Predicting fatalities using newspaper text

Team ConflictForecast
Alexandra Málaga, Hannes Mueller, Christopher Rauh,
and Benjamin Seimon

June, 2024

Abstract

We submit two entries to the challenge, both country and grid-cell level predictions, using our proven ingredients and revised methodologies. Our approach relies on topics derived from summarizing a corpus of over 6 million newspaper articles, alongside historical conflict data in a Random Forest framework. For grid-cell level predictions, we use the locations detected in our corpus. Due to data availability and structure, we employ distinct strategies for the country level and grid-cell level sample forecasts. At the country level, we sample predicted errors based on the predicted probability of conflict and the predicted number of fatalities using a Tweedie distribution. In contrast, for grid-cell level forecasts, we draw samples from percentiles obtained through a Quantile Forest Regression.[1]

1 Country level forecast

The goal is to forecast a sample of conflict-related fatalities for each country and month of the upcoming year based on the information available up to October of the current year. For this submission, we provide predictions for seven years, from January to December of 2018 to 2023, and from July 2024 to June 2025. To achieve this, we undertook the following steps:

1. **Building the corpus.**

[1]This paper documents a contribution to the VIEWS Prediction Challenge 2023/2024. Financial support for the Prediction Challenge was provided by the German Ministry for Foreign Affairs. For more information on the Prediction Challenge please see Hegre et al. (Forthcoming)[1] and https://viewsforecasting.org/research/prediction-challenge-2023.

The text data consists of over 6 million documents from 1989 to the present day. These documents are obtained from Factiva and originate from the New York Times, Washington Post, the Economist, BBC Monitor, and LatinNews. The text is retrieved based on specific criteria defined in an extensive query. In general, a document is retrieved if it contains a country or capital name in the title or lead paragraph. This process results in a corpus where documents are associated with a specific country and month. Therefore, we have a collection of documents that represent the news coverage for each country for every month from January 1989 to the present.

2. **Summarizing corpus into topics.**

   We employ standard natural language preprocessing (NLP) techniques, which include removing punctuation, stop words, and lemmatization. In addition to single words (unigrams), we also take into account common combinations of two or three words (bigrams and trigrams). Any token (whether unigram, bigram, or trigram) that appears in over half of the documents (too frequent) or in fewer than 200 documents (too infrequent) is also removed.

   To condense this text data into a set of features, we rely on a dynamic Latent Dirichlet Allocation (LDA) topic model with 15 topics. This approach allows us to reduce the dimensionality of the text data without imposing assumptions about which elements of the text are most relevant for predicting conflicts. This model returns 15 values for each document, which sum to 1. These values represent the relative proportion of each topic within that document. Consequently, for each country-month, we can calculate the proportion of the news content that is attributed to a specific topic by averaging across the document-topic distribution assigned to that country-month. This process is dynamic because we reinterpret the topic distribution of previous months as new documents become available each month.

3. **Rolling forecast of conflict occurrence and fatality count.**

   In addition to the textual features discussed before, we incorporate features that characterize recent conflict history to predict conflict occurrences for each country-month using a Random Forest Classifier. These historical features encompass the number of fatalities in the past 6 months, 1 year, 5 years, and 10 years, along with the duration in months since the last conflict for each conflict level. We define conflict levels as "any violence" (involving at least one fatality), "armed conflict" (involving more than 0.0005 fatalities per 1000 inhabitants), and "civil war" (involving more than 0.03 fatalities per 1000 inhabitants). Additionally, we include the number of tokens (another textual feature) and population as part of our predictive features.

   A strength of having text data is that there is some hope to be able to predict what Mueller and Rauh (2022, JEEA) call *hard onsets*. These onsets occur outside the conflict trap window and are therefore of particular interest to policy-makers. We can show that in a 12-month-ahead forecast at the country level, we can predict these onsets with a precision that makes it economically viable to use them for preventative

purposes. We will return to these hard cases as they also pose a significant challenge when submitting error distributions to the competition.

Similarly, we predict the number of fatalities for each country-month using Random Forest regressions. In addition to the features mentioned for the classification problem, we incorporate the following variables: the number of fatalities from the current month, the average number of fatalities over the last three months, the number of fatalities per capita from the current month, and the conflict occurrence probability predicted by the Random Forest Classifier. For instance, to predict the number of fatalities for January (s=3) in October of the previous year, we include as a feature the conflict occurrence probability for January (s=3) predicted using information up to October of the previous year. This same approach is extended to 12 models that forecast the fatalities for each month throughout the year (s=3,..., 14) based on data available up to October of the preceding year.

4. **Sampling prediction errors.** We implement a sampling approach based on our forecasts for n-periods ahead, using our predictions on the probability of any violence ($\hat{p}$), the number of fatalities ($\hat{y}$), and the number of fatalities given that there was at least one fatality in the current month ($\hat{y}_1$). This method involves sampling from a Tweedie distribution, with parameters specifically tuned for each period ahead (ranging from 3 months to 14 months ahead), based on training data from 2010 to 2017.

   In our approach, we use a sample size (n) of 1000, although a larger sample size could better capture the likelihood of extreme events. We stress that using a larger sample is a risk-averse method which aims to provide the best performance of the system in the long-run. However, this method might not be the best strategy for winning the competition, as there may not be an outlier occurrence in the true out-of-sample period. The final sample comes from three groups. First, we generate a sample of zeros, corresponding to the forecasted probability of no violence, calculated as $(1 - \hat{p}) \times n$. Next, we draw $(1 - \hat{p}) \times (1 - \hat{p}) \times n$ samples from a Tweedie distribution using the unconditional forecasted fatalities ($\hat{y}$). Finally, we draw the remaining samples, $(1 - \hat{p}) \times \hat{p} \times n$, from a Tweedie distribution based on the forecasted fatalities conditional on observing violence in the current month ($\hat{y}_1$).

2 Grid-cell level forecast

In the first step, we adapt the corpus. We use prepositions to detect locations in the news corpus. We then compute average reporting in a given grid-cell by computing the coordinates of detected locations.

For the grid-cell level predictions, we follow a different strategy due to the availability of data. For each month we have to predict, we train seven separate random forest regression models, one for each of the following samples. Grid cells that have:

1. had no battle death in the last five years and have no battle deaths in their immediate neighboring cells

2. had at least one battle death in the last five years but no ongoing violence and have no battle deaths in their immediate neighboring cells
3. had no battle death in the last five years and have battle deaths in their immediate neighboring cells
4. had at least one battle death in the last five years but no ongoing violence and have no battle deaths in their immediate neighboring cells
5. ongoing violence and no battle deaths in their immediate neighboring cells
6. ongoing violence and battle deaths in their immediate neighboring cells

The seventh model is trained on the full sample. We use grid cells across the entire world in order to train the model to maximize training samples. This approach gives us two predictions for each grid cell: i) one based on the conditional distribution of fatalities in a similar situation and ii) one based on the full sample. We use 200 trees with a depth between 3-5 while minimizing the mean-squared error.

We obtain the distribution of uncertainty of predictions using a quantile forest regressor. Here we draw 99 draws from the 1st to the 99th percentile in steps of one percentiles. The specific models 1-6 together with the predictions from the general model 7 give us a total of 198 draws for each grid cell month.

In terms of grid-cell level predictors we include:

- The discounted past deaths
- Time since last battle death
- Current battle deaths
- Discounted and current neighboring battle deaths
- Local news topics
- Neighboring news topics
- Distance to capital
- Population
- Discounted and current number of riots
- Time since the last riot
- Consecutive months with battle deaths

And from the country level we include:

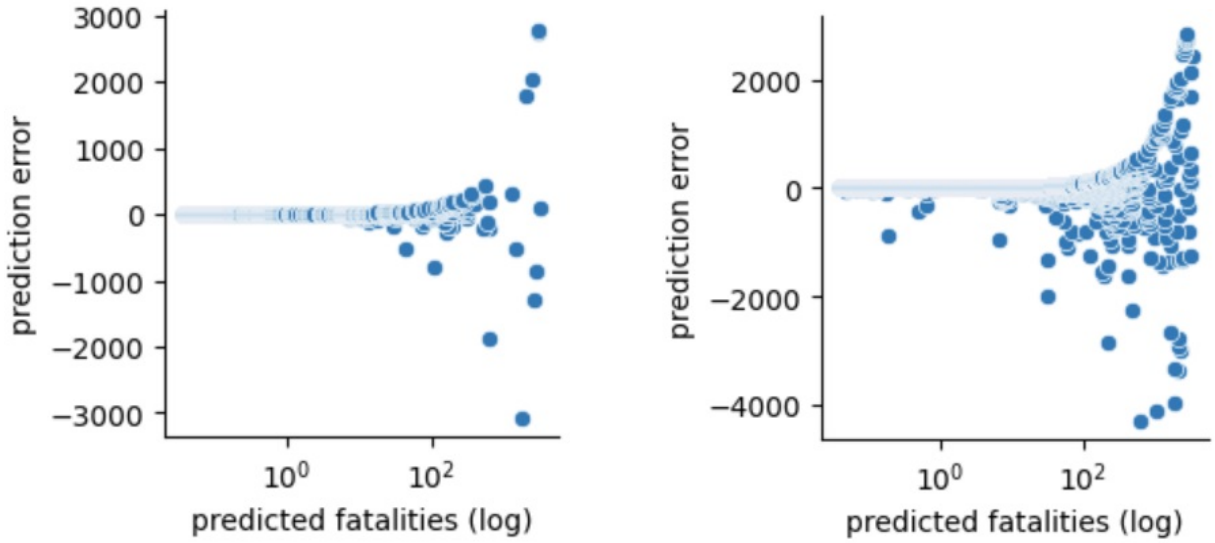
- Discounted and current battle deaths

The predictors included in each model are tailored to its situation in order to speed up computation.

3 Methodological Considerations

An important aspect of armed conflict is the so-called conflict trap. Countries are caught in repeated cycles of violence. When forecasting point estimates the trap shows up as a period of extremely high risk after a recent conflict and very low baseline risk for countries that did not suffer from armed conflict in their recent past. Forecasting conflict without a recent conflict is extremely hard as the baseline risk is so low.

In Figure 1 we show that this also leads to problems when forecasting distributions. In panel A) we show 1,000 draws from our predictions in the past (on the x-axis) and the corresponding prediction error (on the y-axis). Prediction errors here are defined as the predicted value minus the true realization. One of the main features of this figure is very low error values for low prediction values. All errors are positive which means that the prediction seems to be consistently too pessimistic. The reason for this is only revealed once we sample 10,000 times in Figure 1, Panel B. Here the hard onsets in the sample region appear as, relatively large, negative prediction errors for low prediction values. These explain why the model puts a positive prediction on other stable situations as well. We can also see that, on average, the model is good at distinguishing the hard onsets as relatively risky.



(a) 1,000 draws

(b) 10,000 draws

Figure 1: Sampling prediction errors

In conclusion, we consider that a sample size of 1,000 observations may be insufficient to capture the occurrence of hard onsets, which are instances where a country experiences the onset of conflict after an extended period of peace. It is, however, unclear how to react to this when submitting distributions. Should we oversample the hard onset cases leading to the "wrong" distribution on average or do we want to be prudent and pay special attention to the hard onsets?

References

[1] H. Hegre et al. The 2023/24 views prediction competition. *Journal of Peace Research*, XXX, Forthcoming.