Barcelona School of Economics

Master's Degree in Data Science
Specialization in Data Science for Decision Making

"Harnessing Big Data News Media for Conflict Prediction and Anticipatory Decision-Making"

Authors: Giovanna Chaves, Margherita Philipp, Luis Quiñones

Supervisors: Jesús Cerquides and Hannes Mueller

July 2023

ABSTRACT IN ENGLISH (100 words)**:**

Advances in data and computing techniques have opened possibilities for real-time and cost-efficient conflict prediction and early warning capabilities, with news-based data being utilized to generate relevant forecasts. This Master's thesis explores the use of big data news media for conflict prediction and anticipatory decision-making, with a focus on harnessing the Global Database of Events, Language and Tone (GDELT). We investigate the effectiveness of using GDELT events to predict conflict at the country-level by extracting relevant features and comparing the performance of text-based models with different target definitions and time horizons. The results show that GDELT-based features perform well in conflict prediction, particularly in tree-based and LSTM models, indicating the value of using text data for capturing patterns and providing insights into potential conflict events.

ABSTRACT IN CATALAN/ SPANISH (100 words)

Avances en data y computación han abierto la posibilidad de la predicción costo eficiente y en tiempo real de conflictos, con capacidad de alerta temprana, utilizando datos basados en noticias para generar pronósticos relevantes. Esta tesis de maestría explora el uso de medios de comunicación de noticias de 'big data' para la predicción de conflictos y la toma de decisiones anticipadas, con un enfoque en aprovechar la Base de Datos Global de Eventos, Lenguaje y Tono (GDELT). Investigamos la efectividad de utilizar los eventos de GDELT para predecir conflictos a nivel de país mediante la extracción de características relevantes y comparando el rendimiento de modelos basados en texto con diferentes definiciones de objetivo y horizontes temporales. Los resultados muestran que las características basadas en GDELT tienen un buen desempeño en la predicción de conflictos, especialmente en modelos basados en algoritmos 'Trees' y LSTM, lo que indica el valor de utilizar datos de texto para capturar patrones y proporcionar información sobre posibles eventos de conflicto.

KEYWORDS IN ENGLISH: prediction, conflicts, GDELT

KEYWORDS IN CATALAN/ SPANISH: predicción, conflictos, GDELT

Barcelona School of Economics
Barcelona, July 2023

Harnessing Big Data News Media for Conflict Prediction and Anticipatory Decision-Making

Giovanna Chaves, Margherita Philipp, Luis Quiñones

Supervisor: Jesus Cerquides, Hannes Mueller

Master thesis, Data Science

Data Science for Decision Making

BARCELONA SCHOOL OF ECONOMICS

This thesis was written as a part of the Master of Science in Data Science at BSE. Please note that neither the institution nor the examiners are responsible for the theories and methods used, or results and conclusions drawn in this work. All code is publicly available and may be accessed through https://github.com/luisquinonesPR/thesis.

Acknowledgements

We would like to express our heartfelt gratitude to all the individuals who have supported and guided us throughout the completion of this Master's thesis. First and foremost, we thank Hannes Mueller and Jesus Cerquides for their supervision and invaluable feedback, which were instrumental in shaping this thesis. We also thank the team at the United Nations Development Programme (UNDP) and United Nations Children's Fund (UNICEF), notably Sun-joo Lee, Kevin Wyjad and Evan Wheeler, for providing us with this thesis challenge and giving us their time and thoughtful recommendations.

Finally, we also extend our gratitude to Elliot Motte for the availability, constructive criticism, and exceptional suggestions that greatly contributed to the improvement of this work.

To everyone mentioned above and those who have played a role, no matter how small, in this thesis, we are forever grateful for your support and encouragement. Thank you for being a part of this significant milestone in our academic journeys.

Barcelona School of Economics

Barcelona, July 2023

Giovanna Chaves

Luis Quiñones

Margherita Philipp

Abstract

Advances in data and computing techniques have opened possibilities for real-time and cost-efficient conflict prediction and early warning capabilities, with news-based data being utilized to generate relevant forecasts. This Master's thesis explores the use of big data news media for conflict prediction and anticipatory decision-making, with a focus on harnessing the Global Database of Events, Language and Tone (GDELT). We investigate the effectiveness of using GDELT events to predict conflict at the country-level by extracting relevant features and comparing the performance of text-based models with different target definitions and time horizons. The results show that GDELT-based features perform well in conflict prediction, particularly in tree-based and LSTM models, indicating the value of using text data for capturing patterns and providing insights into potential conflict events.

Keywords – GDELT, UCDP, conflict prediction, text as data

Contents

List of Figures

List of Tables

1 Introduction

According to World Bank (2011), more than two thirds of children without access to education, infants dying, and mothers dying during childbirth in the developing world reside in countries that are at risk of or affected by violence. Political and criminal violence have profound and long-lasting consequences, as no low-income fragile or conflict-affected state has achieved any of the Millennium Development Goals. Children in these countries are twice as likely to suffer from undernourishment and three times more likely to be out of school; women are at higher risk of experiencing rape, trafficking, and prostitution, while men face increased rates of morbidity and mortality (World Bank, 2011).

Conflict prediction and forecasting play a crucial role in peace research, offering benefits such as theory testing, policy formulation, and early warning capabilities (Hegre et al., 2017a). Recent advancements in data and computing techniques have fueled research in this area, with the hopes of generating early-warning systems that could allow organizations like UNICEF and UNDP to allocate scarce resources and personnel strategically to prevent or mitigate conflicts risks before they escalate (Bazzi et al., 2019; Celiku and Kraay, 2017).

Traditional conflict indexes often suffer from time delays and revisions, making real-time and cost-efficient measurement of well-being crucial. In this context, news-based data can provide global, accurate, relevant and timely updates about fine-grained events, which have recently started to be exploited in order to generate conflict forecasts.

One such database is the Global Database of Events, Language and Tone (GDELT). This thesis aims to examine the effectiveness of using GDELT events to predict conflict at the country-level globally by extracting relevant features from the data and comparing the performance of text-based models with different target definitions and time horizons.

The rest of the paper is organized as follows. Section 2 provides an overview of existing research on predicting conflict, including the identification of useful covariates, data aggregation challenges, sparse data issues, and the use of text-based sources. In Section 3, the focus shifts to discussing the datasets at hand, data preprocessing and feature creation. Section 4 addresses the methodological framing of the prediction problem and the models used. Section 5 presents our main results, while Section 6 discusses the relevance of these findings, and Section 7 provides the concluding remarks.

2 Background

In this section we highlight key research that has been done in the area of predicting the incidence, onset and escalation of conflict. We discuss what covariates have been deemed useful, the extent to which data has been aggregated at different levels in the dimensions of geography and time, the difficulties of sparse data and conflict traps, and to what extent text-based sources have been used to predict conflict.

The evolution of the conflict prediction literature can be broadly categorized into three generations: the first was interested in accumulating scientific knowledge about war, the second generation introduced game-theoretical models and statistical models using artificial intelligence and machine learning, while the third focused on predicting a long range of political instabilities some time in advance (Hegre et al., 2017a). The present study forms part of this third generation of predictions, and as such it is distinct from the literature seeking causal explanations for past conflicts (Cederman and Weidmann, 2017). While causal relationships are of interest and should inform the inclusion of covariates in models, the forecaster is first and foremost concerned with getting the predictions right rather than knowing exactly why. Sometimes the two strands of research are also at direct odds with each other: causal literature has linked economic and political shocks to intensified violence, but some predictive models barely improve when adding time-varying 'shock' factors such as natural disasters, elections or fluctuations in rainfall (Bazzi et al., 2019).

Factors that have been found to be significant predictors of conflict include ethnic polarization, economic greed or grievances, geographical factors (e.g. terrain), natural resource endowments, as well as the effects of climate change (Celiku and Kraay, 2017). However, many of these factors, alongside GDP, more general levels of economic development, political institutions, levels of democracy or infant mortality are time invariant or very slow moving (Hegre et al., 2013). Thus they are effectively country fixed effects and capture a general level of risk rather than a particularly heightened risk at a given time (Mueller and Rauh, 2018). This is why features that differ not only across different countries but rather vary within a given country are of particular interest in predicting conflict-related targets. Otherwise strong predictive performance of forecasts is

mainly driven by where, but not when, violence is likely to occur (Bazzi et al., 2019).

The most coarse forecasts operate at the state-year level using primarily the aforementioned structural variables (Yonamine, 2013), but any indicators that are typically yearly are unable to detect the timings of tensions escalating and conflict erupting (Chadefaux, 2014). This is why many studies now seek to work at finer grained level of data with daily rather than annual and local rather that country-level data (Yonamine, 2013). In particular, events-data that is frequently compiled from news reports allow for temporarily closer early warnings (Rost et al., 2009). Blattman and Miguel (2010) suggest that a particularly promising avenue for new empirical research is on the subnational scale. While they primarily refer to causal analysis of conflict at the local level, this may also be a relevant direction for prediction tasks, especially as more geographically disaggregated data becomes available. Nuanced histories of violence (e.g. capturing severity and which actors are involved) tend to be the best predictors of hot spots for conflict (Bazzi et al., 2019).

While disaggregation allows for more nuance and thus potentially better chances of picking up on predictive variations in the data, there are also downsides. In particular, we expect greater sparsity of conflict events, i.e. an even more imbalanced dataset. There has been much work on country or region-specific models to predict conflict (Kolusheva et al., 2023), but there may still be gains from allowing models to learn across regions, even when working at reduced level of aggregation. Indeed, some efforts have been made to build more generalizable models that can theoretically be used to make predictions on any country (Wen et al., 2023). Tree-models are particularly conducive to this as they explicitly operate on the basis of finding covariates that helpfully divide up observations independently from country or region fixed effects (Voukelatou et al., 2020). Bayesian multivariate time series have also been used, but for a regional model of the Levant rather than a global model (Brandt et al., 2011), and similarly a Hidden Markov Model has been applied to GDELT events to predict social unrest in a single country (Qiao et al., 2017). Some authors have also employed neural networks, in particular Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM) models (Smith et al., 2018), a Multi-Input LSTM (MILSTM) (Chen et al., 2020), or a Context-Aware Attention LSTM (CALSTM) (Wang et al., 2018).

Alongside generally sparse conflict events, another peculiarity of this forecasting task is that countries get stuck in conflict traps: onset of a new armed conflict in a country substantively increases the long-term expected incidence of conflict (Hegre et al., 2017b), and almost 90% of all conflict onsets happen within two years of a previous conflict in the same place (Mueller and Rauh, 2022). This means that incidence may be a lot easier to predict (Ward et al., 2010) compared to the more challenging problem of detecting early signs of conflict onset in a location that has seen a long period of peace, i.e. that is not currently in the conflict trap.

Using text data as a feature for conflict prediction has been a relatively recent addition to the research agenda. So far, newspaper and media content has mainly been included to supplement models that primarily rely on other features (Kolusheva et al., 2023). An important advantage of newspaper-based data over other event-based data is that they can capture tensions even when a given threshold for conflict was not reached (Chadefaux, 2014). Indeed, while historical variables are powerful for predicting conflict once a country has already descended into violence, such a look into the past is not helpful for 'hard problem' cases of conflict onset outside the conflict trap. Meanwhile text-based features are more likely to predict conflict on the basis for these harder cases (Mueller and Rauh, 2022). Text may also be able to capture proxies for data that would otherwise be hard to come by. For example, information about military spending, diplomatic agreements and other international events may be hard to systematically capture and analyze (Chadefaux, 2014).

There are several different ways in which text data can be used. A simple starting point would be to use a list of keywords and count their frequencies in news sources (Chadefaux, 2014), while a more complex approach would be to generate vectors from news or work with Latent Dirichlet Allocation (LDA) approach to determine a set of topics from a corpus of text (Mueller and Rauh, 2022). Some authors also make use of the Integrated Crisis Early Warning System (ICEWS) (Ward et al., 2010; Chiba and Gleditsch, 2017).

GDELT is another source for news-based data that is yet to be fully explored and therefore the focus of this project. It is a substantial database with word embeddings for news articles, but also auto-generated event data sets. It has many desirable properties, such as global coverage, high density of events (which is especially less likely in human-coded

datasets), detailed geo-coding and future data becomes available in real-time (Yonamine, 2013), although we discuss some potential drawbacks in terms of the dataset's accuracy in the next section. GDELT data has already been used to predict protests, social unrests, domestic political crises and Global Peace Index (GPI) for individual countries or regions (Voukelatou et al., 2020), but there have been fewer attempts to use it for conflict prediction or at a global scale.

Therefore this project contributes to better understanding the extent to which GDELT events can be used to predict conflict, working at the country-level across the globe and aggregating events on a monthly basis. In particular, we explore what features can extracted from GDELT data to see how well a pure text-based model performs compared to models with additional parameters. While, due to time constraints, we are not able to directly predict at the subnational level, we make use of the subnational level data available within GDELT to generate features. Most papers we have encountered decide on a single target to predict. We instead explore the predictive power of a text-only model for different target definitions and different time-horizons for the prediction. Thus we are able to contribute a rich bias analysis to offer insight to where the GDELT events can most add value in the context of conflict predictions.

3 Data

In this section we first describe key aspects of the GDELT and UCDP data that may affect our results. Then we outline how we obtained, preprocessed and joined data from different sources, including the assumptions and modeling choices that were made. Finally, we elaborate on the features created to help predict conflict.

3.1 Comments on GDELT

GDELT (Global Database of Events, Language and Tone) is a real-time, global news monitoring project supported by Google that provides a valuable resource for studying global conflicts as it has information from 1979. GDELT's event database is auto-generated on the basis of news articles, and it gathers information from a diverse range of sources, including news websites, online blogs, social media platforms, and even local newspapers. It employs automated web crawlers and APIs to scrape and retrieve data from these sources. The data collection process is designed to be broad and comprehensive, covering various languages and regions. Nevertheless, the majority of news sources are in English and many come from US-based outlets.

Once the data is collected, it undergoes preprocessing to standardize and prepare it for further analysis. This step involves tasks such as language identification, entity recognition, and de-duplication before moving on to event extraction. The aspects of de-duplication and event extraction can be particularly difficult. The database updates every 15 minutes, but the identical event may be picked up by different sources at different times of day, and there is also a risk of retrospective articles leading to new events being generated although the media are covering the anniversary of a past event. While this may occur, we believe that the effects will not be systematically different across countries and we also believe that the impact is less severe as we ultimately aggregate at the monthly level in the time dimension and at the country level in the geographical dimension.

For a detected event, GDELT ascribes an Actor1 and Actor2 as well as various attributes of the event 'action', i.e. what Actor1 did to Actor2. These are codified on the basis of Conflict and Mediation Event Observations (CAMEO) Event and Actor Codebook. There are 310 different event codes, grouped into 20 root event codes. In order to work with a

manageable number of columns and avoid sparsity, we aggregate at the level of root event codes. Figure 3.1 shows the relative distribution of event codes and also highlights the disparity in data volumes over the years. This is a further caveat for working with this data.

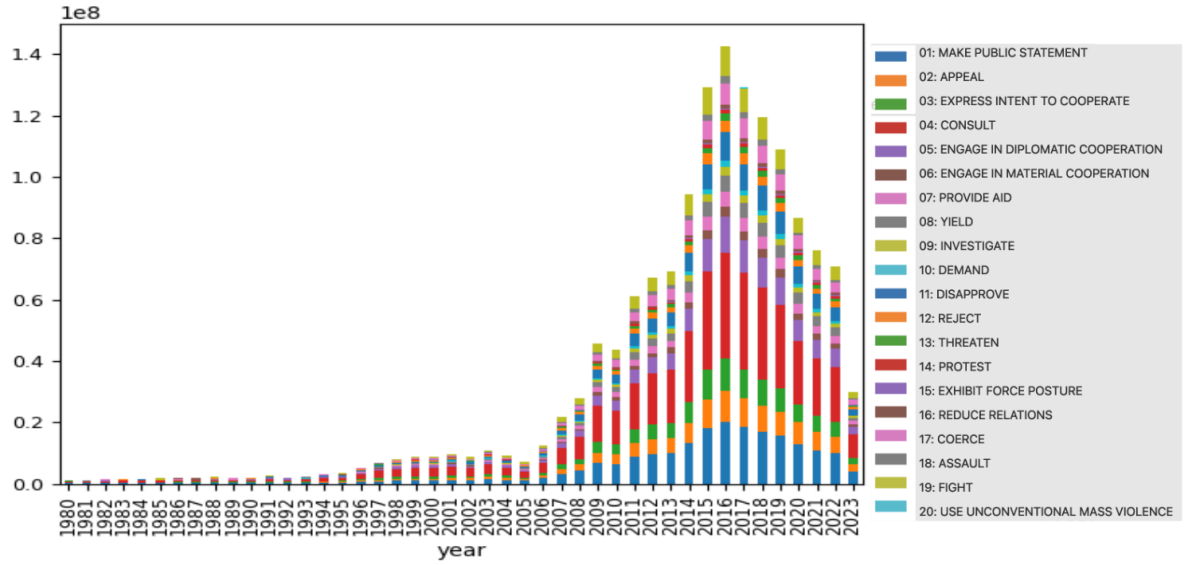

Figure 3.1: Event counts in GDELT for each CAMEO root event code.

We ultimately choose not to train on data before the year 2000. Starting from 2000 gives us certain advantages, including greater data availability and reduced complications related to the aftermath of the USSR and differences in country encoding. We also carefully preprocess event counts to contextualise them within countries and also globally across time.

3.2 Comments on UCDP

What is defined as a conflict is far from being standard in the literature. Various data sources that track figures related to conflict episodes use a particular definition. The Correlates of War (COW) project, for instance, uses the concept based on "sustained combat, involving organized armed forces, resulting in a minimum of 1,000 battle related fatalities," where the latter includes "not only those armed personnel killed in combat but also those who subsequently died from combat wounds or from diseases contracted in the war theatre" (Sarkees, 2010). On the other hand, the Armed Conflict Location & Event Data Project (ACLED) does not predefine what constitutes a conflict, instead

covering six types of events such as “battles”, “violence against civilians” and “protests” (ACLED, 2019). Thus, as highlighted in Hasell (2022), figures for the number of deaths from conflict will differ between sources.

The Uppsala Conflict Data Program (UCDP) counts battle-related deaths, which they define as deaths resulting from “the use of armed force between warring parties in a conflict dyad, be it state-based or non-state.” UCDP records deaths on the basis of dyads, i.e. two actors who are in conflict with each other. Only once a dyad crosses the threshold of 25 deaths in one calendar year does the dyad start to get tracked. Thus a conflict within a so far untracked dyad that leads to 24 deaths in December and no more than 24 deaths in the following year does not enter the UCDP records.

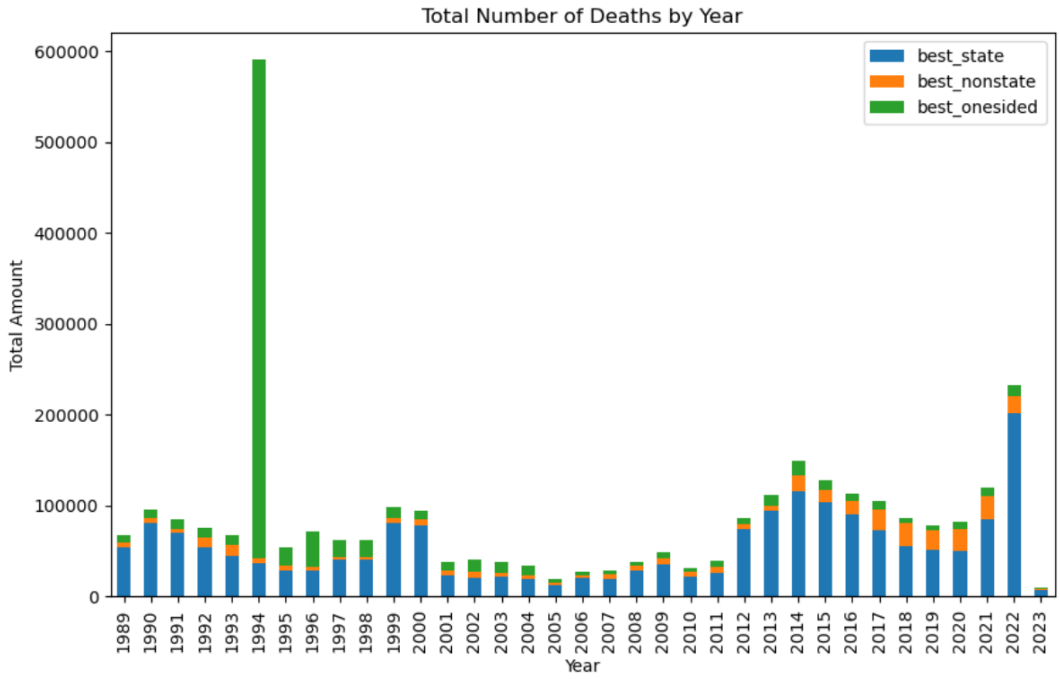

Figure 3.2: Annual deaths in the UCDP database for each category of death.

In the UCDP data, ‘best’ refers to the best estimate of deaths. In Figure 3.2 we show the deaths associated with the three types of violence classifier by UCDP. State violence refers to a ‘contested incompatibility’ that concerns government and/or territory where there is use of armed force between two parties, of which at least one is the government of a state. Meanwhile non-state violence refers to the use of armed force between two organised armed groups of which neither is state, and one-sided violence refers to the use of armed force by the government of a state or by a formally organised group against

civilians. From the figure below we can see that usually at least one state's government is involved in the conflict dyads.

In our analysis we base our definition of conflict on all deaths, thus not discriminating between the perpetrators of violence. We believe this is in line with the needs of organisations such as the UNDP and UNICEF. While a simple within-country correlation analysis showed that state violence was most closely correlated with refugee flows, all conflict comes with a potential need for humanitarian intervention and is thus helpful to predict.

3.3 Obtaining and Merging Data

We work with data from three main sources: the GDELT Events database (GDELT Project, 2023), UCDP's conflict-related death counts (Uppsala University, 2023) and population data from the World Bank (World Bank, 2023). We later added population data for Taiwan and Western Sahara from the World Population Review in order not to drop those countries from the analysis (World Population Review, 2023).

We accessed the UCDP's Georeferenced Event Dataset (GED) both through direct csv downloads and via the API. We work with version 23.1. Up until December 2022, death counts have been finalised, while events for 2023 are so-called 'candidate' events that may still be subject to change.

The GDELT data was the most complex to obtain. The events are stored in compressed csv files under http://data.gdeltproject.org/events/ (1997 until 2005 in annual files, 2005 until March 2013 in monthly files and thereafter in one file per day). The files are too large to all be downloaded and stored at once. Therefore we worked with a script that opened one file at a time, aggregated the data and then appended the extracted data to a dataframe. We aggregated at the Admin1 level, i.e. one level below the national level, as we had hoped that we could match this with the Admin1 information provided in UCDP. However, similar to Yonamine (2013), matching these sub-national units proved to be time and computationally intensive, as well as require a non-negligible number of manual matching, and doing so on a global scale lay beyond the scope of this thesis project. Nevertheless, we were able to make use of the Admin1 level information in GDELT to generate a useful set of features.

Historical global event data research has shown the monthly level to provide better results (Smith et al., 2018), thus both the GDELT and the UCDP data are aggregated at the monthly level by country, and subsequently merged by keeping all GDELT rows and inserting 0s for the UCDP rows when there are no matches. This choice was deliberate and relied on the assumption that the number of relevant deaths tracked in UCDP is our "ground truth". Thus, if no data is available for a country in a particular period, it means there are no conflicts.

3.4 Pre-Processing and Feature Creation

With the exception of Taiwan and Western Sahara, we eliminate 25 countries for which we do not have population data. These are countries that rarely appear in GDELT (e.g. Cook Islands and Pitcairn), as news information for them is spotty. Additionally, there are no battle-related deaths reported in UCDP and a low likelihood of conflict erupting there (e.g. Vatican City and Antarctica). The full list of these countries can be found in Appendix A1. In our analysis we ultimately work with data from January 2000 until March 2023 and we are able to keep 197 countries.

To appropriately forecast, it is important to obtain a sequential set of months for each country within our dataset, starting from its first appearance until March 2023. To achieve this, after the initial mention of a country in GDELT, we inject 0 values for both conflict deaths and GDELT information during any periods where there are gaps between appearances. The underlying assumption is that if there is no available GDELT data, it indicates that nothing significant occurred during that period. It is important to acknowledge the biases of GDELT data, as it may miss smaller-scale events, particularly for periphery countries. Despite this data constraint, we are comfortable proceeding with the analysis because significant events are more likely to be captured internationally.

While we have changing population data between years, two options are considered. The first involves calculating the average population of each country over the training period. However, this approach may lead to underestimating population and overestimating deaths per capita, especially for countries with significant population growth. Since one of our target variables is defined by deaths per 100,000 people, even small fluctuations in population can result in different conflict outcomes (see Appendix A2). The second option

is to work with yearly changes in population. One limitation of this approach is that year-on-year changes can be artificial due to population jumps between December and January. We choose the second approach to ensure our model more faithfully represents reality, but set the population of the entire year to January and linearly interpolate the population in the following eleven months. This allows for a more accurate representation of population changes over time.

Finally, we add our country and month labels as features. The binary encoder allows us to create country dummies without one-hot encoding, which would require creating an extensive number of columns. This captures country fixed-effects which are typically absent in the forecasting literature, usually replaced by time-invariant or slow-moving structural variables (Mueller and Rauh, 2018). Similarly, we construct a *neighbor in conflict* feature that takes value of 1 if at least one of the country's neighbors is in conflict at t_0, and 0 otherwise. This captures potential spatial spillovers; for instance, Carmignani and Kler (2016) find that higher conflict incidence in the neighbourhood increases domestic country involvement in war. At last, we consider that certain conflicts might have a seasonal component, especially when tied to harvest shocks and climate change (Landis, 2014; Guardado and Pennings, 2020). In order to do that, we convert each month to their corresponding sine and cosine values to uncover cyclical patterns.

3.4.1 GDELT Features

Our research aimed to leverage the GDELT dataset to enhance conflict prediction capabilities. To achieve this, we focused on assessing the applicability of GDELT data across different conflict definitions, prediction problems, and forecasting windows, and maximized the utilization of the rich GDELT Events data by constructing numerous features.

As we performed data aggregation during the download process, we calculated the monthly event counts for each root event code at the sub-national, and then country, level. Instead of relying solely on event counts, which could be skewed due to reporting biases in certain countries, we worked with *event shares*. By calculating the shares of specific root event codes within a country and time period, we were able to identify the predominant types of events reported and make comparisons across countries and time. For example, during

periods of conflict, we would expect a higher event share for CAMEO root event code 19, denoting "fighting", compared to other event shares. Voukelatou et al. (2020) also found that the most important event types varied by country profile; in a war-torn country such as Somalia, events related to military engagement, explosives, international involvement and arrests had higher feature importance when predicting conflict. We also keep track of the share of total events in which either the government (represented by government officials, military and/or police forces) or the opposition (insurgents, opposition forces, rebels, or separatist) participate.

In addition, we accounted for the total event count in each country and time period by normalizing them based on the number of events occurring globally during the same month. This normalization process helps mitigate the influence of time on the count of events, as the data reveals a substantial increase in the number of reported events by the mid-2000s (Figure 3.1).

To maintain a more stable representation of a country's event share values over time, we developed event share stocks. This approach is similar to that of Mueller and Rauh (2022) and involves continuously updating the event share values with the current's period shares, while gradually diminishing the impact of previous shares (decaying factor of 0.8). By doing so, we aim to capture the overall profile and trends of a country's event shares, avoiding significant month-to-month fluctuations.

In order to capture similarities between countries based solely on textual information, we employed clustering techniques using the shares and stocks of events and deaths. By applying the elbow method, we determined that 15 clusters provided the optimal separation within the feature space. Countries within each cluster were aggregated together based on identified similarities in the data. Notably, one cluster exclusively comprised the United Kingdom and the United States, while another cluster included countries such as Afghanistan, Bangladesh, Colombia, Iraq, Israel, Lebanon, Pakistan, and the Philippines (Appendix A3).

Finally, while we faced computational limitations that prevented us from forecasting monthly at the sub-national level, we leveraged valuable information from GDELT reporting on Admin1 events by incorporating them as useful features. First we calculated the number of regions within each country that had reported events in GDELT during

a specific period. By comparing this count of regions to previous periods, we might gain insights into the level of activity within the country at that particular time. Then, although we did not have specific data on deaths at the Admin1 level, we replaced the regions with a measure of their potential impact on deaths, specifically when events with root codes 18 (assault), 19 (fight), or 20 (use unconventional mass violence) occurred in those regions. This target encoding process assigned higher numerical values to regions that were associated with higher levels of country-wide deaths, compared to regions that consistently appeared regardless of the death toll. To our knowledge, our study is the first instance of incorporating this GDELT information in such a manner.

3.4.2 Non-text Features

We minimize the use of external information not present in GDELT or UCDP in order to emphasize GDELT's capabilities. This allows us to rely on the binary country dummies to capture permanent variations between countries.

Utilizing the UCDP data, we created variables related to the history of deaths, including battle-related deaths per 100,000 people in the past 6 months, 1 year, 5 years, and 10 years. Similar to the stock of events, we also constructed a decayed stock of deaths. Additionally, we monitored periods of peace by calculating the number of months since the last occurrence of a conflict in each country.

At last, we incorporate data on annual refugee flows by country from UNHCR (UNHCR, 2023), also linearly interpolating them month-by-month. This allows us to capture refugee outflows as well as inflows, which might indicate more subtle shifts and pressures preceding the escalation of full-blown conflict.

4 Methodology

Our main goal was to predict whether a country will be in conflict in the near future using big data news media. As conflict can be defined in different ways, we aimed to explore and experiment with various definitions using GDELT data as our primary predictive features, leveraging machine learning and deep learning methods to develop accurate and robust conflict forecasting models.

In this section, we first discuss how to frame our prediction problem methodologically and the practical implications associated with each choice. Then we define three conflict targets to compare and contrast model performance, and consider forecasting with different short- and medium-term windows. Finally, we establish the tree-based and neural network models and their training and out-of-sample testing procedures.

4.1 The prediction problem

The prediction problem in conflict literature presents challenges due to the absence of a universally accepted definition of conflict. Existing literature has approached this problem from two main perspectives: classification and regression. Most studies have taken the classification approach, categorizing countries as either in conflict or not based on an arbitrary threshold of deaths or violent events. On the other hand, some researchers have tackled this question through a regression problem: Voukelatou et al. (2020) predicts peace based on GPI scores, while Yonamine (2013), Chen et al. (2020), Wang et al. (2018) and Smith et al. (2018) attempt to forecast the number of conflict events happening in the next period.

Both approaches bring benefits and drawbacks to the conversation around conflict forecasts and early warning models. The regression approach allows policymakers the flexibility in determining the factors they deem more relevant when assessing conflict risk and enables forecasting of intensity. However, interpreting the results of regression models, such as the prediction of a specific number of deaths, can be challenging. It becomes difficult to assess the practical implications of such predictions, particularly if large confidence intervals are involved.

On the other hand, the classification approach provides more tangible outcomes. By presenting results in terms of predicted probabilities, policymakers can assess the risk of conflict and compare it to the actual occurrence or non-occurrence of conflict. This allows for a clearer assessment of the model's performance in terms of predicting the likelihood of conflict. However, changing the threshold at which conflict is determined can lead to substantially different predictions.

Another modeling decision arises when considering the types of conflict we aim to predict. Specifically, we need to decide whether we focus on predicting conflict incidence or conflict onset. The first is relatively easier due to what researchers have defined as the "conflict trap" (Mueller and Rauh, 2022), where countries with a history of conflict are more likely to experience conflict in the future. Therefore, modeling conflict incidence can rely heavily on historical conflict data and patterns. On the other hand, predicting conflict onset presents a different challenge as it entails detecting shifts or deviations from the established patterns of peace and stability. Predicting conflict onset necessitates identifying triggers, early warning indicators, or sudden shifts in the underlying factors that contribute to conflict.

In order to fully harness the predictive potential of GDELT data, we adopt a mixed modeling approach that encompasses both regression and classification techniques. Specifically, we treat deaths per capita as a regression problem, aiming to forecast the intensity or magnitude of conflict, as well as address the presence or escalation of conflict as classification problems. By combining regression and classification models, we can effectively capture the multi-faceted nature of conflict and leverage the comprehensive GDELT data to provide a more nuanced understanding of conflict dynamics.

4.2 Defining conflict

As previously mentioned, a key aspect of the classification approach is the threshold selected to define a conflict episode. We depart from the simplistic notion of considering any fatality as an indication of conflict, adopting instead a more nuanced definition that aligns with the perspectives of organizations such as UNDP and UNICEF. Specifically, we classify a country as being in armed conflict if the number of deaths exceeds 0.05 per 100,000 people. This threshold strikes a balance by acknowledging that policy interventions

are more feasible and effective when addressing lower levels of conflict, such as local unrest (Bazzi et al., 2019).

Moreover, we generate a binary measure of escalation by examining the percentage changes in total deaths between the current period and the preceding period. If the observed change exceeds the 75th percentile of changes in the previous 24 months, and the current period's deaths per capita exceed the conflict threshold, we classify it as an escalation. This is in line with Kolusheva et al. (2023), and takes into account small increases in peaceful countries while filtering out minor fluctuations in long-standing conflicts.

4.3 The forecasting horizon

When deciding on the prediction horizon, it is common for studies to focus on forecasting one period ahead. However, it is important to also consider the availability and time lag of the data used as predictive features. In the case of UCDP Candidate data, which is typically released at the end of each month and used to generate the non-text features, predicting conflict for the next month based on a window of only 10 days may not provide sufficient information for organizations such as UCDP and UNICEF. We expand this analysis and predict outcomes not only for the next period but also within 3 and 6 months, which allows for a more comprehensive assessment of the predictive power of the model and provides policymakers with a longer-term perspective on conflict incidence, escalation, and deaths per capita.

4.4 Tree-based models and LSTM

In evaluating the predictive capabilities of GDELT data, we employ three distinct models for each target variable and forecasting window: one incorporates only features derived from GDELT, focusing solely on textual information; the second utilizes non-textual features, such as conflict history and refugee flows; and the third model combines both GDELT features and non-textual features.

To ensure successful model training, it is crucial that the data is structured in a specific manner, with a particular emphasis on sorting. The sorting of data by time period and country is of utmost importance, and failure to adhere to this approach can lead to

training the model on certain countries, validating it on others, and testing it on a select few that it has never encountered before. By sorting the data by time period and country, we preserve the temporal integrity of our panel data, allowing the model to learn from past events before making predictions for the future.

Tree-based models, such as Random Forest and XGBoost, have gained popularity in the conflict prediction literature as they offer improved performance compared to earlier logistic regression approaches. These models are highly interpretable, accommodating both categorical and numerical data, and can be used for both classification and regression tasks. Random Forest, in particular, has been widely used in recent research and has demonstrated strong results across various studies (Bazzi et al., 2019; Celiku and Kraay, 2017; Voukelatou et al., 2020; Galla and Burke, 2018; Mueller and Rauh, 2022). XGBoost, on the other hand, provides additional advantages such as efficient handling of missing data and large datasets. It offers a range of hyperparameters that can be tuned to optimize model performance, including parameters related to learning rate, tree depth, and regularization.

In our analysis, we also employed a vanilla Long Short-Term Memory (LSTM) model, which is a type of Recurrent Neural Network (RNN). LSTMs are well-suited for processing sequential data and have shown promising performance in capturing long-term dependencies, including in GDELT prediction tasks (Chen et al., 2020). Specifically, our model leverages the temporal memory capabilities of LSTMs to effectively process past information and make predictions for future time steps (Smith et al., 2018).

4.4.1 Training the models

The models were trained using data from January 2000 to January 2018 to predict future events, starting with the straightforward case of predicting what would happen in February 2018. To improve model performance, an expanding window approach was employed for tree-based models, which involved making predictions for each subsequent month using data up to the previous month. This allowed the models to incorporate information from the most recent period, enabling them to capture evolving patterns and improve over time (Voukelatou et al., 2020). For the LSTM model, a lag of 1 was introduced to incorporate information from previous time steps, as these recurrent networks have the advantage of

being able to effectively capture temporal dependencies and learn from past information. Additionally, a label encoder was used for the LSTM, as previous attempts with a target encoder led to overfitting issues for the classifier.

To address the significant class imbalance in the dataset, we tuned hyperparameters with a focus on optimizing for recall, as the goal was to capture potential conflict events even if it meant accepting a higher rate of false positives. This approach was motivated by the understanding that missing an impending conflict could have severe consequences, and the organizations utilizing the tool would prefer to receive alerts for potential conflicts, even if some turn out to be false alarms. In order to mitigate the imbalance, the Binary Focal Loss function was used specifically for the LSTM model, which assigns higher weights to the minority class and allows the model to pay more attention to learning patterns and characteristics related to conflict.

For the regression task, the LSTM model had a simpler setup with a ReLU activation function at the output layer instead of the sigmoid function used for classification. Additionally, the country category feature was handled using a target encoder, which helped capture the trends of deaths per capita and enhance the regression predictions.

However, it is important to note a potential limitation in the implementation of the grid search for the tree-based models. Due to time and computational constraints, the optimal hyperparameters were selected based on one round of tuning using only the training period. In future implementations, it would be beneficial to make the code more flexible and grid search in each extended window fit, allowing for a more comprehensive search for optimal hyperparameters. The computational and time costs required for this approach should not be underestimated, though. A grid search in batches could be explored as a compromise between computational efficiency and parameter optimization. It's worth noting that hyperparameters are unlikely to change significantly over time, but periodic tuning can still contribute to improved model performance and adaptability.

5 Analysis

In this section we outline how the three models (text-only, history-based and combined) compare in predicting our three different targets (incidence, escalation and deaths per capita) across the three different time horizons (one month ahead, within the next three months and within the six months) and across the different algorithms employed (Random Forest, XGBoost and LSTM), and how they compare to our naive baseline model. Table 5.1 displays the results for all models described in the previous section. In addition, we provide as baseline a naive "conflict persistance" model that assumes conflict in the future will be the same as it is today (Yonamine, 2013). We can see that the text-only model, which uses only features derived from GDELT, is indeed competitive.

		RF			**XGB**			**LSTM**			**Naive**
		tx	**hs**	**all**	**tx**	**hs**	**all**	**tx**	**hs**	**all**	
Incid.	**f1**	0.96	0.98	0.96	0.97	0.98	0.98	0.91	0.94	0.96	0.88
	f3	0.94	0.98	0.98	0.95	0.98	0.97	0.90	0.94	0.94	0.84
	f6	0.93	0.97	0.97	0.95	0.96	0.96	0.88	0.94	0.93	0.79
Escal.	**f1**	0.90	0.93	0.92	0.89	0.92	0.92	0.81	0.92	0.91	0.60
	f3	0.92	0.95	0.94	0.92	0.95	0.94	0.83	0.94	0.94	0.60
	f6	0.92	0.96	0.95	0.94	0.95	0.95	0.82	0.95	0.95	0.60
Deaths	**f1**	0.958	0.620	0.624	0.861	0.587	0.631	0.809	0.663	0.661	1.169
	f3	0.718	0.491	0.494	0.624	0.405	0.419	0.533	0.486	0.480	0.936
	f6	0.591	0.372	0.375	0.512	0.352	0.340	0.399	0.356	0.352	0.838

Table 5.1: Metrics for all models on the three targets and forecast horizons, as well as a naive model. We report ROC AUC scores for models with *incidence* or *escalation* as target, and RMSE for models forecasting *deaths*. The naive model predicts for $t+1$ the same that happened in period t.

For our first classifier target, conflict **incidence**, the text model performs almost as well as the history-based and combined models in predicting armed conflict at $t+1$, and within 6 periods with the XGBoost classifier. With regards to the second classifier target, conflict **escalation**, our text model demonstrates comparable performance when using the tree-based classifiers (XGBoost and Random Forest), although it does not perform as well with LSTM. It appears that the the history-based and combined models leverage the temporal context more effectively in capturing the sequential escalation patterns LSTM should be able to detect. While text model more clearly under-performs compared to the other two for the **regression** on deaths per capita, it remains substantially better than

the naive model. Using an LSTM to forecast the average deaths within 6 months, the RMSE is 0.399, just slightly above the history-based model at 0.356 and much lower than the naive model at 0.838.

That the naive model is insufficient is unsurprising; although history-based models provide strong predictions, Bazzi et al. (2019) notes that "local violence is not merely autoregressive", and such models perform consistently poorly.

5.1 Predicting conflict incidence

Based on the evaluation results, it appears that the tree-based models perform best for predicting incidence at any of the forecasting horizons. Specifically, even a text-only XGBoost outperforms any LSTM model in terms of ROC AUC score. The performance of all algorithms in predicting incidence declines as the forecasting horizon increases to longer time periods. In this case, history-based models tend to be more robust compared to models that rely solely on GDELT features.

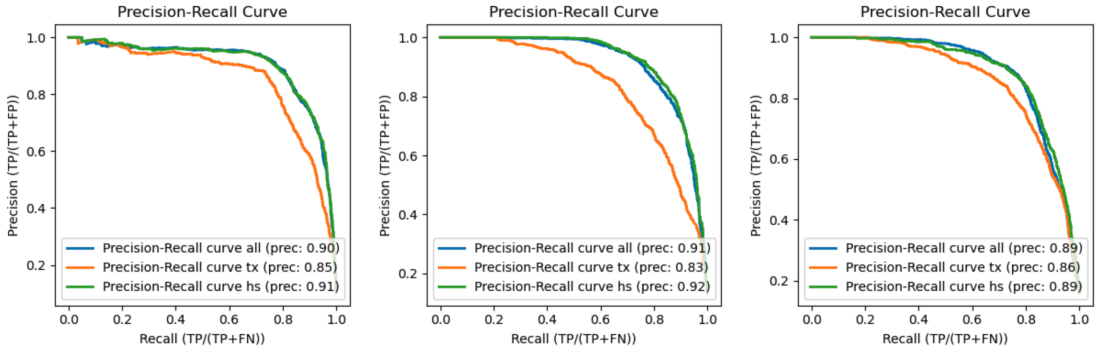

Figure 5.1: Precision-Recall Curves for XGBoost classifier in forecasting conflict incidence one month (left), three months (center), and six months ahead (right).

The recall and precision values for the text model in Figure 5.1 are not as high as those in the combined or history models for XGBoost. However, the fact that the text model can provide substantial precision and recall with text-only variables underscores its potential utility in conflict prediction.

The improvement in the precision-recall curve for the 6-months ahead prediction (t+6) over the 1-month (t+1) and 3-months (t+3) ahead predictions could be attributed to the inherent properties of these measures and the nature of conflict prediction itself. As we extend the prediction window from 1 month to 6 months, we inherently check for

conflict in more time periods, and thus are more likely to observe at least one positive case (conflict) during this extended period. It is equivalent to casting a wider net: the likelihood of catching a 'fish' increases, thereby potentially improving recall.

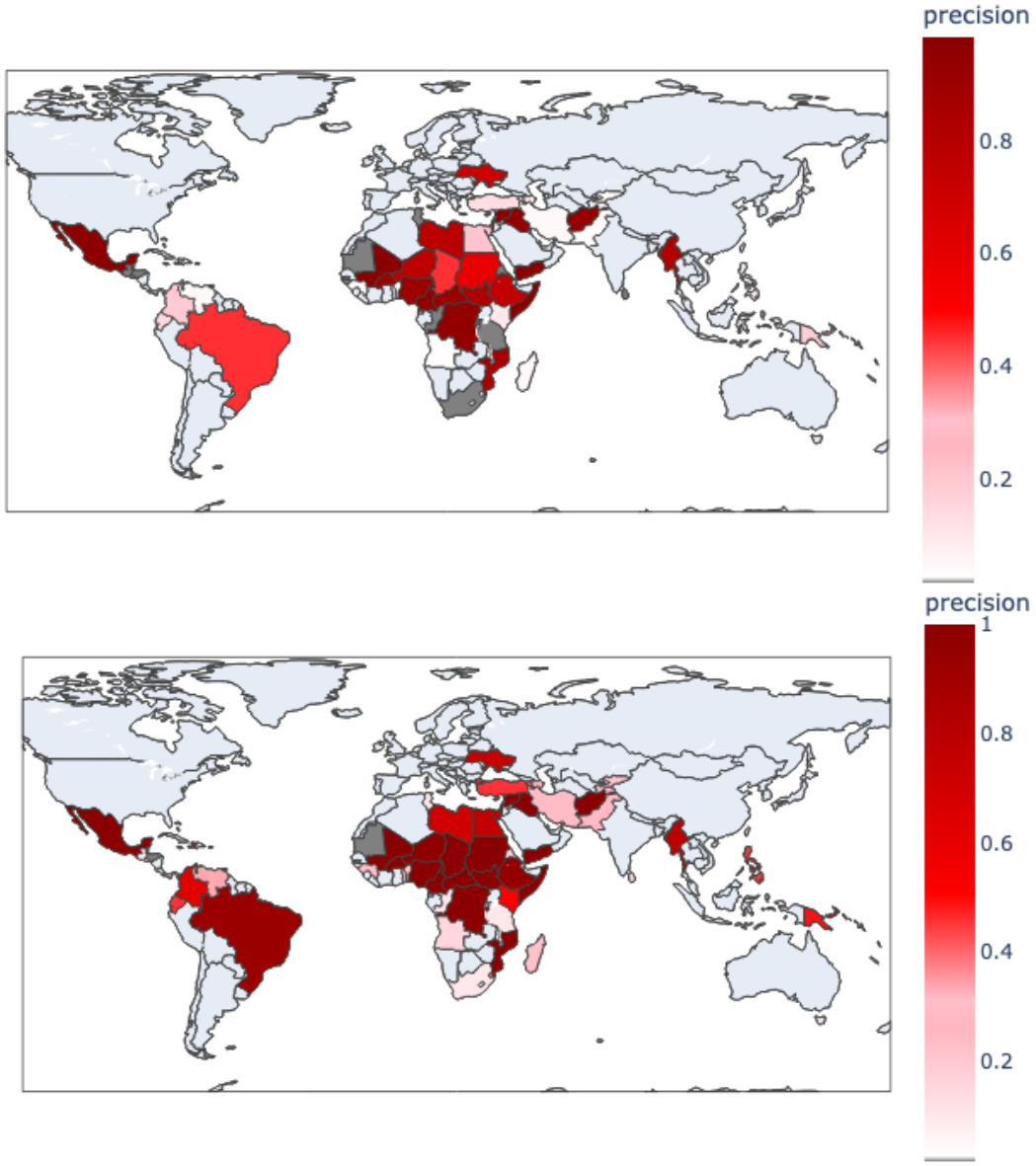

Figure 5.2: Precision of the text-only XGBoost model in forecasting conflict incidence within one month (top) and within 6 months (bottom).

Figure 5.2 illustrates the performance of the XGBoost classifier using only GDELT features when predicting conflict incidence. The top panel represents precision for the next month, while the bottom panel displays the precision for predicting conflict incidence within the next semester. Countries that are not colored indicate that there were no conflicts within the test period, as the precision rate would be zero by construction. Countries shown in dark grey represent cases where the model performed poorly, as it falsely predicted

a higher number of conflict incidents in the next month in Tanzania, Mauritania and South Africa, for instance, compared to the actual occurrences. On the other hand, it was particularly precise with conflict incidence in countries like Mexico, Mali, and the Democratic Republic of the Congo.

In the bottom panel, as the forecasting horizon expands to predict conflict incidence *within* the next semester, the precision tends to increase. This is expected since the model can anticipate that countries with a high likelihood of conflict are more likely to experience another conflict within the next 6 months, even if the exact timing is uncertain. One exception to this is Mauritania. Upon further investigation, we find that Mauritania experienced conflict only in March of 2023. However, none of the XGBoost predictions were able to capture this shift before it actually occurred. Unsurprisingly, this suggests that the there could be limitations in detecting conflict patterns in the 'hard' cases, as Mauritania had been almost 12 years without an armed conflict being detected before it erupted. In contrast, the model seems to perform quite well for countries like Mexico, Mali and the DRC, which are almost always in conflict.

Finally, we present the most important features for the XGBoost classifier when forecasting incidence in the Appendix A4. The feature importance analysis reveals interesting insights into the predictive models. The no-text and combined models prioritize variables related to deaths, including current deaths per capita, deaths over the past 6 months, and the stock of deaths. In contrast, the GDELT-only model sheds light on the relevance of text-based features, such as Admin1 features and shares and stocks of events related to fighting (e.g. event code 19). As the forecasting horizon extends to 3 and 6 months, the significance of deaths over the past 6 months becomes more prominent in the no-text and combined models, alongside the median admin1 'lethality' in the GDELT-only model.

5.2 Predicting conflict escalation

When examining somewhat harder case of predicting conflict escalation, the three classifiers demonstrate comparable performance, except for the text-only LSTM model, which slightly under-performs. Interestingly, the models show only mild differences compared to predicting conflict incidence, and ROC AUC scores tend to increase as the forecasting horizons extend.

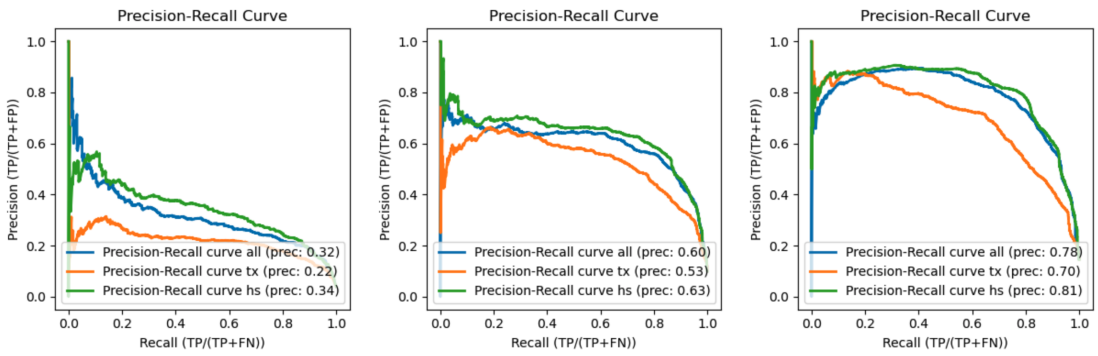

Figure 5.3: Precision-Recall Curves for Random Forest classifier in forecasting conflict escalation one month (left), three months (center), and six months ahead (right).

Figure 5.3 illustrates the precision-recall trade-off for different threshold values in the context of conflict escalation prediction. Notably, as the forecasting horizon extends, the area under the curve increases, indicating improvements in both precision and recall. For instance, in the Random Forest text-only model, the precision rises from 0.27 when predicting one period ahead to 0.69 when forecasting escalations within six months. Importantly, this enhancement does not come at the expense of recall, which also exhibits an increase from 0.67 to 0.88. These results indicate that the classifier not only provides accurate predictions but also captures a significant proportion of actual conflict escalations.

Figure 5.4 provides an overview of the Random Forest classifier's performance in predicting conflict escalations worldwide, using features derived solely from GDELT. Notably, the precision at $t+1$ is generally low, reaching a maximum of approximately 0.3. However, as the forecasting horizon expands to within $t+6$, the precision significantly improves. Similar to the incidence prediction, the model performs poorly in the hard cases of prolonged peace followed by conflict, such as Mauritania. The text-only model using a Random Forest classifier performs poorly once we restrict the analysis of escalation to countries with at least 10 years of peace, varying between 0.58 and 0.66. This is to be expected considering there are only 11 countries that move from long-standing peace to conflict within our test period; however, this model under-performs relative to the history-based and combined models (Figure A5.1 in Appendix A5).

In the case of predicting conflict escalation, a wider range of predictors are considered important in all three models (Figure A4.1). This highlights the distinct nature of the escalation prediction problem, where a diverse set of variables is necessary to anticipate changes in conflict patterns. Notable variables include refugee flows, deaths over the past

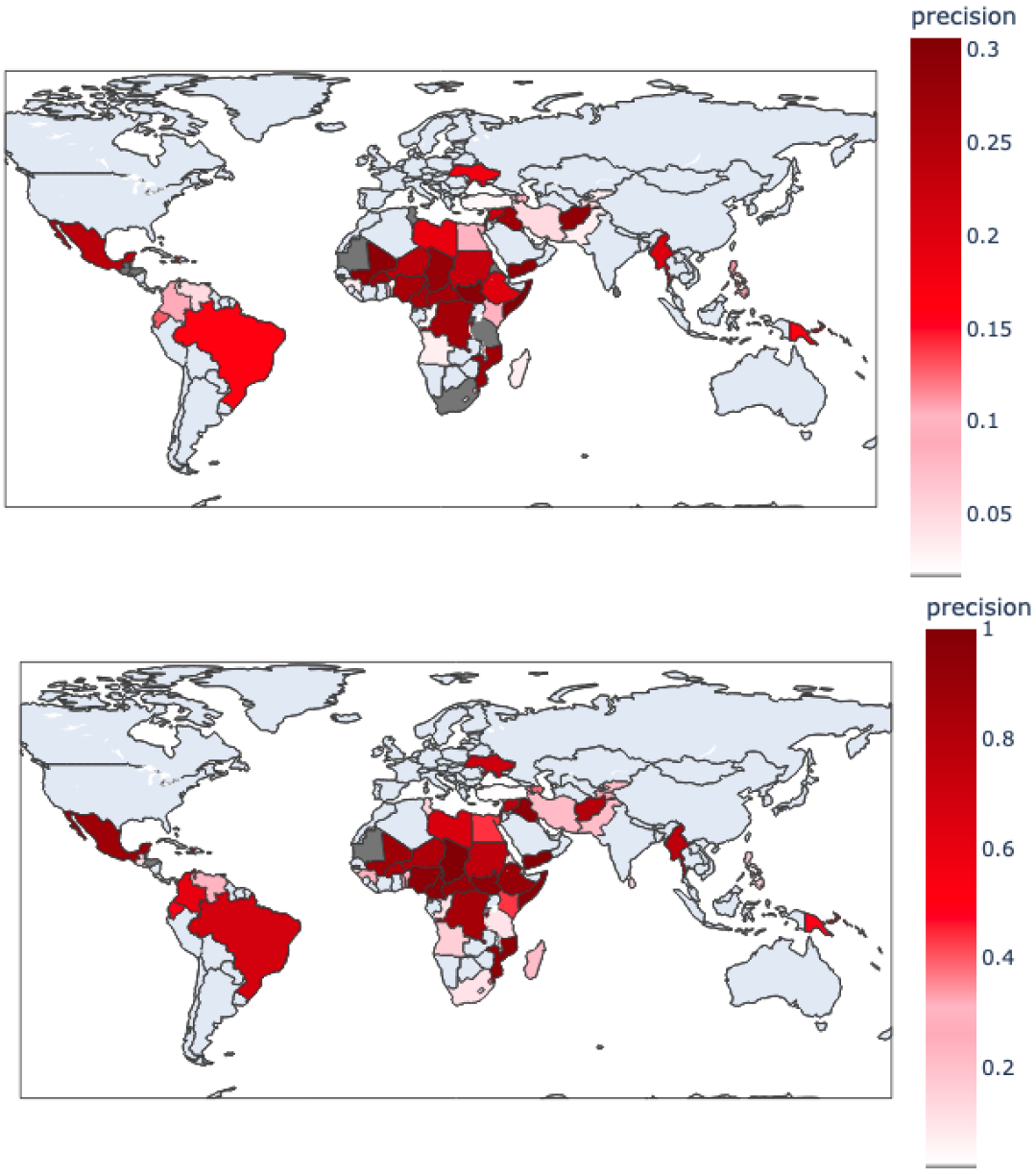

Figure 5.4: Precision of the text-only Random Forest model in forecasting conflict escalation within one month (top) and within 6 months (bottom).

10 years, and other event stocks such as 18 (assault) and 15 (exhibit force posture). The broader range of important predictors underscores the complexity of forecasting escalation and the need for comprehensive data sources and modeling approaches.

5.3 Predicting deaths

The XGBoost regressor outperforms most models when predicting deaths per 100,000 people. However, when restricting our analysis to only text-based features, LSTM is by far the best model. As we extend the time horizon for regression predictions, we observe

an improvement in performance across all algorithms, and this is especially noticeable for the GDELT-only model.

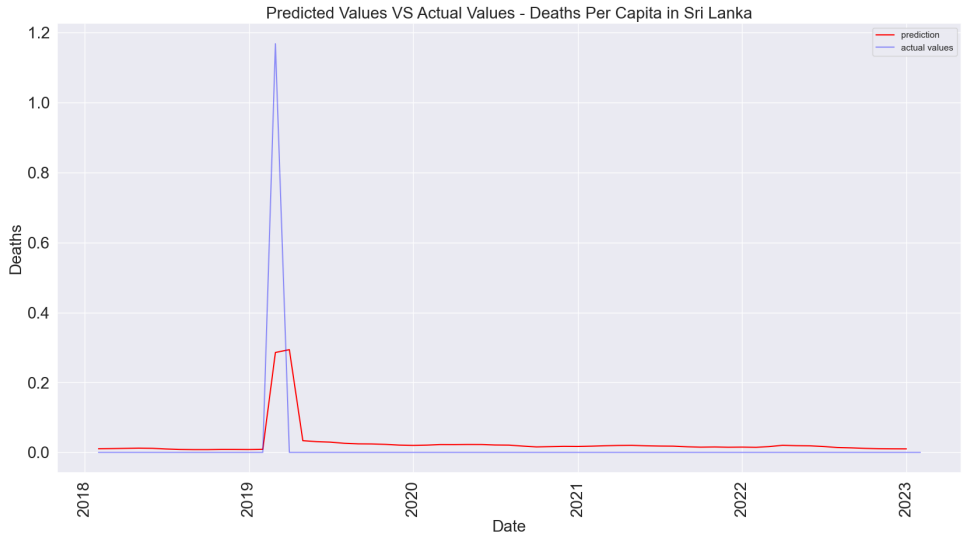

Figure 5.5: Prediction results for the combined LSTM model in forecasting deaths within one month.

Upon closer examination, we find that the LSTM model was remarkably capable of detecting the increase in deaths that occurred due to the Sri Lanka Easter bombings in April 2019 (Figure 5.5). While it underestimated the number of deaths, this is noteworthy considering the prior period of relative peace and stability in the country.

The LSTM text only regressor showed a better performance in capturing the trends of deaths per capita, compared to the other text only machine learning algorithms used in the project. Our LSTM model demonstrated superior performance in terms of following these trends, which is promising, given that due to time constraints we had not optimized our regression parameters.

6 Discussion

Using GDELT data for conflict prediction reveals several key insights. First, while exclusively using GDELT-based features does not outperform the predictive power of a no-text or combined model, it performs relatively well. In particular, it shows strong performance in tree-based models for classification tasks and LSTM models for regression tasks. This demonstrates the value of using text as data to capture patterns and provide insights into potential conflict events, especially over longer time horizons.

The real-time availability of GDELT data is a significant advantage, as it allows for the generation of features to predict the next period with minimal time lag. This timely information can aid decision-making and response efforts by international organizations. In fact, may be overestimating the performance of the history-based and combined models. In our approach, we give the model access to the t_0 value of the target, but in reality, such data may only reach decision makers with a delay and it may initially be imprecise. For example, the UCDP data on conflict related deaths is only published part-way, sometimes even only at the end of the next month. Thus it would be interesting to see if the gap between the text and other models becomes more narrow when restricting visibility of historical variables to t_{-1}.

However, there are also certain challenges and limitations to consider when analyzing these prediction results. Specifically, text-based models exhibit lower performance compared to history models when it comes to predicting the hard cases of escalation, those that occur after long periods of peace. While the no-text, history-based model achieved a ROC AUC score of 0.77 forecasting hard escalations within three months, the GDELT-only model's best score was only 0.66. Although not alarming, this performance gap is disappointing considering the potential of text data to capture nuanced shifts in sentiment.

One possible and interesting extension which was explored but not implemented due to time constrains is the framing of the prediction task as a Hidden Markov Model. In that respect, the goal would be to model sequential changes in hidden states that cannot be observed directly and can give rise to sudden onsets or escalations of conflict. Text information gathered from GDELT could prove to be fruitful in this type of model and useful in predicting the harder cases. To understand whether this approach might be

appropriate we conducted an initial cluster analysis with k-means clustering. This allowed us to see how many clusters - which could represent latent states - each country falls into. The preliminary results of this analysis are in Appendix A6.

Additionally, for the purposes of early warning models that aspire to be used by policymakers, there is a tradeoff between the sophistication of the prediction methodology and its suitability for easy interpretation. One example of this are the LSTM algorithms, for which we were not able to extract feature importance, a challenge when engaging with policymakers; while we can forecast the what, we lack the why. Although achieving both interpretability and high forecasting performance is challenging, efforts should be made to improve one without sacrificing the other (Hegre et al., 2017a).

The high levels of forecasting accuracy achieved by many models is also at odds with the complexity and randomness associated with conflicts (Bazzi et al., 2019). It is unlikely that there will be significant improvements in this respect, as country-level conflict predictions are not meant to be oracles of the future, but rather measure the pulse of the global media conversation and provide early warnings for conflicts (Kolusheva et al., 2023). Forecasts indicate what is likely to happen if no other action is taken, and can be made more effective when combined with policy analyses that assess the causal drivers of specific conflicts (Cederman and Weidmann, 2017).

Finally, media biases can impact conflict prediction based on news data, as they may distort reality or display reporting biases towards violent events (Voukelatou et al., 2020; Cederman and Weidmann, 2017), which can generate discrepancies between observed and real-world violence. Another related concern centers around freedom of the press in autocratic regimes, which could lead to failed predictions. However, prior research suggests that this relationship is not straightforward and predictive performance may not be significantly worse in countries with limited reporting, though it tends to suffer in autocracies (Mueller and Rauh, 2022).

7 Conclusion

The effectiveness of using the Global Database of Events, Language, and Tone (GDELT) as a source for forecasting geopolitical instability appears to be a complex issue. On the one hand, GDELT shows potential: it performs almost as well as combined text-and-history models with certain algorithms, such as tree-based models for classification and LSTM for regression. Importantly, GDELT data is available in real-time, allowing for almost instantaneous feature generation, which is a significant advantage for any forecasting model. A GDELT text-only model may also be effective in detecting spikes in conflict-related text references even when there are no official deaths reported, suggesting greater volatility on the ground than the available conflict data implies.

However, despite these strengths, GDELT text does not clearly outperform more all-encompassing models, and its relative effectiveness depends on the context and specific forecasting horizon. Forecasting geopolitical instability remains a fundamentally challenging task due to its inherent complexity, and GDELT is far from being a one-size-fits-all answer to the quest for reliable forecasts over broad times and spaces.

Concerns have also been raised about the potential biases in GDELT data, given media sources are susceptible to misrepresenting reality and vulnerable to changing political contexts.

Finally, there is the problem of intervention. While GDELT may be used to forecast events, the effectiveness of any policy intervention based on these forecasts is contingent on the knowledge of the drivers of conflict. Theory-free prediction does not necessarily guide intervention, emphasizing the need for careful policy analyses assessing the causal effectiveness of conflict-reducing measures.

As has been demonstrated throughout this study, GDELT shows great potential, especially in its broad global coverage, density, real-time availability and sub-state, geo-coded events. A model that uses only features derived from GDELT is not unsatisfactory when it comes to predicting conflict, and it performs generally well compared to more robust models. However, its use for geopolitical instability forecasting comes with its limitations and challenges. Any use of GDELT for such purposes should be mindful of these considerations, with appropriate adjustments and skepticism applied to the results it yields.

Areas for future research could focus on enhancing the granularity of the data processing techniques used with GDELT. An intriguing avenue to explore is the potential usage of the "IsRootEvent" attribute as a proxy for event importance. This attribute could provide a mechanism for weighting events during data aggregation, possibly leading to improved forecasting models.

References

ACLED (2019). Armed conflict location event data project (acled) codebook. Technical report.

Bazzi, S., Blair, R. A., Blattman, C., Dube, O., Gudgeon, M., and Merton Peck, R. (2019). The promise and pitfalls of conflict prediction: Evidence from colombia and indonesia. *NBER Working Paper Series*, 25980.

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

Brandt, P. T., Freeman, J. R., and Schrodt, P. A. (2011). Real time, time series forecasting of inter- and intra-state political conflict. *Conflict Management and Peace Science*, 28(1):41–64.

Carmignani, F. and Kler, P. (2016). Surrounded by wars: Quantifying the role of spatial conflict spillovers. *Economic Analysis and Policy*, 49:7–16.

Cederman, L.-E. and Weidmann, N. B. (2017). Predicting armed conflict: Time to adjust our expectations? *Science*, 355(6324):474–476.

Celiku, B. and Kraay, A. (2017). Predicting conflict. *World Bank Policy Research Working Paper Series*, 8075.

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

Chen, P., Jatowt, A., and Yoshikawa, M. (2020). Conflict or cooperation? predicting future tendency of international relations. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, SAC '20, page 923–930, New York, NY, USA. Association for Computing Machinery.

Chiba, D. and Gleditsch, K. S. (2017). The shape of things to come? expanding the inequality and grievance model for civil war forecasts with event data. *Journal of Peace Research*, 54(2):275–297.

Galla, D. and Burke, J. (2018). Predicting social unrest using gdelt. In Perner, P., editor, *Machine Learning and Data Mining in Pattern Recognition*, pages 103–116, Cham. Springer International Publishing.

GDELT Project (2023). GDELT Project. https://data.worldbank.org/indicator/SP.POP.TOTL. Retrieved on 01/05/2023.

Guardado, J. and Pennings, S. (2020). The seasonality of conflict. *World Bank Policy Research Working Paper Series*, 9373.

Hasell, J. (2022). Counting world conflict deaths: why do sources differ?

Hegre, H., Karlsen, J., Nygård, H. M., Strand, H., and Urdal, H. (2013). Predicting Armed Conflict, 2010–20501. *International Studies Quarterly*, 57(2):250–270.

Hegre, H., Metternich, N. W., Nygård, H. M., and Wucherpfennig, J. (2017a). Introduction: Forecasting in peace research. *Journal of Peace Research*, 54(2):113–124.

Hegre, H., Nygård, H. M., and Ræder, R. F. (2017b). Evaluating the scope and intensity of the conflict trap: A dynamic simulation approach. *Journal of Peace Research*, 54(2):243–261.

Kolusheva, D., Stoughton, C., and Wheeler, E. (2023). Text is all you need: Predicting conflict escalation using global news content.

Landis, S. T. (2014). Temperature seasonality and violent conflict: The inconsistencies of a warming planet. *Journal of Peace Research*, 51(5):603–618.

Mueller, H. and Rauh, C. (2018). Reading between the lines: Prediction of political violence using newspaper text. *American Political Science Review*, 112(2):358–375.

Mueller, H. and Rauh, C. (2022). The Hard Problem of Prediction for Conflict Prevention. *Journal of the European Economic Association*, 20(6):2440–2467.

Qiao, F., Li, P., Zhang, X., Ding, Z., Cheng, J., and Wang, H. (2017). Predicting social unrest events with hidden markov models using gdelt. *Discrete Dynamics in Nature and Society*, 2017.

Rost, N., Schneider, G., and Kleibl, J. (2009). A global risk assessment model for civil wars. *Social Science Research*, 38(4):921–933.

Sambanis, N. (2004). What is civil war? conceptual and empirical complexities of an operational definition. *The Journal of Conflict Resolution*, 48(6):814–858.

Sarkees, M. R. (2010). The cow typology of war: Defining and categorizing wars (version 4 of the data). Technical report, Correlates of War.

Smith, E. M., Smith, J., Legg, P., and Francis, S. (2018). Predicting the occurrence of world news events using recurrent neural networks and auto-regressive moving average models. In Chao, F., Schockaert, S., and Zhang, Q., editors, *Advances in Computational Intelligence Systems*, pages 191–202, Cham. Springer International Publishing.

UNHCR (2023). UNHCR Refugee Statistics. https://www.unhcr.org/refugee-statistics/download/?url=2bxU2f. Retrieved on 01/05/2023.

Uppsala University (2023). UCDP Conflict Encyclopedia. ucdp.uu.se. Retrieved on 01/05/2023.

Voukelatou, V., Pappalardo, L., Miliou, I., Gabrielli, L., and Giannotti, F. (2020). Estimating countries' peace index through the lens of the world news as monitored by gdelt. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 216–225.

Wang, X., Chen, H., Li, Z., and Zhao, Z. (2018). Unrest news amount prediction with context-aware attention lstm. In Geng, X. and Kang, B.-H., editors, *PRICAI 2018: Trends in Artificial Intelligence*, pages 369–377, Cham. Springer International Publishing.

Ward, M. D., Greenhill, B. D., and Bakke, K. M. (2010). The perils of policy by p-value: Predicting civil conflicts. *Journal of Peace Research*, 47(4):363–375.

Wen, Q., Zhou, T., Zhang, C., Chen, W., Ma, Z., Yan, J., and Sun, L. (2023). Transformers in time series: A survey.

World Bank (2011). *World Development Report 2011: Conflict, Security, and Development*. The World Bank.

World Bank (2023). Population, Total. https://data.worldbank.org/indicator/SP.POP.TOTL. Retrieved on 01/06/2023.

World Population Review (2023). Total Population by Country. https://worldpopulationreview.com/. Retrieved on 01/06/2023.

Yonamine, J. E. (2013). Predicting future levels of violence in afghanistan districts using gdelt. *Unpublished Manuscript*.

Appendix

A1 List of dropped countries

The following countries were dropped from the analysis, primarily due to lack of population data:

Anguilla, Antarctica, Bouvet Island, British Indian Ocean Territory, Christmas Island, Cook Islands, Falkland Islands (Malvinas), French Guiana, Guadeloupe, Guernsey, Heard Island and McDonald Islands, Holy See (Vatican City State), Jan Mayen, Jersey, Mayotte, Martinique, Montserrat, Niue, Norfolk Island, Pitcairn, Réunion, Saint Helena, Ascension and Tristan da Cunha, Saint Pierre and Miquelon, Svalbard, Tokelau, Wallis and Futuna

A2 Notes on choice of population

As discussed in the Data section, we were reluctant to simply take an average of the population value for the respective countries as some countries saw significant changes to their population and the average would depend quite heavily on the time-period for which the average is taken. This is visualised in the two figures below. Population mattes because it affects whether a country meets the per capita threshold for being in conflict. Relatedly, it has been argued that the "statistically significant results for the importance of population size in raising the risk of civil war might be a statistical artifact resulting from the high absolute threshold of battle deaths for civil wars". (Rost et al., 2009; Sambanis, 2004).

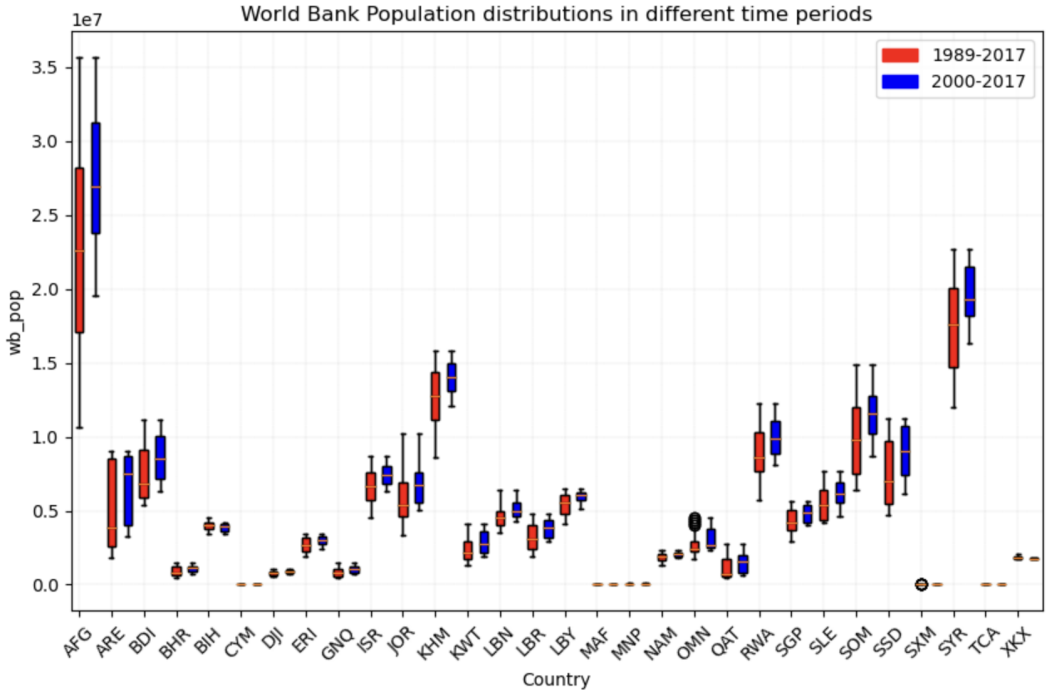

Figure A2.1: The World Bank population distribution changes considerably for different potential training periods.

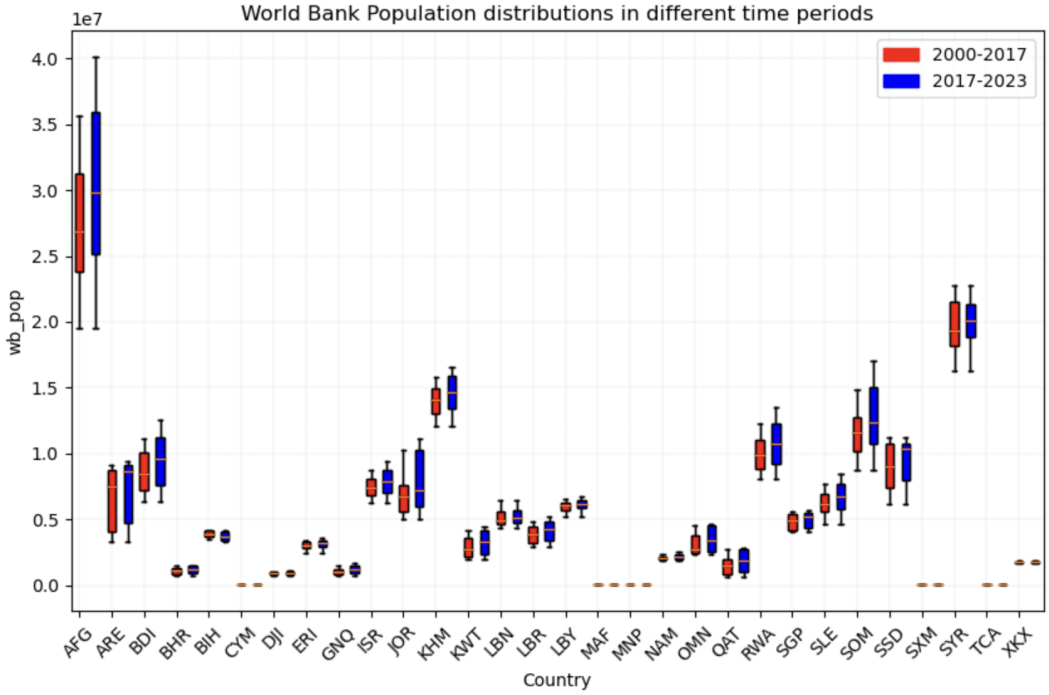

Figure A2.2: The World Bank population distribution changes considerably between the chosen train and test period.

A3 Cluster feature

To obtain the cluster feature, we only used GDELT features and the target, aggregating them at the country-level such that each country would only be part of one cluster (unlike the clustering approach taken in the exploration of potential latent states for a Hidden Markov Model approach - see Appendix A6 below). As we used different and often multiple aggregations per feature (e.g. median and maximum), this resulted in many columns. We reduced these using principal component analysis and selected the first three Principal Components (PC). Using k-means we determined the optimal number of clusters to be 15. The results of plotting the countries, coloured by cluster, along the dimensions of the first three principal components can be seen below.

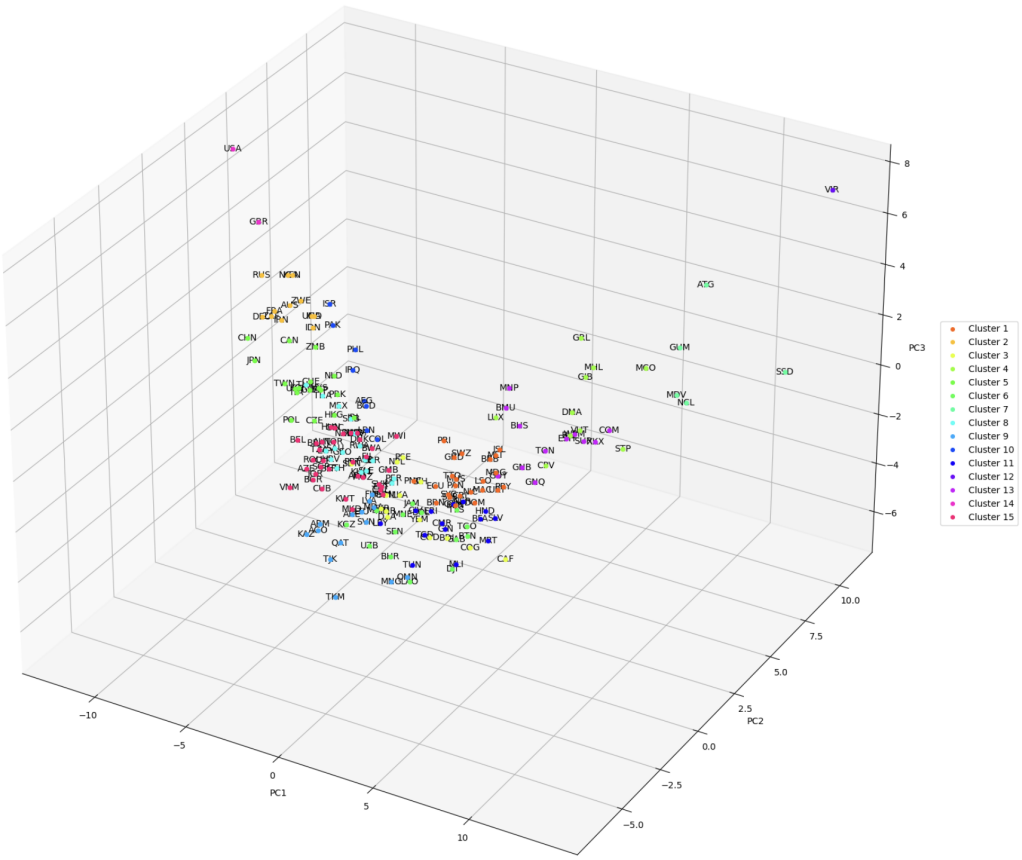

Figure A3.1: Clusters based on GDELT features and the target (encoded within the training set from 2000 to 2017).

A4 Feature importance for tree-based models

Below we plot the feature importance identified by the tree-based algorithms applied to all three models.

A5 The hard cases on escalations

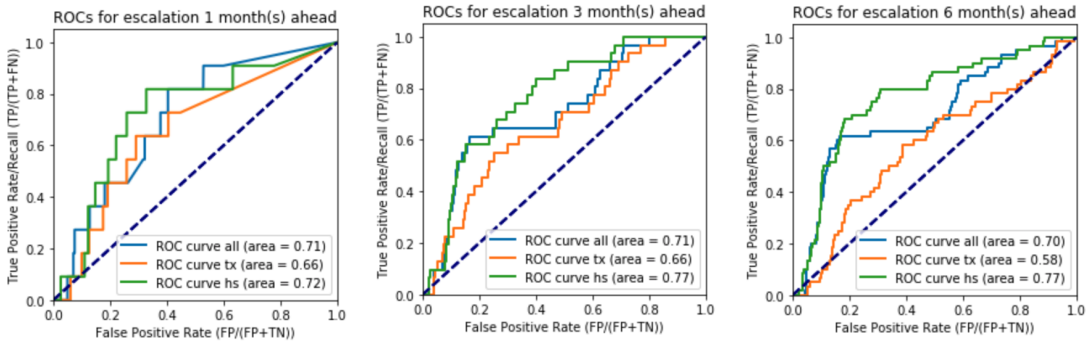

Figure A5.1: ROC Curve and AUC Scores for the Random Forest model forecasting escalation one period ahead (left), within three periods (center), and within six periods (right), restricted to cases in which there has been no conflict in at least 120 months.

A6 Clusters for Hidden Markov Model

When considering a Hidden Markov Model approach to the problem, we initially sought to estimate the number of latent states to encode. The graph below shows that most countries' observations seem to form part of four different clusters, possibly suggesting four latent states. The optimal number of overall six clusters to apply the k-means algorithm to was determined by the elbow method.

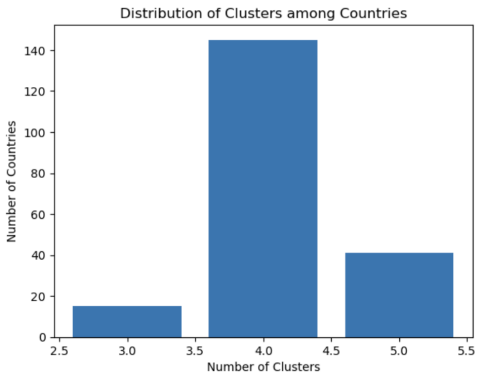

Figure A6.1: Clusters based on features and the target, allowing each observation to be part of a cluster and thus allowing for multiple clusters per country.

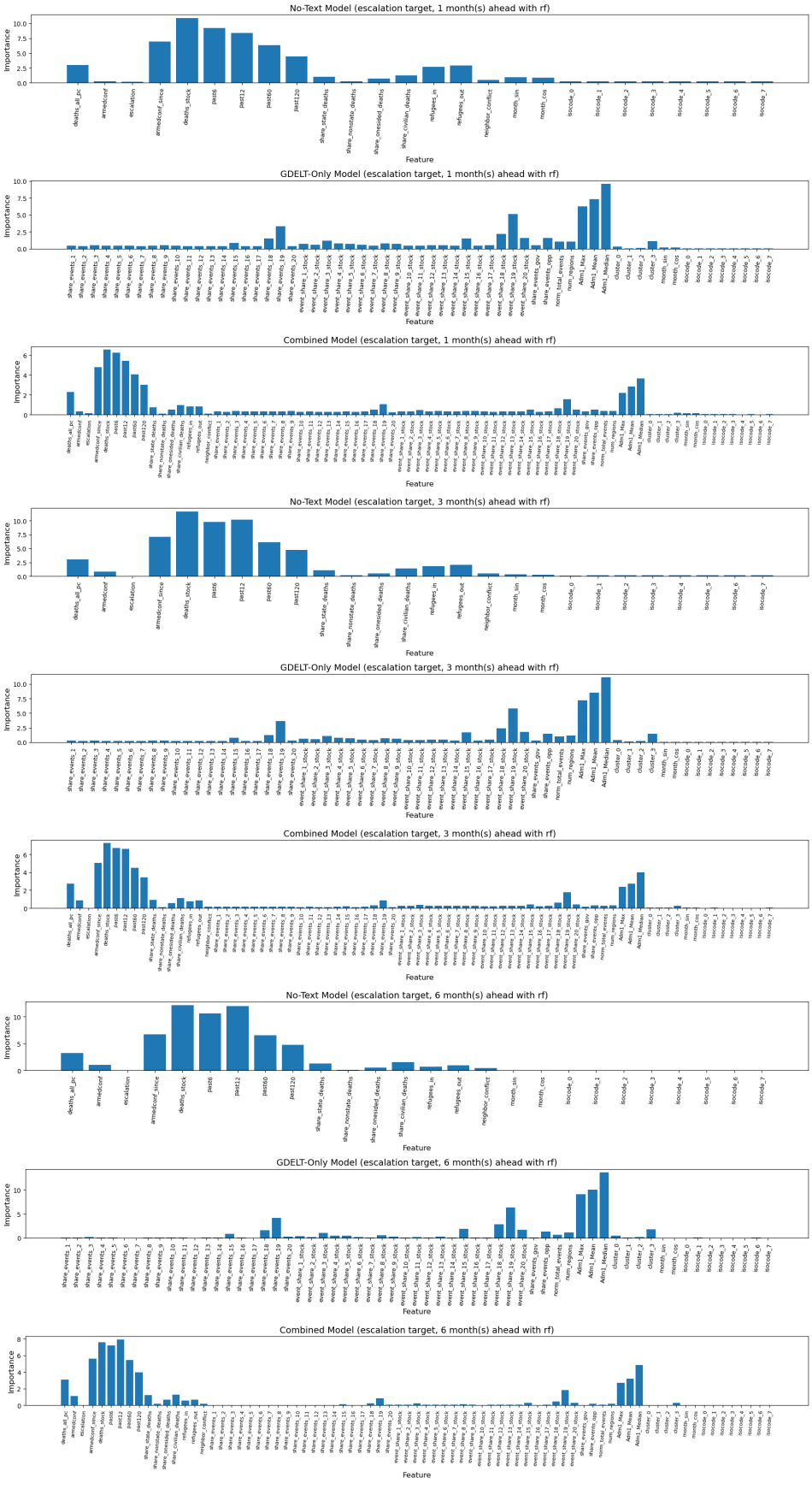

Figure A4.1: Feature importance for the Random Forest model forecasting escalation one period ahead (top), within three periods (middle), and within six periods (bottom)

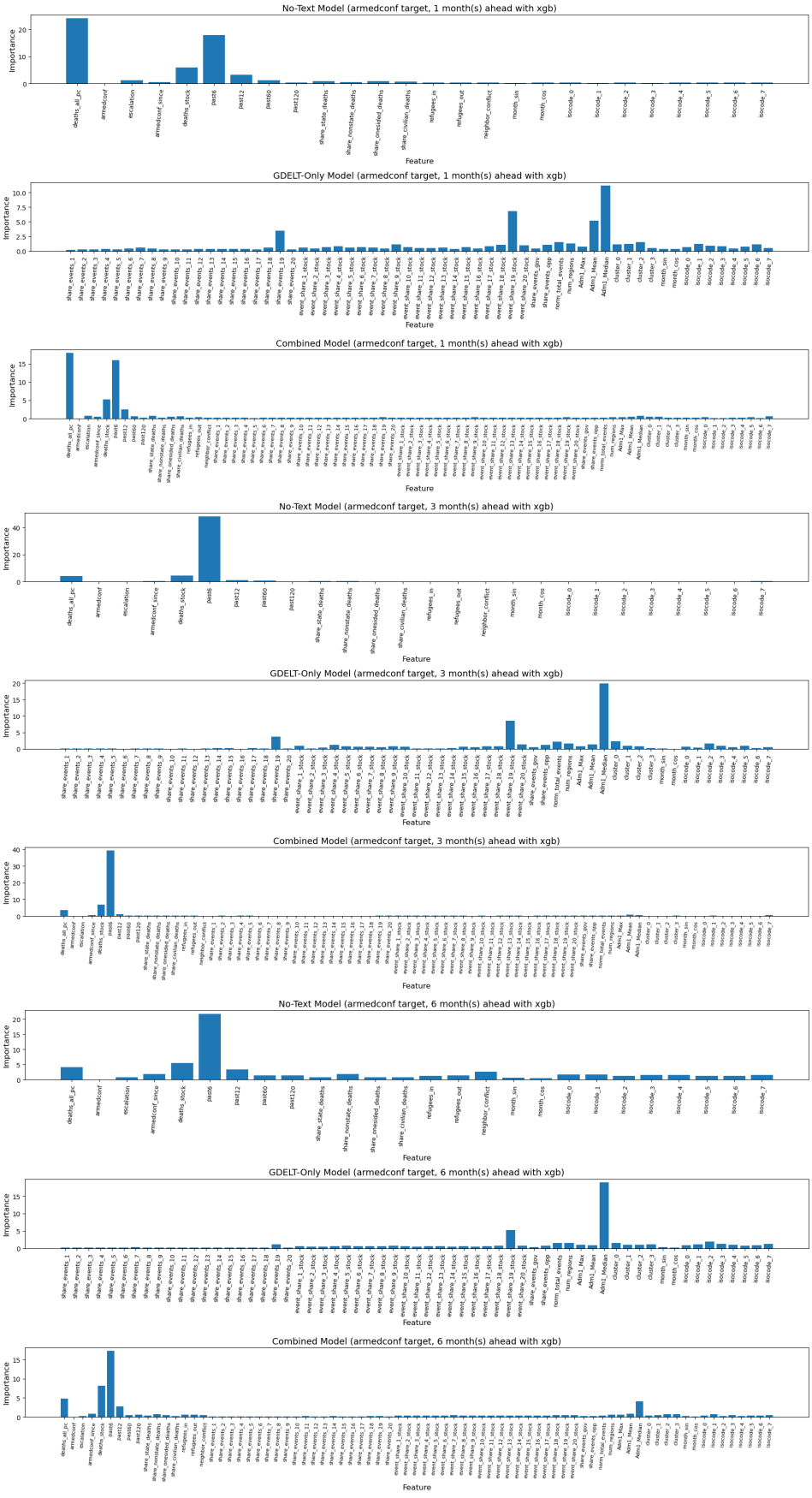

Figure A4.2: Feature importance for the XGBoost model forecasting incidence one period ahead (top), within three periods (middle), and within six periods (bottom)