DISCUSSION PAPER SERIES

DP13748

THE HARD PROBLEM OF PREDICTION FOR CONFLICT PREVENTION

Hannes Felix Mueller and Christopher Rauh

DEVELOPMENT ECONOMICS

ISSN 0265-8003

THE HARD PROBLEM OF PREDICTION FOR CONFLICT PREVENTION

Hannes Felix Mueller and Christopher Rauh

Discussion Paper DP13748
Published 22 May 2019
Submitted 24 April 2019

Centre for Economic Policy Research
33 Great Sutton Street, London EC1V 0DX, UK
Tel: +44 (0)20 7183 8801
www.cepr.org

THE HARD PROBLEM OF PREDICTION FOR CONFLICT PREVENTION

Abstract

There is a growing interest in better conflict prevention and this provides a strong motivation for better conflict forecasting. A key problem of conflict forecasting for prevention is that predicting the start of conflict in previously peaceful countries is extremely hard. To make progress in this hard problem this project exploits both supervised and unsupervised machine learning. Specifically, the latent Dirichlet allocation (LDA) model is used for feature extraction from 3.8 million newspaper articles and these features are then used in a random forest model to predict conflict. We find that forecasting hard cases is possible and benefits from supervised learning despite the small sample size. Several topics are negatively associated with the outbreak of conflict and these gain importance when predicting hard onsets. The trees in the random forest use the topics in lower nodes where they are evaluated conditionally on conflict history, which allows the random forest to adapt to the hard problem and provides useful forecasts for prevention.

JEL Classification: N/A

Keywords: Armed Conflict, Forecasting, Newspaper Text, Machine Learning, Topic Models, Random Forest

Hannes Felix Mueller - h.mueller.uni@gmail.com
Institut d'Analisi Economica, CSIC and CEPR

Christopher Rauh - christopher.rauh8@gmail.com
University of Montreal

Acknowledgements
We thank Elena Aguilar, Bruno Conte Leite, Lavinia Piemontese and Alex Angelini for excellent research assistance. We thank the discussant Michael Colaresi and seminar and conference audiences at the ISA Toronto, INFER conference, the Barcelona GSE, Tokio University, Osaka University, GRIPS, Uppsala University, Quebec Political Economy Conference, University of Montreal, SAEe Barcelona, the German Foreign Office, Geneva University, Warwick University, the Montreal CIREQ workshop on the political economy of development and the Barcelona workshops on conflict prediction. Mueller acknowledges financial support from the Ayudas Fundación BBVA. The authors declare that they have no competing financial interests. All errors are ours.

The Hard Problem of Prediction for Conflict Prevention

Hannes Mueller and Christopher Rauh*

May 21, 2019

Abstract

There is a growing interest in better conflict prevention and this provides a strong motivation for better conflict forecasting. A key problem of conflict forecasting for prevention is that predicting the start of conflict in previously peaceful countries is extremely hard. To make progress in this hard problem this project exploits both supervised and unsupervised machine learning. Specifically, the latent Dirichlet allocation (LDA) model is used for feature extraction from 3.8 million newspaper articles and these features are then used in a random forest model to predict conflict. We find that forecasting hard cases is possible and benefits from supervised learning despite the small sample size. Several topics are negatively associated with the outbreak of conflict and these gain importance when predicting hard onsets. The trees in the random forest use the topics in lower nodes where they are evaluated conditionally on conflict history, which allows the random forest to adapt to the hard problem and provides useful forecasts for prevention.

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

1 Introduction

Civil wars are a serious humanitarian and economic problem. According to data from the United Nations Refugee Agency (UNHCR) in 2017 on average 44,400 people around the world had been forced from home every day, the large majority by armed conflict. Once started, a small armed conflict can quickly escalate and lead to repeated cycles of violence that have the potential to ruin society for a generation. International organizations like the UN, the World Bank, the IMF and the OECD have therefore all identified fragility as a key factor for long-term development. Most recently, this has led to calls for more resources and institutional reforms aimed at preventing civil wars (United Nations and World Bank 2017). Most explicitly this general trend was expressed by the former President of the World Bank Jim Yong Kim on September 21st 2017 (OECD 2018): *"[...], we need to do more early on to ensure that development programs and policies are focused on successful prevention."*

We argue that if academic research is to provide help for prevention then forecasting conflict becomes a justified goal. But while there is a developed literature on this issue in political science, it is completely absent in economics. This article aims to help close this gap by providing a universal forecasting framework for all countries based on a corpus of 3.8 million newspaper articles and a combination of unsupervised and supervised machine learning. We show that this framework is at least as good as any other framework based on more standard data but has the advantage of being available in real-time.

If prevention is the declared goal then the forecast problem we face is one of extremely low baseline risk, i.e. heavily imbalanced classes. The reason is the so-called conflict trap. (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. Fairly high levels of precision in forecasting conflict could therefore be reached by only looking at conflict history. But forecasts which, explicitly or implicitly, rely on a recent conflict history are not useful for the declared

goal of conflict prevention because they cannot predict when countries are at the verge of falling into the conflict trap. If we want to prevent we need to focus on the previously peaceful countries which means we need to predict cases with an extremely low baseline risk. We call this the *hard problem* of conflict prediction. The hard problem is typically not explicitly taken into account when evaluating forecasting models but is of first-order importance for prevention.

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

There is a long history in prediction in economics and for macroeconomic variables like inflation or economic growth it has long been an accepted goal of academic research. But it is also becoming more common for other outcomes as well.[1] However, for conflict the economics literature has mostly focused on understanding causal mechanisms. As a consequence, the literature has made huge strides in understanding the causes of conflict.[2] However, these efforts are often not effective for forecasting. The reason is that causal mechanisms which can be identified need not be good predictors of conflict (Ward, Greenhill and Bakke 2010; Mueller

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

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

and Rauh 2018). However, the policy context for conflict studies is clearly one where forecasts are valued and valuable. In the language of Kleinberg et al. (2015), conflict prediction is at least partly a prediction policy problem. Here, we have an approach in mind in which automated forecasts provide a benchmark for the allocation of resources (and attention) across countries but where the growing micro evidence provides insights about how they should be used.[3]

An additional reason for developing prediction models in academic research is that they are able to provide both measures of risk and measures of forecast errors. In other areas of economics this has already produced important insights.[4] In our application the side-product of the forecast is a quarterly conflict risk measure which allows us to study both false negatives (surprising conflict outbreaks) and false positives (high risk situations that never escalate into conflict). This could provide useful data on unforeseen shocks or help the study of stabilizing factors.

There is a growing interest in the use of text to generate data, i.e. feature extraction.[5] Baker, Bloom and Davis (2016) and Ahir, Bloom and Furceri (2018) use relative frequencies of predetermined keywords positively related to economic uncertainty in order to provide a measure uncertainty for the US and 143 countries, respectively.[6] For our feature extraction we rely on the full text but reduce the dimensionality through the latent Dirichlet allocation (LDA) topic model (Blei, Ng and Jordan 2003). Topic models provide an extremely useful way to analyze text because they do not rely on strong priors regarding which part of the text will be useful. In addition, the LDA is in itself a reasonable statistical model of writing and we show that it is able to reveal useful latent semantic structure in our newstext corpus.[7] We find both positive and negative relationships with conflict risk in the topics. And, perhaps surprisingly, a lot

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

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

[5] See Gentzkow, Kelly and Taddy (2017) for an overview.

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

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

of the predictive performance comes some topics reducing their share before conflict breaks out. While the interpretation of this fact is difficult, the LDA provides at least the possibility to understand the factors that provide predictive power.

Our goal is to maximize forecasting performance and so we cross-validate the model to find the optimal depth and number of trees in the random forest model, i.e. we regularize the model optimally.[8] Our regularization suggests that a model that uses a handful of variables is optimal. The reason lies in the so-called "small n large p" problem we face when forecasting macro events like conflict. The number of cases is limited and so the forecasting problem cannot be simply solved through sophisticated supervised machine learning model. A way forward in these situations is to use theory to build priors regarding the variables and model to use. An alternative, which we follow here, is to use unsupervised learning for dimensionality reduction. This method has a long tradition in macroeconomics (Stock and Watson 2006) but also outside the social sciences, in particular where sample size is small like in medical applications (Mwangi, Tian and Soares 2014).

Compared to previous work, this project advances on several fronts.[9] First, we explicitly take conflict history into account. This treatment of history allows us to evaluate performance for cases without a violent past. To the best of our knowledge our project is the first that evaluates its performance in these crucial, hard cases. Second, we use a new full-text archive of 3.8 million newspaper articles dense enough to summarize topics at the quarterly level for nearly 200 countries. This, in turn, allows us to combine feature extraction using unsupervised learning with supervised machine learning. We show that, over time, the supervised model relies more and more on the extracted features and improves its forecasting performance. This suggests that the supervised learning is actually benefitting from generalizable, subtle signals contained in the extracted features. Finally, we rely on innovations in the estimation of the topic model (Blei and Lafferty 2006) to solve the computational challenges implied by the need

[8] This is not common practice even in research which clearly aims to provide predictions.

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

to re-estimate the topics from millions of articles for every quarter. This makes our method particularly useful for actual policy applications that rely on timely risk updates using vast amounts of text.

In what follows we first explain the importance of the so-called conflict trap for conflict forecasting. We also show that countries seem to transition in an out of the trap so that treating it as a characteristic of the country which is fixed 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, followed by a section in which we show the importance of topics in the machine learning exercise. In Section 6 we present case studies of our risk measure before we conclude.

2 The Hard Problem of Conflict Prediction

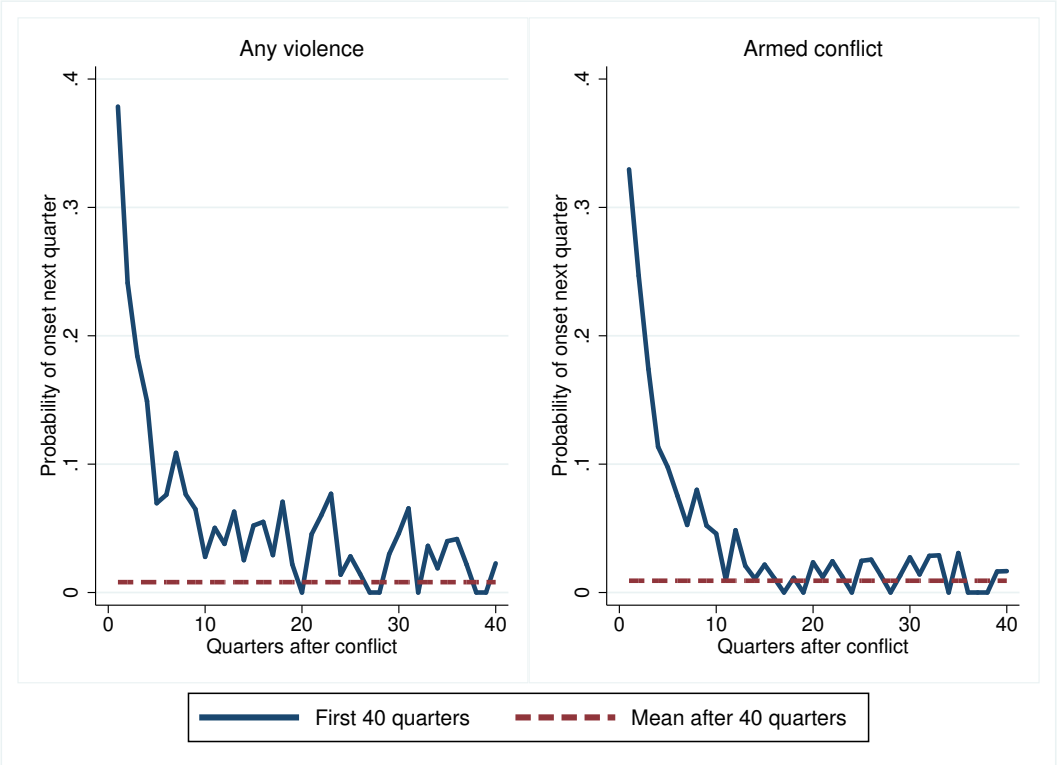
The most encompassing conflict data is provided by the UCDP Georeferenced Event Dataset (Sundberg and Melander 2013; Croicu and Sundberg 2017). We include all battle-related deaths in this dataset and collapse the micro data at the country/quarter level. The data offers three types of conflict which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

The conflict literature often relies on absolute fatality counts to define conflict. However, these are typically defined at the yearly level and it is not obvious how to translate these definitions to the quarterly level. In addition, onsets of intense violence are relatively easy to predict with ongoing less-intense violence. We therefore use two definitions of conflict. The first takes a quarter with one or more fatalities as conflict (any violence), and the second assumes that con-

flict is a quarter with at least 50 fatalities (armed conflict). We only consider onset, i.e. only the quarter conflict breaks out. Subsequent quarters of conflict are set to missing. This is important as predicting outbreaks is much harder than predicting conflict. In our data we have 739 onsets of any violence and 450 onsets of armed conflict.

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

Fig. 1: Likelihood of Conflict Relapse



Note: Figure shows the sample likelihood of conflict relapse after violence ended (at 0) conditional on remaining in peace.

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

Of course, it might be tempting to instead focus on cases that are easier to predict - and indeed this is what the current system of peacekeeping is geared to do. But the conflict trap does not only make prediction outside of it difficult, it is also a serious long-term threat that countries should avoid. Armed conflict has its own dynamics and a simulation of possible futures with the data displayed in Figure 1 indicates that prevention in hard conflict cases could have considerable dynamic payoffs. This is because the expected future outcome when in conflict or post-conflict peace is suprisingly similar but differs dramatically to peace in hard cases. When a country experiences an outbreak of violence in a hard-to-predict scenario its possible future seems to change dramatically. This is much less true for an outbreak of violence post-conflict.

Therefore, our motivation to study the hard problem is prevention. Prevention is of key interest for the international community. All big international organisations treat fragility and conflict risk as key problems.[10] 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.

[10]This is best illustrated by a joint press conference by the UN Secretary-General Antonio Guterres and the President of the World Bank Jim Yong Kim on September 21st 2017 which we cite in the introduction.

3 Simulating the Policy Problem

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

As a basis of our method we use a new unique corpus of 3.8 million documents from three newspapers (New York Times, Washington Post and the Economist) and a news aggregator (BBC Monitor). A text is downloaded if a country name or capital name appears in the title or lead paragraph. The resulting data is described in detail in the Online Appendix. Using the dynamic topic model we derive topic models with $K = 5$, 10, 15, 30 and 50 topics. The reason we choose relatively few topics is to avoid topics adapting to particularly newsworthy cases of conflict, regions or countries. Topic models between 5 and 15 topics only contain topics that can be attributed to generic content like politics or economics, whereas with 30 or 50 topics tend to become specific to certain situations or countries.

Figure 2 shows word clouds for four out of 15 topics estimated on the 2017Q3 sample depicting the most likely terms proportional to their importance in size. The first topic in Panel a) is what we describe as the economics topic. It features terms such as "econom","dollar", and "growth" prominently. Panel b) displays a topic which features mostly terms related to the military. Similarly, Panels c) and d) present other topics related to conflict, which we label terror and violence due terms such as "terrorist" and "kill" being keywords, respectively.

With the estimated topic model we then calculate the share of topics for all countries in each quarter between $1989Q1$ and T. We then use these shares, together with a set of dummies which capture the post-conflict risk, in a random forest model to forecast conflict out of sample.

Fig. 2: Word Clouds of Topics

Note: The word clouds represent the most likely terms of 4 out of 15 topics estimated using all text until 2017$Q3$. The size of a token is proportional to the importance of the token within the topic. The location conveys no information.

In this step we take the perspective of a policymaker who observes all available text and conflict until period T and has to make a forecast for period $T+1$. We summarize the K topic shares and the log of the word count in the vector, θ_{it}. We then train a forecasting model with all data available in T using the model

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

where y_{it+1} is the *onset* of conflict in quarter $t+1$, $\mathbf{h}_{it}$is a vector of dummies capturing post-conflict dynamics and lower levels of violence. The role of machine learning in this step is to discipline the regularization of the model $F_T(\mathbf{h}_{it}, \theta_{it})$. We fix the hyperparameters in the first

sample (1989Q1-2000Q1) through cross-validation. The newest conflicts that break out in the training sample are those that break out in T. Note that this implies that, during training, we only use data for $\mathbf{h}_{it}$ and θ_{it} available until $T-1$. With the resulting model we then produce predicted out of sample values

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

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

We then update our topic model with the news written in the next quarter, add the new information on conflicts, retrain the prediction model, and predict the probabilities of outbreaks in the following quarter. For testing, we thereby produce sequential out-of-sample forecasts, $\hat{y}_{iT+1}$, for $T+1 = 2000Q2, 2000Q3, ..., 2017Q4$. 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.

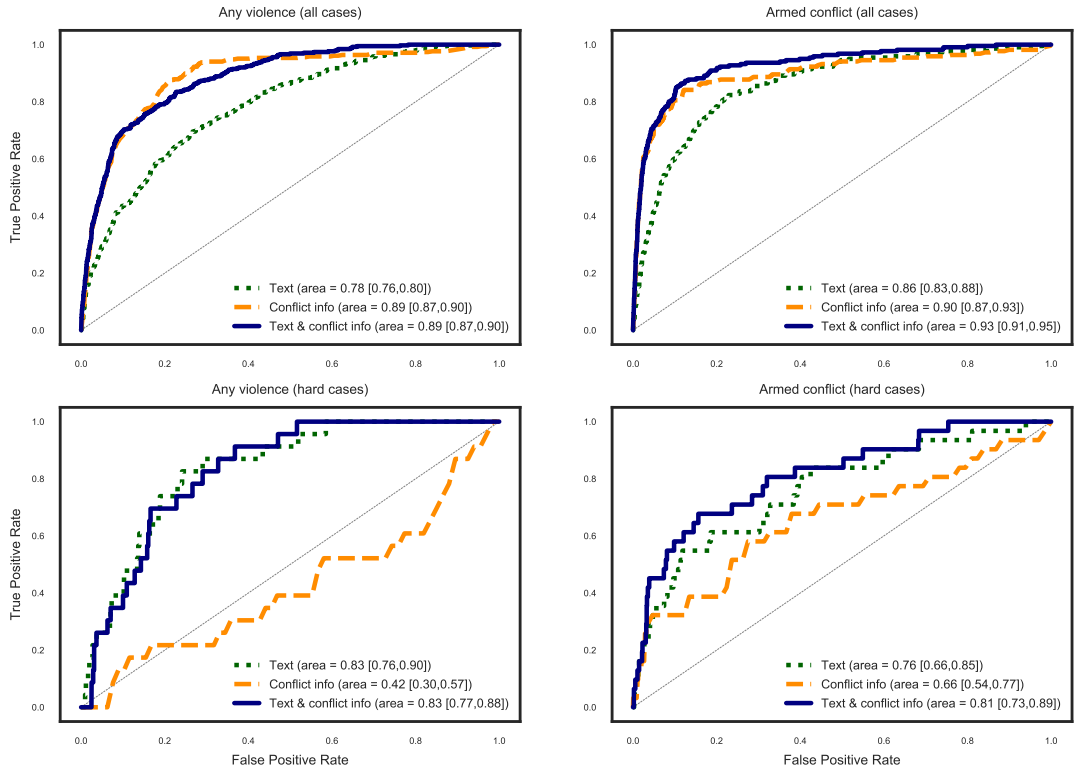
To generate $F_T(.)$ we tested predictions from k-nearest neighbor, adaptive boosting, random forests, neural network, logit lasso regression, and ensembles of all previously mentioned models. The hyperparameters are chosen by maximizing the AUC through cross-validation within the sample up to $2000Q1$.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

4 Solving the Hard Problem

Figure 3 show receiver operating characteristics (ROC) curves for the two cutoffs we analyze, i.e. at least 1 (any violence) and 50 (armed conflict) battle deaths, respectively. ROC curves display the trade-off between true positive rate and false positive rate. False negatives drive the true positive rate (TPR) down. False positives drive the false positive rate (FPR) up. A way to

summarize the performance of the model here is the area under the curve (AUC).

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



Note: The prediction method is a random forest with a tree depth of 7 and 500 trees for any violence (left) and a tree depth of 4 and 425 trees for armed conflict (right). 'Text' contains 15 topics and token counts and 'conflict info' contains 4 dummies indicating the first quarter, quarters 2-4, years 2-5 and years 6-10 after the last conflict and a dummy for the presence of any violence. Top and bottom ROC curves are alternative evaluations of the same forecasting model. Hard cases (bottom) are defined as not having had armed conflict in 10 years. The bottom ROC curves are evaluated only for those cases. Bootstrapped 95% confidence intervals of the AUC are reported in square brackets.

In each panel, we show the forecasting performance of three sets of variables: (i) A model using just topics and word counts as predictors, which is labeled as *text*, (ii) a model using only information about present or previous violence, which is labeled as *conflict info*, and (iii) a model that draws from both. More specifically, conflict info contains four dummies capturing conflict history: an indicator whether there was conflict (i) last quarter, (ii) 2-4 quarters ago, (iii) 2-5 years ago, or (iv) 6-10 years ago. Moreover, for armed conflict the set of predictors

contains a dummy indicating whether any violence is present.

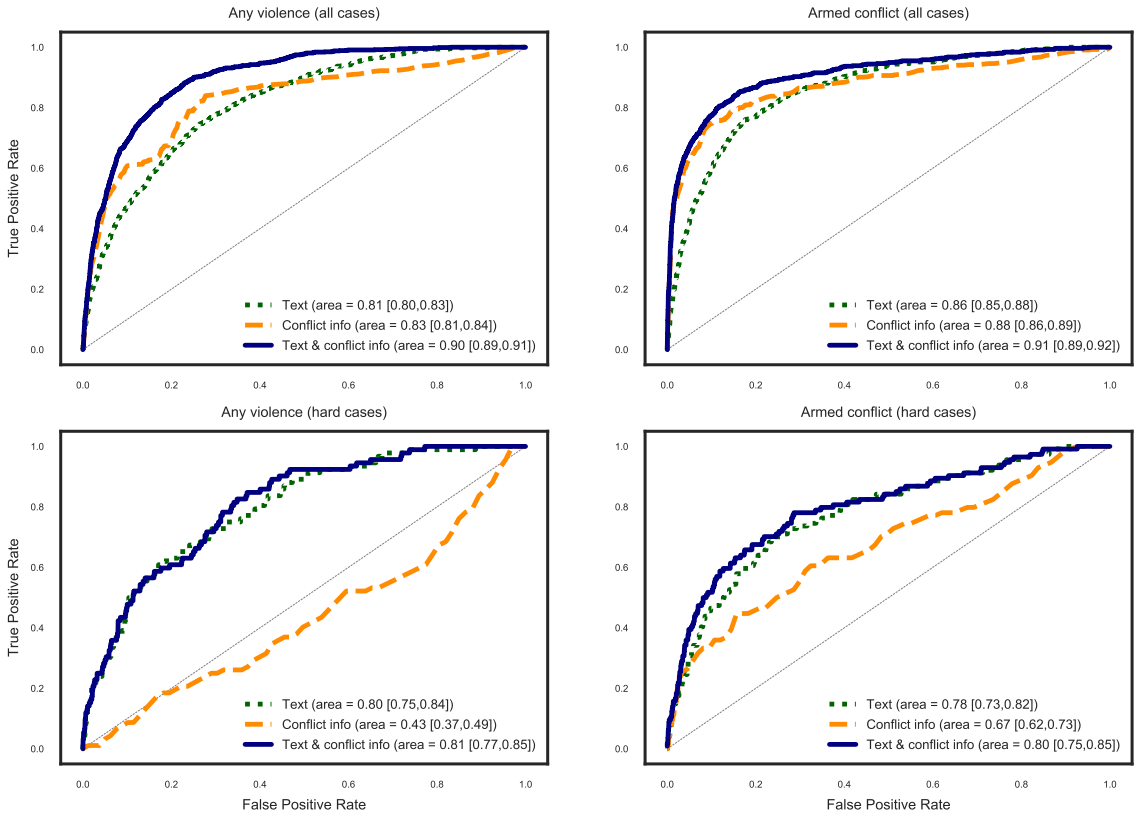
On the top of Figure 3 we see the overall performance of all three models when forecasting any violence (left) and armed conflict (right). Text alone (green dotted line) provides some forecasting power and this forecast is not worse to what is common in the literature when predicting at the quarterly level. But it is clear that the conflict information (orange dashed line), a simple model of five dummies, dominates the text forecast. The combined model reaches an AUC of 0.89 for any violence and an AUC of 0.93 for armed conflict. We generated bootstrapped calculated confidence intervals for the AUC but, even at the lower bound of these confidence intervals, the AUCs of the combined model are relatively high. Importantly, the AUC and could not be improved significantly by adding more variables.

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

For many applications in prevention a prediction of one year ahead prediction will be more useful. In Figure 4 we therefore provide evaluations of a prediction model that considers an onset if conflict breaks out within any of the four following quarters. The predictive performance of the model remains strong. A forecast of onset up to a year ahead produces an AUC of 0.90 for any violence and 0.91 for armed conflict and topics now adds significant forecasting power even in cases with a conflict history. This is due to the fact that an immediate conflict history

is less informative about a renewed outbreak after a year. In other words, the text features now seem to capture some of the post-conflict dynamics.

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

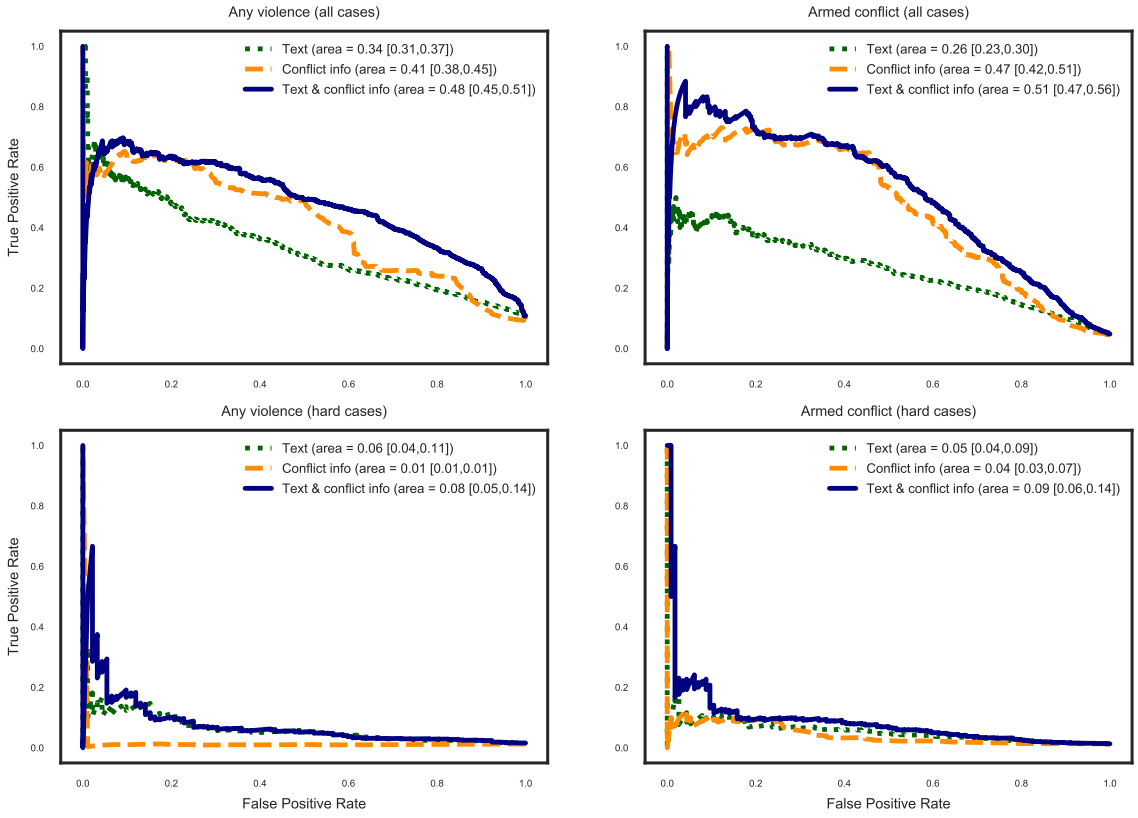


Note: 'Text' contains 15 topics and token counts and 'conflict info' contains 4 dummies capturing time passed since the last conflict and dummies for the presence of lower levels of violence. Hard cases are defined as not having had conflict in 10 years. Bootstrapped 95% confidence intervals of the AUC are reported in square brackets.

Figure 5 shows precision-recall curves for the one-year ahead forecasts. In all four graphs the x-axis displays the true positive rate, while the y-axis summarizes the precision, i.e. the share of alarming situations where conflict actually broke out. On the top we show the results for all onsets. Precision overall is extremely good (around 50-60 percent for any violence and 60-80 percent for armed conflict when the TPR is between 10-50 percent). On the bottom we show results for hard onsets. Precision deteriorates somewhat when predicting hard onsets (around 10-20 percent for any violence and armed conflict when the TPR is between 10-50 percent).

This is due to the extreme imbalance in the hard onset sample. Figure 1 showed the precision one would achieve with a random forecast as a red dashed line. A precision score of over 10 percent is a significant improvement in this context.

Fig. 5: Precision-Recall Curves of Forecasting Violence

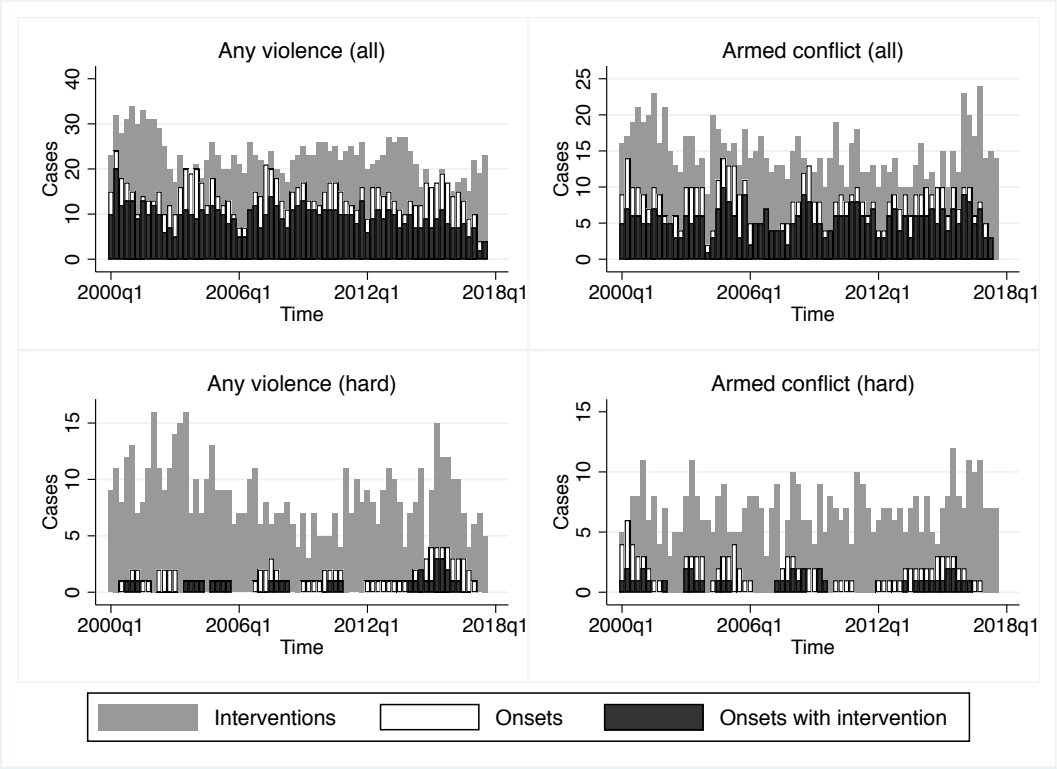


Note: 'Text' contains 15 topics and token counts and 'conflict info' contains 4 dummies capturing time passed since the last conflict and dummies for the presence of lower levels of violence. Hard cases are defined as not having had conflict in 10 years. Bootstrapped 95% confidence intervals of the average precision are reported in square brackets.

Another way to look at precision in this context is to run simulations of how many interventions would be needed to reach a given TPR overall and look at the resulting precision over time. We do this for the year-ahead forecast in Figure 6. Given the overall precision levels displayed in Figure 5 it is reasonable to assume that, given the low precision, policymakers would not want to reach very high levels of TPR for hard onsets but would aim higher in overall cases. We simulated the number of interventions necessary to reach a TPR of 2/3 in all cases and a TPR of

1/3 in hard cases. 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 is then given by comparing the black bars to the grey bars. The TRP is given by comparing the while bars to the black bars and is, naturally, lower for hard onsets.

Fig. 6: Simulating Timing and Frequencies of Interventions Using our Model



Note: The predictions underlying the figure are based on a model using 15 topics and token counts as well as 4 dummies capturing time passed since the last conflict and a dummy for the presence of lower levels of violence. Hard cases are defined as not having had conflict in 10 years. The cutoff for interventions is chosen such that a TPR of 2/3 is reached in all cases and a TPR of 1/3 in hard cases. 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).

To benchmark the performance of text, we compare the predictive power to standard variables in the forecasting literature (Goldstone et al. 2010) including political institutions dummies based on various dimensions of the Polity IV data, number of neighboring conflicts, the newest data on child mortality from the World Bank and the share of population discriminated

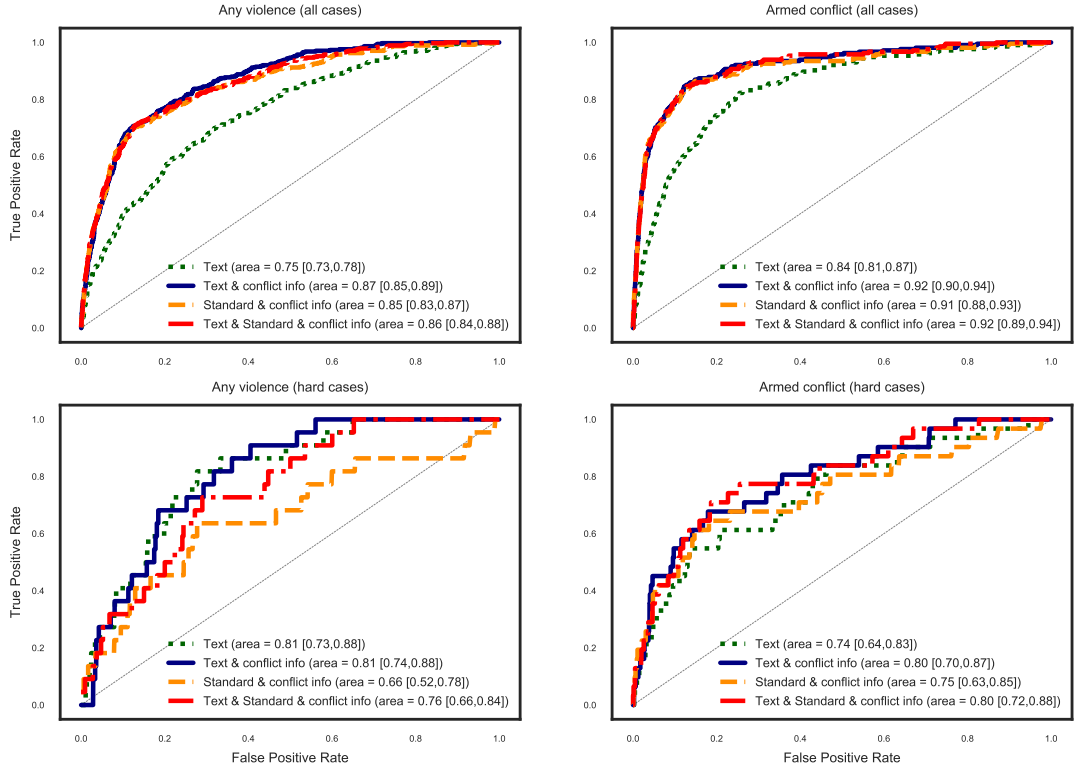
against using data from the Geographical Research On War, Unified Platform (GROWup) (Girardin et al. 2015).

In Figure 7 we compare the performance of the topics to the standard variables related to politics and economics (orange line). Variables such as infant mortality are available for less countries and years. For the sake of comparability, we only use overlapping predictions for evaluation, i.e. country-quarters in which the availability of data allow predictions for both sets of variables.

We find that standard variables never add forecasting power. If anything, a text-based forecast which ignores standard variables is better. Moreover, topics have the advantage that they are based on newspaper text which is available on a daily basis whereas standard variablesare available with lags up to several years, which we do not consider in this comparison. More problematic is the practice of recoding macro variables which means some of the variables in the standard model use future information. Therefore, the presented predictive power of these variables should be considered an upper bound. Other obvious economic variables like GDP growth or levels do not add any forecasting power either.

A detailed discussion of parameters, separation plots, additional results and robustness checks are reported in the Online Appendix. Two findings worth highlighting are: First, more than 40 variables generated from text-based event data do not provide a better forecast than our topics topics. If anything, topics produce higher AUCs. In addition, we find that when forecasting armed conflict, our topics and the event data have some complementarities in the sense that a model with conflict history plus both sets of variables performs better than a model that relies on only one of the two. These findings are in line with the idea that the event data is better able to capture a situation which might escalate, whereas the forecast in the topic model relies only in parts on escalation. Second, we also find that the random forest model performs particularly well when compared to other methods of supervised machine learning like logit lasso regressions. We now turn towards providing a tentative explanation for why our model

Fig. 7: ROC Curves of Forecasting Any Violence (left) and Armed Conflict (right) Compared to Standard Variables



Note: 'Text' contains 15 topics and token counts, 'conflict info' contains 4 dummies capturing time passed since the last conflict and dummies for the presence of lower levels of violence, and 'standard' contains infant mortality, political institutions, share of discriminated population, and neighboring conflicts. Hard cases are defined as not having had conflict in 10 years. Bootstrapped 95% confidence intervals of the AUC are reported in square brackets

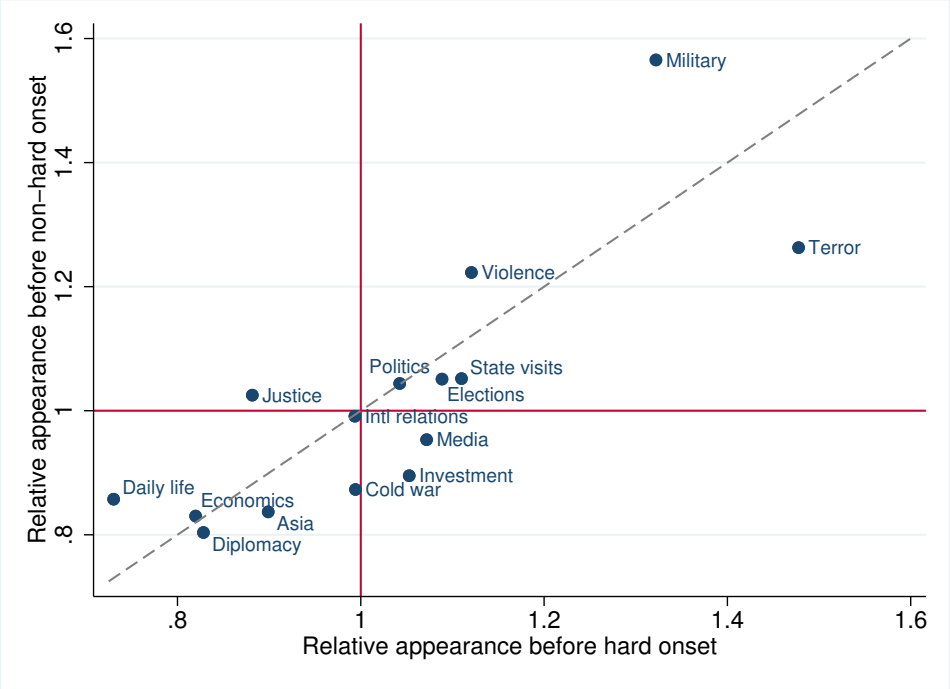
works so well.

5 How Machine Learning Solves the Hard Problem

We use a methodology which is standard in other areas like inflation forecasting or pattern recognition and apply it to the prediction of conflict. First, unsupervised learning is used for feature extraction. Then these features (topics) are used to predict conflict. An important advantage of this approach is that the model can rely on positive and negative associations of specific

topics with violence. In Figure 8 we show the shares of each of the 15 topics in quarters before an onset in hard cases (x-axis) and in onset cases where the country has a conflict history (y-axis) relative to quarters in which there is no onset in the next period. For instance, the military topic is 1.3 times more likely to appear before a hard onset relative to a peaceful quarter but almost 1.6 times more likely to appear before an onset in a country in with a conflict history. Other topics like economics, investment, daily life or justice appear less before onsets. News stories on daily life, for example, are almost 30 percent less likely before a hard onset. In this way the topics provide signals which the supervised learning is able to exploit.

Fig. 8: Topic Shares Before the Onset of Any Violence Relative to Peaceful Quarters



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

The random forest relies on these different associations by combining conflict history and the various topics in decision trees. In the top panels of Figure 9 we show the relative total importance of the topics compared to conflict history in our random forest model. For simplicity,

we combine the importance of the topics in one bar and only distinguish topics by whether they contain tokens that indicate violence prominently. In this way we separate the signals contained in the three positively related topics (violence, terror and military) from the other topics.[11]

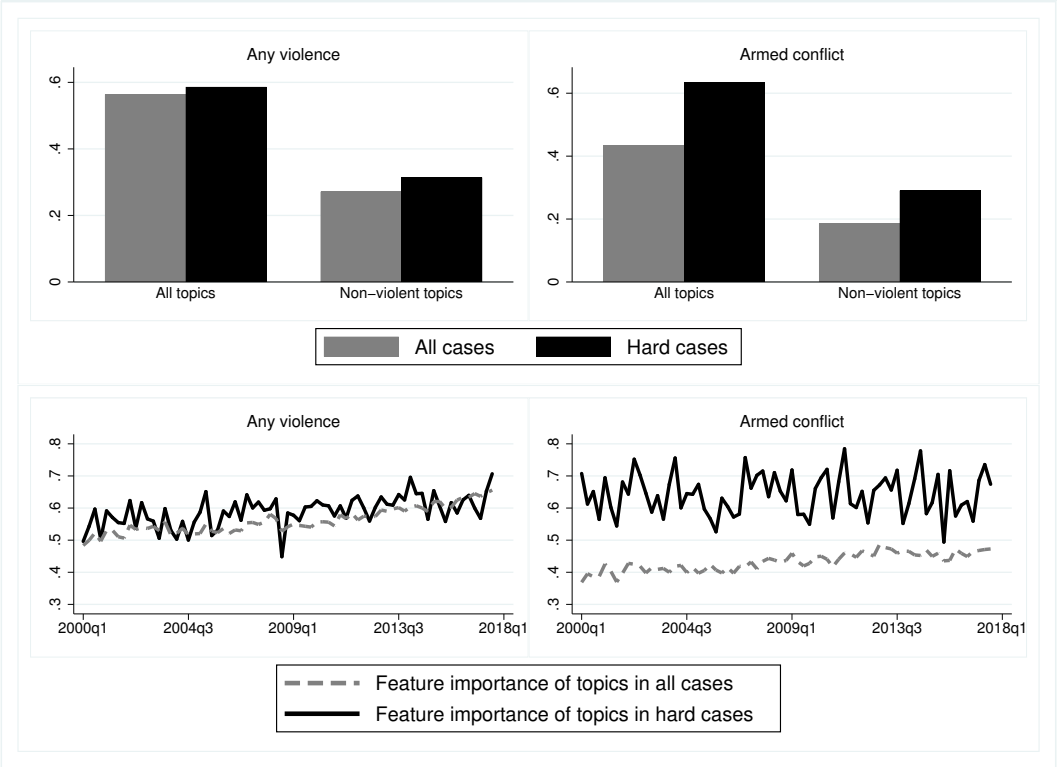
The gray bars in Figure 9 indicate the relative total importance of topics when predicting conflict generally and the black bars indicate their importance in hard cases. Topics provide 50 percent of total importance and non-violent topics around 25 percent. The rest can be attributed to the token count and the variables capturing conflict history or low levels of violence. Importantly, the share increases when predicting hard cases and this increase is partly driven by non-violent topics. In other words, the forecast of hard cases seems to rely to a large degree on the subtle associations displayed in the lower left corner of Figure 8.

But why is the random forest model using this subtle variation better than other models? A detailed analysis of our forecasting models shows that the decision trees tends to pick conflict history at the top of the tree to divide the sample. For example, the dummy indicating the first quarter post-conflict receives the largest importance score by far which means it is used in top nodes of the tree more often. Topics are then introduced in lower branches. In other words, the random forest model is automatically geared towards picking up more subtle risks with topics when conflict history is absent. In this way the forecasting model works around the importance of the conflict trap by conditioning on conflict history and at the same time uses information contained in the text. An additional, even more subtle aspect of this process is that the model uses topics to capture stabilizations in countries with a conflict history. This drives down the false positive rate.

An important aspect of our method is that the sample available to the model is increasing in time. In the bottom panels of Figure 9 we use this fact to show how the total importance of topic changes over time, i.e. with a growing sample. Again, we see that the importance of topics is higher when predicting hard cases. In addition, importance of the topics is increasing

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

Fig. 9: Feature Importance of Topics in Random Forest



Note: The feature importance is calculated on sequential out-of-sample predictions of random forests with conflict history and text using a tree depth of 7 and 500 trees for any violence and a depth of 4 and 525 trees for armed conflict.

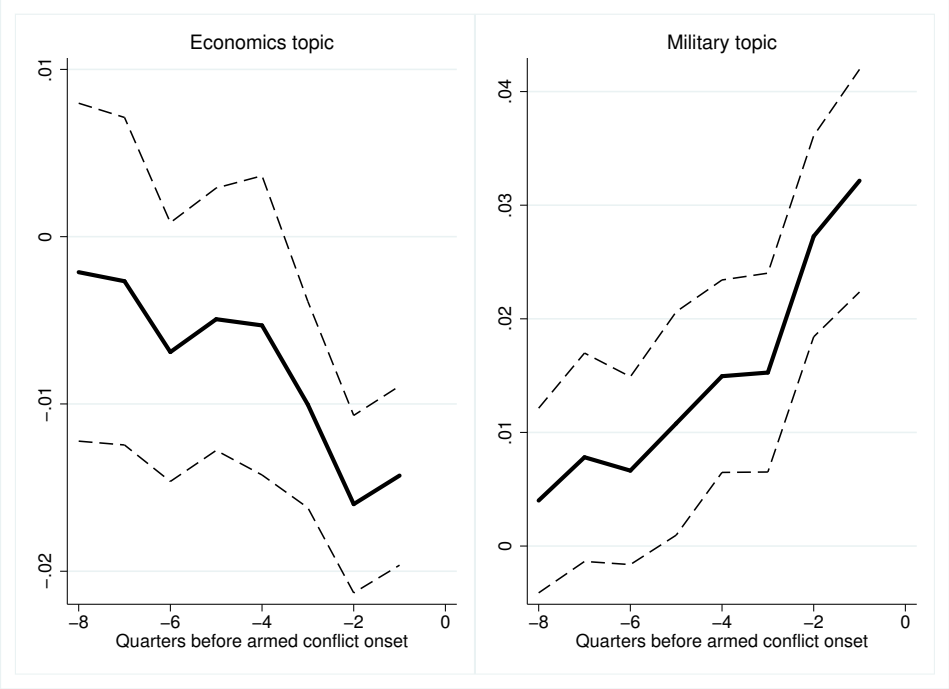
considerably for any violence. This means that the random forest model relies more and more on the text to separate high from low risk. Using cross validation we confirm that the overall predictive performance of the resulting random forest also tends to increase over time. This is surprising given the dramatically changing international context and new instabilities in the period 2000 to 2017.

6 Dynamic Predicted Risk and Case Studies

There are several strong positive and negative associations between different topics and conflict onset. Importantly, these relationships are not only varying between countries but also within countries over time so that meaningful dynamic risk profiles result from our forecasts. To

illustrate the dynamic properties of two topics, Figure 10 shows the movement of the economics (left) and military (right) topic shares and the 95 percent confidence interval before the onset of armed conflict in the full sample. The figure is based on regressions which include country fixed effects so that they are showing the change in topics within countries over time. The increase of reporting on military and the decline of reporting on economics leading up to the outbreak of violence is very clearly visible.

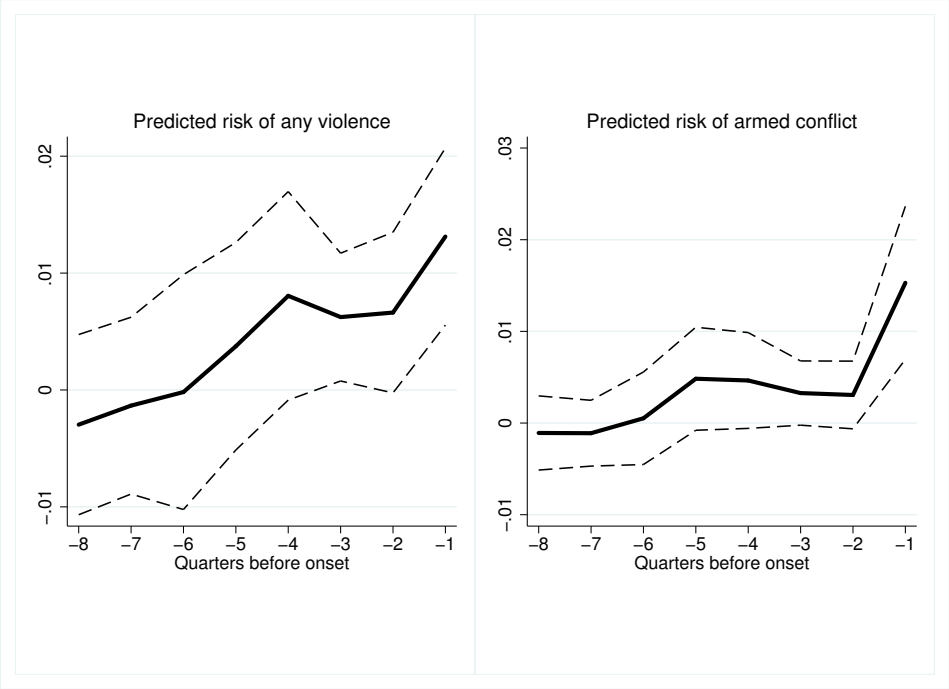
Fig. 10: Share Written on Economics and Military Topic Before Onset of Armed Conflict



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

As a result of such movements, the risk evaluations that come out of our methodology are changing over time. In Figure 11 we show the rolling out-of-sample risk estimates coming out of the text-only model controlling for country fixed effects as onset approaches. The predicted risk is clearly increasing both when forecasting any violence (left) and when forecasting armed conflict (right).

Fig. 11: Predicted Risk Before the Onset of Conflict

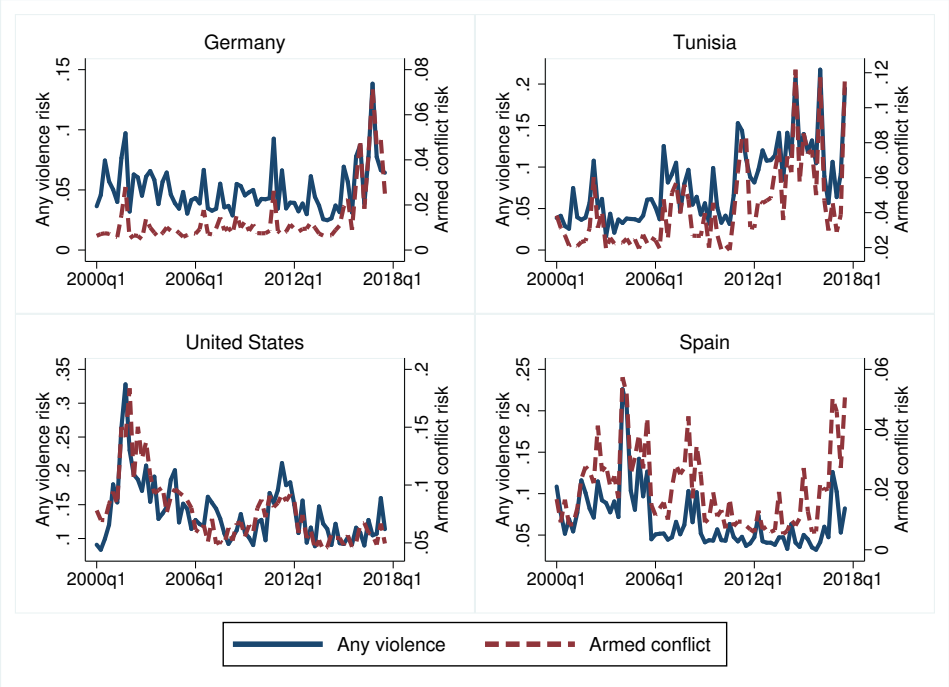


Note: Risk residuals relative to other quarters in the same country are represented by the solid lines. The data is generated through regressions of the out-of-sample predicted risk on dummies for the number of quarters before the onset of conflict and country fixed effects. Dashed lines mark 95 percent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

Given the dynamic movements of estimated risk within countries reflect actual onset risk, it makes sense to treat our risk estimates as data to be analyzed. We picked four countries which illustrate the nature of the risk forecast clearly, which are reported in Figure 13. The red dashed lines report the risk of armed conflict (right y-axis) whereas the blue lines report the risk of any violence (left y-axis). In all cases the risk estimate is for the next quarter.

What is clear from the figures is that risk reacts to violence. For example, the change in risk after the September 11 terror attacks in the United States is visible as a large shock. Similarly, terror attacks in Spain, Germany, and Tunisia all brought changes in predicted conflict risk with them. Also, the figures give a very clear idea of high risk and low risk periods in the respective countries. Germany entered a period of relatively high risk after 2014 and Tunisia after 2010. In the United States risk has fallen to relatively low levels in 2017Q3. Spain had a relatively

Fig. 12: Predicted Risk of Any Violence and Armed Conflict (Case Studies)

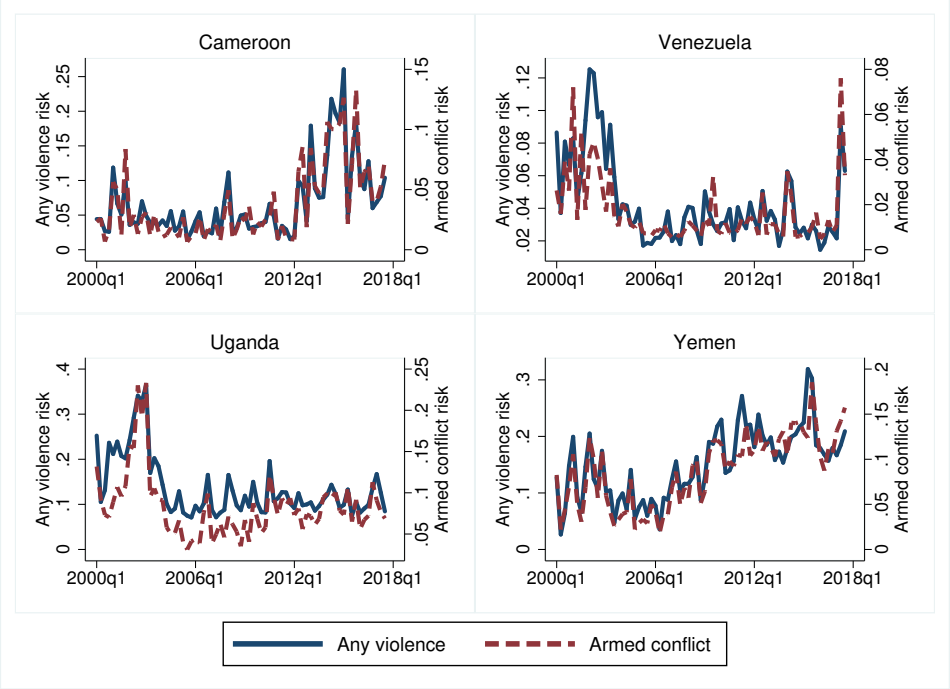


Note: Predictors include 15 topics and token counts as well as 4 dummies capturing time passed since the last conflict and dummies for the presence of lower levels of violence.

calm period but, given its secessionary movements and the central government's response, is recently experiencing higher risk again.

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

Fig. 13: Predicted Risk of Any Violence and Armed Conflict (Case Studies)



Note: Predictors include 15 topics and token counts as well as 4 dummies capturing time passed since the last conflict and dummies for the presence of lower levels of violence.

7 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 supervised and unsupervised 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 method could therefore also be used to predict other policy-relevant events, such as migration flows or economic uncertainty and recessions.

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

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

Appendix

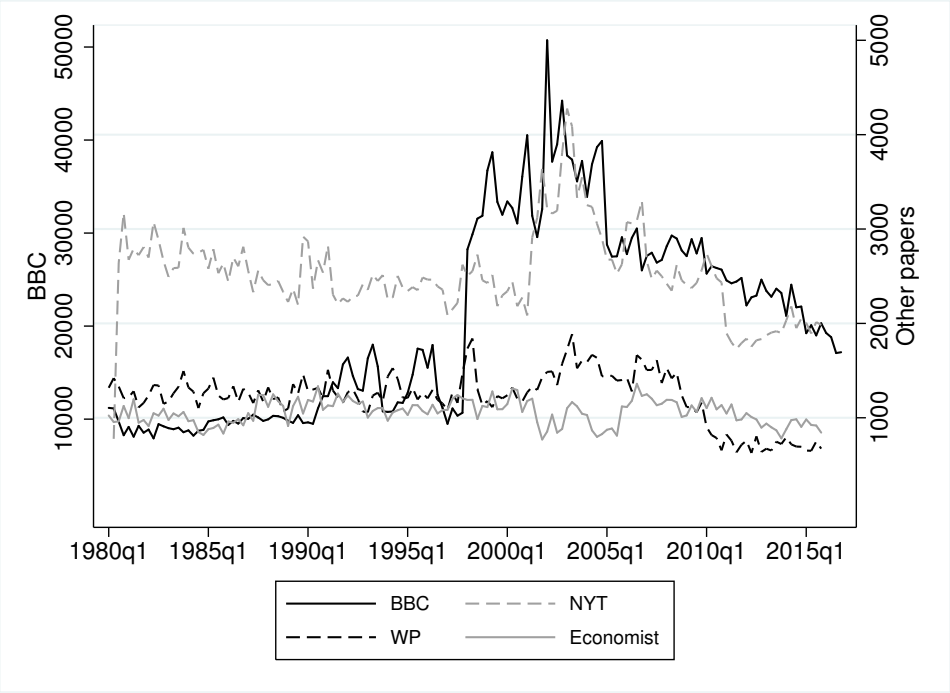
A Data Description

All our text data is downloaded manually from Lexis Nexis. Due to copyright issues the raw newspaper articles cannot be shared. Summarized topics and all other data and codes will be made available upon publication. The key factor in choosing our news sources is that they should be english-speaking, offer as much text as possible and long time-series. We therefore chose the New York Times (NYT), the Washington Post (WP), the Economist and the BBC Monitor (BBC). The latter source tracks broadcasts, press and social media sources in multiple languages from over 150 countries worldwide and produces translations into English.

We download an article if the name of the country or its capital appears in the title of the article. This gives us a panel of articles from all sources for over 190 countries for the period 1989Q1 to 2017Q3. In total we have access to about 700,000 articles from the New York Times, Washington Post and the Economist and 3.1 million articles from the BBC Monitor. This means that the BBC Monitor articles dominate our data. Figure A.1 shows the number of articles we have for each quarter. From this is clear that the number of BBC news available from Lexis Nexis increases around 2000. Part of this increase came from a change in the headlines, which around this period often began with the country name followed by a colon and then a traditional headline. This temporary change does not affect the regional distribution substantially. In any case, the increase in the amount of news is only problematic for our forecasting exercise if the increase or decrease in the number of news somehow affects the share of news written on a specific topic. It would then become impossible to use this data effectively for forecasting as the training of the model would not produce useful forecast in the testing sample.

In Table C.3 we summarize the different sets of predictors we use. The first model is based on our text data. This includes 5, 10, 15, 30 or 50 topic shares and the log of the word count of

Fig. A.1: Number of Articles by Source



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

that quarter. The word count varies between 5 and 1226371 and is log-normally distributed with a mean of 4274 words, while the topics sum to one within each country-quarter. The second model is based on the violence data from GED. It includes dummies for the conflict history and dummies for ongoing low-level violence. Finally, we use the ICEWS event database to generate a quarterly panel between 1995Q1 and 2016Q4. We only use events in which the source and the target of the action were in the same country. We then make a count of all 20 event types on the Conflict and Mediation Event Observations (CAMEO) integer scale and another count of all 20 events on the CAMEO scale that involve the government either as target or as source. In addition, we generate a count of all protest events, the average CAMEO code of events involving the government, and the average CAMEO code of all events taking place in the country.

Table A.1: Sets of Predictors

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

B Discussion of Estimated Topics

In this section we discuss various aspects of the estimated topics. We estimate the topic model repeatedly starting with all text until 2000Q1 and then we update every quarter using a dynamic topic model (Řehůřek and Sojka 2010). We allow the weight variational hyperparameters for each document to be inferred by the algorithm. Before feeding the text to the machine learning algorithm we conduct standard procedures when working with text. We remove overly frequent words defined as stopwords. Then we stem and lemmatize the words before also forming two and three word combinations. Next we remove overly frequent tokens, i.e. those appearing in at least half of the articles. Finally, we also remove rare expressions appearing in less than 100 documents.

In what follows we will focus on topic models estimated in 2017Q3 as these describe all relevant text we use in our forecasting framework. Table B.1 summarizes the top 10 terms in the $K = 15$ topic model estimated in 2017Q3. There are three topics which are clearly related to conflict. Topic 1, which we label the military topic, contains terms like army and military and terms indicating fighting. Topic 8, which we label the violence topic, has terms like province and other location indicators mixed with words indicating fighting.

In Figure B.1 we show the timeline for these two topics in our sample period for Afghanistan, Angola, Iraq and Ukraine. The aim is to show that the two topics give a fairly good idea of the

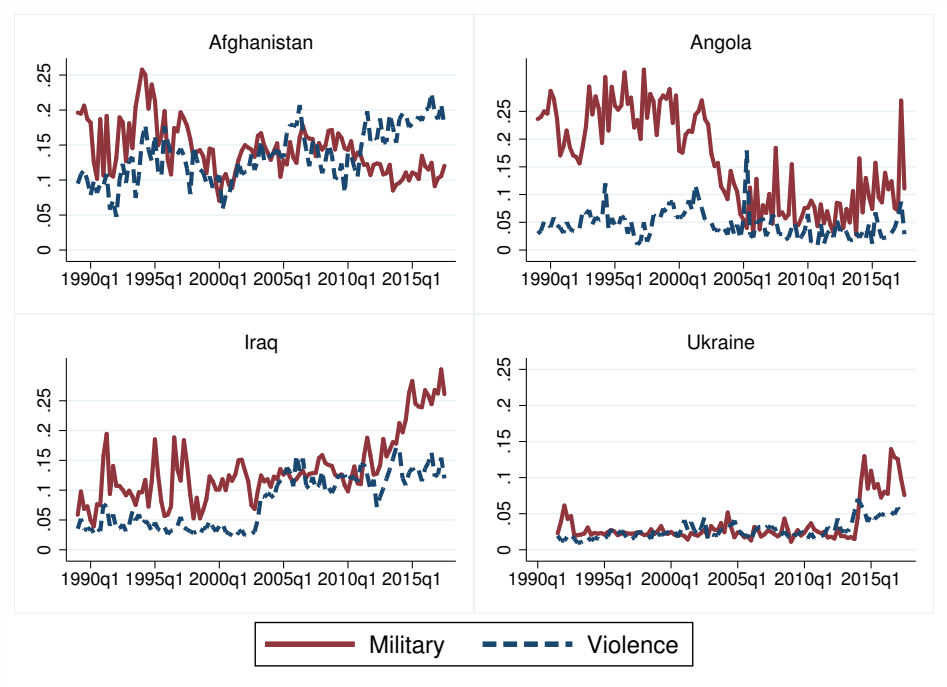
different conflict histories in the respective countries. In the image for Afghanistan we see that both topic shares fluctuate wildly and are very high throughout the period, accounting for between 6 and 21 percent of all news written on Afghanistan. Most recently, more is written on violence. In Angola writing on conflict topics decreases dramatically following the cease-fire in 2002. In Iraq the invasion of 2003 is very clearly visible. In Ukraine the start of the turmoil in 2014 and outright war later is clearly visible in the news text. Of course, such movements in conflict topics are only helpful for forecasting if they anticipate conflict. In Figure B.1 we show this for the military topic.

Table B.1: Top Ten Keywords of 15 Topic Model Using All Text Until 2017Q3

Nr	Label	Keywords
0	Justice	court, case, law, investig, right, human, arrest, offic, human_right, polic
1	Military*	forc, militari, armi, oper, border, arm, group, region, secur, troop
2	Media	ministri, medium, channel, region, inform, servic, say, head, accord, author
3	Daily life	websit, user, school, bit, com, child, time, onlin, woman, work
4	Terror*	islam, terror, terrorist, arab, muslim, group, leader, sharif, say, persian
5	Intl relations	secur, intern, region, relat, cooper, issu, develop, support, import, council
6	Politics	trump, time, polit, war, say, bit, polici, make, want, like
7	Investment	project, compani, trade, develop, oil, energi, gas, product, construct, industri
8	Violence*	attack, polic, kill, video, provinc, citi, secur, district, area, local
9	Diplomacy	beij, cooper, visit, tie, myanmar, bilater, develop, headlin, relat, trade
10	Economics	cent, bank, websit, compani, econom, economi, dollar, billion, busi, financi
11	State visits	meet, talk, agreement, visit, issu, prime, peac, negoti, discuss, offici
12	Cold war*	nuclear, missil, websit, defenc, ukrainian, militari, test, korean, dprk, weapon
13	Elections	parti, elect, opposit, polit, vote, parliament, ial, leader, constitut, candid
14	Asia	sea, korean, sanction, offici, abe, washington, secur, donald, island, militari

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

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



C Prediction Algorithms

To explore the gains from supervised learning we look at five different algorithms specified in Table C.1 which are trained with the available data using a Python implementation (Pedregosa et al. 2011). We standardize the data in order to improve the performance of machine learning algorithms such as neural networks.

Table C.1: Models

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

The five individual supervised prediction algorithms we use are a logistic lasso regression,

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

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

While the final evaluation of our model is carried out strictly out-of-sample, i.e. in the future without using any contemporaneous or future information, the training of the models is performed through cross-validation. More specifically, the method used is k-fold cross-validation, where the training set is split into k smaller sets. For each of the k 'folds' the following procedure is performed: A model is trained using $k-1$ of the folds as training data; using the remaining data a test is carried out by computing our chosen performance measure, the AUC. The performance measure reported by k-fold cross-validation is then the average of the values computed in the loop. Each individual algorithm also requires the specification of hyperparam-

Table C.2: Hyperparameters

Predictors	Random forest	
	Depth	Trees
Any violence		
Text	7	400
Conflict info	4	10
Text & conflict info	7	500
Armed conflict		
Text	5	250
Conflict info	4	50
Text & conflict info	4	425
Civil war		
Text	5	100
Conflict info	1	150
Text & conflict info	2	275

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

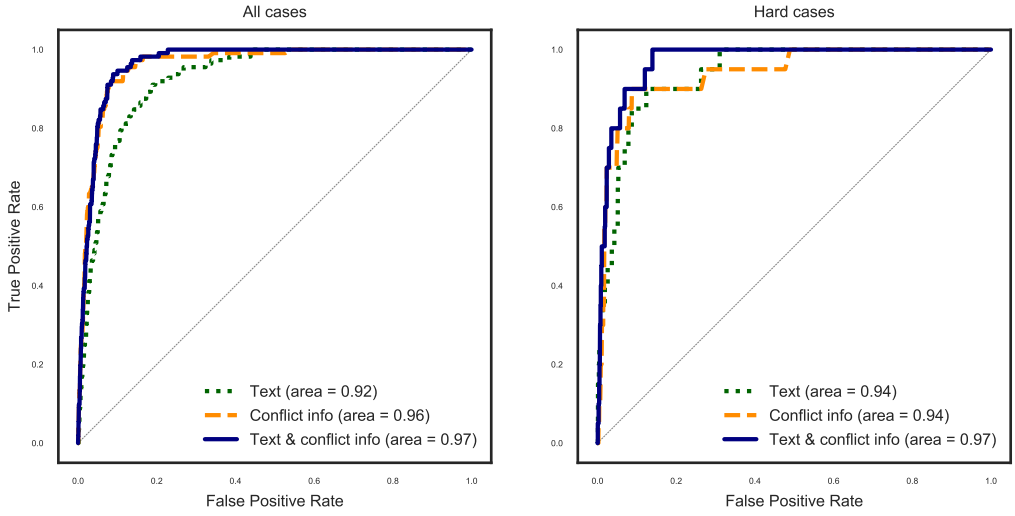
eters by the user. For each set of predictors, we choose these hyperparameters by doing a grid search using the sample until the year 2000 and then selecting the hyperparameters that generate the highest AUC. Note, that this will understate the performance of the forecasting model slightly if more information leads to a deeper or modified model in later years.

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

D Additional Results

In the following section we show additional results, including for a greater cut-off of 500 battle deaths. In Figure C.1 we see that for the very large cutoff in terms of battle deaths, the onset of conflict becomes relatively easy to predict. The harder it is to predict conflict, the more topics add to the forecasting power. In particular, when forecasting the hard cases of any violence, the text-only model provides a relatively good forecast given the difficulty of predicting these events. When predicting civil war, the presence of any violence or armed conflict are powerful predictors, even in the hard cases, which is why it is difficult to augment the prediction of further escalation even with text. However, one should note that text alone also achieves high levels of accuracy for all and the hard cases.

Fig. D.1: ROC Curves of Forecasting Civil War

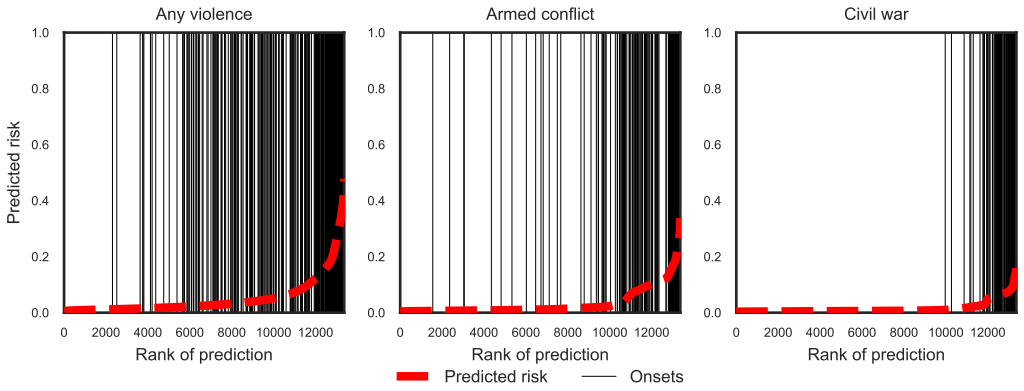


Note: The random forest has a tree depth of 2 and 125 trees. 'Text' contains 15 topics and token counts and 'conflict info' contains 4 dummies capturing time passed since the last conflict and a dummy each for the presence of any violence and armed conflict. Hard cases are defined as not having had civil war in 10 years.

In Figure C.2 we show separation plots for each of the outcomes for predictions using topics and conflict information. The figures order predictions by their rank on the x-axis and plot the predicted level of risk using the red dashed line on the y-axis. The black vertical lines indicate

actual onsets. For all outcomes, onsets tend to be bunched on the right side of the panel where the predicted probabilities are highest. But separation plots have the additional advantage of providing an idea of where the model fails to predict conflict. The 5000 lowest risk observations contain only 19 onsets without clear common features.

Fig. D.2: Separation Plot of Forecasting Violence using Text and Conflict Information

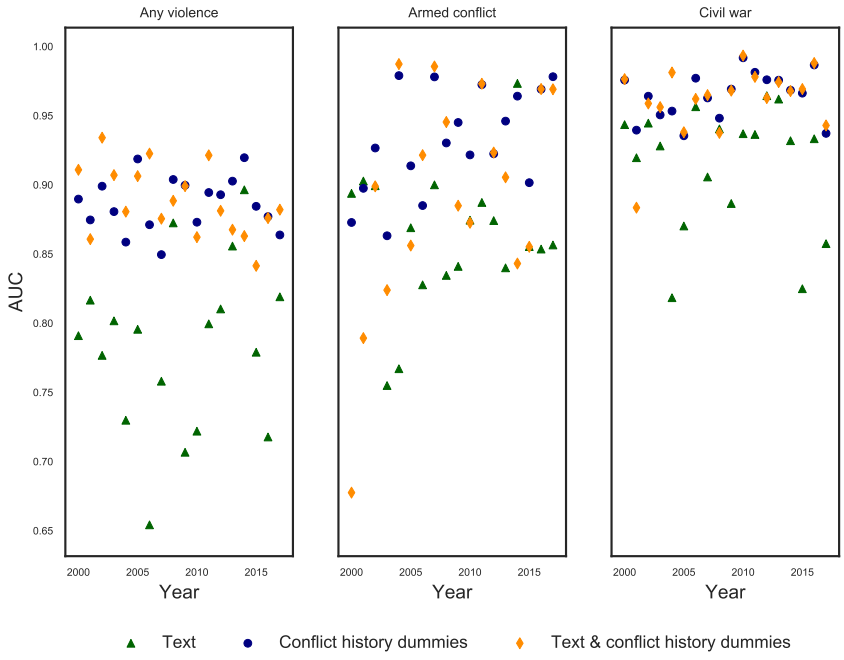


Note: 'Text' contains 15 topics and token counts and 'conflict info' contains 4 dummies capturing time passed since the last conflict and dummies for the presence of lower levels of violence.

In Figure C.3 we show the AUC for ROC curves computed for every single year for each of the three outcomes. We see that the AUCs stay constant or seem to increase slightly over time especially when using text only. We attribute this positive trend to the increase in the training sample over time.

A problem in Figure C.3 is that we have only relatively few onsets, which means that the general trend in model performance is hard to evaluate due to high volatility. In Figure C.4 we therefore show the results of a cross-validation exercise in which we fix the number of trees in the forest but run a gridsearch over the optimal tree depth and record the maximal AUC of this cross validation. The results again suggest a clear upward trend in the cross-validated AUC which is in line with the out-of-sample AUC in Figure C.3. Interestingly, the cross validation AUC is significantly below the true out-of-sample AUC. This is most likely because the folds in the cross validation do not take the panel structure into account and train models on very

Fig. D.3: AUC by Year of Forecasting Violence



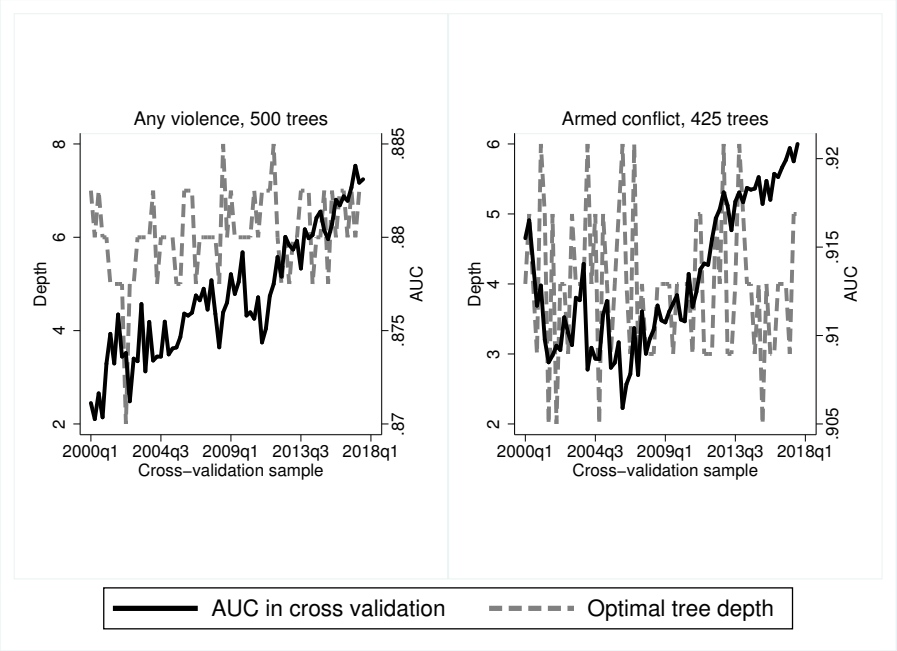
Note: Figures show the overall forecast performance by year.

different parts of the data. Our out-of-sample always uses the most recent past data to predict one quarter or year ahead which is less challenging. The optimal tree depth fluctuates from quarter to quarter but there are no broader trends in the optimal depth.

Overall, Figures C.3 and C.4 provide some evidence against the idea that no generalized forecasting model can be developed as changes in the international context prevent generalization. In both figures forecast performance with text tends to improve with increasing sample size despite a dramatically changing international context and a completely new set of violence onsets.

In Figure C.5 we compare the performance of the topics to events from the Integrated Conflict Early Warning System (ICEWS) database. The ICEWS model we build relies on over 40 event counts. We use 20 counts of all CAMEO event categories that have their target on the territory of the country. We also take the 20 counts involving the government. In addition, we add the average CAMEO scale number of all events. Here, again, we find that topics combined with

Fig. D.4: Cross Validated AUC Over Time



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

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

We show the performance of each of these individual prediction models using text only (Figure C.6) and both text and conflict information (Figure C.7). Across all outcomes it seems that the random forest is the algorithm performing best. But it stands out particularly when predicting any violence with conflict history and text. Here the random forest reaches an AUC of 0.83 whereas the logit lasso only reaches and AUC of 0.68. This is consistent with the idea that the random forest receives an advantage because the model is able to use the information contained in the text conditional on conflict history. This is less important when predicting

Table D.1: Sets of Predictors

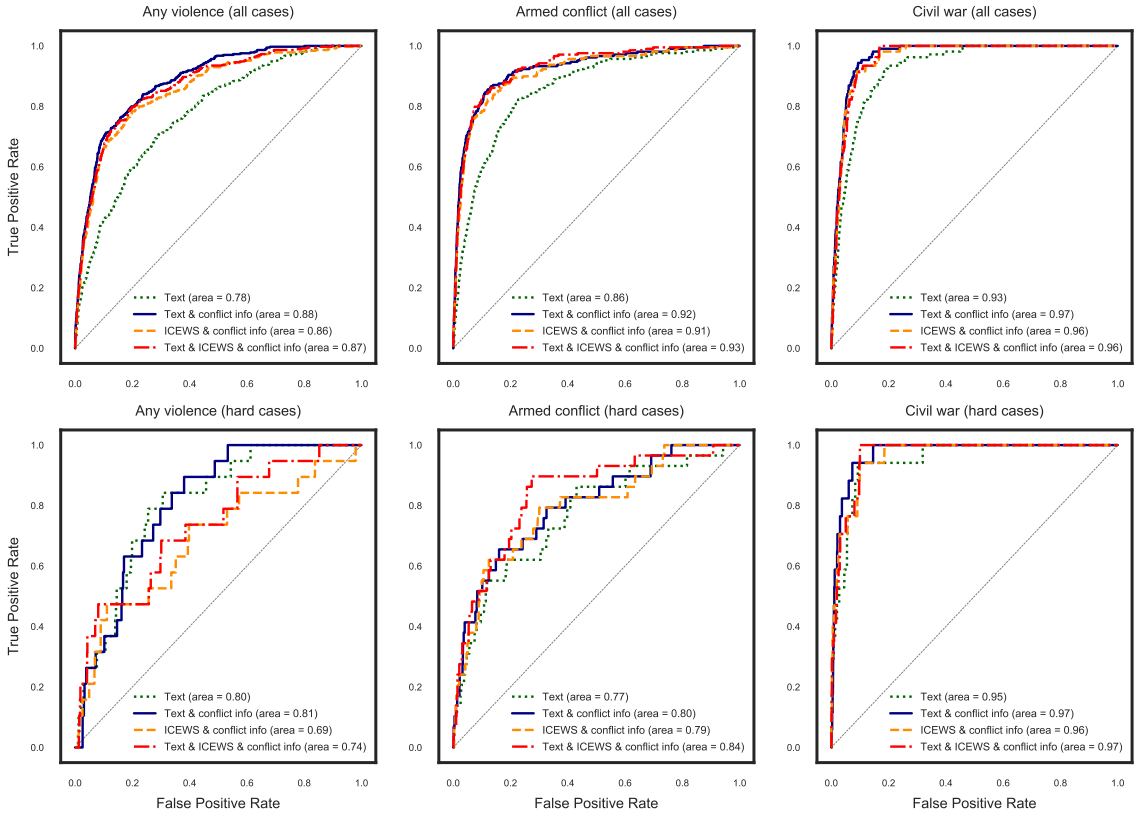
Name	Variables
Topics	Estimated topics using dynamic topic model and total number of tokens
Conflict info	Conflict history and low-level current violence indicators
Standard	Infant mortality, political institutions, share of discriminated population, and neighboring conflicts
ICEWS	Count of all event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

armed conflict as violence escalation is much more important there.

In Figure C.8 we show that the models performance is not specific to 15 topics. The model performs similarly for 5, 10 and 30 topics. For 50 topics, however, the performance starts to become worse, in particular concerning the hard problem, due to overfitting.

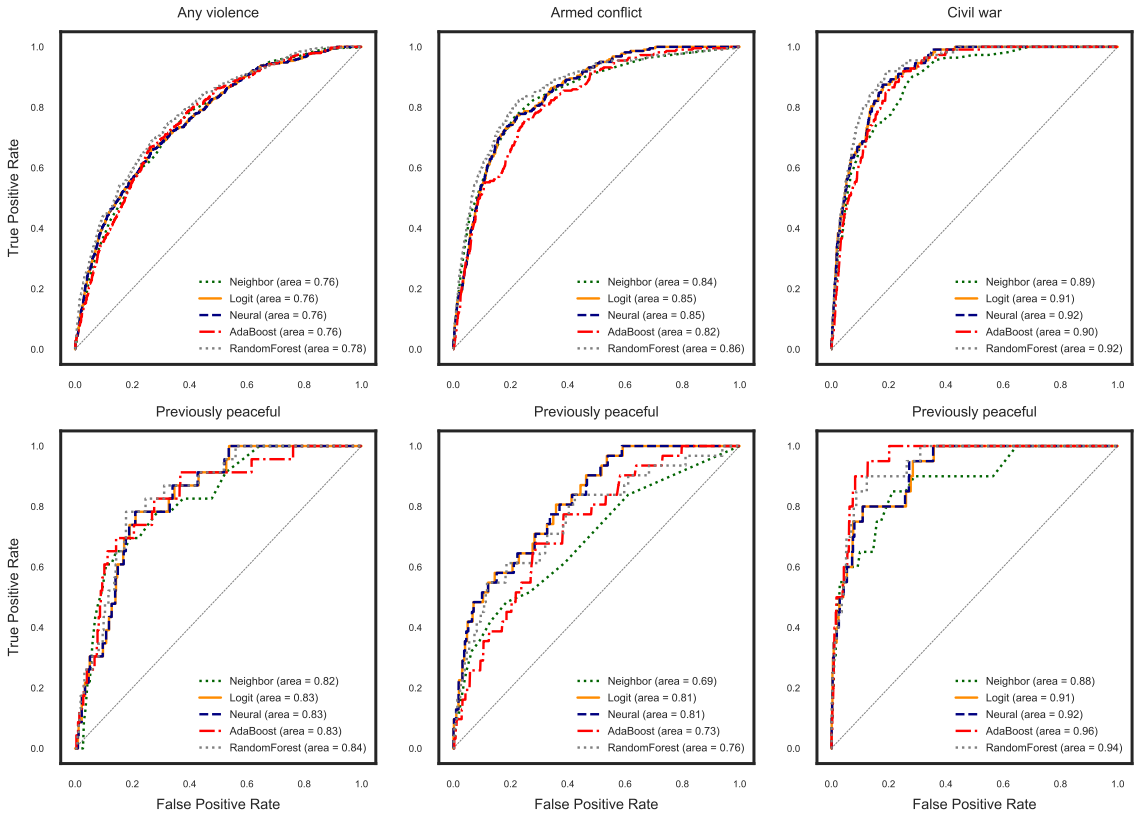
All in all, the Figures paint a consistent picture: Conflict history and present violence are very good predictors of the outbreak of violence. Nonetheless, text summarized by topics adds useful information to predict topics, in particular in countries without current violence or a conflict history.

Fig. D.5: AUC Curves of Forecasting Violence Using Text Compared to ICEWS



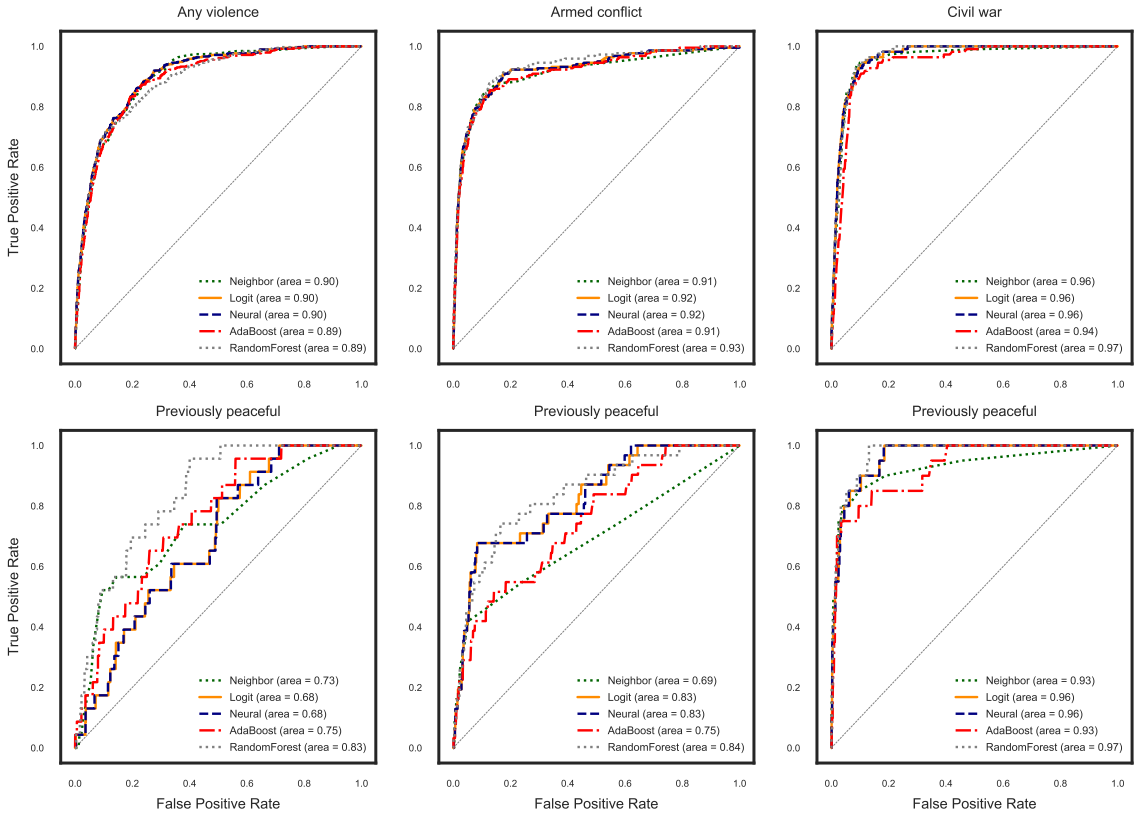
Note: 'Text' contains 15 topics and token counts, 'conflict info' contains 4 dummies capturing time passed since the last conflict and dummies for the presence of lower levels of violence, and 'ICEWS' contains a count of all event types, events involving government, all protests, overall average CAMEO code, and average CAMEO code of events involving the government. Hard cases are defined as not having had conflict in 10 years.

Fig. D.6: ROC Curves of Forecasting Violence with Individual Prediction Models Using Only Text



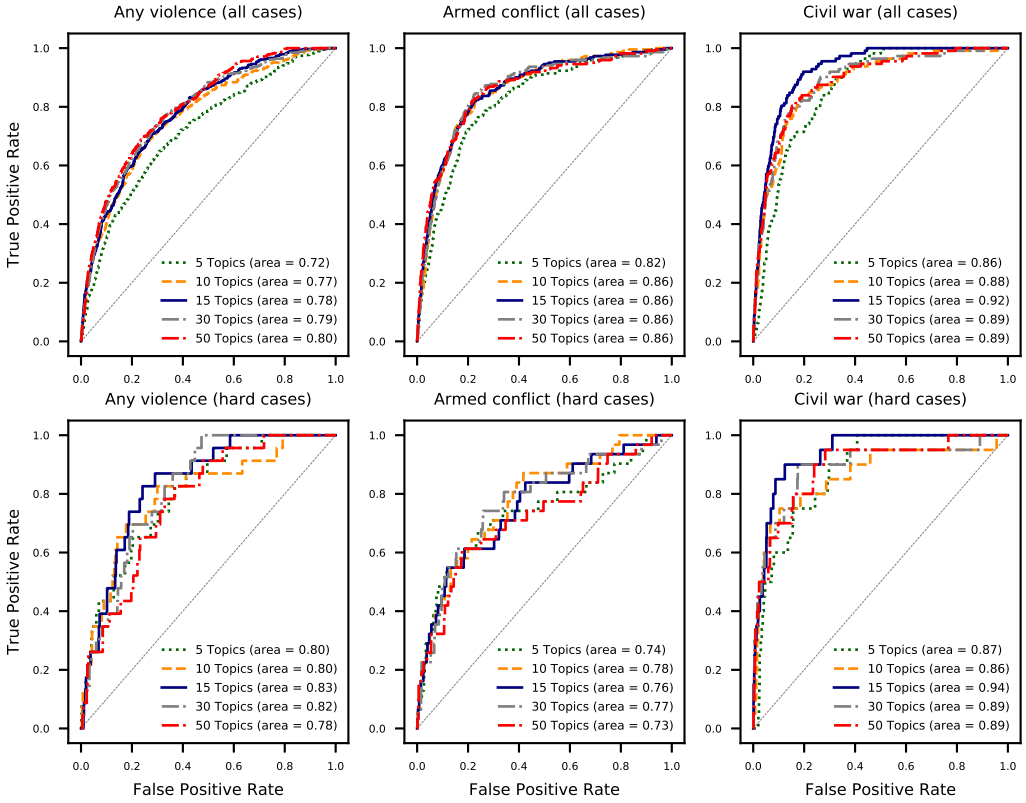
Note: 'Text' contains 15 topics and token counts. Hard cases are defined as not having had conflict in 10 years.

Fig. D.7: ROC Curves of Forecasting Violence with Individual Prediction Models Using Text and Conflict Information



Note: 'Text' contains 15 topics and token counts and 'conflict info' contains 4 dummies capturing time passed since the last conflict and dummies for the presence of lower levels of violence. Hard cases are defined as not having had conflict in 10 years.

Fig. D.8: AUC Curves of Forecasting Violence with Random Forest Using Text while Varying Number of Topics



Note: Predictors include specified number of topics and token counts as well as dummies for time passed since the last conflict and dummies for the presence of lower levels of violence. Hard cases are defined as not having had conflict in 10 years.

References

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

Baker, Scott R, Nicholas Bloom, and Steven J Davis. 2016. "Measuring economic policy uncertainty." *Quarterly Journal of Economics*, forthcoming.

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

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

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

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

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

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

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

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

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

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

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

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 T Kelly, and Matt Taddy. 2017. "Text as data." National Bureau of Economic Research.

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

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

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

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

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

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

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

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

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

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

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

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

OECD. 2018. *States of Fragility 2018.*

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.

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

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

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

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

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.