Early Warning System for Unrest Prediction

Luca Macis [1] Marco Tagliapietra [1]

[1]University of Turin - Department of Economics and Statistics

{*luca.macis, marco.tagliapietra*} *@unito.it*

September 22, 2023

Early Warning System for Unrest Prediction

Purpose of the Tutorial

Primary Reasons:

Sharing the work we've accomplished so far. Summary of two years of efforts and research by show our approach.

Aid individuals

Help people who want to learn about these topics by making the learning process faster.

- Overview of Datasets
- Overview of Early Warning Systems

Conflict Prevention

Why Conflict Prevention?

Conflict prevention holds utmost importance in the realm of international relations and diplomacy. Timely intervention can often lead to the resolution of conflicts before they escalate into more severe and devastating occurrences.

Example

- Hegre et al. (2012) simulated a scenario where the United Nations ceased its peacekeeping efforts after 2001 and found that, compared to the actual number of countries with ongoing peacekeeping efforts, three to four additional countries would have experienced major conflicts by 2013 in the absence of UN's efforts.

Conflict and Unrest Prediction

Why Conflict Prediction?

In this context, conflict prediction and early warning systems play a crucial role by identifying potential risks and threats, offering decision-makers timely information to formulate policies for conflict mitigation and prevention.

The Aid of Data Science

the two possible ways of doing conflict and unrest prediction.

- Identification of potential conflicts and crisis relied heavily on individual diplomatic and political knowledge, intuition, and subjective judgment.
- Recognize the potential of current technology in the field of Data Science.

Choosing the Datasets

The model is a heavily data-driven model.

Type of Datasets

- **Social Datasets:** Twitter, Telegram, ecc...
- **Disaggregated Datasets:** ACLED, GDELT, ecc...
- **Aggregated Dataset:** World Bank, V-DEM, ecc..

Social Dataset

Some of the work on Social Datasets:

- Korkmaz et al. (2015) explores the use of multiple data sources to predict and analyze civil unrest in Latin America.
- Compton et al. (2014) presents a data mining system that generates forecasts of civil unrest incidents in Latin America using public posts on Twitter and Tumblr.

Some of the issues on Social Datasets:

- The effectiveness of Twitter for organizing insurrections has diminished due to prohibitions on violent tweets and the tracking of users by authoritarian regimes Junior et al. (2021).
- There are no notable applications in the field of conflict prediction using Telegram (Khaund et al. (2021))

Disaggregated Dataset

Some of the Diplomatic Datasets:

Armed Conflict Data Location and Event Data Project (ACLED), Uppsala Conflict Data Program (UCDP), Global Database of Events, Language and Tone (GDELT), ...

Some of the work on Diplomatic Dataset:

Armed Conflict Data Location and Event Data Project (ACLED),The Violence & Impacts Early-Warning System (VIEWS), ...

Some of the issues on Diplomatic Datasets:

Come from indirect (newspapers, etc.) and not direct sources. They are mostly disaggregated Dataset.

Armed Conflict Location and Event Data Project (**ACLED**) - History

2005	Created by Clionadh Raleigh
2014	Operates as a US non-profit, non-governmental organization incorporated in Wisconsin
2022	Expanded coverage to the entire world, collecting data in real time and publishing weekly updates

Armed Conflict Location and Event Data Project (**ACLED**) - Founding/Partners

Founding

ACLED receives financial support from the **Complex Risk Analytics Fund**, the **Dutch Ministry of Foreign Affairs**, and the **Tableau Foundation**.
Previously it received funding from the **Bureau of Conflict and Stabilization Operations (U.S)**, the **German Federal Foreign Office**, the **International Organization for Migration**, the **University of Texas at Austin**, and **the European Research Council**.

Partners

It collaborates with more than 50 entities to collect data within different countries worldwide.

Armed Conflict Location and Event Data Project (**ACLED**) - Aim

The Armed Conflict Location and Event Data Project (ACLED) is a disaggregated data collection, analysis, and crisis mapping project. Its main goal is to capture the evolution of disorder around the World.
Disorder can be generated by various types of events, and ACLED tracks these events daily. The types of events that it keeps track of are as follows:

- Battles, Explosions/Remote violence, Violence against civilians, Riots, Protests, Strategic developments.

Armed Conflict Location and Event Data Project (**ACLED**) - Methodology

The ACLED project codes reported information on the type, agents, location, date, and other characteristics of political violence events:

1. Human-coded
2. Wide-ranging network of local data collection partners on the ground
3. Different coverage periods for different regions and countries
4. Weekly update frequency
5. Open Source

Armed Conflict Location and Event Data Project (**ACLED**) - Data structure

Processing the dataset

To process the dataset effectively, data wrangling was necessary in order to obtain clear and functional variables.

- Filtered by date range and sub-event type
- Grouped by *ISO 3166-1 alpha-3 code* and date

This process resulted in a dataset where each row corresponded to a specific day in a country, including the number of events (and related fatalities) of a particular type that occurred on that day in that country.

Global Database of Events, Language and Tone (**GDELT**) - History

History

The first public version of GDELT was launched in 2013.
The GDELT Project was created by Kalev Leetaru along with Philip Schrodt.

Global Database of Events, Language and Tone (GDELT) - Founding/Partners

Collaborations

Initially, GDELT was supported and had contributions by **Political and Policy Sciences at the University of Texas**, the **U.S. National Science Foundation**, the **Peace Research Institute Oslo**, and many others.
It currently has a big partnership with Google and many others: **Google Ideas**, **Google Cloud**, **Google News**, the **Yahoo! Fellowship**, **BBC Monitoring**, the **National Academies Keck Futures Program**, **Reed Elsevier's LexisNexis Group**, **JSTOR**, **DTIC**, and the **Internet Archive**.

Global Database of Events, Language and Tone (**GDELT**) - Aim

GDELT is the largest, most comprehensive, and highest resolution open database regarding diplomatic news.
The GDELT Project is an initiative to construct a catalog of human societal-scale behavior and beliefs across all countries of the world, connecting every person, organization, location, count, theme, news source, and event across the planet into a single massive network that captures what's happening around the world, what its context is and who's involved, and how the world is feeling about it, every single day.

Global Database of Events, Language and Tone (**GDELT**) - Methodology

1. Machine-coded
2. Focuses on NLP and sentiment analysis
3. Frequent updates – every 15 minutes -
4. News from various media sources in over 100 languages

Global Database of Events, Language and Tone (**GDELT**) - Data Structure

The GDELT dataset has 62 attributes divided in 5 categories:

- Event Id and Date Attributes
- Actor Attributes
- Event Action Attributes: *Average Tone, NumArticles, CAMEOCodeDescription, QuadClass and the GoldsteinScale*
- Event Geography
- Data Management Fields

Global Database of Events, Language and Tone (**GDELT**) - Data Structure

Cameo

Conflict and Mediation Event Observations is a framework for coding event data. Every event is linked with a CAMEO code and its description, which refer to a specific diplomatic scenario.

Goldstein Score

The Goldstein Score is an intensity value that range from -10.0 to +10.0, with the aim to associate a "diplomatic value" to the event linked.

QuadClass

the QuadClass attribute groups all events encoded by the CAMEO framework into 4 distinct classes: **Verbal Cooperation, Material Cooperation, Verbal Conflict and Material Conflict.**

Global Database of Events, Language and Tone (**GDELT**) - Data Structure

Processing the dataset

- The dataset was reduced so that only the most significant items were retained ($>$ 10 mentions)
- Grouped by *ISO 3166-1 alpha-3 code* and date, mirroring the approach taken with the ACLED dataset

VARIABLE	TYPE	DESCRIPTION
Date	Datetime.date	Date on which the news occurred
Country	object	Country ISO 3166-1 alpha-3 code in which the events occurred
QC1 ... **QC4**	float64	Number of news items associated with each QuadClass
GoldsteinScale_QC1 ... **GoldsteinScale_QC4**	float64	GoldsteinScale value associated with the news group belonging to the same QuadClass
AvgTone_QC1 ... **AvgTone_QC4**	float64	Average tone associated with the news group belonging to the same QuadClass
NumArticles_QC1 ... **NumArticles_QC4**	float64	Number of mentions received by the news group belonging to the same QuadClass

Uppsala Conflict Data Program (**UCDP**) - History

1979 Was founded at the University of Uppsala in Sweden by Peter Wallensteen

1993 Data on armed conflicts have been published yearly in the Journal of Peace Research since 1993

2004 Creation of the global conflict database (UCDP Conflict Encyclopedia)

2016 The UCDP launched an interactive web-based system for visualising, handling and downloading data.

Uppsala Conflict Data Program (**UCDP**) - Founding/Partners

Founding

Various research foundations and governmental agencies have extended considerable support to the work. **The Swedish Research Council** and **Uppsala University**, the **Bank of Sweden Tercentenary Foundation** and **Sida**.

Partners

The main collaborations are **PRIO (Peace Research Institute Oslo)** and **HSRP (Human Security Report Project)** at the Simon Fraser University in Vancouver.

Uppsala Conflict Data Program (**UCDP**) - Aim

The Uppsala Conflict Data Program is the world's main provider of data on organized violence and the oldest ongoing data collection project for civil war, with a history of almost 40 years.

Its definition of armed conflict has become the global standard of how conflicts are systematically defined and studied.

UCDP produces high-quality data, which are systematically collected, have global coverage and are comparable across cases and countries.

Furthermore, the program is a unique source of information for practitioners and policymakers.

Uppsala Conflict Data Program (**UCDP**) - Methodology

The methodology used by the UCDP for data collection is based on a series of rigorous steps:

1. Definition of Conflicts:
   "An incident where armed force was used by an organised actor against another organized actor, or against civilians, resulting in at least 1 direct death at a specific location and a specific date"
2. Collection of Open Sources: $\rightarrow$ *media reports, government communiqués, UN documents, and non-governmental organization reports.*
3. Triangulation of Sources:
   Comparison of information from several independent sources.
4. Updated Yearly

Uppsala Conflict Data Program (**UCDP**) - Data Structure

The Dataset of the UCDP is the UCDP Georeferenced Event Dataset (UCDP GED). Some of the most important variables:

- Event Id: *A unique numeric ID identifying each event*;
- Year: *The year of the event*;
- Actor A: *The name of Side A in the dyad*;
- Actor B: *The name of Side B in the dyad*;
- Where Coordinates: *Name of the location to which the event is assigned*;
- Date Start: *The earliest possible date when the event has taken place*;
- Fatalities: *The best (most likely) estimate of total fatalities resulting from integer an event.*

Uppsala Conflict Data Program (**UCDP**) - Data Structure

- Type of Violence:
  State Based Conflict: "a contested incompatibility that concerns government and/or territory where the use of armed force between two parties, of which at least one is the government of a state, results in at least 25 battle-related deaths in a calendar year.";
  Non State Conflict: "the use of armed force between two organized armed groups, neither of which is the government of a state, which results in at least 25 battle-related deaths in a year.";
  One Sided Violence: "the use of armed force by the government of a state or by a formally organized group against civilians which results in at least 25 deaths."

Uppsala Conflict Data Program (**UCDP**) - Data Structure

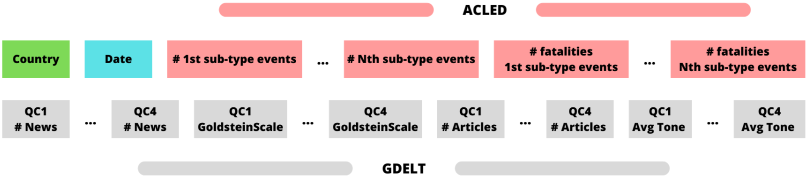
Processing the dataset

Again, to process the dataset effectively, data wrangling was necessary in order to obtain clear and functional variables.

- Creation of three variables, divided by type of violence
- Grouped by country and date

Our Final dataset

The integration of these two processed datasets (ACLED and GDELT) was accomplished by matching the *ISO 3166-1 alpha-3* code assigned to each country and the corresponding date.



Global Conflict Risk Index (**GCRI**) - History

The model was originally developed in 2014 at the Joint Research Centre (JRC) and has since been updated and revised on a yearly basis, in close collaboration with the European External Action Service (EEAS).
Their results are not public, however, the last published paper in which they discuss their system is dated December 8, 2022.

Global Conflict Risk Index (**GCRI**) - Aim

The Global Conflict Risk Index (GCRI) expresses the statistical risk of violent conflict in a given country in the coming 1-4 years and is exclusively based on quantitative indicators from open sources.
The output of the GCRI serves as the quantitative input to the EU conflict early warning framework for identifying countries at high risk of conflict and those whose risk is worsening significantly.

Global Conflict Risk Index (**GCRI**) - Dataset

Uses 22 variables in 6 different fields taken from different Datasets:

Political	Regime type	Democracy	V-DEM
	Regime performance	State capacity	V-DEM
		Repression	V-DEM
		Corruption	V-DEM
Security	History of conflict	Recent internal conflict	UCDP
		Years since last conflict	UCDP
	Current conflict situation	Neighboring conflict	UCDP
		Homicide rate	IHME
Social	Social cohesion and diversity	Female empowerment	V-DEM
		Ethnic exclusion	EPR
		Transnational ethnic ties	EPR
Economy	Development and distribution	GDP per capita, log	World Bank
		Income inequality	WID
		Trade openness	World Bank
		Oil exports	World Bank
	Provisions and employment	Food security	FAO
		Unemployment	World Bank
Geography - Environment	Environment	Droughts	SPEI/CSIC
		Temperature change	FAO
Demographics	Demographics	Population, log	UN
		Youth bulge	UN
		Child mortality	World Bank

Global Conflict Risk Index (**GCRI**) - Methodology

Data Management (1)

By working with a type of data that is already aggregated they do a job of cleaning the data:

- Problems of Missing Data:
  $\rightarrow$ Filling gaps

Global Conflict Risk Index (**GCRI**) - Methodology

Data Management (2)

- Variable Selection:
  - $\rightarrow$ Correlation *(problem if it introduces multicollinearity)*
  - $\rightarrow$ Predictive Performance
  - $\rightarrow$ Advice from experts in the field

Global Conflict Risk Index (**GCRI**) - Methodology

Model Architecture

The original GCRI used **logistic regressions** to model and predict conflict probabilities, and relied on linear regressions for conflict fatalities.

In the last work they conducted a systematic comparison of 14 modelling frameworks.

The Violence & Impacts Early-Warning System (**VIEWS**) - History

The Violence & Impacts Early-Warning System (VIEWS) is a project developed and maintained by the Uppsala Conflict Data Program (UCDP) at Uppsala University in Sweden. Its director and principal investigator is **Håvard Hegre**.
VIEWS started as an European Research Council funded Advanced Grant project that ran since 2017, the goal of which was to build a pilot early-warning system for fatal political violence in Africa.

The Violence & Impacts Early-Warning System (**VIEWS**) - Founding

Founding

The **European Research Council** (ERC) under the European Union's Horizon 2020 research and innovation programme, **Riksbankens Jubileumsfond**, **Uppsala University**, **Peace Research Institute Oslo**, the **United Nations Economic and Social Commission for Western Asia**, the **United Kingdom Foreign, Commonwealth & Development Office**, the **Swedish Research Council**, the **Swedish Foundation for Strategic Environmental Research**, the **Norwegian MFA**, the **United Nations High Commissioner for Refugees**, and the **Complex Risk Analytics Fund**.

The Violence & Impacts Early-Warning System (**VIEWS**) - Aim

It brings together a suite of interrelated projects to study and predict the risk of political violence and its societal impacts. It offers a forecasting system that systematically monitors hundreds of structural factors and complex conflict dynamics and generates monthly forecasts of impending conflicts worldwide-for every country and subnational location up to three years in the future.

The Violence & Impacts Early-Warning System (**VIEWS**) - Dataset

This is the sets of the key features with the corresponding databases:

- **Conflict History** $\rightarrow$ A suite of features capturing the history of conflict in each country and sub-national grid cell, e.g. the number of battle-related deaths per unit and level of analysis, and measures of the temporal and spatial distance to recent conflict events. (UCDP, ACLED)
- **Political Institutions, Democracy** $\rightarrow$ Features that capture democracy indices and the strength of political institutions in each country, such as liberal democracy, rule of law, equality, and the level of exclusion of social groups in politics. (V-Dem)

The Violence & Impacts Early-Warning System (**VIEWS**) - Dataset

- **Development** $\rightarrow$ Measures of development as provided by the World Bank Indicators, e.g. GDP per capita, infant mortality rate, and school enrollment. (WDI)
- **Economic Growth** $\rightarrow$ A feature set focusing specifically on historic and future economic growth, e.g. real GDP growth per year and growth forecasts for the coming years. (The International Monetary Fund World Economic Outlook)
- **News Monitoring** $\rightarrow$ A feature set based on the Mueller & Rauh topic model, which captures conflict risks as drawn from a topic analysis of news media. (Mueller & Rauh (2018))

The Violence & Impacts Early-Warning System (**VIEWS**) - Dataset

- **Climate & societal vulnerability** $\rightarrow$ Feature sets capturing climate extremes and societal vulnerability to climate hazards and other external shocks, e.g. climate extreme indices, reliance on agriculture, crop yields, precipitation, freshwater withdrawal, water management efficiency, and access to renewable resources. (UN-FAO, FAO AQUASTAT, PRIO-GRID, MIRCA, MAPSPAM, SPEI Global Drought Monitor)
- **Natural and Social Geography** $\rightarrow$ A feature set capturing terrain type, distance to natural resources, demography, proximity to cities and country borders. (PRIO-GRID)

The Violence & Impacts Early-Warning System (**VIEWS**) - Dataset

- **Food Security and Access to Basic Needs** $\rightarrow$ Feature sets capturing staple food prices along with measures of food security and access to basic human needs, such as mean food prices, food price inflation, undernourishment, access to clean water, and basic sanitation. (UN-FAO, FAOSTAT)

The Violence & Impacts Early-Warning System (**VIEWS**) - Model

It is continuously developed, tested and iteratively improved, resulting in frequent releases of new models and versions thereof.

Current Model

The current model is known as the fatalities model. For each month in a 3-year forecasting window, it generates predictions for the number of fatalities in impending conflict.

The Violence & Impacts Early-Warning System (**VIEWS**) - Model

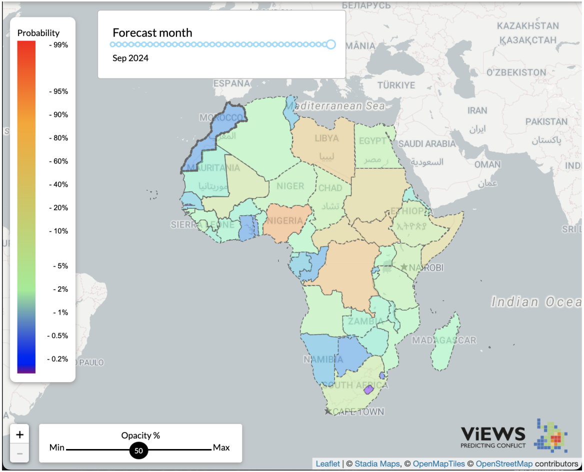
Algorithms

These are the algorithms that have been studied:

- Random forests
- Gradient boosting models
- 'Extreme' gradient boosting
- Light Gradient Boosting
- Hurdle models
- Markov models

→ Ensembling: 'Wisdom Of The Crowd'

The Violence & Impacts Early-Warning System (**VIEWS**) - Output



Introduction

Forecasting Prediction

Estimate of future events or trends through Machine Learning models.

- Monthly estimates (1 to 36 months in advance)

Early Prediction

Prediction of rare events in advance through Machine Learning models.

- The farther we are from an event the weaker are the predictive signals $\rightarrow$ poor predictions
- Daily estimates (1 to 90 days in advance)

Forecasting System

- ViEWS (Uppsala University and PRIO)
- GCRI (European Commission)
- CAST (ACLED)

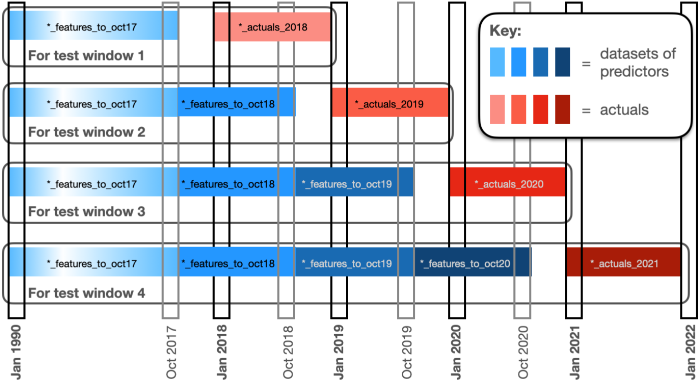
Commonalities: Reliance on open source data
Differences: data collection and processing, models

Our attempt

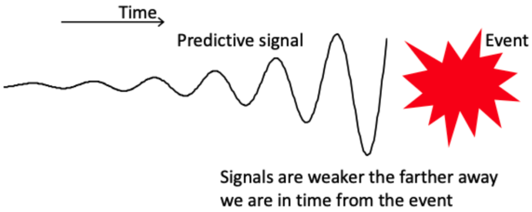
A transformer model for monthly forecasting of the number of battle-related deaths (1-12 months)
→ In development for ViEWS Competition 2023
→ Sharing results after the next 3 months of round table discussions with ViEWS

ViEWS Data

Google: "prediction challenge views"



Early Prediction



Days away from a rare event $\rightarrow$ flat, normal signal
Days near to a rare event $\rightarrow$ strong signal
$\Downarrow$
the more in advance prediction we want
the harder it gets for a model

Early Prediction

Why?

- We can interpret the onset of wars and crises as **rare events** w.r.t. the history of a given Country.
- Useful for diplomats to have an early warning (even of a few days) about the possible emergence of a war/crisis

...But what is the concept of war?

- Qualitatively: political declaration of war/crisis
- Quantitatively: lack of a belligerence index

$\rightarrow$ In addition to the daily variables extracted and grouped from the ACLED and GDELT databases, we introduced a new variable called **Weighted War**

Weighted War

Structure

- Created through an examination of the temporal progression of ACLED variables across more than 50 wars and conflicts (2000-2022)
- The *Weighted war* index encompasses all ACLED variables, with their contributions weighted based on our examination.

Weighted War

Aim

Create a new daily variable called "weighted war" for each country in our data.
→ This variable is intended to serve as an index that quantitatively measures the beginning of a crisis or war.

Data Sources

- a CSV file containing information about wars, conflicts, and violent protests (from now on, **unrest events**)
- collection of CSV files with ACLED data related to various countries

Weighted War

Standardization of ACLED Data

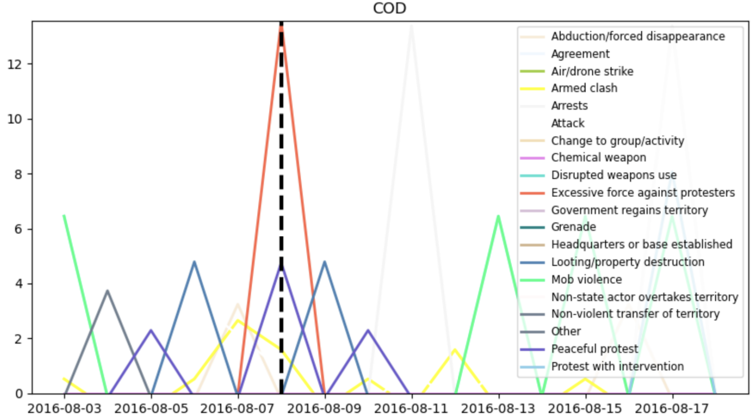
The ACLED data for each country -related to a unrest event under study- is standardized. This standardization ensures that variables with different scales and units are transformed into a common scale (mean=0, variance=1).

Collect ACLED Data related to the wars under study

We cut out a time window around each unrest event under examination. Standardised ACLED data are retained from 5 days before the unrest event until 10 days after (15-day window).

Weighted War

Kamwina Nsapu rebellion, August 8th 2016 (Luba rebels)



Weighted War

Summation of Standardized Data

The standardized data in the window is then summed column-wise. This aggregation helps capture the overall changes in the data's standardized values during this critical period for each variable.

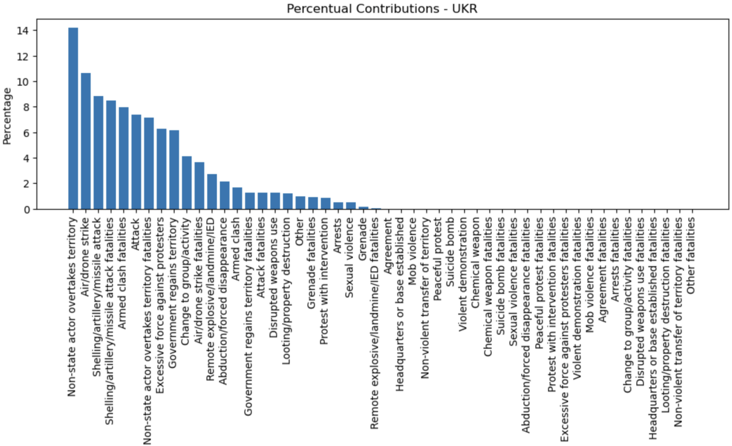
Calculation of Percentual Contribution

For each country and event, the percentage contribution of each standardized variable to the total sum within the specified time window is calculated. This provides insights into which variables have the most significant impact.

What we have now?

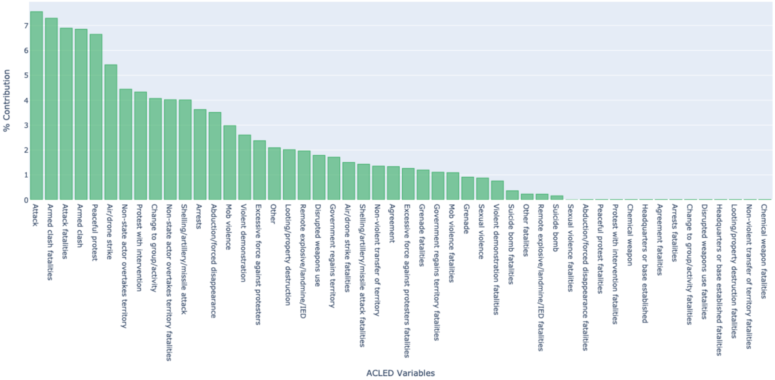
For every war since 2000, we know which variables have varied most in intensity during the start of the war!

Weighted War



Weighted War

Relation between sub-events to the Weighted War variable



Weighted War

Calculation of the Weighted War Index

The final step involves calculating the Weighted War Index for each country and each day (row) in the dataset. This calculation is based on the standardized values of relevant variables and their respective percentage contributions.

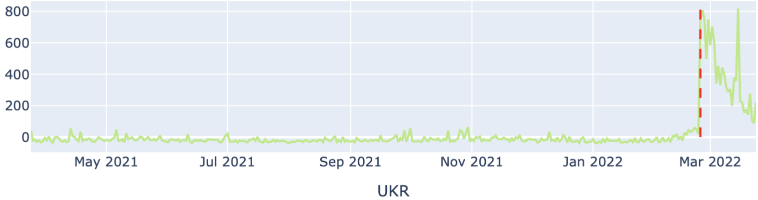
$$ww_{i,j} = av_{1_{i,j}} \cdot cav_1 + av_{2_{i,j}} \cdot cav_2 + \ldots + av_{n_{i,j}} \cdot cav_n \qquad (1)$$

ww: weighted war
av: standardized ACLED variable*
cav: the percentual contribution of the ACLED variable
i: i-th day
j: j-th country
n: total number of ACLED variables

* standardized up to the i-th day

Weighted War

Russian invasion of Ukrain, Feb, 24th 2022



Weighted War

Pro

- Quantitative description of unrest for every country
- Country-specific definitions of unrest events

Cons

- Unique scale of magnitudes per country
- The value 0 refers to the normality of a given country, not global

Target variable implementation

Thanks to the weighted war variable, we can construct our target variable.
What is our goal? Anticipate the start of a war/crisis.

Labelling Data

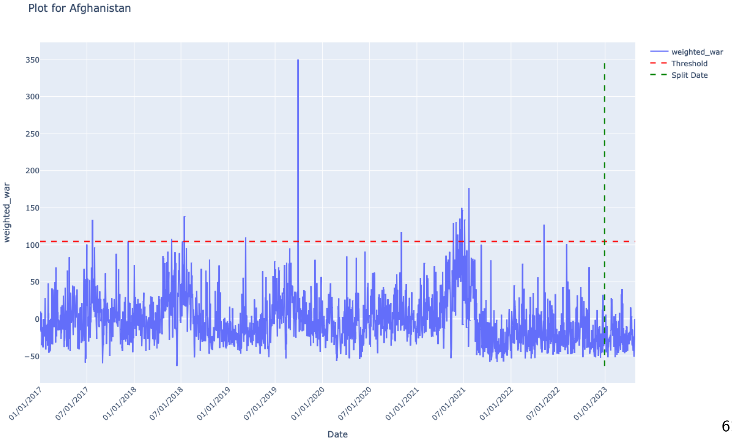
We first need to separate between “normal” days and “unrest” days

What is an unrest day?

Target variable implementation

What is an unrest day?

We define a day as an unrest day when the Weighted war variable reach the top 0.5% value of its entire historical evolution.



Target variable implementation

Labelling process

Introduction of a new binary variable y

- 0: “normal” days
- 1: “unrest” days

This distinction is not sufficient. If we trained the model to distinguish between normal and unrest days, the solution would be trivial.

Instead, we need a model that can distinguish a “normal” day from a “pre-unrest” day.

Target variable implementation

What is a “pre-unrest” day?

We define “pre-unrest” days as the 90 days prior to an “unrest” day.

Labelling process

- 0: “normal” days
- 1: “pre-unrest” days

The goal of the project

For each country, the system was trained and tested to assess its ability to classify days as either “normal” or “pre-unrest”.

Output

0% to 100% probability of being a “pre-unrest” day, i.e. being one of the 90 days before the war

Recap

What is done

Collection of ACLED data related to wars

Weighted war variable implementation

Target variable implementation

What is left

Input pre-processing

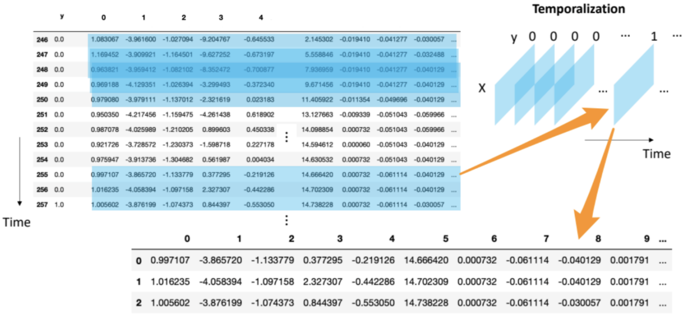
Time-series data splitting

Model implementation

Results

Input pre-processing

Capture Temporal Dependencies



	y	0	1	2	3	4						
246	0.0	1.083067	-3.961600	-1.027094	-9.204767	-0.645533		2.145302	-0.019410	-0.041277	-0.030057	...
247	0.0	1.169452	-3.909921	-1.164501	-9.627252	-0.673197		5.558846	-0.019410	-0.041277	-0.032488	...
248	0.0	0.963821	-3.959412	-1.082102	-8.352472	-0.700877		7.936959	-0.019410	-0.041277	-0.040129	...
249	0.0	0.969188	-4.129351	-1.026394	-3.299493	-0.372340		9.671456	-0.019410	-0.041277	-0.040129	...
250	0.0	0.979080	-3.979111	-1.137012	-2.321619	0.023183		11.405922	-0.011354	-0.049696	-0.040129	...
251	0.0	0.950350	-4.217456	-1.159475	-4.261438	0.618902		13.127663	-0.009339	-0.051043	-0.059966	...
252	0.0	0.987078	-4.025989	-1.210205	0.899603	0.450338	⋮	14.098854	0.000732	-0.051043	-0.059966	...
253	0.0	0.921726	-3.728572	-1.230373	-1.598718	0.227178		14.594612	0.000060	-0.051043	-0.040129	...
254	0.0	0.975947	-3.913736	-1.304682	0.561987	0.004034		14.630532	0.000732	-0.051043	-0.040129	...
255	0.0	0.997107	-3.865720	-1.133779	0.377295	-0.219126		14.666420	0.000732	-0.061114	-0.040129	...
256	0.0	1.016235	-4.058394	-1.097158	2.327307	-0.442286		14.702309	0.000732	-0.061114	-0.040129	...
257	1.0	1.005602	-3.876199	-1.074373	0.844397	-0.553050		14.738228	0.000732	-0.061114	-0.030057	...

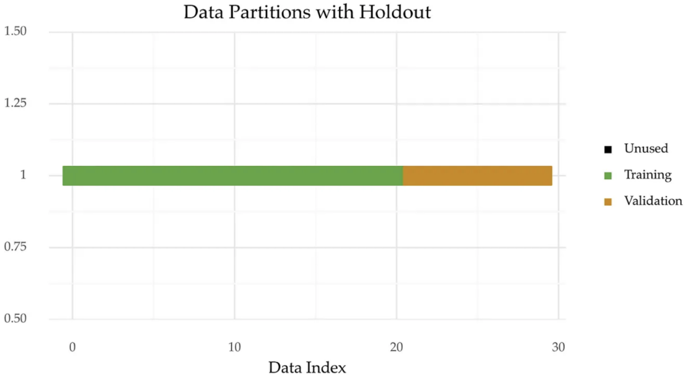
⋮

	0	1	2	3	4	5	6	7	8	9	...
0	0.997107	-3.865720	-1.133779	0.377295	-0.219126	14.666420	0.000732	-0.061114	-0.040129	0.001791	...
1	1.016235	-4.058394	-1.097158	2.327307	-0.442286	14.702309	0.000732	-0.061114	-0.040129	0.001791	...
2	1.005602	-3.876199	-1.074373	0.844397	-0.553050	14.738228	0.000732	-0.061114	-0.030057	0.001791	...

Figure: Temporalization process

Splitting Data for Sequential Validation

What is the right way to split data?



Problem: we do not need to learn a trend but the rupture of the trend → Anomalies need to be in validation and test sets

Splitting Data for Sequential Validation

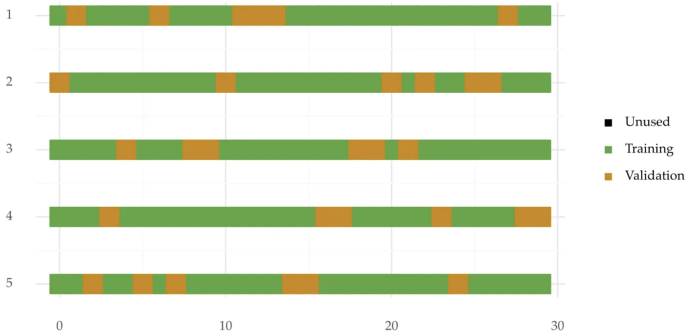
Unique Approach to Time-Series Anomaly Detection

- Addressing the complexity of time-series anomaly detection and the critical role of data splitting.
- Traditional data splitting techniques struggle when anomalies need to be present in both test and validation data.

$\rightarrow$ Implementation of a new splitting function.

```
def split_3d_data(X, y, timesteps, shift):
    ...
return sets
```

How it works



Splitting Data for Sequential Validation

How it works

Splits data into operative, train, test, and validation sets.
Constructing test and validation sets with sliding windows.

Key role in creating validation and test sets

Sliding windows enable us to capture past wars and crisis effectively.

```
def get_slice_idx(val_len, shift, y):
    ...
return possible_val_indices
```

Transformer

Introduced in *"Attention is all you need"* by Vaswani et al. (2017)
Adapted to time-series by Zerveas et al. (2021).

How it works: Attention mechanism

- Time-series data contains patterns and dependencies that change over time
- Capturing relevant temporal dependencies
  - It can dynamically assign different weights to different time steps, focusing more on the relevant information and less on irrelevant or noisy data

Advantages

- Ability to efficiently utilize GPU hardware
- It requires less long time-series

Transformer

The steps behind the operation (1)

- Taking as input a training sample, denoted as $\mathbf{X} \in \mathbb{R}^{w \times m}$, represents a multivariate time series comprising w feature vectors, $\mathbf{x}_t \in \mathbb{R}^m$, where $\mathbf{X} \in \mathbb{R}^{w \times m} = [x_1, x_2, \ldots, x_w]$.
- Reduce the dimensionality of the input vector $\mathbf{x}_t$ from m to d, which corresponds to the transformer model's sequence element representations.
- Incorporating positional encodings.
- Passing the input to the self-attention module
- Flattening the output of the self-attention module

Transformer

The steps behind the operation (2)

- After the flattening, the resulting vector $\bar{\mathbf{z}}$ is further processed through a sequence of feed-forward layers.
- The output of the last feed-forward layer is then fed into the final layer, which applies the sigmoid function.

The sigmoid function maps the output of the last feed-forward layer to a probability distribution between 0 and 1, providing the likelihood of each class for the given input. This enables the model to generate class probabilities, which can be used for classification purposes.

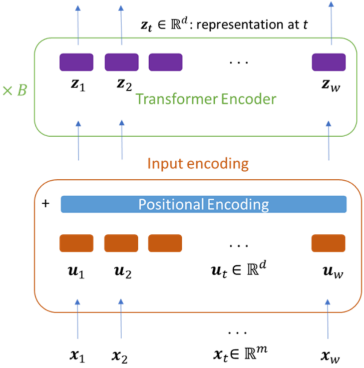
Unsupervised Pre-training

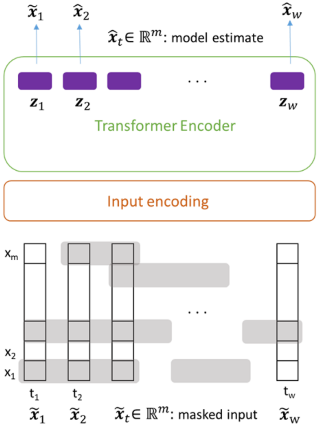
An additional step to improve the model, in accordance with the approach proposed by Zerveas et al. (2021), was to employ unsupervised pre-training to initialize our model.

The pre-training task we adopt is the autoregressive denoising task, wherein a portion of the input is masked, and the model is trained to predict the masked values.

Transformer

A Graphical Representation

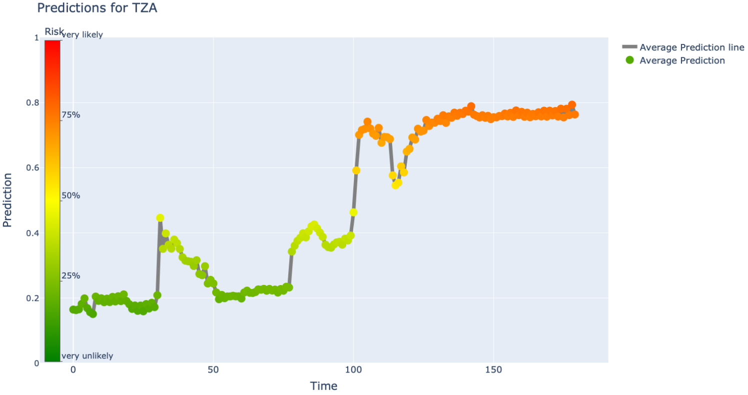




Remember the aim

First of all, let us recall the goal of the project

For each country, the system was trained and tested to assess its ability to classify days as either "normal" or "pre-unrest" within the 90-day window preceding an unrest event.



Evaluation results

Performance Measure: the Area Under the Curve (AUC) and accuracy were used to evaluate the models' effectiveness

Table: Average Performance Measures Across 101 Countries

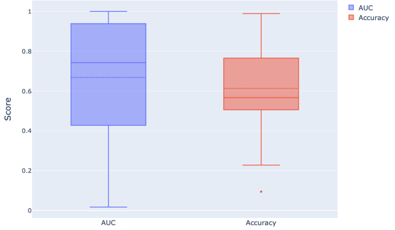
Performance Measure	Median	Mean	Standard Deviatio
AUC	0.74	0.65	0.31
Accuracy	0.57	0.61	0.17

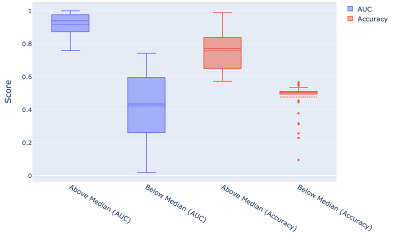
Analyze the results

Disparities of performances

Disparities between the underperforming and outperforming group:

- **Countries above the median**: the mean and median of AUC above 0.9 and the mean and median of accuracy just below 0.8
- **Countries below the median**: results hovering around 0.5 for both AUC and accuracy





Analyze the results

Why the disparities?

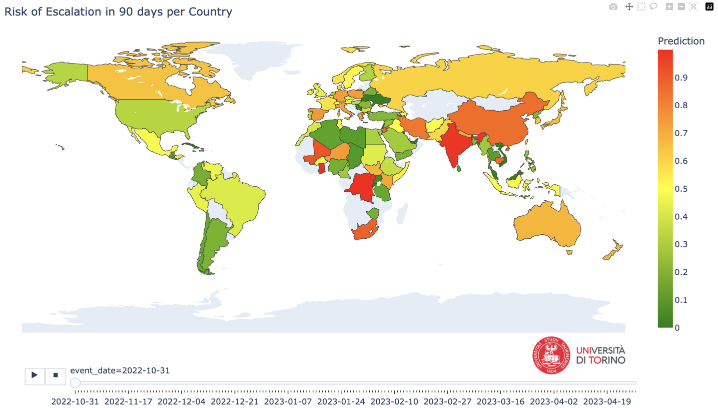
- Lack of data: data collection is uneven and started only a few years ago for some countries.
- Limited ground truth data: all data sourced indirectly (newspapers, social media), no available field-based open dataset.

How to boost underperforming cluster performance?

- Include diverse data types (relational, economic, social, etc.).
- Deeper optimization (high computational cost)
- Apply transfer learning.

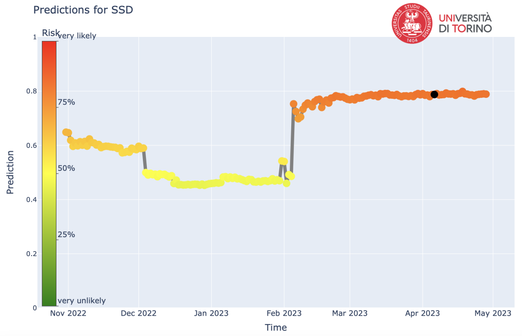
Operational system

Output 1:
Each country is associated with a color that indicates the probability that a day will be one of the 90 days prior to the onset of an escalation, i.e., of an Unrest day.



Operational system

Output 2: Probability, day by day, of being a Pre-unrest day.



April 6, 2023 (black point):

- Attack: 1
- Peaceful protest: 5
- Abduction/forced disappearance: 1

Conclusion

- Overview of sociopolitical field datasets
- Overview of early warning systems
- Challenge: splitting time series data
- New approach in conflict prediction (fully reproducible)
    - Implementation of a transformer model for time series

  $\rightarrow$ Good results already in its first version with very modest computing power
- Future improvements (to assist utilization):
    - Explainability
    - Probabilistic forecasting of disorder events

References I

Compton, Ryan; Craig Lee, Jiejun Xu, Artieda Moncada Luis & Tsai-Ching Lu (2014). Using publicly visible social media to build detailed forecasts of civil unrest.

Hegre, Håvard; Joakim Karlsen, Håvard Mokleiv Nygård, Håvard Strand & Henrik Urdal (2012) Predicting armed conflict, 2010-2050. *International Studies Quarterly 57*: 250–270.

Junior, Manoel; Philipe Melo, Ana Paula Couto da Silva, Fabricio Benevenuto & Jussara Almeida (2021). Towards understanding the use of telegram by political groups in brazil. In: Proceedings of the Brazilian Symposium on Multimedia and the Web WebMedia '21 , 237–244. New York, NY, USA. Association for Computing Machinery (https://doi.org/10.1145/3470482.3479640).

References II

Khaund, Tuja; Muhammad Nihal Hussain, Mainuddin Shaik & Nitin Agarwal (2021). Telegram: Data collection, opportunities and challenges. In: Lossio-Ventura, J. A; Valverde-Rebaza, J. C; Díaz, E & Alatrista-Salas, H (eds) Information Management and Big Data , 513–526. Cham. Springer International Publishing.

Korkmaz, Gizem; Jose Cadena, Chris J. Kuhlman, Achla Marathe, Anil Vullikanti & Naren Ramakrishnan (2015). Combining heterogeneous data sources for civil unrest forecasting. In: Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015 ASONAM '15 , 258–265. New York, NY, USA. Association for Computing Machinery (https://doi.org/10.1145/2808797.2808847).

References III

Vaswani, Ashish; Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser & Illia Polosukhin (2017). Attention is all you need. In: Guyon, I; Luxburg, U. V; Bengio, S; Wallach, H; Fergus, R; Vishwanathan, S & Garnett, R (eds) Advances in Neural Information Processing Systems volume 30. Curran Associates, Inc.

Zerveas, George; Srideepika Jayaraman, Dhaval Patel, Anuradha Bhamidipaty & Carsten Eickhoff (2021). 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 , 2114–2124. New York, NY, USA. Association for Computing Machinery.