Integrating XAI for Predictive Conflict Analytics*

Luca Macis[1,*], Marco Tagliapietra[1,*], Alessandro Castelnovo[2], Daniele Regoli[2], Greta Greco[2], Andrea Claudio Cosentini[2], Paola Pisano[1] and Edoardo Carroccetto[3]

[1]*Department of Economics and Statistics, University of Turin, Via Lungo Dora Siena 100, 10153 Torino, Italy*
[2]*Intesa Sanpaolo S.p.A., C.so Inghilterra 3, 10138 Torino, Italy*
[3]*University of Turin, Via Giuseppe Verdi 8, 10124 Torino, Italy*

Abstract

Predicting global conflicts through data-driven approaches has the potential to aid political decision-makers in formulating more effective and targeted policies. However, high-performance models that derive patterns from data often become highly complex, making it challenging to extract understandable rationales behind their outcomes. In this paper, we suggest integrating a transformer-based Artificial Intelligence Early Warning System (AI-EWS) with integrated gradients, an eXplainable Artificial Intelligence (XAI) technique attributing model predictions to specific features at a given time in the input data, thereby enhancing interpretability. To validate our methodology, we conduct experiments on a prominent geopolitical dataset: ACLED. This dataset provides comprehensive insights into global conflict events, facilitating effective pattern learning and generalization by our model. Leveraging these explainability techniques, our goal is to bridge the gap between complex, high-performance models and the practical needs of policymakers in conflict prevention and resolution. Predictive analytics algorithms in conjunction with an XAI approach can foresee the impact of decisions on various population segments, fostering equity, and inclusion and supporting a data-driven approach, along with a culture of openness and accountability within the public administration.

Keywords

eXplainable Artificial Intelligence, Transformers, Time Series Forecasting, Integrated Gradients, Conflict Prediction, Early Warning System, Public Policy

1. Introduction

Predicting potential conflicts has played a crucial role in the landscape of peace research since Singer's work in the early 70's [1]. This historical backdrop sets the stage for understanding the evolution of conflict forecasting methodologies, encompassing diverse approaches including algorithms for event data coding [2]. Ward and co-authors marked a significant turning point, bringing prediction methodologies into the mainstream of peace research [3]. Subsequently,

Late-breaking work, Demos and Doctoral Consortium, colocated with The 2nd World Conference on eXplainable Artificial Intelligence: July 17–19, 2024, Valletta, Malta

*Corresponding author.

luca.macis@unito.it (L. Macis); marco.tagliapietra@unito.it (M. Tagliapietra); alessandro.castelnovo@intesasanpaolo.com (A. Castelnovo); daniele.regoli@intesasanpaolo.com (D. Regoli); greta.greco@intesasanpaolo.com (G. Greco); andrea.cosentini@intesasanpaolo.com (A. C. Cosentini); paola.pisano@unito.it (P. Pisano); edoardo.carroccetto@edu.unito.it (E. Carroccetto)
0000-0003-0001-304X (L. Macis); 0009-0009-3099-1400 (M. Tagliapietra); 0000-0001-5234-1155 (A. Castelnovo); 0000-0003-2711-8343 (D. Regoli); 0000-0001-6328-1303 (G. Greco); 0009-0000-6643-3894 (A. C. Cosentini); 0000-0003-3611-1346 (P. Pisano); 0009-0005-3620-531X (E. Carroccetto)

various organizations contribute to the field through the development of comprehensive data analysis and interpretation systems. The most current and prominent example is the Violence and Impacts Early-Warning System (VIEWS) [4] developed by Uppsala University; on our end in collaboration with the Italian Ministry of Foreign Affairs we implemented a new AI-EWS employing transformer models, built upon a multi-headed attention mechanism [5]. However, the usage of such sophisticated techniques, if on one side brings the benefit of increased accuracy of conflict predictions, on the other it opens new challenges to be faced. One of such challenges lies in the inherent complexity of such models: if predictions of conflicts are not accompanied by detailed explanations of the choices that led the model to make those predictions, policymakers are unlikely to trust such models to build robust policies and effective actions. Against this background, we introduce XAI approaches — in particular, those based on integrated gradients [6] — to enhance the transparency and comprehension of the AI-EWS's outcomes. In constructing our dataset to predict conflict, we opted for a publicly available disaggregated dataset that is regularly updated: the Armed Conflict Location and Event Data Project (**ACLED**), that collects real-time data on the locations, dates, actors, fatalities, and types of all reported political violence and protest events around the world [7].

2. Related Work

The field of conflict prediction has explored the usage of various machine learning models, including Random Forest [8], naive Bayes classifiers [9] and Neural Networks [10]. Notably, in the realm of time series forecasting, researchers have recently applied transformer architectures to univariate time series forecasting tasks. For instance, Li and co-authors solution showcased superior performance compared to classical statistical methods like ARIMA, as well as recent approaches such as TRMF, DeepAR, and DeepState, on four public forecasting datasets [11]. In our work we extend the application of transformers for multivariate time-series tasks as done by Zerveas and co-authors [12], where they use only the encoder part of the original transformer architecture. We use the same approach with some modification: in particular, we decided to include residual connections between input and output, ensuring that a purely linear model is always a subclass of our model [13]. The reason is that a simple linear models surprisingly may outperform existing sophisticated transformer-based models for long time-series forecasting problems [14]. Due to the intricate nature of the model, it's necessary to exploit XAI approaches to provide trustworthy explanations of its output. Regarding the use of integrated gradients within transformers, a self-attention attribution method was proposed and demonstrated on BERT [15]. Integrated hessians, an extension of integrated gradients, explain pairwise feature interactions in DistilBERT and demonstrating its effectiveness in sentiment analysis [16]. Following this trend, a recent work focuses on applying model-agnostic XAI techniques [17], such as SHAP [18] and LIME [19], to interpret predictions from transformer-based models in mental healthcare monitoring on social networks. The study underscores the social and public importance of explainability for the adoption of AI-based diagnostic systems.

3. Methodology

3.1. Geopolitical data collection and preparation

The value of data selection in defining a data-driven conflict prediction model's performance is recognized. Such models usually are dependent on either social media or diplomatic datasets. Social media datasets, specifically Twitter, were historically utilized due to their convenience and utility in examining factors influencing civil unrest [20, 21]. However, owing to restrictions on violent content and monitoring by authoritarian regimes, their efficacy has been mitigated [22]. Consequently, our study leans towards diplomatic datasets. Our chosen dataset, ACLED, is disaggregated, regularly updated, and emphasizes disorder events. ACLED data has proven valuable in predicting conflict [23], and is publicly accessible via their API[1]. The data chronicles various conflict events with distinct descriptions, location, and time, and ensures reliability through a rigorous verification process. Although coverage periods vary across nations, the detailed structure and transparency foster academic research and informed decision-making. Our research aggregates these data weekly and categorizes them by event types, resulting in a dataset where each observation corresponds to the number of a specific event type within that week in a particular country.

Table 1
Example of the aggregated ACLED dataset. The last column, *Fatalities*, represents the target variable obtained by summing all fatalities recorded in that week.

Country	Date	Armed clash	Air/drone strike	...	*Fatalities*
Afghanistan	2016-12-31	155	13	...	666
Afghanistan	2017-01-07	140	10	...	546

3.2. Transformer Model

Our AI-EWS employs a transformer model that focuses on predicting the number of fatalities across all countries over twelve weeks. Inspired majorly by the Time-series Dense Encoder (TiDE) model prominent in the domain of long-term forecasting [13], our design substitutes the dense encoder conventionally used in TiDE with an attention-based encoder, due to better results with the dataset in use in our study. The model, as in the original TiDE implementation, incorporates residual connections from input to output ensuring the preservation of linear activation, an approach backed by empirical evidence for its efficiency in time-series forecasting [14]. For a comprehensive understanding of the model's operation and data flow, please refer to the detailed explanation provided under Figure 1. Overall, the model's design is geared towards robust long-term prediction while maintaining a fundamental simplicity in its architecture, balancing advanced modeling techniques with practical forecasting reliability. The model is trained using the Negative Log Likelihood (NLL) loss function to optimize its probabilistic forecasts. Additionally, it's worth noting that for each country, 12 weeks were retained for testing, 24 for validation, and the remaining weeks were allocated for training.

[1] Armed Conflict Location and Event Data Project (ACLED); https://acleddata.com/

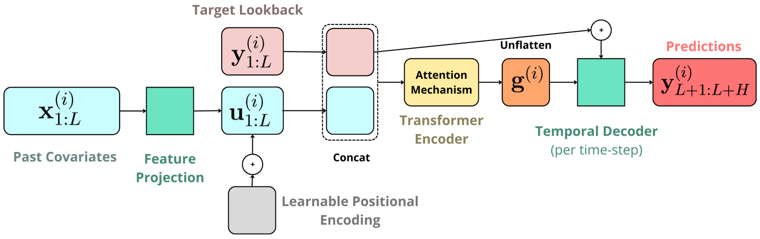

Figure 1: Model Architecture. The model takes as input independent features $x_{1:L}^{(i)}$ and the corresponding target feature, the number of fatalities $y_{1:L}^{(i)}$, both spanning L past time-steps. Initially, the dimensionality of $x_{1:L}^{(i)}$ is reduced using a residual block with linear activation to maintain foundational linearity [13], resulting in a reduced vector $u_{1:L}^{(i)}$. Subsequently, positional encoding is applied. The concatenated vectors $u_{1:L}^{(i)}$ and $y_{1:L}^{(i)}$ are then passed through the attention mechanism of the encoder. The output is unflattened and fed into the temporal decoder, which expects input of shape $dec\_dim \times horizons$, where dec_dim represents the dimension of the decoder, and $horizons$ denotes the number of future time-steps to predict. Predictions are made per time-step, processing one block of $dec\_dim \times 1$ at a time. Additionally, a residual connection is maintained from the original vector $y_{1:L}^{(i)}$ to the temporal decoder to preserve simple linear models. Hyperparameters are shown in the table below.

feat_proj	transf_enc_dim	head_size	num_heads	num_transf_enc	dec_dim	dropout
16	16	16	3	1	256	0.1

3.3. Integrated Gradients

Integrated Gradients (IG) is a technique utilized for attributing predictions of deep neural networks to their input features, facilitating a deeper understanding of their behavior [6]. Addressing the challenge of empirical evaluation inherent in attribution techniques, IG adopt an axiomatic approach. Two fundamental axioms guide attribution methods:

Sensitivity: This axiom dictates that for inputs and baselines differing in a single feature yet yielding different predictions, the differing feature must receive a non-zero attribution.

Implementation Invariance: Attributions should remain consistent for functionally equivalent networks. Networks are functionally equivalent if their outputs coincide for all inputs, despite potential differences in their implementations. Failure to satisfy this axiom may indicate sensitivity to insignificant model aspects. IG computes attributions by following a straight-line path from a baseline input x' to the input x, evaluating gradients along this path. Specifically, IG along the i^{th} dimension for inputs x and x' are calculated as:

$$IG_i(x) = (x_i - x_i') \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i} \, \mathrm{d}\alpha, \tag{1}$$

where $\frac{\partial F(x)}{\partial x_i}$ represents the gradient of $F(x)$ along the i^{th} dimension.
Furthermore, IG adheres to an additional axiom:

Completeness: Attributions sum up to the discrepancy between the output of F at input x and baseline x'. This axiom serves as a sanity check, ensuring the method comprehensively accounts for differences.

In our study, the baseline was established as the mean matrix, a critical decision considering the MinMax scaling applied to our dataset. Notably, employing a baseline filled with mean values for each rescaled feature can assign significance to count features with zero values, especially in a dataset subjected to MinMax scaling. This decision was based on the assumption that the absence of unrest events might hold relevance for the model's prediction. Therefore, we opted for this baseline matrix rather than a zero baseline matrix. The choice of IG is motivated by its transparency, simplifying the comprehension of attributions to input features. While SHAP and LIME, as leading and commonly used methods in XAI, provide in-depth explorations of model behaviors, IG's clear computational approach provides an easier understanding, making it an effective preliminary step before advancing to more complex explanatory methods.

4. Results

This study sets out to forecast the potential number of fatalities from February 13, 2024, to March 30, 2024[2], focusing on 168 countries that have recorded at least one fatality throughout their historical time series[3].

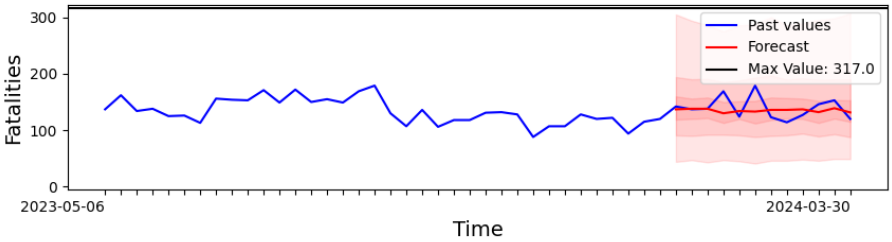

Figure 2: Forecast of Fatalities in Mexico. The y-axis represents the number of fatalities, while the x-axis spans the weeks. The blue line depicts the actual number of fatalities observed over time, providing a reference for the model's performance. The red line corresponds to the median predictions generated by the model. Additionally, the red bands surrounding the median predictions represent a probability band covering the range from the 5th to the 95th percentile.

The primary aim of this research is not to benchmark the prediction accuracy against other leading models but to explore the insights provided by the predictions of AI-EWS. Our investigation is centered around the application of integrated gradients, the chosen XAI methodology, to reveal the reasons behind these forecasts. The analysis initiates by pinpointing the crucial variables influencing the forecasting during the testing period. As depicted in Figure 3, these variables are visualized using boxplots and are arranged in descending order of their influence. For clarity, take the instance of a specific country: a value of 1 marks the highest absolute shift in Integrated Gradients, showcasing that a variable is critically influential in making predictions; a value of 0 suggests no shift, indicating the variable's non-involvement in the prediction model; any value between 0 and 1 highlights the variable's proportional relevance compared to the most impactful variable in that country.

[2] This timeframe spans the most recent twelve weeks of ACLED data available for each country up to April 9, 2024.

[3] This criterion is crucial as predicting deviations from zero fatalities where no prior occurrences exist is statistically improbable. Therefore, countries are assessed individually based on their historical data.

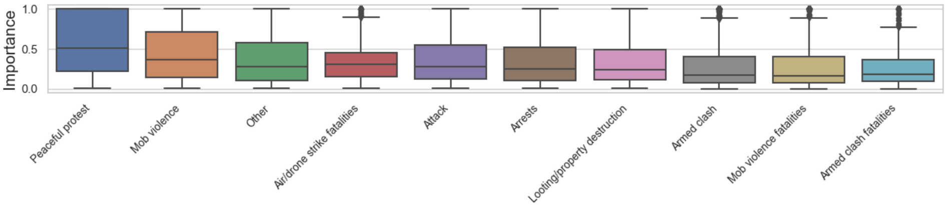

Figure 3: Variable importance across all countries. Only the top 10 most important variables are shown.

Additionally, the relevance of different time intervals within the forecasting model is scrutinized. The model encompasses a lookback period of 48 weeks to integrate the data leading up to a prediction. Figure 4 clarifies the weighting assigned to each subsequent week, arranged chronologically, which assists in understanding the adaptive significance throughout the considered period.

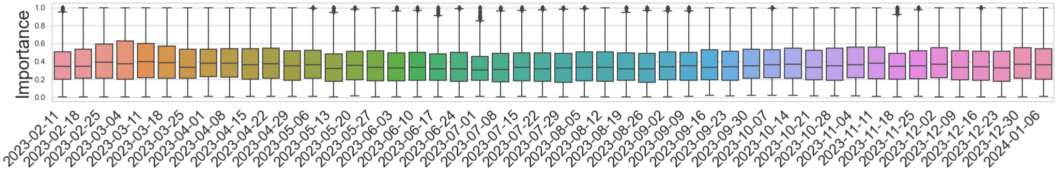

Figure 4: Capturing the significance of time via Integrated Gradients across all evaluated countries.

To summarize, initial findings illuminate prominent patterns concerning the importance of variables and the dynamics of time intervals within the prediction framework. As elucidated in Figure 3, certain variables evidently carry more weight consistently across all analyzed countries. However, a review of Figure 4 portrays a more complex landscape. Although there exists a mild preference for recent weeks, variable importance demonstrates relative uniformity regardless of the elapsed time since the event. This reflection reveals a coherent strategy by the AI-EWS to value variables uniformly, irrespective of their temporal proximity to the predicted event.

Conclusions

In this study, we proposed a novel approach to conflict prediction on a global scale, leveraging advanced transformer models and XAI methodologies. We applied our approach to a comprehensive geopolitical dataset implemented using data obtained from the ACLED API. The transformer model proposed is inspired by its original architecture [5] and incorporates insights from the TiDE model [13]. The AI-EWS goal is to forecast the number of fatalities over a 12-week horizon. This metric provides an immediate and objective index for gauging a country's unrest situation. Integrated gradients were employed as the XAI methodology to enhance interpretability, offering significant insights into how specific features impact the model's predictions and the temporal influence dynamics. Our analysis of the conflict dataset unveiled several key insights. We observed that certain features consistently hold importance across different countries. However, a detailed examination into the importance attributed to varying time frames indicates a subtle preference for recent data, suggesting the AI-EWS maintains consistent variable prioritization regardless of temporal proximity. As we move forward, IG may serve as a foundational tool, enabling clear initial explanations that pave the way for engaging with more advanced XAI techniques in future research while mitigating the complexities often encountered with newer methods. In future studies, to evaluate the comprehensibility of our feature rankings for users such as policymakers, we plan two key activities: *user studies*, for collecting feedback through surveys and interviews to assess their understanding of the model's feature rankings; *usability testing*, where users make decisions based on the model's outputs, evaluating how effectively they can utilize the provided feature rankings. The findings provide valuable insights into the interpretability and performance of advanced machine learning techniques in addressing high-stakes global challenges. As the public sector increasingly relies on AI for decision making, there will be a growing need for mechanisms that can explain AI decisions in a transparent and understandable way. In summary, XAI can make a significant contribution to more responsive and accountable public services. Not only it can deliver accessible and meaningful explanations to non-expert audiences, including the general public and policymakers, but it can also guarantee greater compliance with regulatory evolutions and principles such as fairness, accountability and privacy.

Acknowledgments

This study was funded by the European Union (EU) - NextGenerationEU, in the framework of the GRINS - Growing Resilient, INclusive and Sustainable project (GRINS PE00000018 – CUP D13C22002160001). The views/opinions expressed are solely those of the authors and do not necessarily reflect those of the EU, nor can the EU be held responsible for them.

References

[1] J. D. Singer, The peace researcher and foreign policy prediction, Peace Science Society (International) 21 (1973) 1–13.

[2] P. A. Schrodt, S. G. Davis, J. L. Weddle, Political science: Keds—a program for the machine coding of event data, Social Science Computer Review 12 (1994) 561–587.

[3] M. D. Ward, B. D. Greenhill, K. M. Bakke, The perils of policy by p-value: Predicting civil conflicts, Journal of peace research 47 (2010) 363–375.

[4] H. Hegre, M. Allansson, M. Basedau, M. Colaresi, M. Croicu, H. Fjelde, F. Hoyles, L. Hultman, S. Högbladh, R. Jansen, et al., Views: A political violence early-warning system, Journal of peace research 56 (2019) 155–174.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017.

[6] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, CoRR abs/1703.01365 (2017). arXiv:1703.01365.

[7] C. Raleigh, R. Kishi, A. Linke, Political instability patterns are obscured by conflict dataset scope conditions, sources, and coding choices, Humanities and Social Sciences Communications 10 (2023). doi:10.1057/s41599-023-01559-4.

[8] D. Muchlinski, D. Siroky, J. He, M. Kocher, Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data, Political Analysis 24 (2016) 87–103. doi:10.1093/pan/mpv024.

[9] C. Perry, Machine learning and conflict prediction: A use case, Stability: International Journal of Security & Development 2 (2013) 56. doi:10.5334/sta.cr.

[10] N. Beck, G. King, L. Zeng, Improving quantitative studies of international conflict: A conjecture, The American Political Science Review 94 (2000) 21. doi:10.2307/2586378.

[11] S. Li, X. Jin, Y. Xuan, X. Zhou, W. Chen, Y.-X. Wang, X. Yan, Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 32, Curran Associates, Inc., 2019.

[12] G. Zerveas, S. Jayaraman, D. Patel, A. Bhamidipaty, C. Eickhoff, A transformer-based framework for multivariate time series representation learning, in: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, KDD '21, Association for Computing Machinery, 2021, p. 2114–2124. doi:10.1145/3447548.3467401.

[13] A. Das, W. Kong, A. Leach, S. Mathur, R. Sen, R. Yu, Long-term forecasting with tide: Time-series dense encoder, 2023. arXiv:2304.08424.

[14] A. Zeng, M. Chen, L. Zhang, Q. Xu, Are transformers effective for time series forecasting?, AAAI Conference on Artificial Intelligence 37 (2023) 11121–11128. doi:10.1609/aaai.v37i9.26317.

[15] Y. Hao, L. Dong, F. Wei, K. Xu, Self-attention attribution: Interpreting information interactions inside transformer, 2021. arXiv:2004.11207.

[16] J. D. Janizek, P. Sturmfels, S.-I. Lee, Explaining explanations: Axiomatic feature interactions for deep networks, The Journal of Machine Learning Research 22 (2021) 4687–4740.

[17] A. Malhotra, R. Jindal, Xai transformer based approach for interpreting depressed and suicidal user behavior on online social networks, Cognitive Systems Research (2023) 101186. doi:10.1016/j.cogsys.2023.101186.

[18] S. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, 2017. arXiv:1705.07874.

[19] M. T. Ribeiro, S. Singh, C. Guestrin, "why should i trust you?": Explaining the predictions of any classifier, 2016. arXiv:1602.04938.

[20] G. Korkmaz, J. Cadena, C. J. Kuhlman, A. Marathe, A. Vullikanti, N. Ramakrishnan, Combining heterogeneous data sources for civil unrest forecasting, in: Proceedings of the 2015 IEEE/ACM Int. Conf. on Advances in Social Networks Analysis and Mining 2015, ASONAM '15, ACM, 2015, p. 258–265. doi:10.1145/2808797.2808847.

[21] R. Compton, C. Lee, J. Xu, A. M. Luis, T.-C. Lu, Using publicly visible social media to build detailed forecasts of civil unrest, 2014. doi:10.1186/s13388-014-0004-6.

[22] M. Junior, P. Melo, A. P. C. da Silva, F. Benevenuto, J. Almeida, Towards understanding the use of telegram by political groups in brazil, in: Brazilian Symposium on Multimedia and the Web, WebMedia '21, ACM, 2021, p. 237–244. doi:10.1145/3470482.3479640.

[23] M. Halkia, S. Ferri, M. K. Schellens, M. Papazoglou, D. Thomakos, The global conflict risk index: A quantitative tool for policy support on conflict prevention, Progress in Disaster Science 6 (2020) 100069. doi:https://doi.org/10.1016/j.pdisas.2020.100069.