where for the sake of exhibition we choose a tree depth of three. Each node presents the variable and the associated value chosen for the split, how many cases reach the respective node or leaf, and what case distribution of outbreaks is at the node or leaf. The algorithm settles on the international relations topic stock at the top of the tree and decides to split at the value 0.099. This choice of variable and the corresponding split is determined based on the gains in terms of reducing Gini impurity, i.e. a reduction in entropy.[22]

While at the outset the probability of an onset is 3.6% ($\approx 2112/58117$), after the split it is about 1.9% if the condition is false and 4.1% if the condition is true, i.e. cases with a low international relations topic stock have a higher probability of an onset. The rest of the example tree follows the same logic until we reach the leafs, i.e. the terminal nodes. As a result of these classifications each case receives a probability according to the distribution in the leaf. Then this probability is averaged across all trees, which in our model using *text & history* are 400 trees with a depth of eight. Appendix Figure C.1 presents an example tree of depth eight to illustrate the complexity that the forecast model reaches.[23]

FIGURE 6. Simplified example tree of depth 3



Notes: Decision tree of depth tree. Each node specifies the chosen predictor and the corresponding cutoff. If the condition is fulfilled one moves down the left branch, if not down the right. The nodes and leafs further specify the number of cases that reach that node and the distribution across non outbreaks and outbreaks.

A way of gauging the importance of variables to the predictive performance is by looking at which position of the tree the predictor is chosen how often. Typically, a predictor chosen at the top of a tree contributes more to discriminating between

22. Gini impurity can be understood as the likelihood of an incorrect classification of a case if that case were randomly classified according to the overall distribution. Basically, a maximum Gini impurity means that half of the cases face an onset with a probability of 1, whereas the other half with probability 0. A low Gini impurity indicates that shares of cases are very imbalanced. For instance, if we know that amongst the remaining 100 cases 1 will have an outbreak with probability 0 and 99 with probability 1, then by randomly predicting a probability of 1 would lead to an accurate classification. See Appendix C.1 for more details.

23. In Appendix Figure C.2 we present the feature importance of each predictor, which is the average contribution of a predictor to the reduction in Gini impurity.

cases than one chosen towards the bottom (or not at all). In Appendix Figure C.3 we present the frequency of the positions at which each predictor is chosen.[24] We find that the predictors which contribute most to the reduction in Gini impurity, are also the ones that are most likely to be chosen towards the top of the tree. For example, the predictor indicating the months since the last conflict receives the largest importance score and is more often used in top nodes of the tree. Topics, even those that are almost never chosen in the higher branches of the tree, are often used in lower branches. In other words, the random forest model is automatically geared towards picking up more subtle risks with topics when conflict history 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 displayed in Figure 5. A subtle aspect of this process is that the model uses topics to capture stabilizations in countries with a conflict history. News on the economy, for example, vary dramatically across countries and time. In stabilizing countries these and similar news stories on trade or finance appear more and more, and this information is used in lower branches to indicate low risk.[25] The random forest can therefore use the text features to drive down the false positive rate.

A random forest allows for a fine-grained way of plotting the contribution of each predictor, and whether the variable increases or decreases the predicted likelihood of an outbreak for each observation in a SHapley Additive exPlanations (SHAP) beeswarm plot (Lundberg and Lee 2017).[26] Figure 7 shows on the y-axis the 18 top predictors in terms of predictive performance sorted by their average contribution to the forecast. Variables with relatively high (low) values are light (dark), while the horizontal location displays whether the effect of that value caused a higher (right) or lower (left) prediction. Variables capturing how many months have passed since the last conflict are the most important predictors, followed by the terror topic. The topics contributing most of the predictive power are a mix of topics related to violence (e.g. military), regional topics (e.g. China), and activity (e.g. trade). While the majority of topics are positively associated with the onset of violence, others are not (e.g. the China and the transport topic). In Appendix Figure C.5 we present the extent too which the peace agreements topic is related to the outbreak of violence depending on how much time has elapsed since the last fatality, exemplifying why a random forest, which effectively builds interactions into its prediction, is particularly useful for this purpose.

A key insight from this analysis is that the random forest can use the news text features to produce something like a contextual awareness which is obvious to human observers but hard to quantify objectively. Peace agreements seem, for example, highly indicative of risk which could be because peace agreements are only necessary in situations in which severe armed conflicts have come to an end but where renewed outbreaks are likely.

5. Integrating Forecasting and Prevention

5.1. A Prevention Cost Function

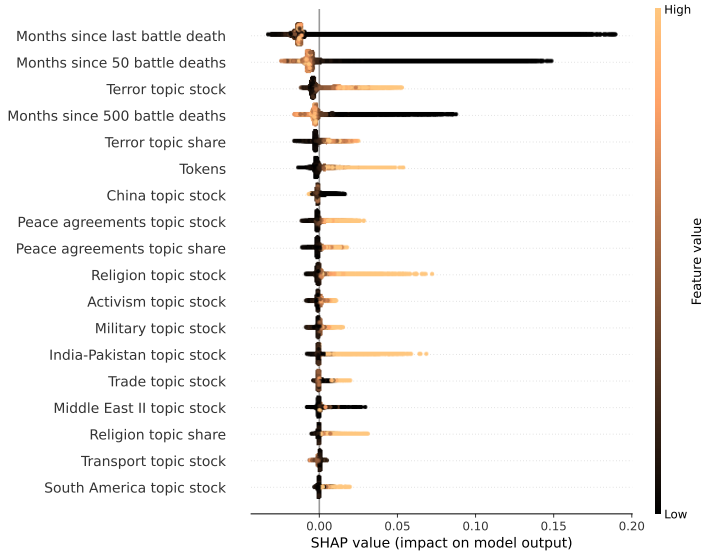
Building on the findings in the standard forecast framework we now derive a way to interpret the model output $\hat{y}_{iT+1}$ for the prevention problem. We propose the simplest possible, static framework that would allow a policymaker to derive

24. In Appendix Figure C.4 we present the relative frequencies.

25. Examples are provided in Appendix Figure B.1.

26. The SHAP value, adapted from the Shapley (1953) value, captures the impact of having a certain value for a given predictor in comparison to the prediction made if that predictor took the mean value.

FIGURE 7. SHAP beeswarm plots of predictors



Notes: The vertical location indicates the ranked average importance of the 18 most important predictors with the predictor with highest mean absolute SHAP value at the top. Each dot represents a case, i.e. a country-month. A dark (light) dot indicates that the predictor takes a relatively high (low) value for the case. The horizontal location shows whether the effect of that value caused a higher or lower prediction.

an optimal cutoff c^*. Our framework assumes a policy rule by which prevention interventions are taken for risk values above a threshold $\hat{y}_{iT+1} > c^*$. However, actions are taken under uncertainty regarding the realizations of conflict outbreaks, which means that interventions cause a cost regardless of whether the forecast was correct or not. This imposes a cost of false positives.

At the same time we assume that interventions do not always work even if they are implemented in a situation with an impeding conflict onset. This allows us to illustrate how the prediction problem interacts with the much larger literature on causal drivers of armed conflict. Research on causal effects allows policymakers to implement the correct policies in situations which are in danger of escalating into conflict, good forecasts save costs because they allow the policymaker target these situations better.

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

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

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 an economic crisis, crime or an epidemic wave. To simplify, we provide a static framework but the cost function could be adapted to a dynamic decision problem in which a policymaker takes repeated preventive actions over time. 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 first 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:

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

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 $p\,(V_D - V_P)$.

Note, how the framework integrates the forecasting problem into the existing literature on the causes of conflict. The forecast is used to flag some cases as positives (likely outbreaks) and others as negatives (likely peace). If this framework is used to decide on interventions, a positive triggers an intervention which needs to be designed based on research on the causes of conflict. The better the research on causes, the higher will be p. The forecast framework also plays an important but more subtle role here. This is best illustrated with the cutoff c above which the forecast framework declares a situation a positive and triggers an intervention. A decrease in the cutoff c generates more positives and less negatives. In Equation (3) this will lower the number of FN_c and TN_c and at the same time increase both TP_c and FP_c. If the forecast is very good it will identify cases which were a FN_c and convert them into TP_c (anticipated outbreaks). This conversion generates a benefit of $p\,(V_D - V_P) - I$. But the lower the cutoff the more errors the forecast will make. This means it will convert more and more TN_c into FP_c (false warnings) which increases costs by I. As observations are ordered by the fitted values, $\hat{y}_{T+1}$, performance is first high and then falls so that the problem is well-behaved. We will show this now.

5.2. The Optimal Intervention Threshold

As an illustration of this integrated prevention framework, we now assume some, relatively conservative, parameters and plot the resulting cost curves to derive the cutoff that minimizes total expected costs in Equation (3). In Appendix D.1 we provide detailed explanations for all the parameter estimates we make.

Assume that the outbreak of a conflict leads to total discounted costs of US$100 billion and that the discounted cost of peace is normalized to US$0. Good estimates of prevention costs and effectiveness are hard to come by - mostly because there is currently no quantitative measurement of prevention efforts at the macro level outside countries with a conflict history. A key cost factor is whether military interventions are necessary so that we would expect interventions in countries without a recent conflict to be much cheaper.[27] We assume that prevention costs for all cases are US$1 billion per month of intervention. This is a high intervention cost per month and we assume that this makes prevention effective in one out of four interventions ($p = 0.25$). To illustrate the role of research on causes in this framework, we also show what happens if interventions are much less effective ($p = 0.05$). We do not assume that interventions are less costly or more effective in hard problem cases.

We plug these values in to get the cost weights in Equation (3).[28] We then use our *text & history* forecast to evaluate the number of true positives, false positives, true negatives and false negatives at different cutoffs. For a sensitivity analysis with respect to changes in V_D at different levels of effectiveness see Appendix D.2.

There are two ways to illustrate the result of this cost minimization. First, it is possible to draw isocost curves in the FPR/TPR-space of the ROC curves.[29] In this space, the slope of isocost curves is given by

$$\frac{\partial TPR}{\partial FPR} = \frac{Cost_{FP} - Cost_{TN}}{Cost_{FN} - Cost_{TP}} \frac{N}{P},$$

where N/P is the number of negatives to positives in the sample (class imbalance implies $N/P > 1$). Higher costs are captured by isocost lines that are shifted to the right. The ROC curve is then our possibility frontier and the higher the AUC, the lower the costs that can be reached. Costs are minimized when the ROC curve runs tangential to the isocost curve representing the lowest possible cost. The slope of the isocost line becomes steeper when N/P increases or when $Cost_{FP} - Cost_{TN}$ increases relative to $Cost_{FN} - Cost_{TP}$. Intuitively, high costs of false positives relative to true negatives drive policymakers to be more conservative. High costs of false negatives compared to true positives will make policymakers less conservative. Keep in mind that, ceteris paribus, changes in N/P do not alter ROC curves. However, our cost function introduces the term N/P through the isocost curves. If N/P increases, precision falls, the number of false positives increases and a policymaker will need to become more conservative to minimize costs - the slope of the isocost curve increases which leads to a lower cutoff value and lower true and false positive rates.

Changes in the destructiveness of civil wars change the slope of the isocost curve. If damage, V_D, increases the term $Cost_{FN} - Cost_{TP}$ in the denominator of the isocost curve increases and the slope of the isocost curve falls, making the decision maker less conservative. A fall in effectiveness has the same effect. We illustrate this in Figure 8 where we show what happens with effective and ineffective interventions. As discussed above, the falling effectiveness increases $Cost_{TP}$ and this increases the slope of the isocost curve. The false positive rate at which minimal costs are reached is indicated through a vertical dotted line. Note, how this line moves to the

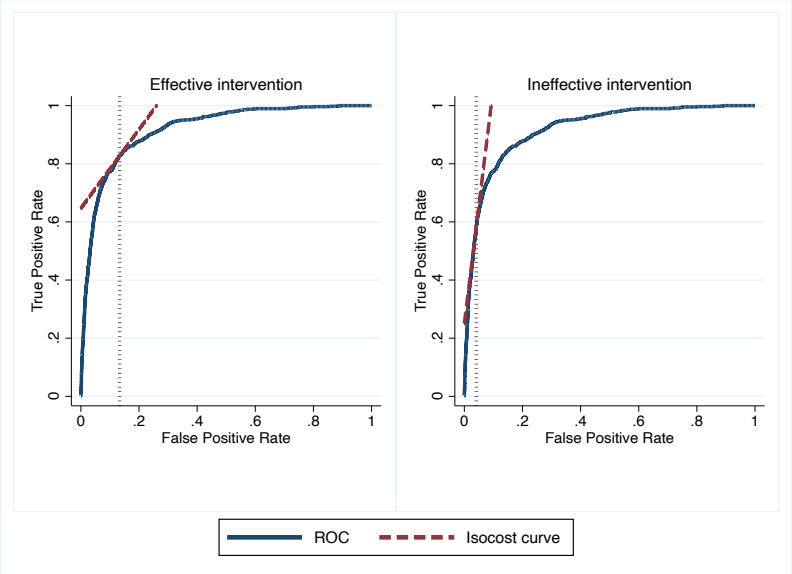
27. See the discussion in Chalmers (2007).

28. We then have that true positives cost $pV_P + (1-p)V_D + I = 76$ billion US$ with effective interventions ($p = 0.25$) and US$96 billion with ineffective interventions ($p = 0.05$). False positives cost the intervention costs, US$1 billion. The most costly outcome are false negatives with US$100 billion and the least costly outcome are true negatives with 0 costs.

29. For a level of costs, $Cost$, isocost curves are given by $TP/(TP+FN) = \frac{Cost_{FP}-Cost_{TN}}{Cost_{FN}-Cost_{TP}} \times \frac{FP}{FP+TN}\frac{N}{P} + \frac{\frac{N}{P}*Cost_{TN}+Cost_{FN}-\frac{Cost}{P}}{Cost_{FN}-Cost_{TP}}$ where P is the total number of positives and N is the total number of negatives.

left when interventions are less effective as shown in the right panel. Intuitively, the loss imposed by less effective interventions needs to compensated by a lower false positive rate. Whereas a false positive rate of 13% is acceptable with effective interventions this falls to just 3% when interventions are ineffective. The true positive rate falls at the same time from close to 80% to around 50%. The respective cutoff values at these true positive rates is the optimal cutoff.

FIGURE 8. ROC curves and isocost curves



Notes: Cost curves are calculated from the sample 2005m1-2020m8. The figure contrasts two scenarios for preventive interventions using Equation (3) and the out-of-sample forecast information of the *text & history* history model shown in the ROC curve from Figure 3. In this space isocost curves appear as straight lines. The left isocost curve assumes interventions are relatively effective ($p = 0.25$) and the right figure assumes interventions are relatively ineffective ($p = 0.05$). Intervention costs are US$1 billion and if an outbreak is prevented this saves damages of US$100 billion.

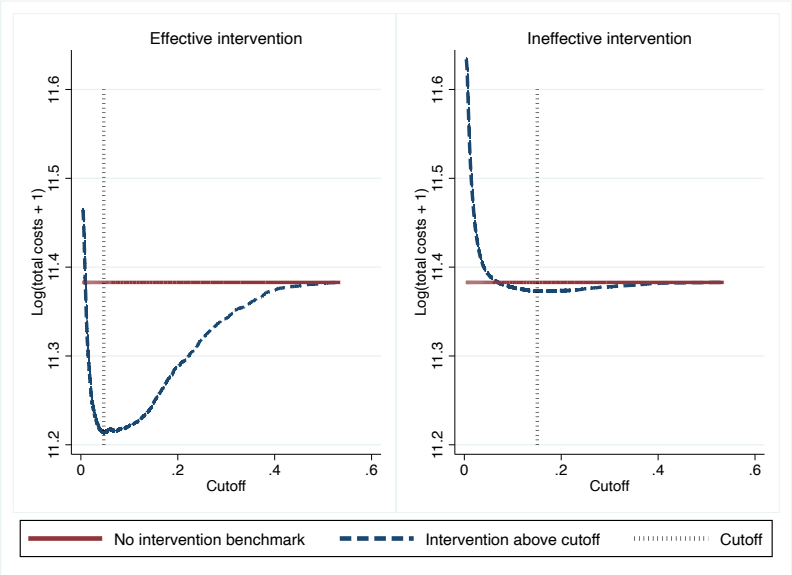
Another way to illustrate the optimal cutoff values is by plotting the cost functions themselves and finding a minimum. The numbers together with the cost weights give cost values for each given cutoff when they are plugged into Equation (3). Figure 9 illustrates the resulting cost function for effective and ineffective interventions. On the x-axis of the figure, we show the cutoff c and on the y-axis we plot total costs. In order to benchmark the costs with intervention, we also show the costs without any interventions as a solid horizontal line. The cost function (dashed 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 ($p = 0.25$). 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 costs. At a cutoff of $c = 0.047$ the cost curve takes a minimum. At this point, an increase in the cutoff increases total 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 of Figure 9 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 with interventions are now

much higher so that cost savings with prevention are lower.[30] With ineffective interventions the required precision needs to be higher, the number of interventions needs to fall and the optimal cutoff therefore increases to a value which is now closer to $c = 0.2$.

FIGURE 9. Cost curves with effective and ineffective interventions



Notes: Cost curves are calculated from the sample 2005m1-2020m8. The figure contrasts two scenarios for preventive interventions using Equation (3) and the out-of-sample forecast information of the *text & history* history model. The left cost curve assumes interventions are relatively effective ($p = 0.25$) and the right figure assumes interventions are relatively ineffective ($p = 0.05$). Intervention costs are US$1 billion and if an outbreak is prevented this saves damages of US$100 billion.

This trade-off between policy effectiveness and the intervention cutoff is a special feature of the framework we propose here. Effectiveness p, a causal concept, influences the way the policymaker interprets the forecasts coming out of the forecasting model. Less effective policy tools make the policymaker more conservative in her forecast. The usefulness of our forecast model therefore critically depends on whether conflicts can be de-escalated and at which costs. But the reverse is also true; optimal intervention policy will react to the precision of the forecast. Conflicts break out after only 3% of all peaceful months. Without targeting, prevention therefore produces 33 times more false positives, FP_c, than true positives, TP_c, and this ratio becomes much worse in cases without a recent conflict history. This can prohibit even cheap, effective interventions.

In our example, the likelihood of an onset after an intervention is 16% and the cost functions indicate that this makes prevention efforts cost effective. 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 have led to about 25 interventions each month assuming effective interventions and around 10 when we assume ineffective interventions. When we assume effective interventions we get hundreds of interventions before outbreaks

30. Cost savings compared to no intervention are US$1 trillion with ineffective interventions and US$13 trillion with effective interventions. In Appendix D.3 we provide a detailed discussion of the use of an intervention framework in practice, and present intervention thresholds and damage estimates under alternative parameterizations.

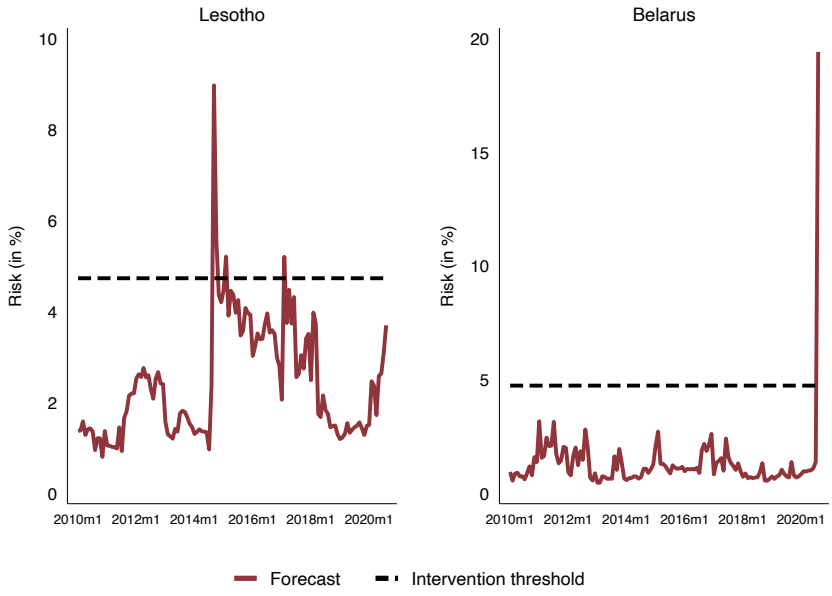
without a recent conflict history. Thus, despite the low baseline risk, prevention would still have been economically feasible under our assumptions.

5.3. Case Studies

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.[31] We illustrate the nature of the risk forecast based on the text model for four countries in Figures 10 and 11, and put our risk estimates in a relationship with the optimal cutoff derived in the previous section. The solid lines report the risk from the forecast model. The forecast is always one month ahead so that the value for 2015m4, for example, is for an outbreak in 2015m5. We show the optimal intervention cutoff as a dashed line.

Figure 10 shows the cases of Lesotho and Belarus. Both of these countries represent failures in the sense that Lesotho did not have a violent onset in 2014 according to UCDP, whereas Belarus had an onset in 2020 which was not anticipated. However, the figures are still meaningful in that they show the variation of risk that is triggered by a deep political crisis in Lesotho in 2014, including gunshots being fired on 30 August in what was coined a coup attempt. Our forecast model clearly picks this up in an episode of heightened risk which stretches out for quite a long time. The case of Belarus represents a false negative as the UCDP codes armed conflict for August 2020 which our model did not anticipate. This is reflected in a dramatic increase of risk as Belarus now has a recent conflict history.

FIGURE 10. Examples of a false positive and a false negative



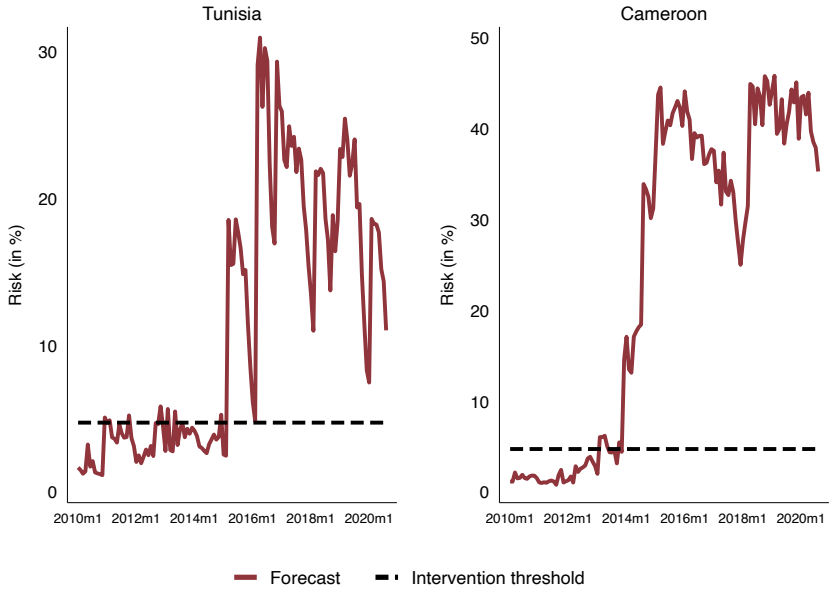
Notes: Predictors include 30 topics, token counts and three conflict history variables. The intervention threshold is taken from Figure 9.

Figure 11 shows the cases of Tunisia and Cameroon. Both of these countries represent partial successes in the sense that conflict risk crosses the intervention threshold for a long period before political violence actually broke out according to UCDP. Whereas these are not correct anticipations of conflict onsets a month before

31. We provide similar outputs for more than 170 countries through the webpage conflictforecast.org.

the outbreak, this suggests that if preventive efforts have long-term effects it might have helped prevent the outbreak of armed conflict in these countries.

FIGURE 11. Examples of anticipation of violence



Notes: Predictors include 30 topics, token counts and three conflict history variables. The intervention threshold is taken from Figure 9.

6. 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 optimal tree depth is quite deep and there is some reliance on text despite the first-order importance of conflict dynamics. 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 for problems with a restricted number of training cases.

Work by Bazzi et al. (2019) and Hegre et al. (2019) suggest that the prediction of violence at the subnational level is possible. However, here the hard problem becomes even more urgent. It is relatively straightforward to use machine learning to track and predict conflict dynamics for ongoing violence. With fine-grained data this includes tracking spatial dynamics which can be used to predict outbreaks of violence in close-by locations. But it is much harder to predict spatially and temporally isolated outbreaks of violence. Low precision becomes such a big problem that there is doubt whether spatially disaggregated onsets can be predicted usefully (Cederman and Weidmann 2017). Our research suggests some moderate optimism is in order. We show that predicting onsets is possible at the country level even if they are not directly related to ongoing conflict dynamics. If there is a way to combine our forecast with the identification of typical break points in space, the prediction of new outbreaks at the subnational level will become possible. Population centers, peripheries or mining locations, for example, might all become high risk depending on a specific news topic context. However, in the context of forecasting conflict onset locations in space and time, it might be necessary to adjust labelling techniques or performance measures. It is clearly less of a failure to have missed the exact location of an onset than having missed it completely (Greene et al. 2019). Moreover, given that currently we only include articles with country or capital names in the title, we miss out on important regional developments such as Northern Ireland, Kashmir, Catalonia, Palestine, or the Sahel region.

Forecasting models like ours also provide objective risk evaluations for countries which never experienced the outbreak of armed conflict. This is not only potentially useful for policymakers but it has the advantage of providing the basis for research on prevention itself. Moreover, despite our predictors not being able to receive a causal interpretation, our forecasts can provide hints for future research concerning where to search for drivers of conflict. The fact that some conflicts are harder to forecast than others might 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 that 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 intervention 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. 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 the former case the publication of the forecast would render it wrong.

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." Tech. rep., National Bureau of Economic Research.

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." Tech. rep., 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.

Berman, Nicolas and Mathieu Couttenier (2015). "External shocks, internal shots: the geography of civil conflicts." *Review of Economics and Statistics*, 97(4), 758–776.

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.

Berman, Nicolas, Mathieu Couttenier, and Raphaël Soubeyran (2021). "Fertile ground for conflict." *Journal of the European Economic Association*, 19(1), 82–127.

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 Jeannie Annan (2015). "Can employment reduce lawlessness and rebellion? A field experiment with high-risk men in a fragile state." Tech. rep., National Bureau of Economic Research.

Blattman, Christopher, Julian C Jamison, and Margaret Sheridan (2017). "Reducing crime and violence: Experimental evidence from cognitive behavioral therapy in Liberia." *American Economic Review*, 107(4), 1165–1206.

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." In *Proceedings of the 23rd international conference on Machine learning*, pp. 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.

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

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.

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

Greene, Kevin, Håvard Hegre, Frederick Hoyles, and Michael Colaresi (2019). "Move It or Lose It: Introducing Pseudo-Earth Mover Divergence as a Context-sensitive Metric for Evaluating and ImprovingForecasting and Prediction Systems." Tech. rep., Presented to the 2019 Barcelona GSE Summer Forum, workshop on 'Forecasting political and economic crisis: Social science meets machine learning.

Hanczar, Blaise, Jianping Hua, Chao Sima, John Weinstein, Michael Bittner, and Edward R Dougherty (2010). "Small-sample precision of ROC-related estimates." *Bioinformatics*, 26(6), 822–830.

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, Marie Allansson, Matthias Basedau, Michael Colaresi, Mihai Croicu, Hanne Fjelde, Frederick Hoyles, Lisa Hultman, Stina Högbladh, Remco Jansen,

et al. (2019). "ViEWS: a political violence early-warning system." *Journal of peace research*, 56(2), 155–174.

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.

Hörner, Johannes, Massimo Morelli, and Francesco Squintani (2015). "Mediation and peace." *The Review of Economic Studies*, 82(4), 1483–1501.

Institute for Economics and Peace (2017). "Measuring Peacebuilding Cost-Effectiveness." Tech. rep., IEP Report 46.

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.

Lo, Adeline, Herman Chernoff, Tian Zheng, and Shaw-Hwa Lo (2015). "Why significant variables aren't automatically good predictors." *Proceedings of the National Academy of Sciences*, 112(45), 13892–13897.

Lundberg, Scott M and Su-In Lee (2017). "A unified approach to interpreting model predictions." In *Advances in neural information processing systems*, pp. 4765–4774.

Meirowitz, Adam, Massimo Morelli, Kristopher W Ramsay, and Francesco Squintani (2019). "Dispute resolution institutions and strategic militarization." *Journal of Political Economy*, 127(1), 378–418.

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

Milante, Gary, Hannes Mueller, Robert Muggah, Katherine Aguirre, Caitriona Dowd, Clionadh Raleigh, Jacob Shapiro, and Carlos Vilalta (2020). "Forecasting the Dividends of Conflict Prevention from 2020-2030." Tech. rep., New York: CIC.

Mueller, Hannes (2017). "How Much Is Prevention Worth?" Tech. rep., World Bank.

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." Tech. rep., 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). english"Software Framework for Topic Modelling with Large Corpora." In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pp. 45–50, ELRA, Valletta, Malta. http://is.muni.cz/publication/884893/en.

Rohner, Dominic and Mathias Thoenig (2020). "The Elusive Peace Dividend of Development Policy: From War Traps to Macro-Complementarities." *Annual*

Review of Economics.
Rossi, Barbara and Tatevik Sekhposyan (2015). "Macroeconomic uncertainty indices based on nowcast and forecast error distributions." *American Economic Review*, 105(5), 650–55.
Shapley, Lloyd S (1953). "A value for n-person games." *Contributions to the Theory of Games*, 2(28), 307–317.
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 (2017a). "The Toll of War : The Economic and Social Consequences of the Conflict in Syria." Tech. rep., World Bank, Washington, DC.
World Bank (2017b). "United Nations and World Bank leaders call for stronger international efforts to prevent violent conflict." Tech. rep., Press release.

Appendix

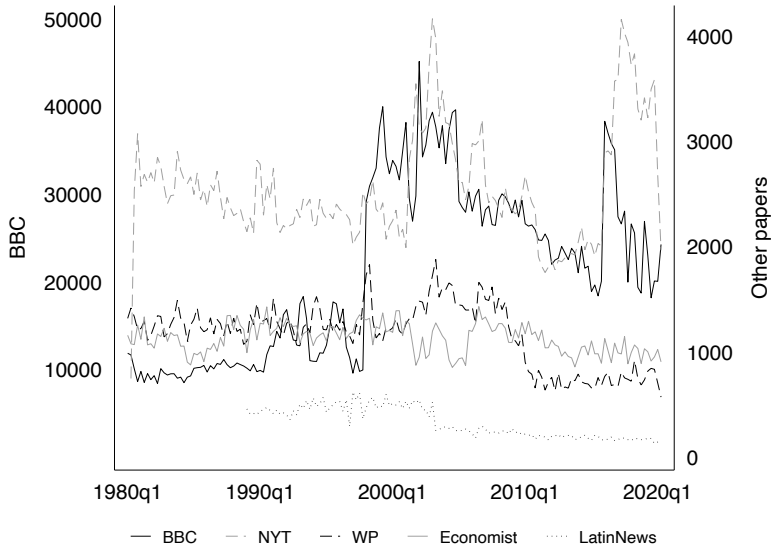
Appendix A: Data Description

All our text data is downloaded manually from LexisNexis and Latin News. 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 monthly 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 monthly panel of articles from all sources for over 190 countries for the period 1989 to 2020. In total we have 4,258,192 articles of which 413,918 articles are from the New York Times, 203,249 from the Washington Post, 41,607 from Latin News, 203,436 from the Economist, and 3,478,592 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 LexisNexis 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 would be most problematic for our forecasting exercise if the increase or decrease in the number of articles somehow affects the share of news written on a specific topic, which does not appear to be the case.

FIGURE A.1. Number of articles by source



Notes: 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.

In Table A.1 we summarize the different sets of predictors we use. The first model is based on our text data. This includes 7, 15, or 30 topic shares and the log of the word count of that month. The word count varies between 5 and 1,226,371 and is log-normally distributed with a mean of 4,274 words, while the topics sum to one within each country-month.

TABLE A.1. Sets of predictors

Name	Variables
Topics	Estimated topic shares and stocks using dynamic topic model and total number of tokens
Conflict history	Months since last fatality, since last occurrence of 50 fatalities, and since last occurrence of 500 fatalities
Standard	Infant mortality, political institutions, share of discriminated population, and neighboring conflicts based largely on Goldstone et al. (2010)
Commodities	More than 50 commodity prices with constant commodity export weights based largely on Bazzi and Blattman (2014), summarized into four measures by extracting the first factor of all export weights, of those related to energy (oil, gas, coal), minerals (e.g. gold, iron, copper), and agriculture (e.g. banana, tea, lumber)

Appendix 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 2005m1 and then we update every month using a dynamic topic model (Řehůřek and Sojka 2010). We allow the a-priori weight variational hyperparameters for each document to be inferred by the algorithm, and α, the a-priori belief for the each topics' probability, is set to the

default of $\frac{1}{K}$. Before feeding the text to the machine learning algorithm we conduct standard procedures when working with text. We remove overly frequent 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 200 documents. Table B.1 summarizes the top 10 terms in the $K = 30$ topic model estimated in 2020m8. The striking feature of the topic model is, 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, terror and economics, in our sample period for Brazil, Angola, Iraq and Ukraine. Two features are remarkable. First, as expected, news on economics disappear rapidly as news on violence surge. Brazil, a country which was characterized by a lot of news on economics 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 economics 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, economics 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.

TABLE B.1. Top ten keywords of 30 topic model using all text until 2020m8

Nr	Label	Keywords
1	Justice	court, case, investig, polic, arrest, offic, prison, charg, justic, prosecutor
2	Middle East	lebanes, annex, prime, arab, hezbollah, occupi, say, territori, plan, bank
3	Economics	cent, increas, rate, econom, economi, worker, price, percent, number, product
4	Current events	situat, time, issu, need, import, possibl, polit, say, decis, posit
5	Industry	compani, oil, product, gas, industri, energi, suppli, technolog, firm, oper
6	Nuclear treaties	sanction, nuclear, missil, weapon, secur_council, secur, intern, deal, launch, agreement
7	Military	militari, forc, defenc, armi, arm, troop, command, oper, arm_forc, secur
8	Urban	citi, local, water, region, area, district, project, resid, mayor, flood
9	Intelligence	offici, secur, administr, week, intellig, washington, hous, nation_secur, foreign, white
10	Russia	websit, region, head, quot, channel, jul, republ, rossiya, jun, territori
11	Transport	flight, air, ship, sea, airport, plane, aircraft, passeng, airlin, port
12	Peace agreements	peac, agreement, secur, talk, intern, region, process, support, polit, statement
13	Daily life	like, time, work, famili, home, woman, life, open, way, citi
14	State visits	foreign, ministri, meet, visit, offici, foreign_ministri, affair, prime, diplomat, issu
15	Health	health, case, pandem, test, virus, infect, hospit, medic, patient, death
16	International relations	cooper, develop, relat, region, econom, support, intern, project, import, secur
17	Politics	parti, elect, polit, opposit, vote, ial, candid, leader, democrat, support
18	Religion	islam, majli, mosqu, religi, god, muslim, leader, offici, prayer, republ
19	Middle East II	say, region, statement, base, uae, moham, arab, pro, quot, support
20	Activism	protest, right, human, law, human_right, social, polic, activist, facebook, freedom
21	Finance	bank, dollar, fund, financi, money, financ, busi, debt, loan, budget
22	Parliament	committe, council, member, meet, parliament, law, ministri, deputi, administr, feder
23	China	beij, parti, communist, citi, communist_parti, mainland, offici, wuhan, central, global
24	South America	los, migrant, refuge, moral, lopez, juan, repatri, social, announc, crisi
25	Terror	attack, kill, polic, group, forc, secur, provinc, area, milit, local
26	Media	medium, newspap, approxim, onlin, channel, video, page, daili, publish, say
27	Trade	trade, agreement, deal, sign, econom, export, busi, good, myanmar, tariff
28	Political power	like, power, say, time, way, war, make, polit, long, chang
29	India-Pakistan	word, daili, languag, taliban, prime, chief, issu, parti, languag_daili, sindh
30	Borders	border, minsk, cross, church, region, citizen, correspond, servic, resid, visa

Note: The labels are arbitrary and have no influence on the prediction model.

The topic model is also fine grained enough to pick up on specific events in developed countries. In Figure B.2 we show the politics topic share in the UK (top) and the US (bottom). Spikes are clearly visible around general elections. In the UK one can even tell when the Scottish independence and the Brexit referendum took place, and for the case of Brexit there is also a clear spike in the month it was announced.

FIGURE B.1. Terror and economics topic shares across time and countries

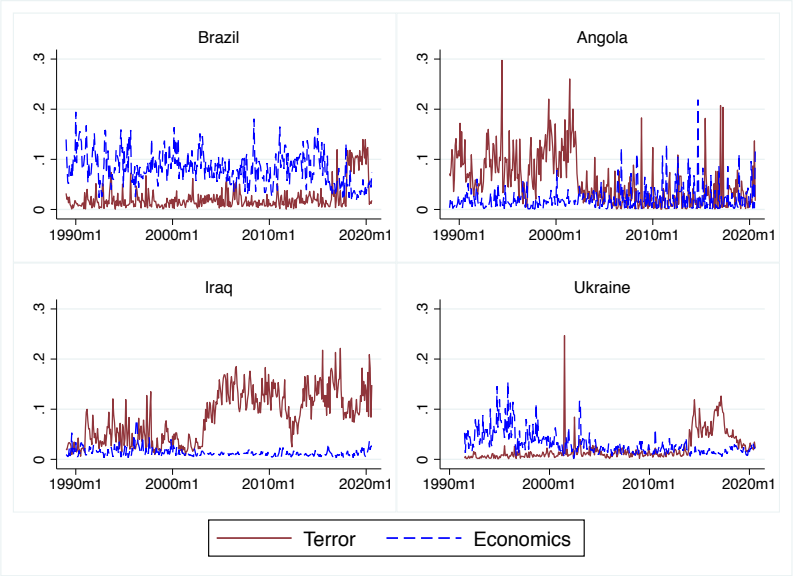
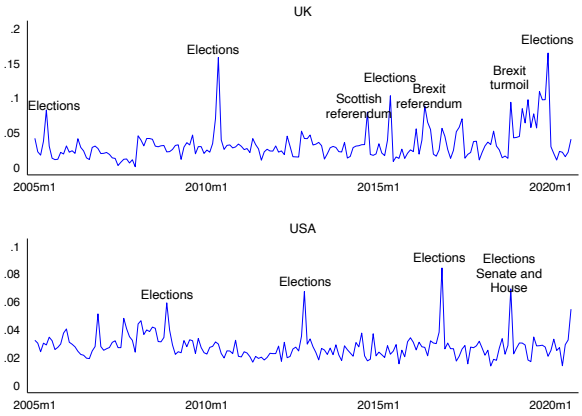


FIGURE B.2. Politics topic shares in the UK (top) and the US (bottom)



Finally, in Figure B.3 we exhibit an example of how the topic model adapts to new situations. The left panel shows the word cloud of the health topic with the text up until January 2020, featuring expressions such as “medic”, “doctor”, and “hospit” prominently. On the right we present the same topic in April where we can see that due to the Covid19 pandemic expressions such as “outbreak”, “virus”, “pandem”, and “quarantin” have become more dominant. This exemplifies how the topic model can absorb a new situation by categorizing in a manner which appears intuitive to a human.

FIGURE B.3. Health topic over time

a) 2020m1 b) 2020m4

Notes: The word clouds represent the most likely terms of one of the 30 topics estimated using all text until $2020m1$ (left) and $2020m4$ (left). The size of a token is proportional to the importance of the token within the topic. The location conveys no information.

Appendix 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.

TABLE C.1. Models

Technique	Brief description
Logit lasso	Linear estimation of log-odds using lasso regularization
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

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 choice of the hyperparameters 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 2005 and then selecting the hyperparameters that generate the highest ROC-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 C.2. Hyperparameters

			Minimum sample	
Predictors	Depth	Trees	split	leaf
Text	7	600	18	6
Conflict history	7	200	18	6
Text & conflict history	8	400	6	6

Note: Hyperparameters chosen through cross-validation within the sample before year 2005.

C.1. Gini Impurity

Imagine we have data of C classes and $p(i)$ is the probability of picking class i. Then the Gini impurity can be computed as

$$G = \sum_{i=1}^{C} p(i)(1 - p(i))$$

For the time being imagine we have green and blue balls, so $C = 2$ and $p(1) = p(2) = 0.5$, i.e. there is an equal amount of green and blue balls. In this case $G = 0.5 * 0.5 + 0.5 * 0.5 = 0.5$. Next imagine we are able to split the data in a way that we are left with $p(1) = 0.8$ and $p(2) = 0.2$. Then $G = 0.8 * 0.2 + 0.2 * 0.8 = 0.32$, i.e. a reduction in Gini impurity of 0.18. When a random forest is looking to split at a given node, it looks to choose the split leading to the maximum reduction in Gini impurity. If we are left with only balls of one color, then the Gini impurity is 0.

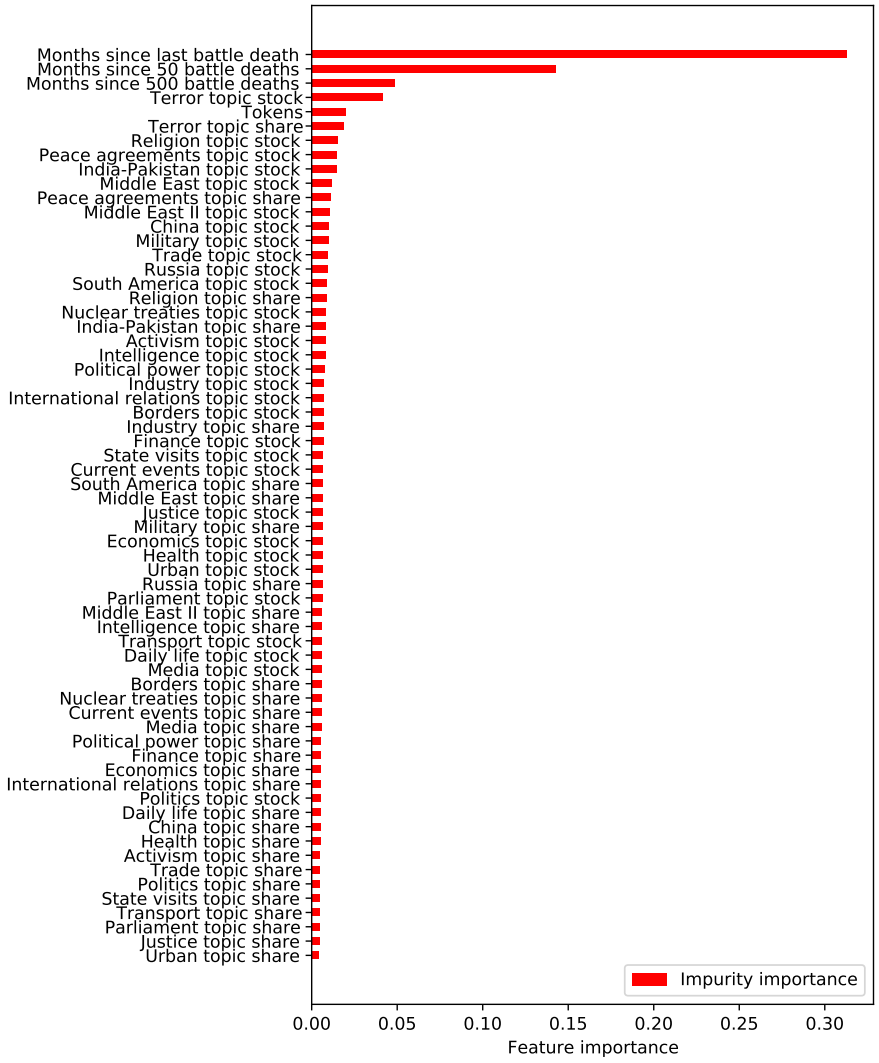
Now in our exercise we down have green and blue balls, but outbreaks of conflicts. See Figure 6 for a fictitious example tree and Figure C.1 for an actual tree out of the random forest we use.

C.2. Random Forest Details

FIGURE C.1. Example tree of depth 8

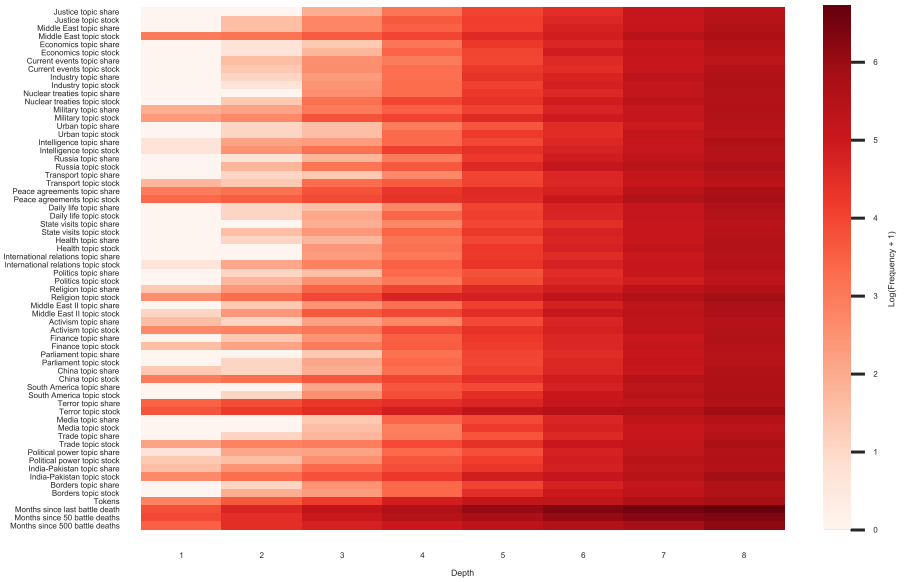
Notes: Decision tree of depth 8 of which 400 are used for predictions relying on *text & history*.

FIGURE C.2. Feature importance of predictors



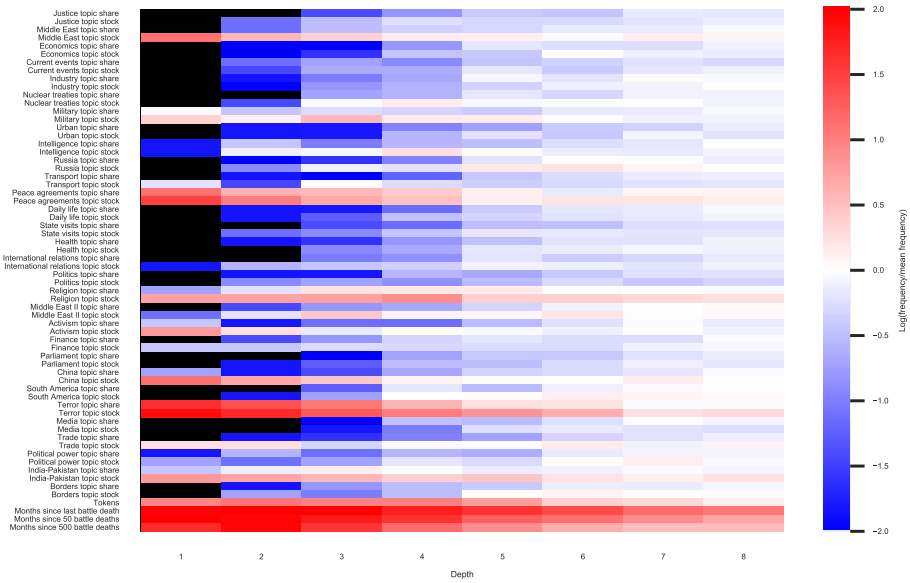
Notes: The bars indicate how much each predictor contributes to the overall reduction in Gini impurity.

FIGURE C.3. Frequency each predictor is chosen at each depth of tree across forest



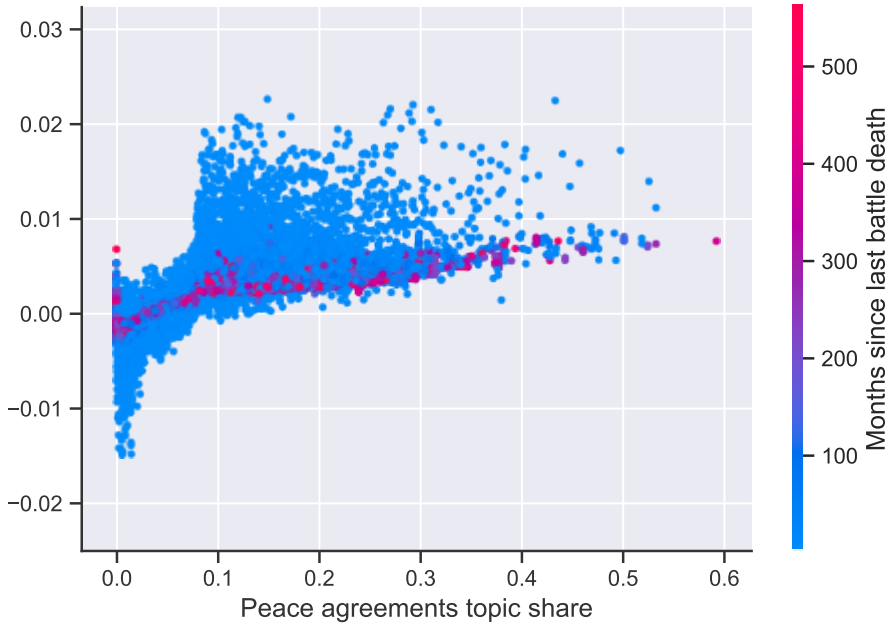
Notes: The x-axis displays the depth of the tree and y-axis the predictors. The darker the shade of a cell, the more often the respective predictor is chosen at this depth. The legend on the right indicates the log(frequency+1) represented by the shade.

FIGURE C.4. Relative frequency each predictor is chosen at each depth of tree across forest



Notes: The x-axis displays the depth of the tree and y-axis the predictors. The darker the shade of red (blue) the shade of a cell, the more (less) frequent the corresponding predictor was chosen at this depth relative to the mean number of predictors at this depth. The legend on the right indicates the log(frequency/mean frequency) represented by the shade. If a cell is white, the predictor was chosen with average frequency. If a cell is black, the predictor was never chosen at this depth.

FIGURE C.5. SHAP values of all individual predictors depending on time since last battle death



Notes: The y-axis present the SHAP value for the variable specified on the x-axis. The color of each dot indicates the the number of months that have passed since the last battle death. A blue dot indicates a short time period, while a red dot indicates a long time period as can be seen in the legend on the right.

Appendix D: The Intervention Framework in More Detail

D.1. Intervention Framework Assumptions

In this section we derive the assumptions made on the cost function in detail. Here we rely heavily on the discussion in Mueller (2017) and Milante et al. (2020). For simplicity, we use the yearly numbers from these reports. Transforming into monthly estimates by dividing them by 12 would lead to analog outcomes. We always take a conservative approach assuming that intervention costs are high, interventions are ineffective, and prevented damage is low.

D.1.1. Intervention Costs. We know little about the costs of prevention. One of the few studies that looks explicitly at prevention is Chalmers (2007). He estimates the costs of hypothetical prevention packages for several case studies in which conflict was arguably imminent. Chalmers proposes relatively large expenditures at a minimum of around US$1 billion per intervention per year. For example, the prevention package proposed for the Balkans in 1989 would include diplomatic engagement, debt relief, and economic assistance made conditional on peace talks and de-escalation. According to Chalmers, such a prevention package would cost US$15.4 billion spent over 15 years. According to Institute for Economics and Peace (2017), peace could be effectively built by spending between US$16.4 and US$20.3 billion per year in 31 conflict-affected countries. This is between US$520 million and US$650 million per country per year. However, these numbers come from post-conflict situations in countries like Rwanda, Afghanistan, or Iraq, where building peace should be more complicated than in countries that do not emerge from civil war.

In any case, it is clear that much less is currently spent. For the years 2016–2017, the Department of Political Affairs (DPA), the UN organization most clearly responsible for prevention, has requested a total of US$50 million to cover its priority areas of engagement. With an engagement in five countries, this would leave only US$5 million per country per year. In addition, most of this spending actually flows into countries which the framework presented above would have categorized as in conflict or recovery, not high-risk. This lack of engagement in high-risk countries is also visible in the overall official development assistance (ODA) data. Mueller (2017) shows that only about US$250 million of ODA flows into high-risk countries on average. Finally, the Institute for Economics and Peace (2017) report goes through a detailed analysis of different parts of ODA and categorizes some of it as peacebuilding. In this way the report can show that US$60.3 billion were spent on peacebuilding in 31 conflict-affected countries between 2002 and 2013. This is about US$130 million per country/year.

Based on these numbers, Mueller (2017) proposes a range of additional costs of prevention of US$100 million, US$500 million, and US$1 billion per year per intervention in high-risk situations. In the pessimistic scenario he assumes that each intervention costs US$1 billion to stay close to the Chalmers estimates. In the neutral scenario, he assumes that prevention costs US$500 million per year, which is close to the IEP estimate of what should be spent. In the optimistic scenario, he assumes prevention costs only an extra US$100 million per year. In other words, we assume that a targeted increase of ODA resources by 40%, from US$250 million to US$350 million, would suffice to lower the likelihood of conflict.

We follow Mueller (2017) and assume costs of US$1 billion per month for all interventions. When simulating the costs of interventions in hard onsets we assume US$100 million per month.

D.1.2. Conflict Damages. There is a large literature on the costs of conflict and so there are many numbers available. Most recently, the World Bank estimates that

from 2011 until the end of 2016, the cumulative losses in GDP from the Syrian civil war were US$226 billion (World Bank 2017a). 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.

Mueller (2017) shows that the impact on economic growth is one of the main determinants of the cumulative costs of conflict. Post-conflict growth does typically not recover the lost output and therefore an output gap opens up which cumulates over time. Assumptions on the growth impact therefore have a huge impact on the final loss. The report shows that an interval around -3.9% (between -2.6 and -5.2%) growth per year in conflict is realistic. Take a country like Syria with a GDP or US$70 Billion pre war. If the country experienced a civil war of four years than this would lead to a cumulative loss of over US$200 billion after only twenty years of growth at 1%. This does not include the cost of life, cost to health, or the trauma of displacement for the affected population. It also does not include the benefit to donors saved in humanitarian aid, peacekeeping or reconstruction. In a follow-up report Milante et al. (2020) derive that preventing 10 countries from falling into the conflict trap would save over US$3 trillion, i.e. around US$300 billion per country.

We take a lower bound from this assuming that damages are US$100 billion per month.

D.1.3. Effectiveness. The impact of prevention on the likelihood of conflict is a major determinant of its effectiveness. The Institute for Economics and Peace (2017) calculates the benefits from peacebuilding by looking at a scenario in which conflict will certainly break out. This implies that the effectiveness of prevention is assumed to be close to 100 percentage points, from certain conflict to (almost) certain peace. Chalmers (2007) assumes lower leverage but still proposes that in such a scenario, there is a 50 to 80% likelihood that conflict will be prevented. What often goes unappreciated is that these levels of effectiveness require the policymaker to be relatively sure that conflict will break out without a prevention effort. However, when undertaking prevention, one deals in probabilities, which are often small, and initiatives, which have an uncertain influence on outcomes. This is exactly what our cost framework picks up automatically through the separation of true and false positives.

We have no way to causally infer the impact of prevention on the likelihood of conflict. Mueller (2017) makes three assumptions in the three scenarios he simulates. He assumes that the likelihood of escalation is reduced by 25, 50, or 75%, respectively.

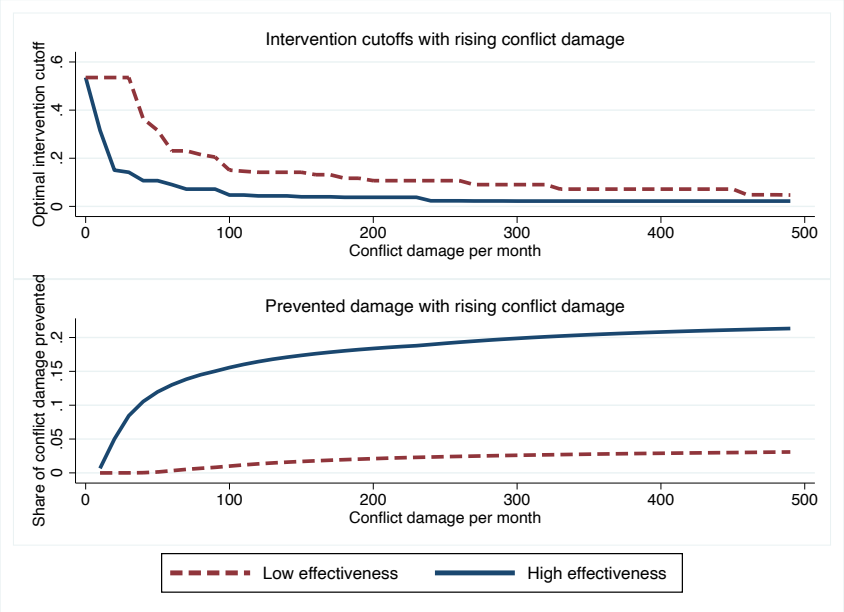
We again take a lower bound and assume that effectiveness is 25% in preventive efforts. To show how results change with changing parameters we also change effectiveness to just 5%.

D.2. Alternative Damage Assumptions

Costs and intervention effectiveness could depend on a range of country characteristics, e.g. rich vs poor countries, democracy vs autocracy, or large vs small population. Importantly, conflict damages and the effectiveness of prevention interact in the cost function. Therefore, we conduct a sensitivity analysis and in Figure D.1 show how the intervention threshold (top) and prevented damages (bottom) change with rising conflict damage under the scenario of low (dashed) and high (solid) effectiveness of interventions. A low level of effectiveness is defined as $p = 0.05$ and high as $p = 0.25$. In the top panel, we see that with rising conflict damages the intervention threshold and the difference in the threshold between low and high effectiveness decrease. However, in the bottom panel we can see that this small

difference in the intervention threshold still translates into large differences in the share of damage prevented due to the gap in effectiveness.

FIGURE D.1. Intervention cutoffs (top) and prevented damage (bottom) with rising conflict damage



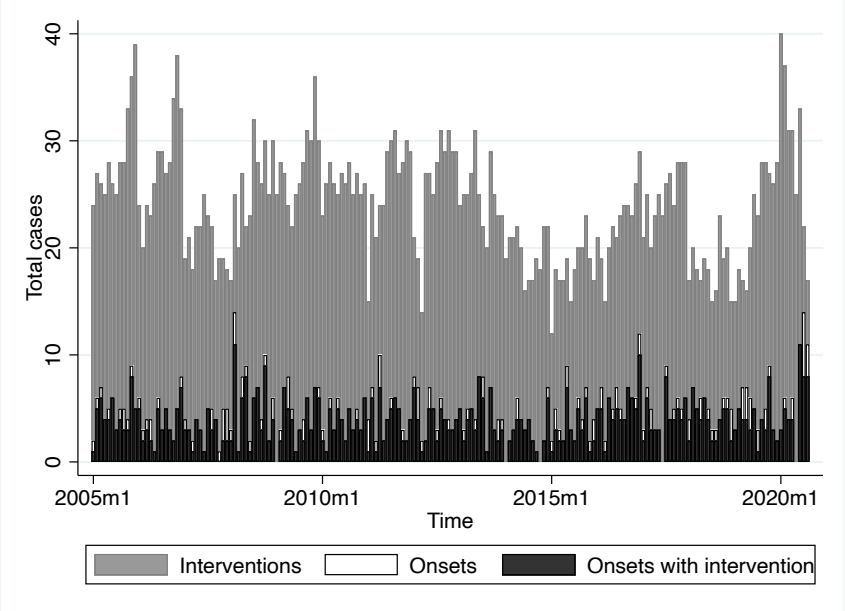
Notes: The x-axis displays varying levels of conflict damage per month of conflict, and the y-axis the resulting intervention cutoffs (top) and share of conflict damage prevented (bottom) for low effectiveness (dashed) and high effectiveness (solid). A low level of effectiveness is defined as $p = 0.05$ and high as $p = 0.25$.

D.3. 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 and low effectiveness respectively and calculate the optimal cutoff for all cases of any violence. The optimal cutoffs and resulting cost functions are displayed in the main text.

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 optimal cutoffs derived under the assumption of effective intervention from Figure 9. 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.

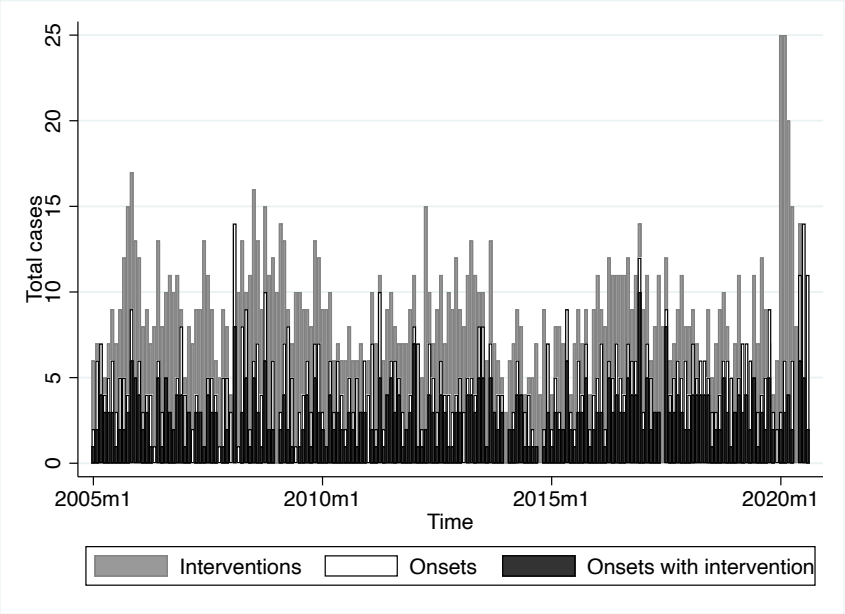
FIGURE D.2. Simulating timing and frequencies of effective interventions



Notes: The predictions underlying the figure are based on a model using 30 topic shares and stocks, token counts, the time that has passed since the last conflict month, the last conflict month with at least 50 fatalities, and the last conflict month with at least 500 fatalities. The cutoff for interventions is chosen to minimize costs as displayed in Figure 9. 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).

In Figure D.3 we show the optimal interventions over time when assuming ineffective interventions. The different policy prescriptions for effective and ineffective interventions becomes immediately clear from a comparison between Figures D.2 and D.3. To minimize costs, the policymaker would intervene before most onsets of conflict if she applied the model under the assumption that interventions work 25% of the time. This would lead to over 20 interventions in most months, i.e. a high level of activity. This high number of interventions implies that only a few outbreaks of violence are not covered by an intervention. Remember, that these are mostly 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.

FIGURE D.3. Simulating timing and frequencies of ineffective interventions



Notes: The predictions underlying the figure are based on a model using 30 topic shares and stocks, token counts, the time that has passed since the last conflict month, the last conflict month with at least 50 fatalities, and the last conflict month with at least 500 fatalities. The cutoff for ineffective interventions is chosen to minimize costs as displayed in Figure 9. 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).

Ineffective interventions would be used much less common, around ten per month, and the share of onsets that would be covered by intervention falls dramatically. Many onsets happen without an intervention. Costs and intervention effectiveness could depend on a range of country characteristics, e.g. rich vs poor countries, democracy vs autocracy, or large vs small population. Therefore, we conduct a sensitivity analysis and in Figure D.1 show how the intervention threshold (top) and prevented damages (bottom) change with rising conflict damage under the scenario of low (dashed) and high (solid) effectiveness of interventions. A low level of effectiveness is defined as $p = 0.05$ and high as $p = 0.25$. In the top panel, we see that with rising conflict damages the intervention threshold and the difference in the threshold between low and high effectiveness decrease. However, in the bottom panel we can see that this small difference in the intervention threshold still translates into large differences in the share of damage prevented due to the gap in effectiveness.

Appendix E: Additional Results

In the following section we show additional results.

E.1. The Role of Political Institutions

In this section we explore the possibility that political institutions might affect reporting and, in this way, affect the quality of our forecast. In Table E.1 we show regressions with the log number of tokens on the left hand side to show how reporting changes with political institutions. Higher scores indicate "more democracy" here and we exclude the category of full democracy (score=5). The regressions clearly show that reporting is lower in countries with political institutions which are not full democracies. However, the results also indicate a kind of U-shape pattern where partial autocracies and partial democracies feature least in our news sources. One explanation could be that the BBC monitor was set up during the cold war to keep an eye on iron curtain countries and, perhaps, sees its role in reporting from autocracies.

In Table E.2 we look into whether the lower reporting (news repression) has an effect on forecast performance. We find contradictory results which are robust on the exact definition of what we assume are country/years with repressed news. In the full sample we tend to find that performance is lower in countries with repressed news, whereas we find that in the hard onsets predictions are better in countries with repressed media. One explanation could be that in the context of violence, the government represses news more and prediction becomes hard. But before violence has broken out, news on things like a deteriorating economic or political situation can come out.

TABLE E.1. Reporting and institutions

	(1)	(2)	(3)
Democracy score = 4	-0.0559**	-0.430***	-0.272***
	(0.0240)	(0.0230)	(0.0185)
Democracy score = 3	-0.386***	-0.609***	-0.557***
	(0.0200)	(0.0194)	(0.0147)
Democracy score = 2	-0.484***	-0.785***	-0.598***
	(0.0204)	(0.0202)	(0.0158)
Democracy score = 1	0.0314	-0.132***	-0.0207
	(0.0267)	(0.0263)	(0.0180)
Democracy score = 0	0.169***	-0.510***	0.0126
	(0.0291)	(0.0280)	(0.0263)
Any violence		1.451***	0.482***
		(0.0165)	(0.0162)
Time fixed effects	Yes	Yes	Yes
Population control	No	No	Yes
Observations	57,966	57,966	56,916
R-squared	0.062	0.160	0.439

Notes: Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Democracy coding follows Goldstone et al (2010) XXX with low codes indicating less democratic (full autocracy = 1, partial autocracy = 2, partial democracy with factionalism = 3, partial democracy without factionalism = 4, full democracy (excluded) = 5). Democracy score = 0 indicate weakly institutionalized months/years and missings.

E.2. Alternative Ways of Showing Model Fit

The AUC has the big advantage of being comparable across classification problems. It is, however, highly problematic in the context of unbalanced classes like in the forecast of conflict onset. We therefore show a set of alternative ways of illustrating model fit in this section.

One of the most intuitive ways of showing model fit are so-called separation plots. In Figure E.1 we show separation plots using topics and conflict history. The figures order predictions by their rank on the x-axis and plot the predicted level

TABLE E.2. Forecast AUC performance and institutions

Panel A			
		Repressed	Not repressed
All cases	Text & history	0.909	0.916
	Text	0.791	0.868
Hard cases	Text & history	0.798	0.799
	Text	0.809	0.791

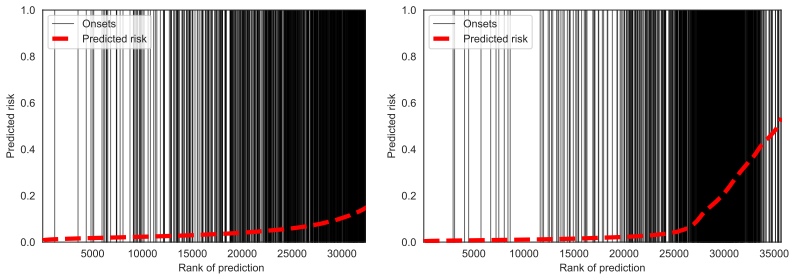
Panel B			
		Repressed	Not repressed
All cases	Text & history	0.908	0.923
	Text	0.795	0.880
Hard cases	Text & history	0.802	0.789
	Text	0.793	0.798

Notes: In Panel A repressed are those countries with a democracy score of 2 or 3. According to E.1 these are the countries with least news coverage in our model, controlling for conflict and population. In Panel B repressed are those countries with a democracy score of 1, 2 or 3. According to E.1 these are the countries with the lowest democracy score or the least news coverage in our model, controlling for conflict and population.

of risk using the red dashed line on the y-axis. The black vertical lines indicate actual onsets. 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 10,000 lowest risk observations contain only 9 onsets without clear common features.

FIGURE E.1. Separation plot of forecasting any violence using text and conflict history

a) Text b) Text & history



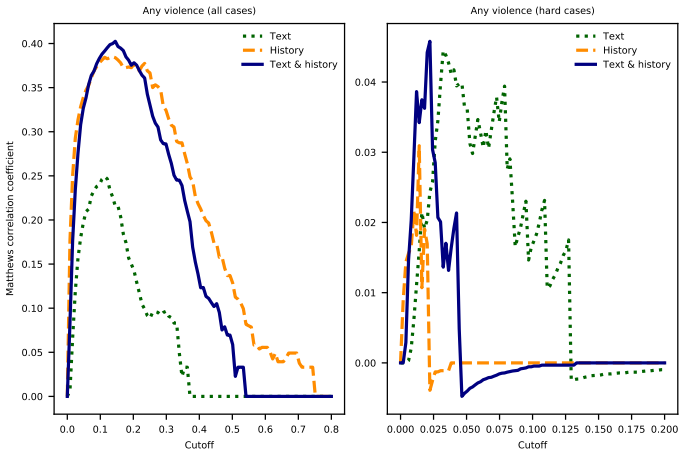
Notes: Note: The prediction method is a random forest. 'Text' contains 30 topic shares and stocks as well as token counts and 'history' contains three variables: the time that has passed since the last conflict month, the last conflict month with at least 50 fatalities, and the last conflict month with at least 500 fatalities. 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.

Another way of measuring predictive performance for binary classification is the Matthews Correlation Coefficient (MCC) which takes into account the balance ratios of the four confusion matrix categories (true positives, true negatives, false positives, false negatives) at a given cutoff c in the following manner (see, e.g., Hanczar et al. 2010):

$$MCC_c = \frac{TP_c \times TN_c - FP_c \times FN_c}{\sqrt{(TP_c + FP_c)(TP_c + FN_c)(TN_c + FP_c)(TN_c + FN_c)}}. \quad \text{(E.1)}$$

The MCC is bound between -1, for only mistakes, and +1, for perfect classification. In contrast to accuracy it doesn't only capture the predictive performance in terms of positives but also in terms of negatives. In Figure E.3 we plot the MCC for all onsets (left) and hard onsets (right) for different cutoffs on the y-axis. For all cases the MCC peaks below a cutoff of 0.2 and reaches values of up to 0.4 for the model combing text and history. For hard onsets, we see history has the lowest MCC and that the achieved overall MCCs are rather low, which is not surprising given the difficulty of the task.

FIGURE E.3. Matthews correlation when forecasting any violence for all onsets (left) and hard onsets (right)



Notes: The prediction method is a random forest. 'Text' contains 30 topic shares and stocks as well as token counts and 'history' contains three variables: the time that has passed since the last conflict month, the last conflict month with at least 50 fatalities, and the last conflict month with at least 500 fatalities. Left and right panels show alternative evaluations of the same forecasting model. The left Matthews correlations, computed as in Equation (E.1), show all cases. Hard cases are shown to the right. Hard cases are defined as not suffering fatalities in ten years.

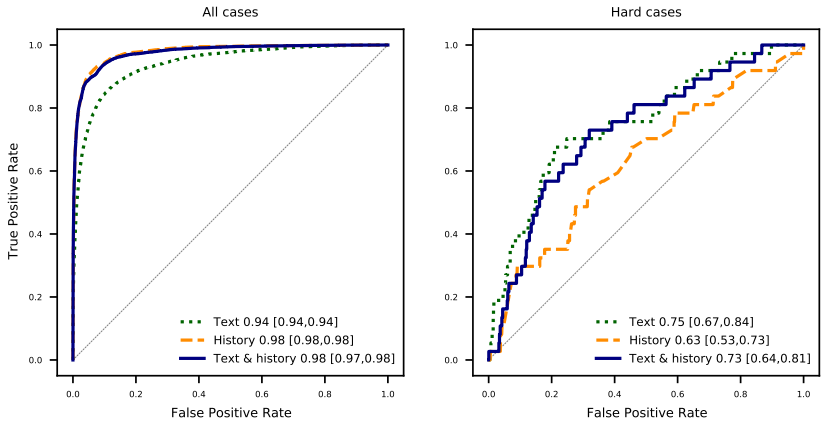
Another way to show the fit of the model in unbalanced datasets is precision which we discuss below when we show results for different algorithms. However, our main way of showing the quality of the prediction is through the cost function.

E.3. Predicting Incidence and Specific Types of Conflict

Policymakers might not only be interested in where conflict is about to break out but also where it might be persist. Therefore, in the mode presented in Figure E.4 we do not set subsequent conflict episodes to missing but include them as positives in both the training and validation sample. Further, we add the current number of fatalities to the history predictors. We see that this simplifies the task and the AUC increases to a staggering 0.98 for the history and *text & history* model. It is also encouraging that

the model achieves a similar performance as our benchmark model for hard onsets despite not being trained to predict only onsets.

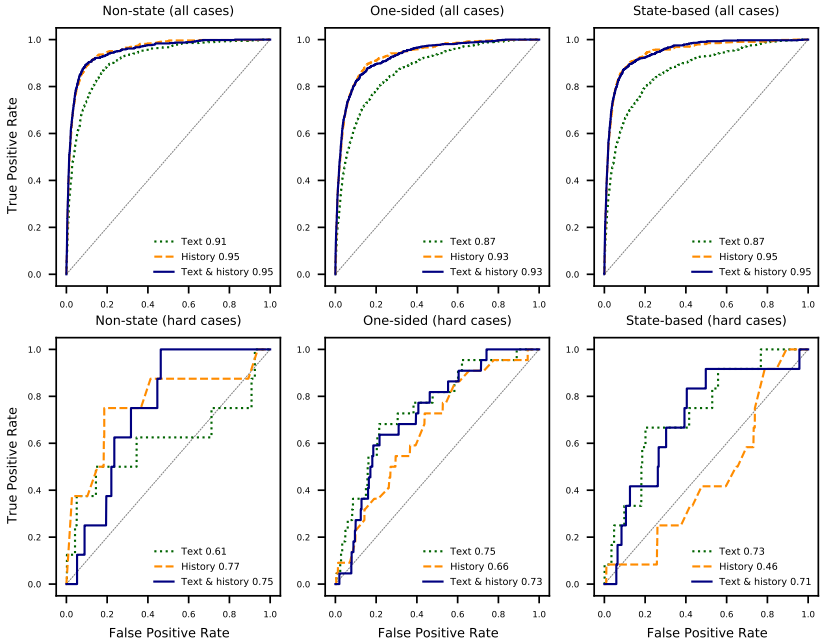
FIGURE E.4. ROC curves of forecasting any violence for all incidences (left) and hard onsets (right)



Notes: The prediction method is a random forest. 'Text' contains 30 topic shares and stocks as well as token counts and 'history' contains four variables: the time that has passed since the last conflict month, the last conflict month with at least 50 fatalities, the last conflict month with at least 500 fatalities, and the number of current fatalities. Left and right panels show alternative evaluations of the same forecasting model. The left ROC curves show all cases. Hard cases are shown to the right. Hard cases are defined as not suffering fatalities in ten years. The numbers in the legends represent the respective area under curve (AUC) with bootstrapped 95% confidence intervals in square brackets.

In Figure E.5 we look at how our predictors perform for any fatality related to specific types of conflict definitions. In the left panels, we look at non-state conflict defined by UCDP as the use of armed force between two organized armed groups, neither of which is the government of a state. The middle panels show one-sided violence, defined as the use of armed force by the government of a state or by a formally organized group against civilians. The right panel covers state-based armed conflict, defined as a contested incompatibility that concerns government and/or territory with the use of armed force between two parties, of which at least one is the government of a state. We follow our usual convention of only predicting onsets, while coding subsequent incidences of the same type of fatality as missing. A case is considered hard if the country has not experienced the specified sort out fatality within the last ten years. To the *history* model we add the months passed since the occurrence of the specific type of fatality. Overall performance in the texthistory model is extremely robust. However, the relative performance of the text and history models vary. For one-sided and state-based violence we see the usual pattern with *history* providing great performance across all cases, but the *text* model dominating when evaluating hard cases. For non-state violence we see a surprising result according to which the conflict history performs better than text, even for hard cases. One explanation is that non-state violence takes place in a context of general fragility and is therefore particularly well predicted by conflict history. In addition, our searches for country and capital names could lead to some state-centrism which might make text less effective when predicting non-state conflict.

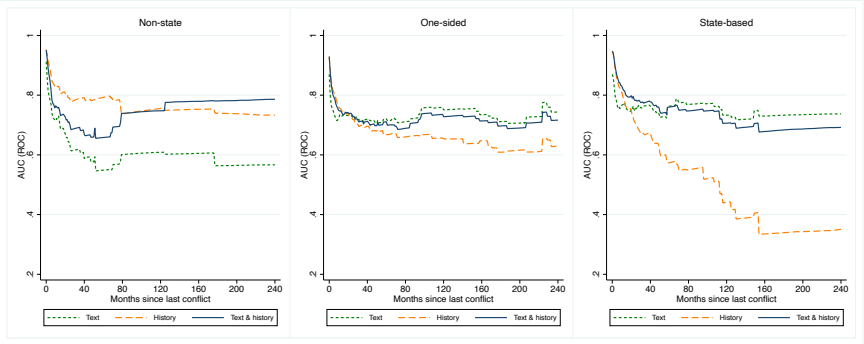
FIGURE E.5. ROC curves of forecasting different types of conflict for all onsets (left) and hard onsets (right)



Notes: The prediction method is a random forest. 'Text' contains 30 topic shares and stocks as well as token counts and 'history' contains four variables: the time that has passed since the last conflict month, the last conflict month with at least 50 fatalities, the last conflict month with at least 500 fatalities, and since the last conflict month of the respective type. Left and right panels show alternative evaluations of the same forecasting model. The left ROC curves show all cases. Hard cases are shown to the right. Hard cases are defined as not suffering fatalities in ten years. The numbers in the legends represent the respective area under curve (AUC) with bootstrapped 95% confidence intervals in square brackets.

In Figure E.6 we further break down the predictive performance for the three different classification of fatalities, i.e. non-state, one-sided, and state-based violence depending on how many months have passed since the last fatality of each type. The patterns emerging in Figure E.5 are confirmed, with conflict history performing well for non-state violence even after long periods without non-state fatalities, while for one-sided and in particular for state-based violence the AUC of the 'history' model drops rapidly. In contrast, the *text* model continues to provide considerable levels of predictive performance even after long periods without the occurrence of the specified type of fatality.

FIGURE E.6. AUC scores with fading conflict history for different types of conflict



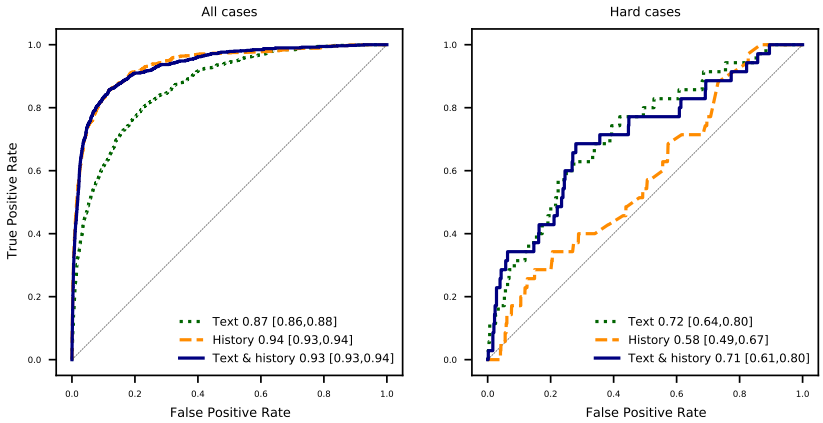
Notes: Figure shows the AUC for onsets that happened in the aftermath of a previous conflict episode. Onsets that occur within a window of months displayed on the x-axis are excluded in the evaluation. At a value of 20, for example, all outbreaks are excluded that occur within 20 months of the previous conflict episode. The prediction method is a random forest. *Text* contains 30 topic shares and stocks as well as token counts and 'history' contains three variables: the time that has passed since the last conflict month, the last conflict month with at least 50 fatalities, and the last conflict month with at least 500 fatalities. Left and right panels show alternative evaluations of the same forecasting model. All three lines show evaluations of the same forecasting models in changing samples.

E.4. Longer Forecast Horizons

For many applications in prevention a prediction of one year ahead is more desirable. In Figures E.7 and E.8 we therefore provide evaluations of a prediction model that considers an onset if conflict breaks out within any of the three or twelve following months, respectively. In order not to count the same onset multiple times, in the evaluation of the models' performance we only keep the month of the first coded onset of each outbreak.

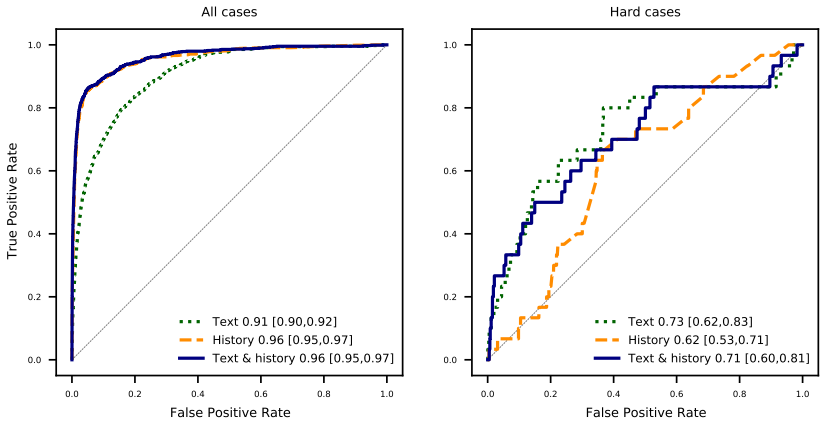
The predictive performance of the model remains strong. Forecasts of onset up to a quarter and year ahead both produce an AUC of 0.93 and 0.96 for any violence, and topics still add significant forecasting power for the hard cases.

FIGURE E.7. ROC curves for predictions of onset within next quarter



Notes: The prediction method is a random forest. 'Text' contains 30 topic shares and stocks as well as token counts and 'history' contains three variables: the time that has passed since the last conflict month, the last conflict month with at least 50 fatalities, and the last conflict month with at least 500 fatalities. Left and right panels show alternative evaluations of the same forecasting model. The left ROC curves show all cases. Hard cases are shown to the right. Hard cases are defined as not suffering fatalities in ten years. . Hard cases are defined as not having had conflict in ten years. The numbers in the legends represent the respective area under curve with bootstrapped 95% confidence intervals in square brackets.

FIGURE E.8. ROC curves for predictions of onset within next year

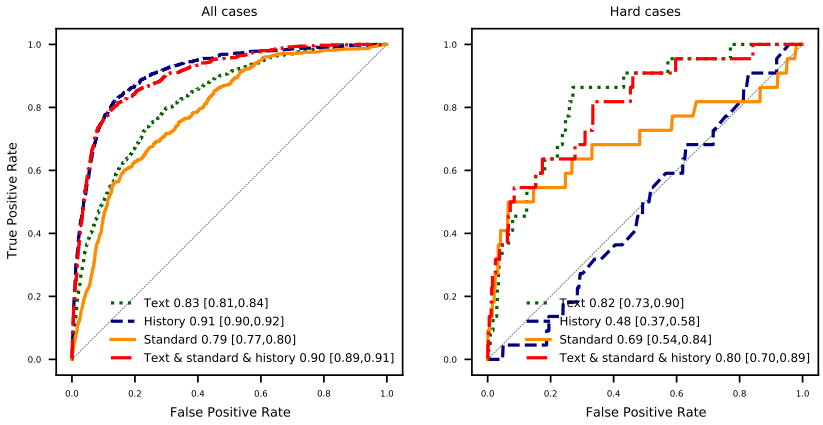


Notes: The prediction method is a random forest. 'Text' contains 30 topic shares and stocks as well as token counts and 'history' contains three variables: the time that has passed since the last conflict month, the last conflict month with at least 50 fatalities, and the last conflict month with at least 500 fatalities. The left ROC curves show all cases. Hard cases are shown to the right. Hard cases are defined as not suffering fatalities in ten years. Hard cases are defined as not having had conflict in ten years. The numbers in the legends represent the respective area under curve with bootstrapped 95% confidence intervals in square brackets.

E.5. Robustness to Alternative Forecast Models, Forecast Methods and Topic Models

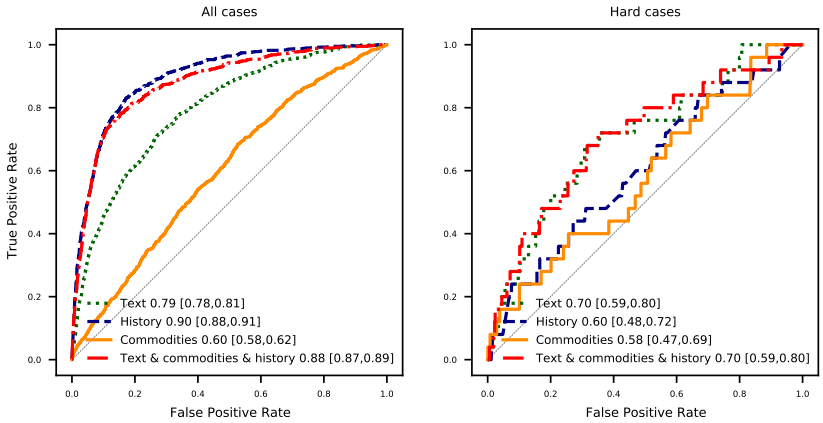
In Figures E.9 and E.10 we compare predictions using other variables typically used in the literature. For the sake of comparison, we only look at overlapping samples. For a further discussion of the variables and findings see Section 4 in the main text. Neither set of predictors performs particularly well for hard cases and adding predictors to the text & history model does not improve the performance either.

FIGURE E.9. ROC curves of forecasting any violence compared to standard variables



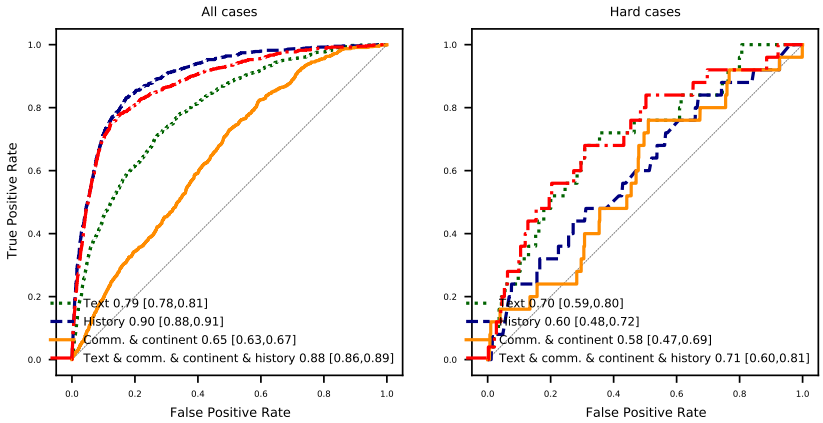
Notes: The prediction method is a random forest. 'Text' contains 30 topic shares and stocks as well as token counts, three variables: the time that has passed since the last conflict month, the last conflict month with at least 50 fatalities, and the last conflict month with at least 500 fatalities, and 'standard' contains infant mortality, political institutions, and neighboring conflicts. Hard cases are defined as not having had conflict in ten years. The numbers in the legends represent the respective area under curve with bootstrapped 95% confidence intervals in square brackets.

FIGURE E.10. ROC curves of forecasting any violence compared to commodity prices



Notes: The prediction method is a random forest. 'Text' contains 30 topic shares and stocks as well as token counts and 'history' contains three variables: the time that has passed since the last conflict month, the last conflict month with at least 50 fatalities, and the last conflict month with at least 500 fatalities, and 'commodities' contains four summary measures of 50+ commodity export weights interacted with the monthly mean price for the commodity. Hard cases are defined as not having had conflict in ten years. The numbers in the legends represent the respective area under curve with bootstrapped 95% confidence intervals in square brackets.

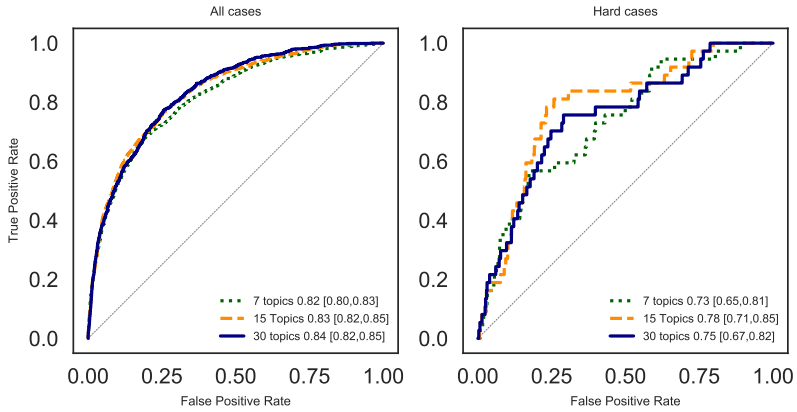
FIGURE E.11. ROC curves of forecasting any violence compared to commodity prices and continent fixed effects



Notes: The prediction method is a random forest. 'Text' contains 30 topic shares and stocks as well as token counts and 'history' contains three variables: the time that has passed since the last conflict month, the last conflict month with at least 50 fatalities, and the last conflict month with at least 500 fatalities, 'commodities' contains four summary measures of 50+ commodity export weights interacted with the monthly mean price for the commodity, and 'continent' are continent fixed effects. Hard cases are defined as not having had conflict in ten years. The numbers in the legends represent the respective area under curve with bootstrapped 95% confidence intervals in square brackets.

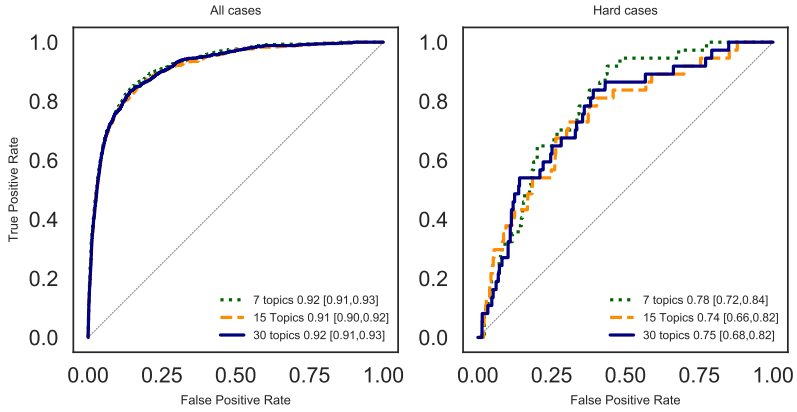
In Figures E.12 and E.13 we present compare the results of a model featuring 7 or 15 topics to our benchmark model of 30 topics. While the model using 30 topics performs slightly better, it also becomes clear that the predictive performance is not specific to summarizing the text into 30 topics.This strengthens that notion that text summarized into topics is a powerful predictor of conflict.

FIGURE E.12. ROC curves of forecasting violence with varying number of topics using only text



Notes: The prediction method is a random forest. 'Text' contains 7, 15, or 30 topic shares and stocks as well as token counts. Hard cases are defined as not having had conflict in ten years.

FIGURE E.13. ROC curves of forecasting violence with varying number of topics using text and history

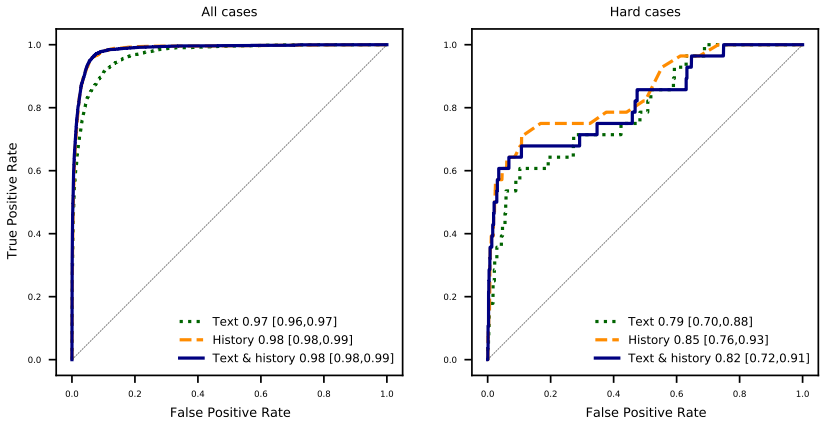


Notes: The prediction method is a random forest. 'Text' contains 7, 15, or 30 topic shares and stocks as well as token counts. Hard cases are defined as not having had conflict in ten years.

In Figure E.14 we show results for higher cutoffs. 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.

We show the performance of each of different prediction models using text only (Figure E.15) and both text and conflict history (Figure E.16). Across most dimensions it seems that the random forest is the algorithm performing best, which stands out particularly when predicting the hard cases of any violence with conflict history and text. Here the random forest reaches an AUC of 0.79 whereas the logit lasso only reaches and AUC of 0.74 using text only and dropping to 0.64 with *text & history*. This is consistent with the idea that the random forest receives an advantage

FIGURE E.14. ROC curves of forecasting armed conflict with at least 50 battle deaths

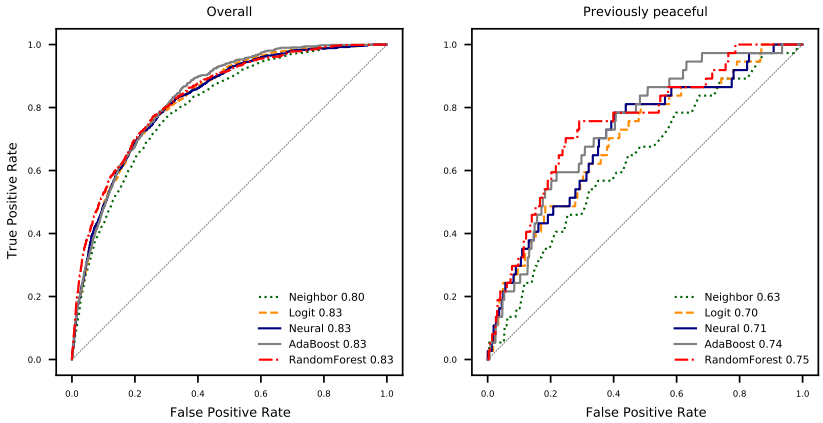


Notes: The prediction method is a random forest. 'Text' contains 30 topic shares and stocks as well as token counts and 'history' contains four variables: the time that has passed since the last conflict month, the last conflict month with at least 50 fatalities, and the last conflict month with at least 500 fatalities, and the most recent number of fatalities. Hard cases are defined as not having had armed conflict in ten years.

because the model is able to use the information contained in the text conditional on conflict history.

One exception is the AdaBoost model, i.e. adaptive boosting, which also relies on decision trees and therefore is related to the random forest. The reason we do not use the AdaBoost algorithm as the benchmark model is because of the predicted probabilities, which albeit performing great in terms of ranking, are distributed largely between 0.4-0.5, which is not only unrealistic, but also achieves a lower performance in terms of precision as can be seen in Figure E.17.

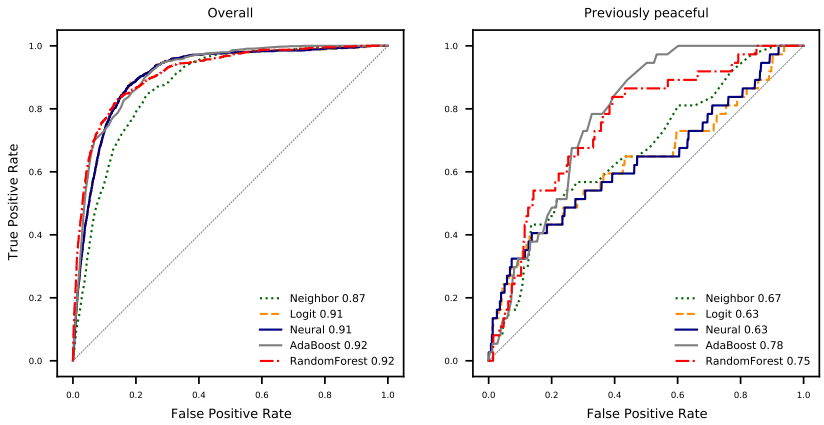
FIGURE E.15. ROC curves of forecasting violence with different algorithms using only text



Notes: The prediction method is a random forest. 'Text' contains 30 topic shares and stocks as well as token counts. Hard cases are defined as not having had conflict in ten years.

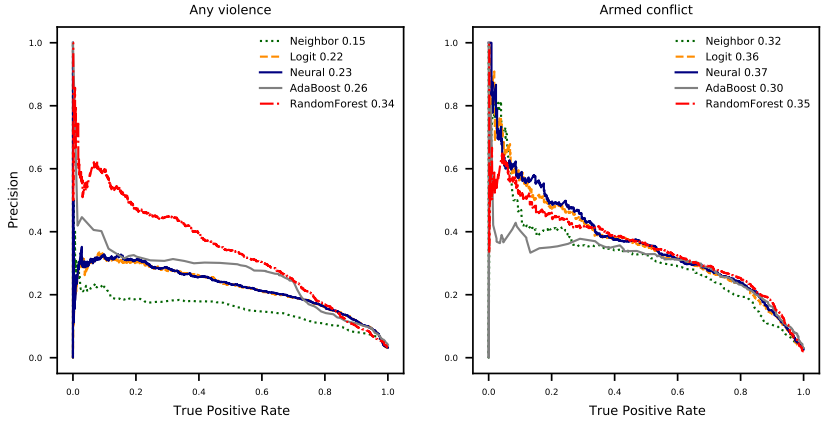
All in all, the figures paint a consistent picture: Conflict history and is a very good predictor of the outbreak of violence. Nonetheless, text summarized by topics adds useful information to predict topics, in particular in countries without current violence or a conflict history.

FIGURE E.16. ROC curves of forecasting violence with different algorithms using text and history



Notes: The prediction method is a random forest. 'Text' contains 30 topic shares and stocks as well as token counts and 'history' contains three variables: the time that has passed since the last conflict month, the last conflict month with at least 50 fatalities, and the last conflict month with at least 500 fatalities. Hard cases are defined as not having had conflict in ten years.

FIGURE E.17. Precision curves of forecasting violence with different algorithms using text and history



Notes: The prediction method is a random forest. 'Text' contains 30 topic shares and stocks as well as token counts and 'history' contains three (four) variables: the time that has passed since the last conflict month, the last conflict month with at least 50 fatalities, and the last conflict month with at least 500 fatalities (and the most recent number of fatalities when predicting armed conflict).