Text Is All You Need

Predicting Conflict Escalation Using Global News Content

Diana Kolusheva Chad Stoughton Evan Wheeler

The task of forecasting conflict escalation is of great interest to policy makers, executives, and social scientists. In this article, we propose a novel, text-first approach to computational conflict prediction using contemporary natural language processing and machine learning techniques. We outline three methods to extract generalizable conflict signals directly from a corpus of over 2.5 million document embeddings derived from publicly available news sources. The first uses text data enriched by country context features, the second uses text data conditioned only on the country label itself, and the third extracts a signal from the text alone, with no additional country or region-specific features. We demonstrate that signals suggesting impending conflict can be extracted directly from text sources, and we outline potential avenues for further research that we believe may yield models that can provide early warnings to decision makers to facilitate strategic planning, resource allocation, and disaster mitigation.

Introduction

Motivation

This project was undertaken in response to the acute needs of the UNICEF Emergency Operations Team. In 2021, UNICEF responded to 483 new and ongoing humanitarian situations in 153 countries (natural disasters, socio-political crises, health emergencies, nutrition crises, etc.) in addition to its global COVID-19 response efforts. This represents a sharp increase since 2018, and the number of disasters continues to climb. With the rising threat of the climate emergency, global shifts toward authoritarianism, and increasing levels of global conflict, innovative approaches are necessary to scale the reach of UNICEF's impact to meet the increasing needs of vulnerable children. Automated means of identifying and escalating potentially negative situations as they unfold—or before they unfold—can help UNICEF's emergency response functions be more proactive and efficient in responding quickly to people in need.

Objective

Our practical objective was to develop a tool to support UNICEF in conflict identification and response. Our research objective was to demonstrate that a generalizable, global conflict prediction model is viable, and that such a model can be built using primarily public text data.

Conflict forecasting

Predicting conflict has long been of interest to researchers and practitioners of international relations, and statistical and computational methods to do so can be found in the literature as early as 1944. Modern methodologies that resemble those we outline here began to appear in the literature in the early 2010s, when researchers began using holdout datasets for model validation (O'Brien 2010). This shift followed work by Ward et al. that sharply criticized the use of statistical significance tests on in-sample prediction to validate conflict models (Ward, Greenhill, and Bakke 2010). Our methodology closely mirrors that which emerged in the following decade, including that described in more contemporary work, such as Mueller and Ruah (H. Mueller and Rauh 2022). In essence, this methodology involves using time-lagged input data to predict a conflict label that corresponds to a date some time after the input data would have become available. The model is then validated using an unseen holdout set, typically at the end of the date range available in the data (Hegre, Vesco, and Colaresi 2022). This overcomes the problem of overfitting observed in earlier methods, and provides a principled starting point from which different models can be compared.

Today, much of the work in this area has continued to focus on country or region-specific models to predict conflict. This introduces a "small-data problem" where relatively few observations of conflict are available for a given country (H. Mueller and Rauh 2022). Some efforts have been made to overcome this by building generalizable models that can theoretically be used to make predictions on any country (Wen et al. 2023). We followed the latter approach in developing our model, hypothesizing that the larger number of observations available to a global model will allow it to extract more generalizable signals of conflict, and that such a model will be better able to predict unforeseen risks in countries with little available training data.

Methodology

The principal difference between our approach and that of prior work is in our choice of predictive features, which we extract directly from text embeddings of publicly available news coverage. While newspaper and media content has been used in conflict prediction models before, it is generally only included as a supplement to models that primarily rely on other features. Despite this, models that include text features have been shown to outperform equivalent models that do not (H. Mueller and Rauh 2022).

In contrast, our approach is text-first, making predictions primarily or exclusively from text, with few or no other features added. To the best of our knowledge, no conflict prediction model described in the literature makes use of text embeddings. Previous efforts have included statistical analysis of text, and unsupervised Latent Dirichlet Allocation (LDA) for topic extraction (H. Mueller and Rauh 2022). These approaches to Natural Language Processing (NLP) have relegated text data to a peripheral role in conflict prediction.

In order to elevate text data to the center-stage of our prediction efforts, we drew from prior work in other fields. Our training data was structured such that each document is represented as an embedding vector, and multiple documents are combined with a label to create an aggregate picture of recent events.

We developed three approaches for extracting conflict signals from text embeddings. The first approach used supervised dimensionality reduction and a feed-forward neural network conditioned with country context characteristics. In the second approach, we used a compact dataset of monthly embedding averages per country per month and applied statistical classification models such as K-Nearest Neighbors and XGBoost. In the third approach we stacked multiple article-level embeddings and applied transformer-based models to extract a signal directly from the text corpus. See the Figure 5, Figure 6, and Figure 8 for details of modeling approaches.

Defining Conflict Escalation

Our task of modeling conflict escalation required a precise definition of what conflict escalation means. We began with raw event and casualty counts from the Armed Conflict Location & Event Data Project (ACLED), whose team of researchers monitor thousands of sources in 20 languages to generate a dataset containing monthly totals of conflict events and their related fatalities (Raleigh et al. 2010). ACLED defines these events as incidents involving armed groups with political objectives as well as demonstrations, riots, and protests. With these totals in-hand, we needed to create a target to label our training data. Prior work has treated this as a binary classification task, asking whether or not conflict escalated in a given month, and we adopt this approach as well. (Hannes Mueller and Rauh 2022)

In contrast to prior work, however, we do not treat the simple presence of a fatality due to conflict as sufficient to indicate conflict escalation. To achieve our objective of creating a generalizable conflict model, we required a metric that was consistent across varying local conditions. Small changes in the absolute number of monthly fatalities can have widely different implications depending on the country and time period.

To address this, we derived a metric designed to highlight relative “spikes” in political violence. Calculating whether a spike occurred in a given month is a two-step process. First, we calculated the trend in violence for a given month, defined by the slope of an ordinary least squares regression of the number of ACLED-documented fatalities that occurred in the target month, and the preceding two months. This trend allowed us to estimate whether violence was increasing or decreasing in the target country at any given time. Next, we calculated whether that slope was in the 75^{th} percentile of such slopes in the target country for the preceding two years. This produced a binary metric that is sensitive to small increases in political violence in peaceful countries, while also filtering out small fluctuations in long-standing conflicts.

Figure 1 shows how this metric captures conflict fluctuations in Ukraine. Prior to the Russian invasion of February 2022, small fluctuations were able to trigger a spike. However, much larger fluctuations in the period following the invasion are filtered out, as the spike formula adjusts to the higher level of baseline conflict. The increase in violence during the Ukrainian counter-offensive in the late summer again triggers a spike, and this reflects significant changes in the situation on the ground.

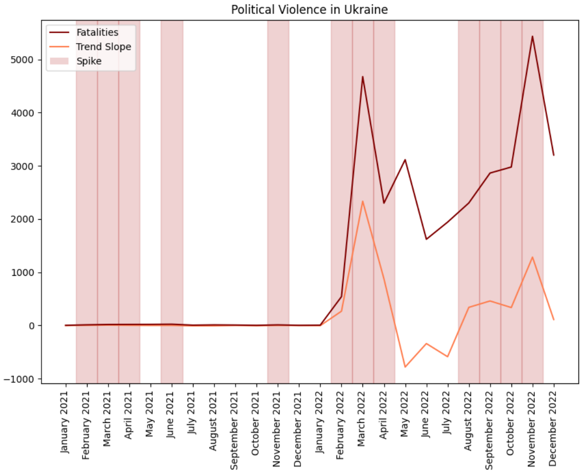

Figure 1: The Spike metric applied to Ukraine

News Content Data

The Global Database of Events Language and Tone (GDELT) Project provides public APIs of current events derived from news articles scraped from sources around the world (Leetaru and Schrodt 2013). With updates every fifteen minutes, GDELT’s Event Database documents physical events such as diplomatic interactions, social unrest, and military action. Approximately 60 attributes are assigned to each event from GDELT’s processing and analysis, including named entities involved, geolocation details, categorization among 300 event types, as well as metrics related to average ‘tone’ of articles about the event and potential severity of event impact. In 2021, GDELT introduced its Global Similarity Graph Document Embeddings dataset, which precomputes 512-dimension document-level embeddings of their scraped news articles using the Universal Sentence Encoder (“Announcing the Global Similarity Graph” 2021) (Cer et al. 2018).

The GDELT Dataset has been used for a wide variety of research purposes, including for conflict prediction (Chen, Jatowt, and Yoshikawa 2020).

While we believed this to be the best dataset available for our purposes, prior research has revealed some limitations that we must consider. GDELT's coverage over-represents the western world, particularly the United States and Europe. GDELT does represent Russia, India and China better than other event datasets, but it continues to underrepresent other parts of the world, particularly the global south. English is the most common language in the GDELT dataset, comprising more than 40% of the articles in its corpus (Kwak and An 2021).

Training Dataset

Our training data was derived from these GDELT and ACLED datasets. GDELT news article embeddings served as our independent variable, while the "spike" metric derived from ACLED's conflict data served as our target variable. The spike metric was used to label article data on a three-month time lag, allowing us to make predictions of future events.

GDELT's news article embeddings exist only from 1 January 2020, limiting the time period for which training data can be drawn. Availability of ACLED's monthly event and fatality totals varies by country, but event and fatality totals are available for all countries since at least 1 January 2020. Our training dataset thus ran from January 2020 through October 2022. Data from November 2022 through March 2023 was used for model validation and testing.

Conflict Prediction Using Text and Country Features

Our first approach employed a feed forward neural network model. Uniform manifold approximation and projection (UMAP) was used to reduce dimensionality of GDELT article embeddings in a supervised manner (Sainburg, McInnes, and Gentner 2021). Our spike variable was used to condition the dimension reduction process of the training data to encourage separation of the variable classes in the reduced embedding space. The fit UMAP model was applied to the validation and test data without supervision. These reduced, 128-dimension embeddings, along with country context indicators, were used as inputs to a feed forward neural network with a single hidden layer of 32 neurons. The model's 4577 parameters were trained for 3 epochs.

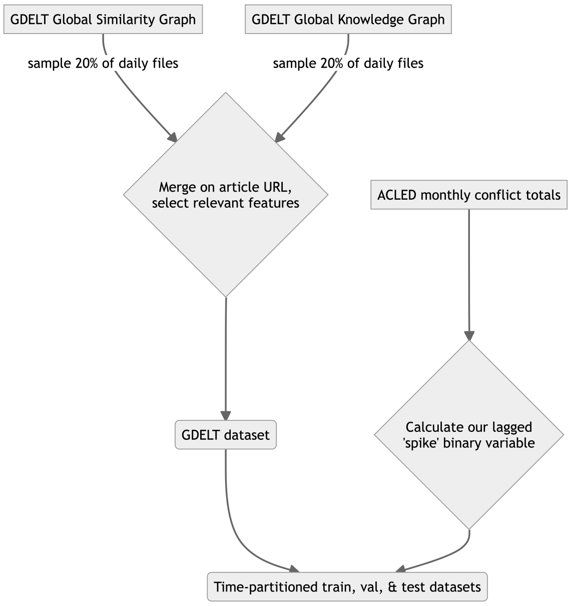
GDELT Global Similarity Graph
GDELT Global Knowledge Graph
sample 20% of daily files
sample 20% of daily files
Merge on article URL, select relevant features
ACLED monthly conflict totals
GDELT dataset
Calculate our lagged 'spike' binary variable
Time-partitioned train, val, & test datasets

Figure 2: Overview of data generation process for our conflict dataset

Country context inputs consisted of two numerical indicators ('Under-five mortality rate 2019' and 'Adolescent population 2020 Proportion of total population (%) Total' from UNICEF's State of the World's Children report (UNICEF 2021)) and three one-hot-encoded categorical indicators (categories of Human Development Index ranges (Nations Development Programme) 2022), categories of gross national income per capita (Fantom 2016), and categories of Fragile States Index ranges (Nate Haken 2022)).

Conflict Prediction Using Text and Country Labels

In this approach we prepared the dataset by averaging news article embeddings per country,

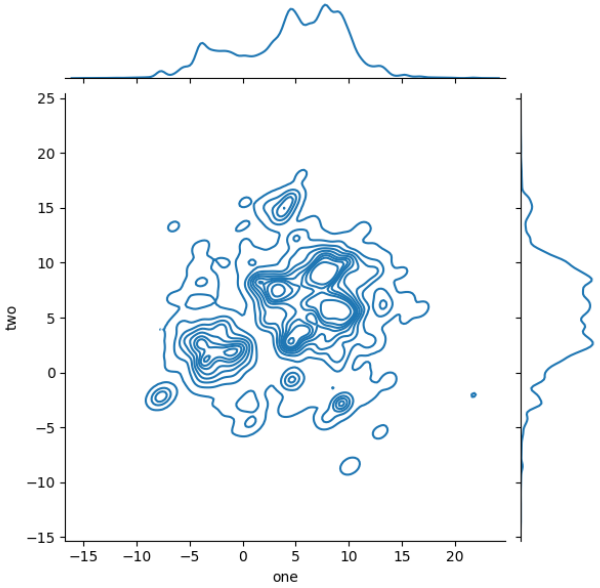

Figure 3: Density of embeddings resulting from supervised UMAP reduction to 2 dimensions

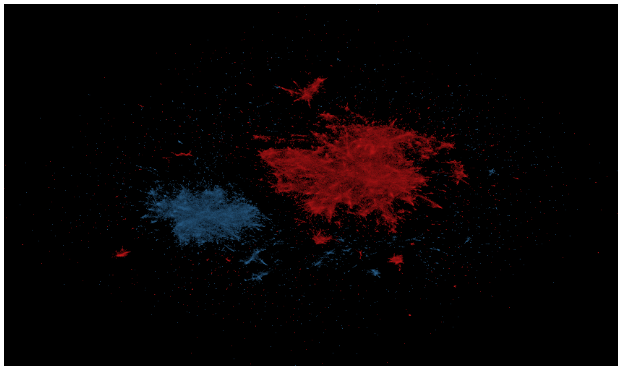

Figure 4: Scatterplot of 2-dimension UMAP-reduced articles with spike labels

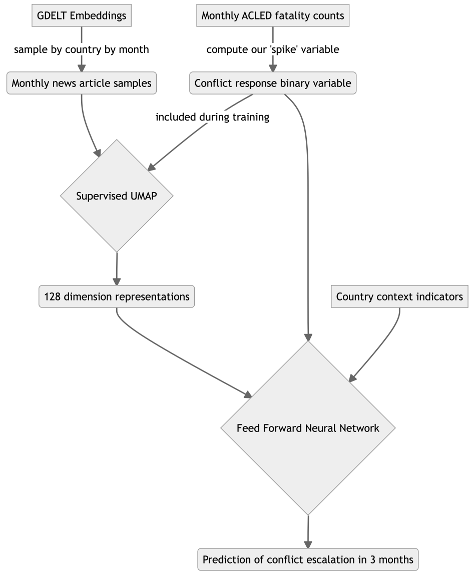
GDELT Embeddings
Monthly ACLED fatality counts
sample by country by month
compute our 'spike' variable
Monthly news article samples
Conflict response binary variable
included during training
Supervised UMAP
128 dimension representations
Country context indicators
Feed Forward Neural Network
Prediction of conflict escalation in 3 months

Figure 5: Feed forward neural network model architecture

per month. We also included one-hot encoded columns for countries, but no other country features were added. Since we only had access to embeddings starting in January 2020, the total length of this dataset was just a little over 6,000 samples (training, validation, and test sets combined). Given the high dimensionality of the data (512-dimensional embedding vectors + 700 dummy variables for countries), this dataset suffered from the "curse of dimensionality". This was handled by performing Principal Component Analysis; we found that reducing the number of dimensions to represent 80% of variance (~169 components with our selected train-test cutoff) improved the overall performance. Another issue was a class imbalance: conflict spike was present in only 14% of the training samples. We used random over-sampling technique to upsample the positive examples.

The model used in this approach is a stacked model that averages probabilities independently predicted by a K-Nearest Neighbors model and an XGBoost model. We used standard open-source versions of the underlying models. The intention behind averaging the predicted probabilities was to reduce the impact of individual model weaknesses.

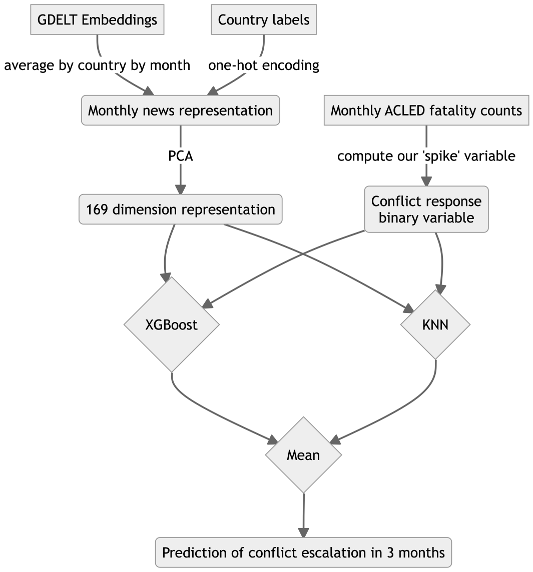
GDELT Embeddings
Country labels
average by country by month
one-hot encoding
Monthly news representation
Monthly ACLED fatality counts
PCA
compute our 'spike' variable
169 dimension representation
Conflict response binary variable
XGBoost
KNN
Mean
Prediction of conflict escalation in 3 months

Figure 6: Statistical model architecture

Conflict Prediction Using Text Alone

Our third approach used a pair of transformer models to extract information directly from the GDELT embedding corpus (Vaswani et al. 2017). Architecturally, both models are encoder-only transformers, but for simplicity, we will refer to the first of these models as “the encoder” and the second as the “classifier”.

The purpose of the encoder was to reduce the dimensionality of the embedding corpus by analyzing a subset of article embeddings and returning a single vector of substantially reduced size. This is roughly analogous to the role of a convolutional layer in a CNN. The classifier could then be used to make final predictions on the resulting, distilled dataset.

To prepare the training dataset, our full sample of the GDELT corpus was split into a single $512 \times n$ matrix for each country and month where n is the number of articles published in that month related to that country. Each of these matrices was then split into several random subsamples of 50 articles each. Samples for countries in which $n < 50$ were padded to 512×50. This process had two effects on training. First, by limiting the size of each sample to 50 articles, we reduced the dimensionality of each sample. Simultaneously, by splitting the corpus into multiple unique training examples per country per month, we were able to substantially increase the size of the training dataset.

This new dataset offered major advantages in combatting the curse of dimensionality, however, it also presented a distorted view of the world which over-represents countries that receive more media coverage and underrepresents those that receive less, amplifying the biases already present in the GDELT dataset. Despite this limitation, the dataset was sufficient to train the encoder, which was fit to the modified dataset with the time-shifted spike metric as its target. Once training was complete, the final layer of the encoder was removed, exposing the penultimate layer of 16 dense neurons. The output of these 16 neurons would be used in the classification stage.

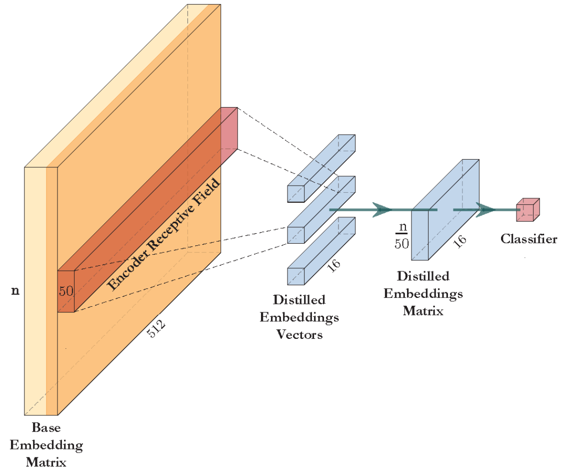

Figure 7: Reducing Embedding Dimensionality with Transformer Encoder

With the encoder’s training complete, we returned to the original base dataset. Each country-month matrix was again split into 512×50 matrices. Each of these samples could then be passed into the encoder, which distilled it into a $length-16$ vector. Each vector was then stacked, resulting in a single training example for each country and month of size $16 \times \frac{n}{50}$. This produced a new, distilled, and undistorted dataset which could then be used to train the classifier using the same spike metric as its target. In contrast to our earlier approaches, the dual-transformer model was trained purely on news article text embeddings. The model was not conditioned by country, and no additional features were added beyond the text corpus itself.

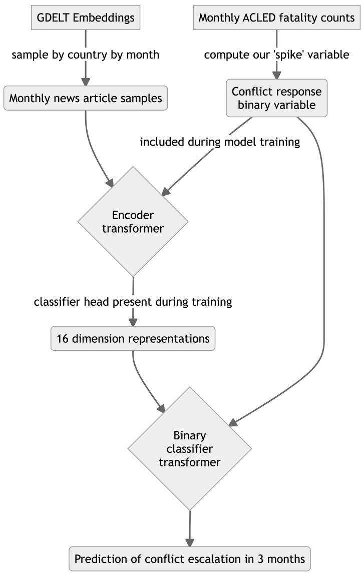
GDELT Embeddings
Monthly ACLED fatality counts
sample by country by month
compute our 'spike' variable
Monthly news article samples
Conflict response binary variable
included during model training
Encoder transformer
classifier head present during training
16 dimension representations
Binary classifier transformer
Prediction of conflict escalation in 3 months

Figure 8: Transformer model architecture

Results

We evaluated the models against a set of common performance metrics (see Figure 9, Figure 10, Figure 11). Despite conceptual differences in dataset creation and modeling techniques, all three of the described above approaches produced comparable results. Each of the models outperformed the other two in a subset of these metrics.

metric	FFNN	XGBoost	Transformer
False Discovery Rate	0.73	0.83	0.80
False Negative Rate	0.59	0.28	0.53
False Positive Rate	0.19	0.51	0.30

Figure 9: Model metrics (smaller values are better)

metric	FFNN	XGBoost	Transformer
Cohen Kappa Score	0.18	0.09	0.11
F1 Score	0.33	0.28	0.28
Precision Score	0.27	0.17	0.20
Recall Score	0.41	0.72	0.47
Roc Auc Score	0.67	0.69	0.64
True Negative Rate	0.81	0.49	0.70

Figure 10: Model metrics (large values are better)

Our target was defined differently from related work which made direct comparison less meaningful. The closest approach used a presence or absence of any armed conflict related fatality as their binary target (Hannes Mueller and Rauh 2022). They focused on ROC_AUC as their metric and were able to achieve 0.83 on overall dataset and 0.75 on hard cases (sudden conflict in peaceful countries) with their text based (LDA topic extraction) model. Our best ROC_AUC score is 69% (achieved by text + country labels XGBoost/KNN model).

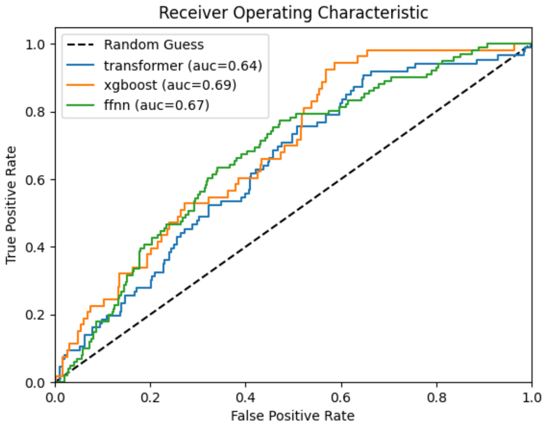

Figure 11: Receiver Operator Characteristics

We assume that UNICEF decision makers would appreciate a tool that flags all potential rising conflicts even if sometimes it has false positives. At the same time, if the false positive rate is too high the signal would become too noisy for the tool to be useful. Given these assumptions and the class imbalance of the original dataset, Recall (ratio of correctly identified conflict spikes to the total number of true spikes) and Cohen Kappa Score (overall agreement between predictions and true labels adjusted for class imbalance) seem to be most relevant metrics. The relative ranks of the models are the opposite for these two metrics with each of the two models with additional country data ranking best on one and worst on the other and the text-only model always being in the middle.

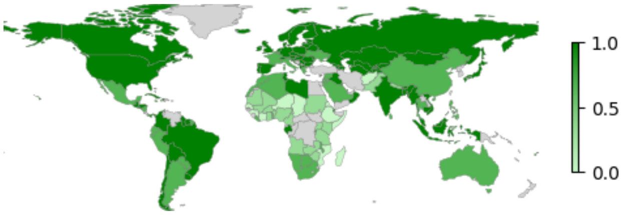

Figure 12: FFNN model accuracy on test data

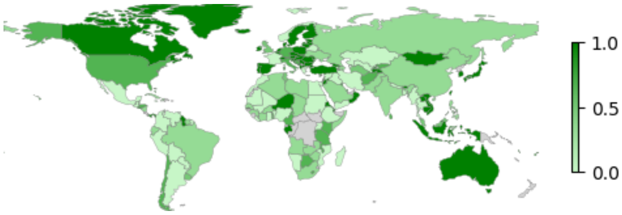

Figure 13: XGBoost model accuracy on test data

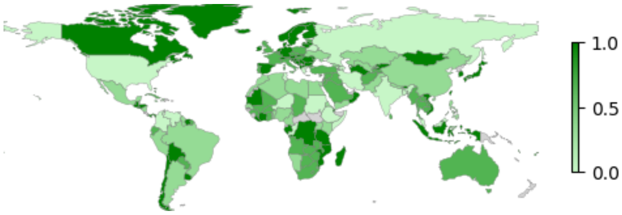

Figure 14: Transformer model accuracy on test data

Further investigation showed that while overall numeric metrics look similar, different models appeared to perform better in different regions of the world—see Figure 12, Figure 13, and Figure 14. While performance is often the main objective to optimize for when developing prediction models, our practical goal was to develop a model that can assist UNICEF in their decision making. With that in mind, cost of deployment is also an important factor for model evaluation. Average embeddings dataset and the corresponding statistical model fit in memory while working with article level embeddings in two other proposed methods requires additional data engineering steps to be processed in batches.

We think that all three approaches produced promising results and the model choice depends on the user's goals, selected performance metrics and available resources.

Ethical Considerations

This work carries significant ethical considerations, especially if it were to be operationalized. The first and most apparent is bias in the training data. All of our models showed varying performance by region, and over-reliance on such models by decision makers could lead to resources being diverted away from regions that need them, but which were underrepresented in the training data and thus had less accurate results.

If models like these were to become widely used, they may be vulnerable to adversarial attacks. Large volumes of fake news content could be used to overwhelm the true signal from news articles to obscure the intentions of malicious actors, or to cause panic.

Finally, and perhaps most concerning, is the potential for conflict prediction models to deliver self-fulfilling prophesies. If policy makers become convinced of impending conflict in their own or neighboring countries because of the predictions of an AI model, they may take actions that make conflict more likely.

Reducing these risks requires emphasizing that these are models, not oracles. They are measuring the temperature of the global media conversation, not gazing into the future. Decision makers need to understand that though models like these can be useful tools, they are not a substitute for careful deliberation and dialogue. At their best, these models can provide an early warning for conflict, and if the right actions are taken, then they can be made to be wrong.

Future work

We foresee several ways to extend this work, and there are exciting opportunities for both further research and operationalizations.

Future research directions

Future research could focus on scaling the methods outline. Due to time and resource constraints, our work used only 20% of the daily embeddings available through GDELT, and increasing the scale of training data could yield better results. We also believe that continued experimentation with data sampling and aggregation approaches, as well as model tuning, could produce models that outperform those outlined here without additional data.

Finally, we believe deeper research into regional disparities in model performance is needed. Understanding the limitations of global media coverage for conflict prediction will be critical to any effort to operationalize this work.

Future practical directions

Building an application with a user interface could be a natural extension of this research. We envision a dashboard where a user can view a world map where countries are colored according to the estimated probability of escalating conflict. Countries with highest risk can be flagged or shown to a user in a separate view. In addition to the visual part of this application, model deployment for easy inference is an important project.

Building a more robust and automated data processing pipeline is another important direction. This would involve designing either a batch or streaming system that can reliably collect data, preprocess and aggregate it into required format, train the model and make predictions. The system should handle computational resources and cost constraints.

Conclusions

In this work we explored the possibility of predicting future geopolitical conflict escalation at a country level using a global collection of news publications. We compared the usage of text-only data and data enriched with country labels and social indicators. We experimented with different ways of aggregating the data and applied both statistical machine learning and deep learning models. The results suggest that text embeddings provide a strong enough signal to make such predictions. Different data sampling and modeling approaches produced comparable results and each of the proposed models performed best at some of the common numeric metrics. The neural network-based models appear to generalize better across geographies while the statistical model provided a size/cost of deployment advantage.

While this work did not produce a single best model for conflict escalation predictions, it demonstrated the possibility of such predictions from primarily text data. We provided a clear definition of a binary conflict escalation target and described several approaches for data aggregation and modeling. We hope this work will contribute to both UNICEF operations and political science research. The performance results can serve as a benchmark for future research projects in this field.

References

"Announcing the Global Similarity Graph." 2021. The GDELT Project. https://blog.gdeltproject.org/announcing-the-global-similarity-graph/.

Cer, Daniel, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, et al. 2018. "Universal Sentence Encoder." https://arxiv.org/abs/1803.11175.

Chen, Peng, Adam Jatowt, and Masatoshi Yoshikawa. 2020. "Conflict or Cooperation?: Predicting Future Tendency of International Relations." In *SAC '20: The 35th ACM/SIGAPP Symposium on Applied Computing, Online Event, [Brno, Czech Republic], March 30 - April 3, 2020*, edited by Chih-Cheng Hung, Tomás Cerný, Dongwan Shin, and Alessio Bechini, 923–30. ACM. https://doi.org/10.1145/3341105.3373929.

Fantom, Umar, Neil; Serajuddin. 2016. "The World Bank's Classification of Countries by Income." *Policy Research Working Paper;No. 7528. © World Bank*. http://hdl.handle.net/10986/23628.

Hegre, Håvard, Paola Vesco, and Michael Colaresi. 2022. "Lessons from an Escalation Prediction Competition." *International Interactions* 48 (4): 521–54. https://doi.org/10.1080/03050629.2022.2070745.

Kwak, Haewoon, and Jisun An. 2021. "Two Tales of the World: Comparison of Widely Used World News Datasets GDELT and EventRegistry." *Proceedings of the International AAAI Conference on Web and Social Media* 10 (1): 619–22. https://doi.org/10.1609/icwsm.v10i1.14763.

Leetaru, Kalev, and Philip A Schrodt. 2013. "Gdelt: Global Data on Events, Location, and Tone, 1979–2012." In *ISA Annual Convention*, 2:1–49. 4. Citeseer.

Mueller, Hannes, and Christopher Rauh. 2022. "The Hard Problem of Prediction for Conflict Prevention." *Journal of the European Economic Association* 20 (6): 2440–67. https://doi.org/10.1093/jeea/jvac025.

Mueller, H., and C. Rauh. 2022. "Using Past Violence and Current News to Predict Changes in Violence." Cambridge Working Papers in Economics 2220. Faculty of Economics, University of Cambridge. https://ideas.repec.org/p/cam/camdae/2220.html.

Nate Haken, Juliette Gallo-Carelli, Daniel Woodburn. 2022. "Fragile States Index 2022 - Annual Report." *Fund For Peace*. https://fragilestatesindex.org/2022/07/13/fragile-states-index-2022-annual-report/.

Nations Development Programme), UNDP (United. 2022. "Human Development Report 2021-22." *UNDP (United Nations Development Programme)*. http://report.hdr.undp.org.

O'Brien, Sean P. 2010. "Crisis Early Warning and Decision Support: Contemporary Approaches and Thoughts on Future Research." *International Studies Review* 12 (1): 87–104. http://www.jstor.org/stable/40730711.

Raleigh, Clionadh, rew Linke, Håvard Hegre, and Joakim Karlsen. 2010. "Introducing ACLED: An Armed Conflict Location and Event Dataset." *Journal of Peace Research* 47 (5): 651–60. https://doi.org/10.1177/0022343310378914.

Sainburg, Tim, Leland McInnes, and Timothy Q Gentner. 2021. "Parametric UMAP Embeddings for Representation and Semisupervised Learning." *Neural Computation* 33 (11): 2881–2907.

UNICEF. 2021. "The State of the World's Children 2021." *United Nations Children's Fund*. https://www.unicef.org/reports/state-worlds-children-2021.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. "Attention Is All You Need." https://arxiv.org/abs/1706.03762.

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): 363–75. https://doi.org/10.1177/0022343309356491.

Wen, Qingsong, Tian Zhou, Chaoli Zhang, Weiqi Chen, Ziqing Ma, Junchi Yan, and Liang Sun. 2023. "Transformers in Time

Series: A Survey." https://arxiv.org/abs/2202.07125.