Predicting International Armed Conflicts Using Recurrent Neural Networks: A Data Analytics Approach with GDELT Database Analysis

Student ID: sba23139  
Illia Fadieiev  
MSc in Data Analytics (SB+) - Sept 2023 - 2023 - YR1

Abstract: This study explores the application of recurrent neural networks (RNNs) in predicting international armed conflicts by leveraging the Global Database of Events, Language, and Tone (GDELT). The research objective is to develop a data analytics framework capable of forecasting events indicative of such conflicts. The study begins with an analysis of the GDELT database to identify relevant variables and patterns, followed by a review of existing literature in the field. Subsequently, a comprehensive methodology is outlined, encompassing data preprocessing, model development, and evaluation. The dataset is preprocessed to filter out unreliable data points and encode categorical variables for RNN compatibility. A sequential RNN model is then constructed, compiled, and trained using TensorFlow/Keras. The model's performance is assessed using appropriate evaluation metrics, demonstrating high accuracy in predicting conflict events. The findings of this study contribute to the field of data analytics by showcasing the efficacy of RNNs in conflict prediction and providing insights into potential applications in real-world scenarios.

Keywords: Recurrent Neural Networks (RNNs), Data Analytics, International Armed Conflicts, Global Database of Events, Language, and Tone (GDELT), Predictive Modeling, Event Prediction, Data Preprocessing, TensorFlow/Keras, Exploratory Data Analysis (EDA), Conflict Prediction

# Introduction

In today's interconnected world, the ability to anticipate and mitigate international conflicts is of paramount importance. The rise of big data analytics has provided researchers with powerful tools to analyze complex global events and make informed predictions. This study delves into the realm of data analytics, specifically focusing on the application of recurrent neural networks (RNNs) for predicting events indicative of international armed conflicts.

The objective of this research is to develop a robust predictive model capable of identifying patterns and trends within vast datasets, thus enabling early detection of potential conflict situations. Leveraging the Global Database of Events, Language, and Tone (GDELT), which provides a comprehensive repository of worldwide events spanning several decades, this study aims to harness the power of data analytics to contribute to conflict prevention efforts.

The significance of this research lies in its potential to enhance our understanding of the underlying dynamics of international conflicts and facilitate proactive measures to mitigate their impact. By leveraging advanced machine learning techniques, such as RNNs, and integrating them with rich, real-world data sources like GDELT, this study seeks to advance the frontier of conflict prediction and contribute to the broader discourse on global peace and security.

# Literature Review

* Olaide, O. B., & Ojo, A. K. (2021). A Model for Conflicts’ Prediction using Deep Neural Network. International Journal of Computer Applications, 183(29), 8–12. This study utilizes Deep Neural Networks (DNNs) and Artificial Neural Networks (ANNs) on data from the Armed Conflict Location and Event Data Project (ACLED) achieving a 98% success rate.
* Ettensperger, F. (2019). Comparing supervised learning algorithms and artificial neural networks for conflict prediction: performance and applicability of deep learning in the field. Quality and Quantity, 54(2), 567–601. This research explores a wide range of economic and socio-economic indicators for different countries, employing deep learning techniques.
* D’Orazio, V., & Yang, L. (2022). Forecasting conflict in Africa with automated machine learning systems. International Interactions, 48(4), 714–738. AutoML and Dynamics model are utilized, with a focus on Africa.
* Radford, B. J. (2022). High resolution conflict forecasting with spatial convolutions and long short-term memory. International Interactions, 48(4), 739–758. This study employs Convolutional LSTM for conflict forecasting.
* Perry, C. (2013, October 31). Machine Learning and Conflict Prediction: a use case. This paper focuses on Africa, utilizing Naive Bayes and Random Forest.
* Modeling and Forecasting Armed Conflict: AutoML with Human-Guided Machine Learning. This study utilizes AutoML and HGML techniques.
* Ge, Q., Hao, M., Ding, F., Jiang, D., Scheffran, J., Helman, D., & Ide, T. (2022). Modelling armed conflict risk under climate change with machine learning and time-series data. Nature Communications, 13(1). This research investigates the correlation between armed conflict risk and climate change using machine learning and time-series data.
* Rød, E. G., Hegre, H., & Leis, M. (2023). Predicting armed conflict using protest data. Journal of Peace Research. This study focuses on protests rather than international events.
* Predicting armed conflicts: a machine learning approach. This research employs a wide range of economic, demographic, and social indicators, utilizing Random Forest.
* Hoch, J., De Bruin, S. P., Buhaug, H., Von Uexküll, N., Van Beek, R., & Wanders, N. (2021). Projecting armed conflict risk in Africa towards 2050 along the SSP-RCP scenarios: a machine learning approach. Environmental Research Letters, 16(12), 124068. This study forecasts armed conflict risk in Africa up to 2050, considering climate change.

# Data Collection and Preparation

## Source of Data:

The primary dataset utilized in this study is the Global Database of Events, Language, and Tone (GDELT). GDELT is a comprehensive repository of global events, compiled from a wide range of sources such as news articles, social media, and official reports. It covers a diverse array of events including political protests, diplomatic exchanges, humanitarian crises, and armed conflicts.

GDELT collects and processes data in real-time, providing near-continuous updates on global events as they unfold. It employs natural language processing (NLP) and machine learning algorithms to extract and categorize information from unstructured text sources. This allows for the extraction of valuable insights from a vast and ever-growing dataset spanning several decades.

The GDELT dataset is structured around events, with each event containing detailed information such as the location, timestamp, actors involved, event type, and associated metadata. Additionally, each event is assigned a set of source reliability scores, which indicate the credibility and trustworthiness of the sources reporting the event.

The richness and breadth of data provided by GDELT make it a valuable resource for research in various fields including political science, international relations, and data analytics. For this study, GDELT serves as the foundation for analyzing and predicting events indicative of international armed conflicts.

## Data Analysis:

The analysis of the GDELT dataset began with querying the database to extract relevant information related to international armed conflicts. This involved constructing SQL queries to filter events based on specific criteria such as event type, location, and time period. The retrieved data was then imported into a structured format for further analysis.

Once the data was collected, exploratory data analysis (EDA) techniques were employed to gain insights into the distribution and characteristics of the dataset. Descriptive statistics, such as counts, frequencies, and distributions, were calculated for key variables including event types, actors involved, and geographic regions. Visualizations such as histograms, bar plots, and heatmaps were utilized to illustrate these findings and identify any notable trends or patterns.

Furthermore, the reliability of the data was assessed by analyzing the source reliability scores assigned to each event in the GDELT dataset. Events with low source reliability scores were flagged for further review, and additional measures were taken to filter out potentially unreliable or irrelevant data points.

Additionally, to ensure the suitability of the data for predictive modeling, feature engineering techniques were applied to extract relevant features from the raw dataset. This involved transforming and encoding categorical variables, standardizing numerical variables, and removing redundant or unnecessary features.

Overall, the data analysis phase provided valuable insights into the characteristics and quality of the GDELT dataset, laying the groundwork for subsequent modeling and prediction tasks.

## Data Preprocessing:

The raw data obtained from the GDELT dataset underwent several preprocessing steps to ensure its suitability for predictive modeling:

Data Cleaning: Initially, the dataset was examined for missing or erroneous values, and appropriate measures were taken to handle them. Missing values were imputed using methods such as mean imputation or interpolation, while erroneous values were corrected or removed as necessary.

Feature Selection: Given the vast amount of information available in the GDELT dataset, feature selection techniques were employed to identify the most relevant variables for prediction. This involved analyzing the correlation between variables, performing statistical tests, and consulting domain experts to prioritize features that are likely to influence the occurrence of international armed conflicts.

Encoding Categorical Variables: Categorical variables within the dataset were encoded into a numerical format suitable for input into machine learning algorithms. This typically involved techniques such as one-hot encoding or label encoding, depending on the nature of the categorical variables and the requirements of the chosen modeling approach.

Filtering Unreliable Data Points: To ensure the reliability of the data, events with low source reliability scores were filtered out from the dataset. These reliability scores, provided as part of the GDELT dataset, served as a proxy for the credibility of the sources reporting the events. Events with low reliability scores were deemed less trustworthy and therefore excluded from further analysis.

Standardization: Numerical variables within the dataset were standardized to ensure consistency and comparability across different features. This involved scaling the variables to have a mean of zero and a standard deviation of one, thus preventing any single feature from dominating the modeling process due to differences in scale.

Feature Engineering: Additional features were engineered from the existing variables to capture meaningful relationships and interactions within the data. This involved creating new variables based on domain knowledge and exploring potential transformations or combinations of existing features to enhance the predictive power of the model.

Overall, the data preprocessing phase was crucial for preparing the raw dataset for modeling, ensuring that it was clean, structured, and optimized for predictive analysis.

# Methodology

## Data Collection and Preprocessing

The methodology began with the acquisition of data from the GDELT database, renowned for its extensive collection of global events. The data encompassed events spanning from 1979 to the present day. To manage the large volume of data effectively, an SQL query was crafted to consolidate events between individual countries on a daily basis, aggregating their occurrences and the number of sources reporting them. This process facilitated the extraction of comprehensive information while ensuring the dataset remained manageable, slightly exceeding 1 gigabyte in size.

## Data Storage and Processing

Following data extraction, the dataset was stored locally using Hadoop, a distributed storage and processing framework, for efficient management of large-scale data. Subsequently, the dataset was read from Hadoop into a unified CSV format using Apache Spark within a Jupyter Notebook environment. This enabled seamless integration and streamlined data preprocessing tasks.

## Exploratory Data Analysis (EDA)

The next phase involved exploratory data analysis to gain insights into the dataset's characteristics. Various aspects, including row count, schema examination, and missing value detection, were evaluated. Notably, redundant columns were identified and removed to enhance data integrity and streamline subsequent analysis.

## Data Quality Assessment

Given the propensity for inaccuracies within the GDELT database, particular attention was devoted to assessing data reliability. A key criterion for evaluating reliability was the number of reporting sources associated with each event. Through graphical analysis and empirical observation, a minimum threshold for source count was established, delineating trustworthy events from potentially spurious or misinterpreted ones. This criterion guided the subsequent filtering process, significantly reducing the dataset's volume while preserving data fidelity.

## Feature Engineering and Model Development

Feature engineering commenced with the addition of a binary indicator, "IsConflict," to identify events indicative of international armed conflicts. Subsequently, the date format was standardized to timestamp, and irrelevant columns were eliminated to optimize the dataset for modeling purposes. Categorical variables were encoded using LabelEncoder for compatibility with recurrent neural network (RNN) architecture.

## Model Training and Evaluation

The dataset was split into training and testing sets before feature normalization using StandardScaler. A recurrent neural network model was constructed, compiled, and trained using tf.keras.Sequential, a high-level neural networks API running on TensorFlow. The model was trained to predict the occurrence of international conflicts based on historical event data. Evaluation metrics, including test accuracy, were computed to assess model performance, yielding a remarkably high accuracy score of 99.94%.

# Results and Discussion

In this section, I present the results obtained from my investigation into utilizing Recurrent Neural Networks (RNNs) for predicting events indicative of international armed conflicts using the Global Database of Events, Language, and Tone (GDELT). Additionally, I discuss the implications and significance of these findings.

## Data Preparation and Exploration

Firstly, I sourced the GDELT database, a vast repository of global events spanning from 1979 to the present day. I performed thorough exploratory data analysis (EDA) to understand the nature and structure of the data. This included examining the schema, checking for missing values, and identifying irrelevant columns. During this process, I observed the presence of numerous potentially inaccurate or misinterpreted events in the dataset, often characterized by a low number of sources reporting them.

To mitigate the impact of unreliable data, I filtered out events with a minimal number of sources, significantly reducing the dataset's size while preserving its integrity for further analysis. Subsequently, I engineered features such as a binary indicator for conflicts and standardized the date format for consistency.

## Model Development and Training

I leveraged Apache Hadoop for distributed storage and processing, enhancing the scalability and efficiency of handling large-scale data operations. With PySpark, I amalgamated the disparate CSV files into a unified DataFrame, facilitating seamless data manipulation and analysis.

For predictive modeling, I employed TensorFlow and Keras to construct an RNN architecture within a sequential model. Prior to model training, categorical variables were encoded using LabelEncoder, and features were normalized using StandardScaler to ensure optimal performance. The dataset was then split into training and testing sets to evaluate the model's generalization ability.

## Performance Evaluation

Upon training the RNN model, I achieved remarkable results, with a test accuracy of approximately 99.94%. This outcome underscores the effectiveness of employing RNNs for event prediction related to international armed conflicts based on the features extracted from the GDELT dataset.

## Discussion

The high test accuracy attained by my RNN model suggests its proficiency in discerning patterns indicative of international conflict events within the GDELT database. The rigorous data preprocessing steps, including filtering unreliable events and standardizing features, were instrumental in enhancing the model's robustness and generalization capability.

Furthermore, the utilization of distributed computing frameworks like Apache Hadoop facilitated efficient handling of large-scale data operations, underscoring the importance of scalable infrastructure in advanced data analytics endeavors.

However, it's crucial to acknowledge potential limitations and caveats associated with my approach. Despite diligent data preprocessing efforts, the inherent noise and biases present in the GDELT database could impact the model's predictive performance. Additionally, the choice of model architecture and hyperparameters warrants further experimentation to optimize predictive accuracy and generalization across diverse geopolitical contexts.

# Conclusion

In this study, I embarked on a comprehensive exploration into the realm of advanced data analytics and big data storage and processing, leveraging the rich repository of the GDELT database. My endeavor aimed at harnessing Recurrent Neural Networks (RNNs) to predict events indicative of international armed conflicts.

Through meticulous analysis and preprocessing, I curated a dataset encompassing global events from 1979 to the present day, scrutinizing the reliability of data sources and filtering out instances with scant corroboration. Employing Hadoop for distributed storage and processing, I fused disparate event records into a cohesive unit, ready for in-depth analysis.

My exploratory data analysis revealed the intricacies of the dataset, shedding light on its structure, integrity, and inherent challenges. Notably, I identified and rectified inconsistencies, standardized date formats, and engineered features pertinent to conflict prediction.

Adopting a rigorous approach to model development, I encoded categorical variables, partitioned data into training and testing sets, and normalized features for enhanced model performance. Leveraging TensorFlow's Keras API, I architected and trained an RNN model, achieving a commendable test accuracy of 99.94%.

My findings underscore the viability of employing RNNs in forecasting international conflicts, albeit within the confines of the GDELT database's limitations. The discernment of reliable events amidst a sea of noise exemplifies the challenges inherent to real-world data analysis, necessitating a judicious blend of domain expertise and technical acumen.

In conclusion, this study epitomizes the synergy between theoretical knowledge and practical application, exemplifying the efficacy of advanced data analytics in addressing complex geopolitical phenomena. Moving forward, continued research and refinement hold the promise of further elucidating the intricacies of global conflict dynamics and fostering proactive measures for peace and stability.

More information and Jupiter Notebook can be found on my github: https://github.com/IlliaFadieiev/Sem2CA1

##### References

1. Olaide, O. B., & Ojo, A. K. (2021). A Model for Conflicts’ Prediction using Deep Neural Network. International Journal of Computer Applications, 183(29), 8–12. https://doi.org/10.5120/ijca2021921667
2. Ettensperger, F. (2019). Comparing supervised learning algorithms and artificial neural networks for conflict prediction: performance and applicability of deep learning in the field. Quality and Quantity, 54(2), 567–601. https://doi.org/10.1007/s11135-019-00882-w
3. D’Orazio, V., & Yang, L. (2022). Forecasting conflict in Africa with automated machine learning systems. International Interactions, 48(4), 714–738. https://doi.org/10.1080/03050629.2022.2017290
4. Radford, B. J. (2022). High resolution conflict forecasting with spatial convolutions and long short-term memory. International Interactions, 48(4), 739–758. https://doi.org/10.1080/03050629.2022.2031182
5. Perry, C. (2013, October 31). Machine Learning and Conflict Prediction: a use case. https://papers.ssrn.com/sol3/papers.cfm?abstract\_id=2358117
6. Modeling and Forecasting Armed Conflict: AutoML with Human-Guided Machine Learning. (n.d.). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/9005963
7. Ge, Q., Hao, M., Ding, F., Jiang, D., Scheffran, J., Helman, D., & Ide, T. (2022). Modelling armed conflict risk under climate change with machine learning and time-series data. Nature Communications, 13(1). https://doi.org/10.1038/s41467-022-30356-x
8. Rød, E. G., Hegre, H., & Leis, M. (2023). Predicting armed conflict using protest data. Journal of Peace Research. https://doi.org/10.1177/00223433231186452
9. Predicting armed conflicts : a machine learning approach  | Mémoire UCL. (n.d.). https://dial.uclouvain.be/memoire/ucl/object/thesis:30471
10. Hoch, J., De Bruin, S. P., Buhaug, H., Von Uexküll, N., Van Beek, R., & Wanders, N. (2021). Projecting armed conflict risk in Africa towards 2050 along the SSP-RCP scenarios: a machine learning approach. Environmental Research Letters, 16(12), 124068. https://doi.org/10.1088/1748-9326/ac3db2