IT TOOLS AND METHODS FOR MANAGING MIGRATION FLOWS

Details of the big data used in predicting country instability and model description

Haithem Afli & Adam Zebrowski
Munster Technological University

Mehwish Alam & Yiyi Chen
FIZ-Karlsruhe

February 2022

"This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement Nº 882986".

Deliverable Factsheet	
Title and number	*Details of the big data used in predicting country instability and model description (D3.3)*
Work Package	WP3
Submission date	28 February 2022
Authors	Haithem Afli & Adam Zebrowski (MTU), Mehwish Alam & Yiyi Chen (FIZ)
Contributors	Zsolt Kardkovacs & Fouad Shammary (MTU)
Reviewers	Tobias Heidland (IfW) & Andre Groeger (UAB)
Dissemination level	PU (Public)
Deliverable type	R (Report)

Version Log			
Issue Date	**Version**	**Author**	**Change**
03/02/2022	v0.1	Haithem Afli & Adam Zebrowski (MTU), Mehwish Alam (FIZ)	First version sent for review.
21/02/2022	v0.2	Haithem Afli & Adam Zebrowski (MTU), Mehwish Alam & Yiyi Chen (FIZ)	Second version corrected after first review
25/02/2022	v0.3	Haithem Afli & Adam Zebrowski (MTU), Mehwish Alam & Yiyi Chen (FIZ)	Final version sent to Coordinator
28/02/2022	V0.4	Cristina Blasi (UAB)	Coordinator final review and layout adjustments. Version ready to be submitted.

Abbreviations

ACLED: Armed Conflict Location and Event Data
AI: Artificial Intelligence
API: Application Programming Interface
BiGRU: Bidirectional Gated Recurrent Unit
BiLTSM: Bidirectional Long Short Term Memory
BUL: Brunel University London
CA: Consortium Agreement
CART: Classification and Regression Tree
CEPS: Centre for European Policy Studies
CERTH: Ethniko Kentro Erevnas kai Technologikis Anaptyxis
CNN: Convolutional Neural Networks
CSO: Civil Society Organization
DARPA: Defense Advanced Research Projects Agency
EASO: European Asylum Support Office
EC: European Commission
EMT: EUMigraTool
ESS: Exploitation Strategy Seminar
EU: European Union
FIZ: Fiz Karlsruhe–Leibniz-Institute fur Informationsinfrastruktur GMBH
FRONTEX: European Border and Coast Guard Agency
GA: Grant Agreement
GDELT: Global Database of Activities, Voice, and Tone
GPS: Global Positioning System
GRU: Gated Recurrent Unit
GSR: Gold Standard Report (GSR)
ETM: Embedded Topic Model
FAIR: Findable, Accessible, Interoperable, Reusable
HRB: Horizon Results Boosters
HTML: HyperText Markup Language
IAI: Istituto Affari Internazionali
ICEWS: Integrated Crisis Early Warning Systems
ICT: Infocommunication Technology (Information and Computer Technologies)
IFW: Institut für Weltwirtschaft
IMF: International Monetary Fund
IOM: International Organization for Migration
IPR: Intellectual Property Rights
JRC: Joint Research Centre
KB: Knowledge Base
KER: Key Exploitable Result
LTSM: Long Short Term Memory

MTU: Munster Technological University
NGO: Non-Governmental Organization
OPE: Occupy Demonstration Case
PWG: Policy Working Group
SPEED: Social, Political, and Economic Event Database
SYPE: Survey of Young People in Egypt
TFEU: Treaty on the Functioning of the European Union
TRC: Terracom AE
TRL: Technological Readiness Level
UAB: Universitat Autònoma de Barcelona
UB: Users Board
UN: United Nations
UNHCR: United Nations High Commissioner for Refugees
URL: Universal Resource Locator
VAR: Vector Autoregression
WP: Work Package

Executive Summary

This deliverable summarises the ITFLOWS Social Listening System, and the methodology used to analyse the GDELT dataset as an instability prediction component of the ITFLOWS EUMigraTool (EMT) policy making and sentiment monitoring module which is described in Deliverable 9.4 Key Exploitable Result No. 3. In this document, we discuss the foundations and the technical details of our implementation of the ITFLOWS Social Listening System to monitor social tensions, unrests, and to detect or to calculate the probability of future events, or escalations.

The present deliverable describes the experimental conflict case model established with GDELT's data-driven limitations. The proposed model integrates and identifies any level of conflict creation or de-escalation, including internationalised contentious intervention. Using country-level actor-specific incident datasets that signal potential triggers of violent conflict such as riots, bombings, or election-related activity, the model calculates the incidence of material conflict incidents. It is based on the idea that as the number of material confrontation cases rises, so does the amount of material and verbal cooperation. The machine learning method used to simulate conflict events is the random forest model. This model is well-suited for using time series data to characterise, process, and predict near-future events.

Using typical yield curve variables and news information, this work is one of the first to study the behaviour of government yield spreads and financial portfolio decisions in the light of political instability. We assume that these new measures would be able to catch and forecast interest unrest trends, especially during periods of turbulence. Overall, the deliverable shows how to derive political unrest metrics from a large-scale database like GDELT to catch possible regional intentions. We are also using random forest to evaluate a range of prediction models, ranging from traditional economic approaches to innovative machine learning techniques.

Potential protesters may feel affected as a result of their weak personal economic condition, which is based on real-world data rather than skewed data. Second, as the country's educational standard increases, more voters would be able to link their personal circumstances to government outcomes. Finally, since citizens in repressive

countries use protests as a substitute for elections, they are likely to be egotistical in their decision to take risky and costly action. While any of these processes needs to be investigated further, the analytical implications of this analysis will help researchers better understand the social consequences of data falsification. The findings also have implications for the political economy of demonstrations, especially in authoritarian regimes.

In this document, Chapter 1 introduces the input data source (GDELT dataset) and its main characteristics, and it also specifies the research goals based on this data source. In Chapter 2 we summarise the efforts made and problems identified by others in this field of area, including ethical issues. We present our solution to process the GDELT data source to convert and to prepare it into a widely usable and cleaned data mine. Using the cleaned data, Chapter 4 describes how our models are designed and built by detailing their implementations. Concluding remarks regarding our models and implementations are summarised in Chapter 5.

Table of Contents

List of Figures

1. Introduction

Country instability is a global issue, with unpredictably high levels of instability thwarting socio-economic growth and possibly causing a slew of negative consequences. As a result, uncertainty prediction models for a country are becoming increasingly important in the real world, and they are expanding to provide more input from 'big data' collections, as well as the interconnectedness of global economies and social networks. This has culminated in massive volumes of qualitative data from outlets like television, print, digital, and social media, necessitating the use of artificial intelligence (AI) tools like machine learning to make sense of it all and promote predictive precision [1].

The Global Database of Activities, Voice, and Tone (GDELT Project) records broadcast, print, and web news in over 100 languages every second of every day, identifying the people, locations, organisations, counts, themes, outlets, and events that propel our global community and offering a free open platform for computation on the entire world. The main goal of our research is to investigate how, when our data grows more voluminous and fine-grained, we can conduct a more complex methodological analysis of political conflict. The GDELT dataset, which was released in 2012, is the first and potentially the most technologically sophisticated publicly accessible dataset on political conflict.

This work aims to advance expertise in this field by proposing a nation's instability prediction model based on GDELT datasets supplemented with broader socio-economic data from the World Bank, Correlates of War, and other sources. This data has already been classified by A.I. algorithms, and it serves as a foundation for the study of multiple events [2]. This research aims to provide insight into the causes and complexities of uncertainty by developing a hybrid predictive model based on Eurasian countries using deep learning algorithms and Bayesian inferences. As a result, integrating event-driven data with more extensive economic and social benchmarks would offer a more granular and accurate way of predicting nation volatility and civil unrest [3].

1.1. Use of GDELT in other Fields

The GDELT architecture must be able to accurately store and access millions of measurements per record due to the increasing number of themes and emotions measured from each post. In addition, an increasing number of queries seek to find macro-level trends at scale across the whole archive. In-database execution is needed because even routine questions may require sophisticated algorithms to be applied to terabytes of data. In terms of comprehension and usability of the results, the complexity, growth rate, and analytic load present specific challenges. Because of GDELT's varied user population and application areas, access habits are inconsistent queries that can access many of columns in a single analysis, obviating the need for a typical indexed database [4].

More specifically, unlike other fields, the quantity and consistency of evidence on political conflict remained essentially unchanged for millennia. Historically, scientific evidence has been sparse. While biologists and physicists have been able to perform studies, researchers interested in the nature of large-scale political conflict have not been able to do so. Furthermore, complex observational data has become challenging to obtain on a wide scale because it would have needed unfeasibly large quantities of highly trained and disciplined human resources, as well as the equipment to witness, register, archive, and then transmit data about political activities [5]. In today's world, studies to learn about political conflict are still difficult (which is actually a positive thing). Still, our ability to capture, store, and analyse observational data has increased exponentially. Such research is made possible by the rapid and relatively recent boom in both the amount and accuracy of data on political conflict [6].

However, the accuracy of economic predictions and now-casting models remains a problem since global economies are subject to a variety of shocks that make forecasting and now-casting practices extremely challenging in the short and medium-term. In this regard, the use of recent Big Data technologies to improve forecasting and now-casting for a wide range of economic and financial applications holds a lot of promise.

The number of tools open to quantitative social scientists has grown dramatically in recent years. ICEWS and GDELT are global databases that have been used to develop statistical models for a wide range of cases, including international and domestic crises,

revolutions, revolts, and religious and ethnic abuse. Techniques used include discriminant analysis, HMMs, Bayesian time series forecasting, and Vector Auto Regression (VAR) [7].

Each record in GDELT contains details about a single event. We use the following attributes from a case to build our models: MonthYear, Actor1Type, Actor2Type, RootEventCode, AvgTone, and where Actor1Type and Actor2Type store the role of the actors participating in the event, RootEventCode $\in \{1, \ldots, 20\}$ determines whether the case is cooperative or incompatible, the captures the event's effect on a country's security, while the AvgTone is a subtle measure of an event's significance that acts as a proxy for its impact [8].

1.2. Detecting and Monitoring Social and Political Events

In social science research, social and political events detection is a well-known, critical and difficult activity. Detecting the occupy demonstration case (OPE), which typically campaigns against social, political, and economic injustice, is of special importance. People protest against problems that affect their life and for which they believe the government (local, state, or national) is responsible during the demonstration (e.g., unfavourable election, poor infrastructure, etc.). Identifying the participants' contact patterns and estimating the likelihood of an OPE will serve as a guide for government emergency response. In recent years, computational scientists have had access to a plethora of data tools [9]. Among them, the Global Dataset of Incidents, Place, and Tone (GDELT), built on open-source data, comprised over 300 million machine-coded events in near real-time (e.g. every day, every fifteen minutes). This dataset has been used to find trends in a variety of incidents, including domestic political conflicts, natural disasters, racial and religious conflict, and more. In a graph-based approach, a monthly time window was used to forecast domestic political crises, which is a much too large resolution for real-time information systems.

We used the GDELT dataset, which includes machine-encoded archives of international events originating from news reports, to collect bilateral sequences of inter-country events and a Bayesian standard mining algorithm to find norms that best represented

the observed behaviour. A statistical study found that a probabilistic model with explicit normative reasoning outperformed a reference probabilistic model in terms of data matching. The Global Archive of Events, Language, and Tone (GDELT) is a continuously updated geopolitical activity database of over half a billion records. The most recent version, GDELT 2.0, is free and open source, and it is updated every 15 minutes. The database contains an incidents table of 60 attributes for each incident (such as the type of the accident and the countries involved), and it has been used for research such as estimating future levels of unrest in Afghanistan, Pakistan, Thailand among others, and identifying protest activities around the world [10].

1.3. Research Aim and Objectives

The project's aim is to better understand the current state of political uncertainty, considering the dynamics of a quickly evolving and increasingly intertwined global climate on nation instability levels [11]. This would be supported by the use of Bayesian inference to provide an in-depth learning approach capable of dealing with such a dynamic and changing stability environment. The massive data sources from GDELT, the World Bank, and the Correlates of War on individuals, locations, incidents, and actions can be combined with A.I. research capability to promote a systemic view of the global stability that considers the different networks and aspects that can affect insecurity [12].

The main research objectives for this study are proposed as follows:

- To look at countries that have a lot of contextual data that has been gathered over a long period of time. Wide data density, precise Geo-coding, Sub-state, latitudinal, and longitudinal data are all supposed to be crucial background specifics.
- To assess the benefits and shortcomings of GDELT machine-coding in order to ensure the precision and credibility of the final predictive model [13].

- To create a model that uses deep learning, Bayesian techniques, and random forest techniques to enable successful intra-county and location-specific analyses, as well as predictive validity in the context of country stability.

- Examine cutting-edge implementations of deep learning approaches to big data in forecasting, as well as best practices and novel predictive analytics applications and processes [14].

2. Literature Review

2.1. Introduction

This review will critically discuss and evaluate existing literature on the application of deep learning techniques for predictive analysis. This will examine in particular such Artificial Intelligence (AI) techniques successfully applied to predicting country instability based on GDELT and associated datasets, providing large amounts of complex contextual data from diverse sources including web, print and broadcast news media in over 100 languages across the world [1]. AI models were found to be quite capable of accurate predictive analysis and obtaining an early indication of country instability from limited information [2, 3]. Such models have become increasingly important considering that classical predictive models have found it increasingly difficult to forecast instability due to evolving mechanisms and factors influencing country instability, such as increased global interconnections of economies, trade and communications [4, 5, 6]. It is important to discuss the concept of uncertainty and complexity management in AI systems and their importance in developing more sustainable and effective prediction systems. The GDELT datasets to be used in this research, which will be complemented by World Bank, Correlates of War, etc., have been categorised using AI techniques. A review of the underlying AI algorithms employed in the categorisations will be undertaken as well as examining the associated ethical and human-machine issues pertaining to machine learning and AI techniques and algorithms.

2.2. Uncertainty and Complexity in AI Models

Prediction systems can significantly benefit from increased capability to handle uncertainty and complexity, since these qualities are inherent and widespread in real-world settings. The ability to cope with these is therefore critical in developing superior intelligence in machine learning applications [7]. Different from risk, which can be evaluated statistically, uncertainty is characterised by information that may be incomplete, random, ambiguous or inconsistent [7, 8]. Complexity, on the other hand, involves the complicated or intricate interaction of different parts of a system. While the capability for increased computational or processing power may be helpful in decision-

making involving complexity, uncertainty requires some level of generalisation, which provides the ability to be able to understand and potentially cope with unknown situations and new data [9]. Deep learning algorithms have gained a significant following in the machine learning field for their superior performance concerning generalisation while also having the capacity to handle complex abstractions in analysing vast amounts of diverse big data [10, 11].

2.3. Deep Learning Models

Deep learning is a sub-branch of machine learning based on artificial neural networks of multiple hierarchical processing layers consisting of non-linear processing units applied to modelling complex high-dimensional input-output data [12]. Artificial Neural networks, as shown in Figure 1, represent computational models imitating biological neural networks. An artificial neural network consists basically of a single layer on input, processing and output elements. Each element of the network has its own neurons or processing nodes connected by links that have associated numeric weights depending on the activation function used for the network. The input neurons of the model receive inputs and sum them up based on the weighted sum of inputs (Figure 2) and distribute them to the processing layer, which repeats the process to deliver the output [13, 14].

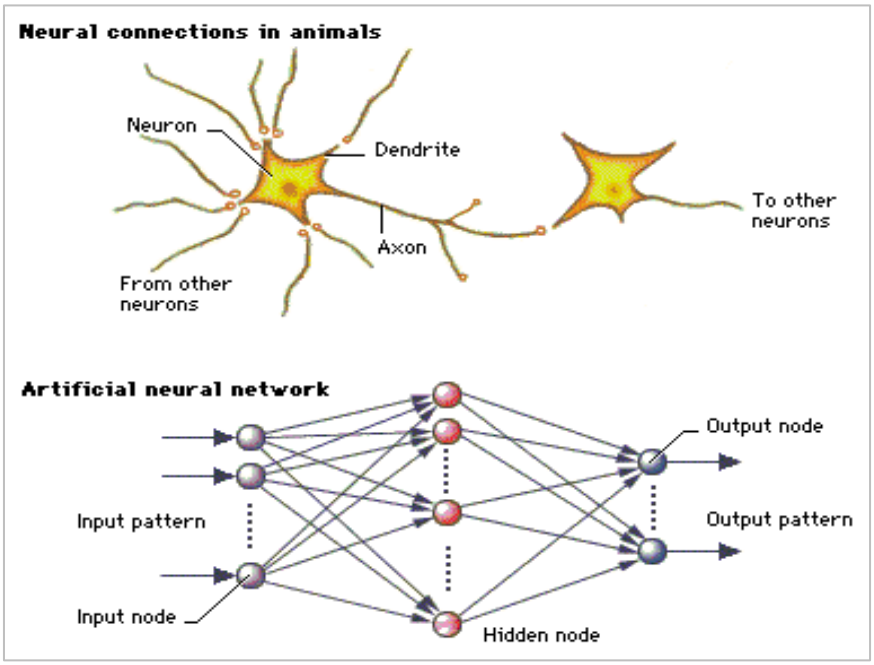

Figure 1. Illustration of Neural Networks
Sources: [15, p. 38]

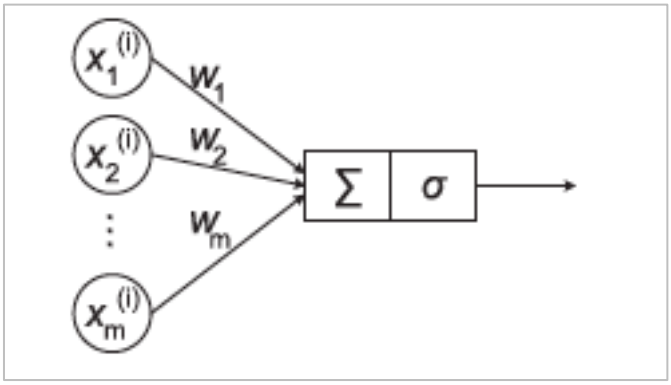

Figure 2. Illustration of artificial neural network neuron summation based on inputs x (i) 1 , . . . , x (i) m , weights w 1 , . . . , w m and activation function σ. _denotes the summation over all inputs. Source: [13, p. 631]

Deep neural network algorithms utilise several layers in the artificial neural network to model complex non-linear high-dimensional output models, with commonly used activation functions including sigmoidal functions (cosh and tanh), rectified linear units (ReLU) or heavyside gate functions [12]. ReLU's have been found to be particularly useful in rapid dimension reduction. The logic behind this is that most real-life datasets, such as images, consist of different levels of features, with the lower-level ones serving as building blocks for the higher-level ones. This thinking is supported by empirical evidence in neuroscience and biology: a study of light-adapted eyes of anaesthetised cats simulated with spots of lights of various shapes confirmed that the projection of the retinas on the cortex occurred in an orderly manner, organising from simple to complex categories [16]. Another study examined the effectiveness of deep neural network models in predicting single neuron responses in primate visual cortex areas and showed that the system predicted the responses of the neurons to a high degree of accuracy [17].

Developed empirically over the years based on heuristic trial-and-error construction instead of theoretical explanations, deep neural networks have capitalised on the recent explosion of computing power and increased availability of big data for the development of highly effective statistical prediction models solving problems in diverse fields such as computer vision, natural language processing, speech recognition and reinforcement learning [18]. A number of computerised personal assistants, including Apple's Siri, Amazon's Alexa and Microsoft's Cortana feature deep neural networks, while facial recognition applications for payment systems and

recommendation systems such as those used by Amazon and Netflix also utilise deep learning networks [13]. Major advantages of these networks over other traditional statistical systems include the following [12]:

1) All data that is possibly relevant to the prediction problem can be included in input data, increasing flexibility. A key advantage of deep neural network systems is that its weight matrices are matrix-valued, allowing the predictor flexibility to discover non-linear features inherent in the data [12]
2) Complex interactions and nonlinearity are easily incorporated into the system due to its multi-layer structure. This enables it to represent complex non-linearity through the composition of several non-linear functions. [19]
3) Overfitting is more easily avoided than in other, traditional, models. Overfitting represents the model learning the training data too well, including associated noise, which in turn leads to unreliable generalisation [12, 20].
4) Fast, scalable computational frameworks are available for processing such as PyTorch and TensorFlow

Commonly used deep learning architectures, as shown in figure 3, include convolutional neural networks (CNNs), recurrent neural networks (RNNs), long-short-memory (LSTM) and neural Turing machines (NTM), which incorporate accelerated learning processes [12, 21]

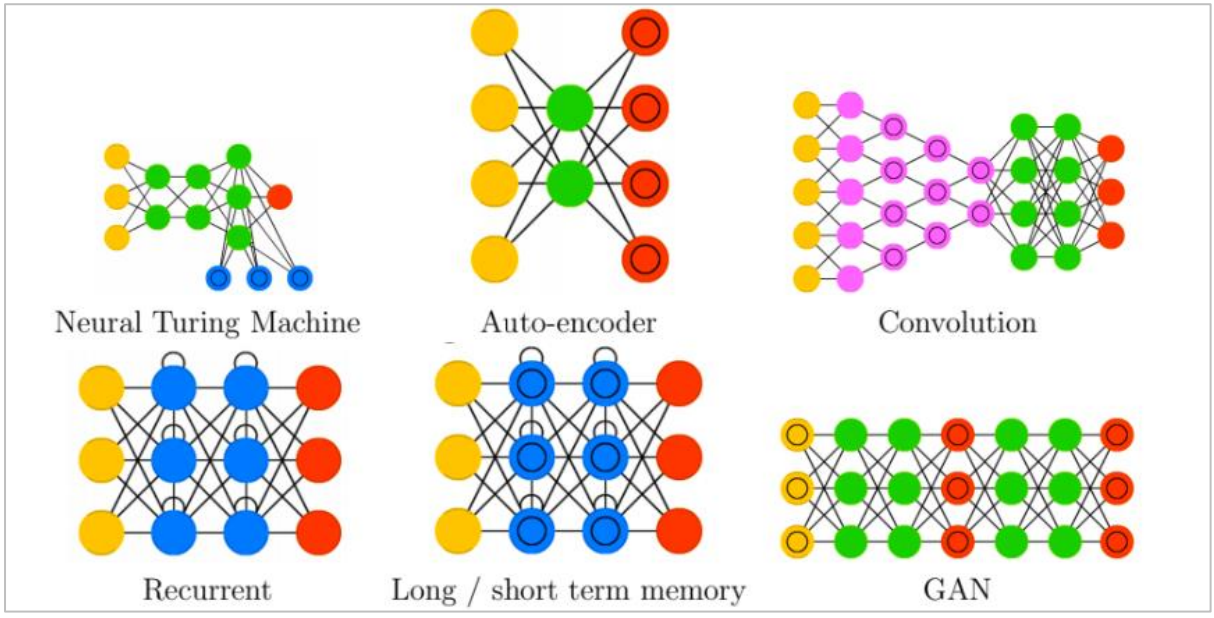

Figure 3. Commonly Used Deep Learning Architectures
Source [12, p. 5]

2.3.1. Challenges and Limitations

The most common issues with deep neural networks (DNNs) are overfitting and requirements for high computation time. The overfitting is mostly due to the added layers of abstraction which enable the system to model rare features and dependencies in the training data [21]. Overfitting challenges can be avoided by implementing various regularisation methods during training, such as batch normalisation, which aims to improve optimisation by introducing noise, dropout, which aims to remove some dimensions randomly, helping to break the rare dependencies, as well as weight decay and sparsity [21, 12, 22]. Extensive dimensions and layers that the model sweeps through can cause challenges. Processing time can be sped up through mini-batching by computing gradients on several training samples at the same time instead of individually, weight pruning to build smaller and faster equivalents of larger networks, heuristic-driven efficient architecture design and methods of automatic architecture search, replacing human heuristics design [23].

2.3.2. Bayesian Deep Learning

Connecting Bayesian probability theory with deep learning has been shown to improve its efficiency and minimise challenges of overfitting and computational efficiency. Dropout techniques, which randomly reduce certain DNN connection weights to zero to improve computing performance and reduce overfitting, have been shown to be identical to approximate inference in Bayesian Modelling. The inherent capabilities of Bayesian probability theory also help to improve the capability of DNNs to handle uncertainty and learning from small data domains [24, 25]. The introduction of Bayesian inference has been shown in studies to speed up the process of adjusting the model, dimensions and weights in the adaptation of more sparse or fit-for-purpose deep neural networks, requiring less computation and increasing performance. An empirical study of an adaptive Bayesian (AEB) sparse deep learning method adapting MNSIT with shallow convolutional neural networks (CNNs) increased its compression performance and improved the system's resistance to adversarial attacks [26]. Another study based on Parsimonious deep Neural Networks that combined Bayesian non-parametrics with a forward model selection strategy provided adaptive hidden layers whose number of

active units are automatically adjusted based on actual data, therefore reducing excessive dimensionality which consumes computation time [25].

2.4. Country Instability Prediction Based on GDELT and Related Data with Neural Networks and Bayesian Theory

Existing literature shows that the application of machine learning, Bayesian inference and deep neural networks have been applied to GDELT and related data for predicting country instability with varying results. Working with a Bayesian theoretical framework and a big open-source dataset, a fused-model approach was used to increase the ability to improve prediction precision for civil unrest — especially with information from heterogenous sources, such as using datasets from multiple countries when compared to models using information from only an individual country [27]. A convolutional neural network method was applied to moderate-resolution imagery in Nigeria to generate measures of poverty and development in developing countries. Findings from this study indicate that CNN-based methods could be used to estimate low-cost development related information which, however, does not replace traditional analysis [28]. On the other hand, a comparative study of machine learning techniques showed the random forest algorithm outperforming the Naive Bayes algorithm [29]. In summary, there appears to be very few studies available utilising the recently increased capabilities of deep learning networks to provide high performance predictions due to increased availability of big data and computational capacity [12, 13]. This provides a gap that this project will explore in its research.

2.5. Limitations of GDELT Data

It will be important when working with GDELT and related datasets to understand the limitations that are likely to influence results and which need to be addressed in the design of the research methodology. One of these is the fact that GDELT's categorised information presents a black box due to the use of non-transparent algorithms for organising available information. As a result, it is not very clear how many media outlets are monitored in the datasets [30]. The theme classifications provided can also be difficult to contextualise. As a result, the UK office of National Statistics has declared, for example, that this makes the identification of UK-based disasters from GDELT unreliable

as well as numeric information reporting deaths associated with the disaster of interest. It is therefore important that activities are undertaken to augment the quality of the information obtained from GDELT. In addition to the more diverse and nuanced data introduced by the use of World Bank and War Correlations, external validity tests using parallel data will be important. This could also include the use of maps and satellite images, as used in the empirical model construction of future conflict in Afghanistan [31].

2.6. Ethics and Human Computer Interaction AI Issues

Considering the relatively vague mechanism presented by deep neural networks built primarily on heuristics and the relatively immature stage of development in the area, it is important to consider how this impacts the research from an ethical and legal perspectives. This also applies to the use of GDELT datasets, which are primarily proprietary and non-transparent [32]. Steps need to be taken when constructing algorithms to reduce potential legal exposure due to hidden biases embedded either in datasets used or within the elements of the AI model. This includes ensuring diversity in the provision of training data — especially with the proposed use of information from a hybrid of European countries to avoid unexpected false results and in-built discrimination [33, 34].

Human-computer interaction issues are also of interest, as it is not yet clear why deep learning networks work so well, since much of their development over decades have been based on mimicking nature and using experience and trial and error to improve the system. One of the most important aspects of human-to-human interactions is the explainability of actions and at least an understanding of underlying reasons. These are important for creating sustainable intelligent systems which can provide high confidence of reliability and protection from ethical and legal issues [35]. It is thus important that these issues are addressed to avoid over-reliance on inexplicable internal workings of an AI system which may lead to an unexpected and significant black swan failure due to its black box mechanism.

3. Experiment Preparation and Design

3.1. Introduction

The proposed work is focused on pragmatist research philosophy, which is a philosophy that combines both positivist and interpretivist positions and promotes the study of multi-modal data. This is suitable for the ITFLOWS goals since the underlying GDELT data comprises both quantitative and qualitative data in a rich and structured format, necessitating a pragmatic methodology driven by the data's requirements and limitations rather than defined conceptual concepts. A pragmatic theory, in particular, promotes a deductive approach to the construction and testing of models and forecasts (see Deliverable 6.2 of ITFLOWS), an approach that is appropriate for the study's predictive goals and objectives [15]. The deductive method would also allow for the evaluation of models that can be used to explain relationships between parameters, as well as the extrapolation of data from GDELT datasets to predict country instability.

Simultaneously, the pragmatic approach seeks to move beyond the constraints of current datasets and add more detail in order to have the most context possible. Furthermore, primary research including telephone interviews with political scientists and other academics may as well be used to help sense-checking of the proposed models to ensure that they are consistent with the needs and preferences of the wider political science field.

It would be important to analyse both quantitative and qualitative data as part of the hybrid approach to data, however only qualitative data is considered in this work.

In general, the quantitative data would be secondary data obtained from the GDELT initiative, the World Bank, and other sources, with deep learning techniques used to create quantitative representations of the rich, qualitative data within these datasets. These can be used to gain insights into the existence of important environmental variables, as well as to promote the use of Bayesian techniques in the testing and refinement of predictive models. This is a powerful deductive reasoning and interpretation approach that allows for repeated testing and refinement of experimentally derived models to ensure their incremental degrees of predictive value and precision in a given study background. Simultaneously, the model's predictive

utility may be measured using mathematical research methods such as regressions and correlations to ascertain the validity of each individual variable and how well it correlates to the model's overall outcomes. The incorporation of these quantitative analysis methods, as well as their introduction into the deductive reasoning paradigm, would help to provide insight into each element in the model, allowing the model to be refined and developed in line with the deep learning methodology to support higher standards of potential validation and insight.

3.2. GDELT event database

The GDELT Event Database tracks over 300 types of physical activity around the world from demonstrations and protests through peace appeals and international exchanges. As a result, the current archive contains more than 2.5 terabytes of data per year. In terms of absolute figures, it has over a quarter billion data [16]. The website, which is powered by Google Jigsaw, contains data from 1979 to the present and is updated every 15 minutes as of April 1, 2013. In other words, documents are constantly being added to the database. The enormous number of event logs - more than any other event dataset - provides a new insight on this field of study. So far, few studies have attempted to use GDELT to forecast civil conflict, and only a few researchers have used GDELT to make forecasts.

The Global Database of Incidents, Language, and Tone (GDELT) is a modern CAMEO-coded dataset that contains geolocated events from 1979 to the present with global scope. The information was gathered from news accounts from around the world [17]. This dataset currently contains regular analysis of incidents contained in news stories released that day. The CAMEO taxonomy divides case forms into four categories: verbal cooperation and material cooperation (numbers 1 to 10) and verbal conflict and material conflict (numbers 11 to 20). Furthermore, each case has 32 separate responsibilities for the players, such as police forces, government, and military. Each record in GDELT contains details about a single incident. We use the following attributes from a case to build our models: "MonthYear, Actor1Type, Actor2Type, RootEventCode, AvgTone, and GoldsteinScale, where Actor1Type and Actor2Type store the role of the actors participating in the event", RootEventCode $\in \{1, \ldots, 20\}$ identifies

whether this incident is cooperative or contradictory, AvgTone is a subtle indicator of an event's magnitude that serves as a surrogate for its influence, and the GoldsteinScale captures the event's impact on a country's stability.

3.3. Background study

GDELT is tested to see whether it will meet the above requirements. It is a Google Jigsaw-supported case archive that's open to the public. It includes data from various large news and broadcasting agencies. The Tabari scheme, in particular, collects events from each article and stores them in an extended version of the dyadic CAMEO format, which is a conflict and resolution case taxonomy. Protests, wars, peace appeals, terrorist threats, crime, and other incidents are examples of known events. Additional software determines the position of each case, using a method similar to that used to map Wikipedia, as well as the sound, which is determined by the tonal algorithm. Many references to the same incident in one or more stories from the same newswire are merged into a single event log, but not through newswires. Data is published on a regular basis, with historical data dating back to 1979.

The development of new variables derived from the 0 GDELT event is used to calculate GPI. Official GPI variables, in particular, are reconstructed by mapping them to GDELT results, which provide similar details [6]. For example, the official GPI variable "Number of imprisoned populations per 100,000 persons" is recreated by the GDELT case categories "Arrest, detain; lawful or extrajudicial seizures, detentions, or imprisonments" and "Threatened with persecution" and is simply called "jailed" for simplicity. Nine new variables are derived from the GDELT event log, as a count of events correlated with at a country and year level, normalised to the total number of events at a country and year level, after diligent mapping. Correlation analysis is used to test the new variables and their association with the GPI official ranking. The preliminary study is carried out without making any distinctions between countries or years. The developed GDELT variables and the official GPI score have significant correlations, according to the results. The Pearson's correlation coefficients for the variables "conventional weapons" and GPI, as well as the variable "effects of extremism" and GPI, are r=0.41 and r=0.35, respectively.

3.4. Dataset

The GDELT has become a real-time archive that updates every 15 minutes and records news media events from all over the world in over 100 languages. This database contains information on over 300 different forms of physical events around the world, ranging from demonstrations and marches to city diplomatic exchanges. Approximately 60 details are captured from each particular case, including details about the venue, the people involved, and the action that was utilised. These events are then coded in the Conflict and Mediation Case Observations (CAMEO) format, which is an event coding system designed specifically for the study of third-party mediation in international conflicts.

The current research uses the data from three countries, which are Egypt, Sudan and Pakistan, however the approach can be extended to any country covered by the GDELT data acquisition. The root level category EventBaseCode is the lowest level category. For example, the code 1411 (demonstrate or rally for leadership change) is subsumed by the base code 141 (demonstrate or rally, not otherwise specified), and the base code is subsumed by the root code 14. (PROTEST). The ActionGeo CountryCode corresponds to the event's venue, which is a two-character FIPS10-4 country code. The overall tone of all records containing one or more references of an occurrence is called AvgTone. The AvgTone scores vary from -100 (extremely negative) to +100 (extremely positive). The tone of an event can be viewed as a signal that provides information about the event's effects. Furthermore, the AvgTone score only applies to the first news article mentioning the incident, which means that if an event is mentioned in many news stories, the AvgTone score will not be changed. It's also worth noting that tone can be viewed with care because it's not a test of emotion.

3.5. Data pre-processing

The first part of the work is filtering and pre-processing GDELT data so that it can be in feature engineering and labelling. The following process is applied to the raw GDELT data:

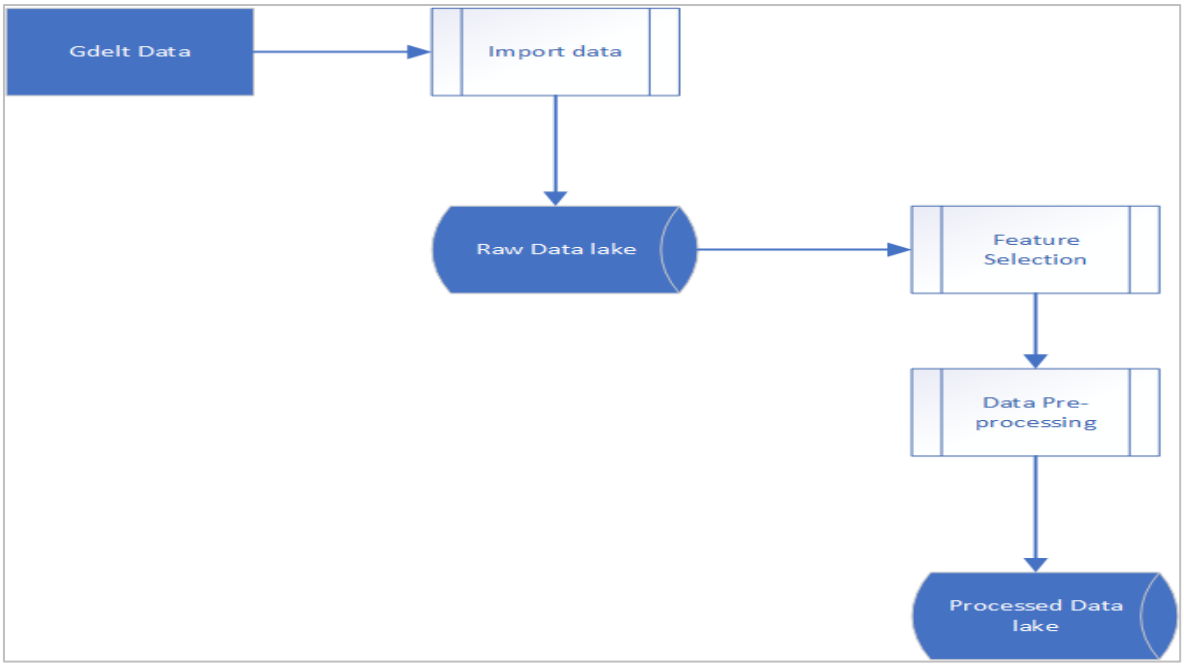

Figure 4. Data pre-processing flow

The data to be imported is stored in 1 master file that contains all pointers to 15-minute intervals payload files. Those files are zipped, hence during the import process those are unzipped as well as divided into yearly categorisation. At this stage all raw files are imported. Any subsequent import jobs run only on new data. The output of the raw import is depicted below:

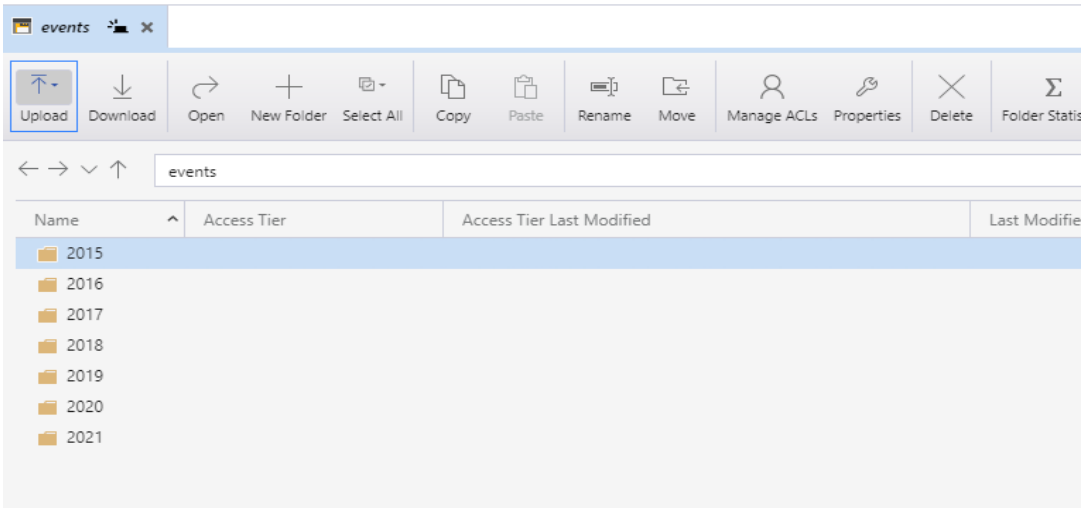

Figure 5. Pre-processed data storage

And for the purpose of this work is stored in a data lake hosted in Microsoft Azure cloud. Next steps consist of filtering the raw data as well as arranging information per country basis. Out of the following fields only small subset is taken (17 relevant fields):

```
schema = StructType([
  StructField( 'GLOBALEVENTID',  StringType(), True) ,
  StructField( 'SQLDATE', StringType(), True) ,
  StructField( 'MonthYear', StringType(), True) ,
  StructField( 'Year', StringType(), True) ,
  StructField( 'FractionDate',  StringType(), True) ,
  StructField( 'Actor1Code',  StringType(), True) ,
  StructField( 'Actor1Name', StringType(), True) ,
  StructField( 'Actor1CountryCode', StringType(), True) ,
  StructField( 'Actor1KnownGroupCode', StringType(), True) ,
  StructField( 'Actor1EthnicCode', StringType(), True) ,
  StructField( 'Actor1Religion1Code', StringType(), True) ,
  StructField( 'Actor1Religion2Code', StringType(), True) ,
  StructField( 'Actor1Type1Code', StringType(), True) ,
  StructField( 'Actor1Type2Code', StringType(), True) ,
  StructField( 'Actor1Type3Code', StringType(), True) ,
  StructField( 'Actor2Code', StringType(), True) ,
  StructField( 'Actor2Name', StringType(), True) ,
  StructField( 'Actor2CountryCode', StringType(), True) ,
  StructField( 'Actor2KnownGroupCode', StringType(), True) ,
  StructField( 'Actor2EthnicCode', StringType(), True) ,
  StructField( 'Actor2Religion1Code', StringType(), True) ,
  StructField( 'Actor2Religion2Code', StringType(), True) ,
  StructField( 'Actor2Type1Code', StringType(), True) ,
  StructField( 'Actor2Type2Code', StringType(), True) ,
  StructField( 'Actor2Type3Code', StringType(), True) ,
  StructField( 'IsRootEvent', StringType(), True) ,
  StructField( 'EventCode', StringType(), True) ,
  StructField( 'EventBaseCode', StringType(), True) ,
  StructField( 'EventRootCode', StringType(), True) ,
  StructField( 'QuadClass', StringType(), True) ,
  StructField( 'GoldsteinScale', StringType(), True) ,
  StructField( 'NumMentions', StringType(), True) ,
  StructField( 'NumSources', StringType(), True) ,
  StructField( 'NumArticles', StringType(), True) ,
  StructField( 'AvgTone', StringType(), True) ,
  StructField( 'Actor1Geo_Type', StringType(), True),
  StructField( 'Actor1Geo_FullName', StringType(), True) ,
  StructField( 'Actor1Geo_CountryCode', StringType(), True) ,
  StructField( 'Actor1Geo_ADM1Code', StringType(), True) ,
  StructField( 'Actor1Geo_ADM2Code', StringType(), True) ,
  StructField( 'Actor1Geo_Lat', StringType(), True),
  StructField( 'Actor1Geo_Long',  StringType(), True) ,
  StructField( 'Actor1Geo_FeatureID', StringType(), True) ,
  StructField( 'Actor2Geo_Type', StringType(), True) ,
  StructField( 'Actor2Geo_FullName', StringType(), True) ,
  StructField( 'Actor2Geo_CountryCode ',StringType(), True) ,
  StructField( 'Actor2Geo_ADM1Code', StringType(), True) ,
  StructField( 'Actor2Geo_ADM2Code', StringType(), True) ,
  StructField( 'Actor2Geo_Lat', StringType(), True),
  StructField( 'Actor2Geo_Long', StringType(), True) ,
  StructField( 'Actor2Geo_FeatureID', StringType(), True),
  StructField( 'ActionGeo_Type', StringType(), True) ,
  StructField( 'ActionGeo_FullName', StringType(), True) ,
  StructField( 'ActionGeo_CountryCode', StringType(), True) ,
  StructField( 'ActionGeo_ADM1Code', StringType(), True) ,
  StructField( 'ActionGeo_ADM2Code', StringType(), True) ,
  StructField( 'ActionGeo_Lat', StringType(), True) ,
  StructField( 'ActionGeo_Long', StringType(), True) ,
  StructField( 'ActionGeo_FeatureID', StringType(), True) ,
  StructField( 'DATEADDED', StringType(), True),
  StructField( 'SOURCEURL', StringType(), True)
])
```

Figure 6. GDELT data structure

The output of this phase is set of files, one per country, that contain the relevant information to the analysis descried in this work.

3.6. Feature engineering

After the data is pre-processed the next tasks of feature engineering and data labelling is performed. This process is described in the diagram below:

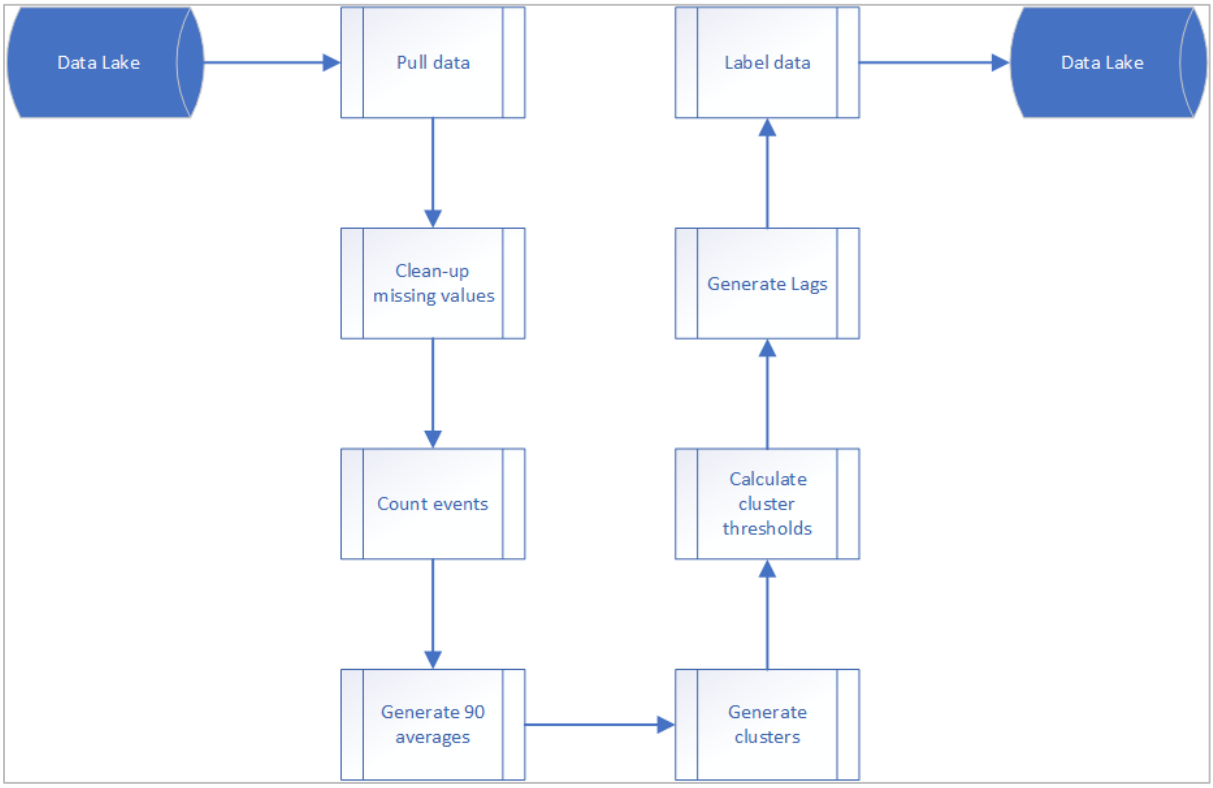

Figure 7. Feature generation flow

Data is pulled from the initial data lake and cleaned up for missing values. Given that GDELT is using machine algorithms to segregate the events, some of the data may be missing at this point. The next important step is to generate the count of events per day. GDELT creates one file every 15 minutes, hence there is a need to bring that data to daily granularity level.

Next important task is to acknowledge that social unrest does not happen on a single day and that process usually builds up gradually over an extended period of time. For this reason, the 90-day moving average is calculated as a base for clustering the events over time.

At this stage the mean clustering technique (MCT) is used to analyse the daily events. For the purpose of this work only 1 CAMEO code is used (#14), however that could be extended in the future work to not include root codes, but specific ones from 10, 11, 13 and 14 categorisations [18]–[20].

It is important to mention that the cluster calculations do not operate on the single days, but on the intervals. The in 3 to 7-day intervals were considered and the best results were obtained using 3-day intervals.

The key part of feature generation is to calculate the cluster thresholds for the particular events. The following formula is used to calculate [20]:

$$\theta = MCTBarComp + 2.576 \times \sqrt{-((MCTBar - MCTBarComp)^2)}$$

This value is probably the most important one in feature generation as it will be used to segregate the events into possible unrest or not.

Where MCBarComp is calculated in the following manner:

$$MCTBarComp = -(MCT / MCTBar)$$

One the primary clustering is obtained; the next step is to add lags to the feature list to enable step ahead predictions. A lagging indicator is an apparent or detectable element that varies after the associated economic, financial, or market variable changes. Trends and shifts in trends are confirmed by lagging measures. Lag generation for rolling averages and future predictions is a naive and effective technique in time series forecasting. It can be used for data preparation, feature engineering, and even directly for making predictions. Hence, lags can be used to predict a variable in the sequence. Given that lags are interpolated from the existing data, they are prone to rapidly decreasing accuracy. For the purpose of this work the lag of maximum 7 days was used.

3.7. Labelling

Once the preliminary processing is performed with respect to arrangement through the dates and clusters, the work can move to labelling the data. At this point it is wort mentioning that the analysis can be performed both on binary labelled data (focus of this work) as well as the threshold analysis (δ). Labels are generated based on cluster distance thresholds based on the following:

MCT_Comp > δ -> unrest s assumed, otherwise it is not.

Once this is executed, the processed and labelled data is then stored into a new CSV file for further analysis.

3.8. Classification

The classification algorithm used here is a Random Forest algorithm, which is an ensemble learning-based supervised machine learning algorithm. The random forest algorithm incorporates many algorithms of the same kind, such as multiple decision trees, to create a forest of trees, thus the term "Random Forest." Both regression and classification tasks will benefit from the random forest algorithm. In the case of a regression problem, each tree in the forest predicts a value for Y for a new record (output). The final value can be determined by averaging all of the expected values from all of the trees in the forest. Alternatively, each tree in the forest predicts the group to which the new record belongs in the case of a classification query. Finally, the current record is given to the group that receives the greatest number of votes.

When you have both categorical and numerical elements, the random forest algorithm works well. When data has missing values or has not been scaled well, the random forest algorithm works well (although we have performed feature scaling in this article just for the purpose of demonstration). The code below extracts data for the countries mentioned below. We don't know which includes the data for which country right now. As a result, the first step removes both the code and the country name. The second step is focused on locating the region.

The Random Forest algorithm must be trained so that it would be able to learn the patterns between the parameters and its values. Since the labelled data has been created, it will be used for training the classifier. The data is split into training dataset and testing dataset based on the dates. When Pakistan is considered, it contains 2222 instances of data out of which 65% of the data is split into the training data, while the remaining 35% is designated as testing data. The data split is based on the following principles:

- Events from 2015-2018 are used as training data
- Events from 2019 are used as test data
- Events from 2020 and 2021 are discarded to eliminate potential anomalies based on the outbreak of COVID-19 pandemic

The trained data is sent to the classifier, where it learns the data and generates a training model. The random forest classifier now uses this training model to evaluate the algorithm and predict the stability of the testing dataset. The algorithm is evaluated with respect to accuracy and mean absolute error.

4. Implementation and Results

4.1. Dynamics of Social Unrest

The aim of this section is to investigate the dynamics of civil war that occurred in Eurasia. We will do so by illustrating how social unrest and state policy react to local and foreign shocks using the Impulse Response Analysis of the VAR model. This study will be carried out on a regional and sub-regional scale (West-Central-East Eurasia). The Eurasian region includes Bulgaria, Moldova, Ukraine, Belarus, Serbia, Georgia, Armenia, Azerbaijan, Kazakhstan, Kyrgyzstan, and Uzbekistan [21].

This group of countries was chosen because it represents a wide variety of social unrest periods, from minor civil movements (such as the Rose, Orange, and Tulip Revolutions) to more conflict-like events like political, religious, and ethnic conflicts (Karabakh conflict in Armenia, Abkhazia conflict in Georgia, etc.). We also removed Russia from consideration due to its geographic dominance and the border-shifting dynamics of world events [22]. The control area will be the Middle East and North Africa, to which the findings will be compared. The MENA area includes Algeria, Libya, Egypt, Tunisia, Morocco, Syria, Israel, Jordan, Turkey, Iran, Bahrain, UAE, Saudi Arabia, Qatar, Oman, and Iraq. Both, we conclude, have a rich and comparable case history that can be studied using the civil unrest taxonomy described earlier. To cast events, we just used a timeline from January 1995 to February 2015 [23].

Prior to 1995, social events such as the breakup of the Soviet Union and the emergence of Post-Soviet states heavily influenced the results (1991-1994). The study's results will be described in terms of the following concepts, which will help to systematise the stylised details of social unrest interaction and State response in the region and are broadly consistent with the three social unrest paradigms [24]:

1. Stronger shocks are registered by more volatile indexes, because a society's instability can be measured in terms of the frequency and resiliency of the shocks it generates. This concept is analogous to the philosophy of social inertia that was previously discussed.

2. The strength and resiliency of the responses in the rest of the shock ladder variables, such as vindication triggering protests and protest causing violence, was measured in terms of a. Escalation potential, or the severity and resiliency of

the responses in the rest of the shock ladder variables. b. As the reaction progresses down the shock ladder, it has the potential to become self-reinforcing or feedback powerful (protest creating vindication, conflict creating protest and vindication) [25]. Both concepts are linked to the principle of interdependence of unrest variables in the Lifecycle Theory of Unrest.

3. Spill over of social instability dynamics, as shown by similar sign responses of unrest variables to shocks in neighbouring variables (contagion of unrest) or replication of the State Response to Civil Unrest within countries (mimic of policy response) [26].

4. Policy should be used to uphold the rule of law. It may be coercive (repressive) or deceptive (repressive) (accommodative). For coercive/cooperative behaviour, the State Response Index has a positive/negative symbol. As previously stated, the definition of this concept corresponds to a broad body of research on the rewards for state repression or accommodation [27].

5. The government's willingness to uphold the rule of law, or how it reacts to minimise the amount of unrest (exhausting protests or ending conflicts, for instance) We believe that these concepts are adequate to characterise the most important stylised features of unrest dynamics in Eurasia, and that they are compatible with the core theoretical rationalist of social unrest as defined [28].

Men and women's overall confidence in people increased by around ten percentage points on average. Around 18% – 19% of people said they had increased their confidence, while 7% – 8% said they had decreased their trust. The results of multivariate regression show that men's confidence increased overall in areas where some event occurred, but women's trust decreased; nevertheless, neither estimate is statistically significant. A curvilinear relationship arises based on the number of events, with men's confidence increasing dramatically for the first 2-4 events before declining as the number of events increases. Similarly, with more than five cases, women's confidence rises at first, but then drops dramatically. Using the occurrences of any deaths during the demonstrations or disturbances as a metric, no results were observed. Uncertainty levels rose by 0.5 points on average for men and 0.5 points on average for women between 2009 and 2014 [6].

The incidence of any event increased men's and women's recorded uncertainty, but the difference was not statistically significant. As compared to no events, the occurrence of only one event dramatically increased men's uncertainty, with the effect size decreasing as the number of events increased; similarly, the occurrence of 2-4 events increases women's uncertainty, with the effect size decreasing as the number of events increased. There is a substantial increase in men's insecurity in places where all incidents happened without a death; the result is also positive for women, but not statistically significant. Human resources are a valuable resource [29]. Both human capital outcomes were subjected to regression analysis. Education is quite important. On-time school grade completion has decreased significantly for both men and women: in 2009, 70% of men were enrolled in the appropriate grade for their age; in 2014, only 22% were in the appropriate grade; similarly, in 2009, 65 percent of enrolled women were on-time; in 2014, only 24 percent were on-time. Half of all school-aged men and 44% of all school-aged women were behind in their grades. The incidence of any demonstration or riot is correlated with a lower probability of delayed school completion for both men and women, according to regression findings, but this association is not statistically important. Nearly all effects are near zero and not important for the number of accidents and events with deaths [30].

In 2009, the vast majority of students (72 percent-74 percent) missed at least one day of school, a figure that has decreased by 14 percentage points for men and 9 percentage points for women over time. For men, there are no consequences of protests and disturbances on the risk of being absent. With the incidence of 2- 4 events, there is a substantial improvement in the probability of absence for women, which stays positive with 5-19 events before decreasing again. The number of days away from school decreases over time, with men missing an average of 4.2 days and women missing an average of 2.8 days. For any indicator of protest and riot cases, no substantial effects are estimated. Individual answers to the SRQ-20 mental wellbeing index products [31].

Young men and women are more likely to experience headaches, poor appetite, poor sleep, nervousness, trouble thinking straight, feeling unhappy, and difficulty making decisions. In general, young people have a higher prevalence of favourable responses to each item than young men, resulting in a mean overview ranking of 2.2 for young men and 4.9 for young women in 2009, indicating that young women's mental health is

categorically worse (i.e. higher mental health score). The mental wellness summary index's changes over time show that, on average, everyone's mental welfare improves (i.e., the index score decreases) [32]. While women's relative improvement in the overview index is higher than men's, closing the gender divide, women's mental health continues to deteriorate in 2014, with the average index score falling to 1.6 for men and 2.7 for women. Men's mental health scores jump (i.e. worsen) when there is a demonstration or riot, and when there is a death, mental health scores increase even more. Women's symptoms are normally the polar opposite of men's. There are no numbers that are scientifically accurate [33].

We exclude women from further study of these outcomes due to the low recorded prevalence of smoking, drinking, and substance use among women. Men's smoking rates rise the most over time (27.6 percent to 35.5 percent), while 12.4 percent quit and 20.3 percent start. Overall, only 2.2 percent and 4.1 percent of men self-report drinking and using drugs, respectively, and only 2.2 percent and 4.1 percent start drinking or using drugs for the first time. The predicted results of demonstrations and riots are near zero and negligible for smoking and narcotics, according to regression estimates. Protests and riots, on the other hand, have a negative cumulative impact on the probability of drinking, and this effect is substantial with the frequency of any incident and 5-19 incidents [34].

4.2. Moderation Analyses to Better Understand Which Populations of Youth Drive the Results

Exposure to protest increased young men's reports of feeling unsure about the future in both the urban and rural samples. The effect sizes in the urban sample were greater in magnitude. There were no consistent major effects estimated for the other outcomes for the rural and urban samples (available on request). We also look at indicators of protest engagement as an impact moderator. Just 6% of men and 2% of women participate in some kind of political action. Between 2011 and 2013/2014, 2.1 percent of people said they participated in demonstrations. We believe that youth participation rates are underreported as a result of the tense political environment that prevailed during the fieldwork for the 2014 Survey of the Young People in Egypt (SYPE) [35]. Young people recorded much higher rates of involvement in political engagement of their siblings (6.4

percent), parents (5.4 percent), and close friends (14.3 percent) during the same period, indicating a clear indicator of this possible response bias. As a result, we construct and regression specifications for binary and categorical outcomes using logistic linking functions [36].

Since 2011, Egypt has seen a drastic rise in the incidence of civil instability, a once-in-a-century historical phase in which a generation of young people is moving to adulthood. In this paper, we explore whether exposure to political violence – as calculated by living in a neighbourhood where demonstrations and/or disturbances have occurred – has an effect on these young people's mental wellbeing, educational investments, or risky behaviour. The preliminary findings do not support our initial hypothesis that exposing young people to protest activities will lead to more negative outcomes [37]. Some of the effects are predicted, but others are not, and only a few are statistically relevant. Exposure to protest events does have an impact on affective outcomes, especially among young men's anxiety about the future. The amount of protest or riot incidents that occurred in the district, our estimate of the degree of exposure to political unrest, also shows some variance in the results. Men and women experience different results, as predicted. The next step in the study would be to look at more subgroup disparities, such as those between urban and rural youth, as well as to integrate a measure of young people's own involvement in demonstrations from the SYPE to see if this has a mediating impact on their outcomes. Exposure to political instability having no impact may also be due to the time frame in which the SYPE data was collected [38].

At a minimum, we would like outcome data obtained at each stage of regime change over the longer duration of political transition in order to quantify the effects of exposure. Young people's reactions to the January 25th revolution, for example, may have been somewhat different from their reactions to President Morsi's regime change, which occurred shortly before the SYPE was launched. Exposure to the demonstrations that brought about these various regime shifts may have been a promising or inspiring experience for certain times and sub-groups, leading to optimistic responses or better mental wellbeing – at least in the short term [30]. Furthermore, except events recorded in non-English media, which may be more locally relevant, may bias events in the ACLED. Estimates could be skewed against a null impact if ACLED events reflect higher-profile events reported by international media. Due to the lack of GPS coordinates in the

SYPE info, we are also unable to calculate proximity to protest events. Since the scale of Qisms/markaz (which translates to centre) varies significantly, the youth in the SYPE are likely to live at varying distances from any events that did occur in their region. In the final version of the manuscript, we will take these variables into account, as well as the additional studies listed above [39].

4.3. Current Event Coding Projects

Two recent data collection efforts have significantly increased the reach of news services used and provided global event data coverage for a variety of events. Furthermore, since these datasets are modified in real time, they can be used for real-time convict analysis. With DARPA's support, the Integrated Crisis Early Warning Systems (ICEWS) expanded earlier event coding structures by incorporating a wide variety of news sources. Lockheed-Martin currently maintains the ICEWS data, and a portion of it was recently made available to the public on Harvard's Dataverse. One of the problems with ICEWS is that the code used to generate event data from news reports is proprietary. As a public and more open version of ICEWS, the GDELT was developed. It was widely praised when it was released in April 2013. One of Foreign Policy magazine's authors was named one of the top 100 global thinkers [40]. Unfortunately, controversy over how it received several of its news resources stifled academic interest in the project, prompting some of its co-authors to withdraw their support. The data, on the other hand, is still used to analyse foreign events, is still relevant in public policy circles, and has recently been integrated into Google's services. Although the legal concerns concerning GDELT are unclear (one author's request for clarification was met with an ambiguous answer that it's a touchy subject), it appears that one of the project's key developers might have used proprietary tools purchased by the University of Illinois' Cline Center for creating the SPEED dataset.

It should be noted that we only compare GDELT to other event data projects using aggregate historical data, and we only use publicly accessible papers for our validity study. We have not accessed any copyrighted materials purchased by the Cline Center or GDELT in any way. There is no pending litigation involving the use of GDELT that we are aware of, and the data has since been re-established on the network. Government

agencies, Google, and other publications have all used it, and it continues to generate reports for Foreign Policy and other publications. Nonetheless, the GDELT controversy highlights the importance of providing a corpus that event data teams can openly use. A coordinated effort, led by a government agency or consortium, to collect such a corpus would help to avoid the usage restriction issues that prohibit all but the most well-funded teams from working on event data initiatives, as well as the legal uncertainty that sparked the GDELT debate. Unlike several text research ventures, such as reviewing Shakespeare's collected works, copyright concerns loom large in news article analysis. The GDELT case highlights how difficult it can be to navigate some of these problems [30].

4.4. An Interpretation of GDELT Sources

We noticed that, in addition to relying on an English-language corpus, the GDELT is also heavily dependent on a few sources within this corpus. Our review of the news outlets mentioned by GDELT reveals that the majority of the events in GDELT are contributed by a limited number of domains. More than 80% of the events in our experiment collection are contributed by the first 10% of domains. The distribution of events to realms is power law. The majority of domains produce a small number of events, but certain sites generate a large number of them [41].

4.4.1. Event Deduplication

Duplication of events is a significant issue. When an incident is recorded by multiple media sources, GDELT often encodes it as multiple events. When one or more high-profile accidents occur, the duplicates would invariably offer an innate approximation of the actual events. When incidents gain more public coverage, the crisis worsens. Even when multiple events are extracted from a single URL, there are always duplications in some cases [42].

Furthermore, successive records for the same occurrence can be encoded as multiple events in some cases. We suggest a framework for event deduplication based on correlation between events. If the similarity of two occurrences exceeds a certain threshold, we consider them to be duplicated. Each event has features such as

eventDate, ActorCode, ActionType, ActionGeoFullName, and event sentences. We create a time frame and then group all of the events in that time window by their locations. The similarity for each pair of events in the position group is then computed, and the duplicated events are removed using a greedy technique. With a similarity threshold of 0.8, we get 113,932 specific events out of 178,987 total [43].

4.4.2. Reliability

The first set of studies looks at event data accuracy, especially whether programmes with ostensibly identical coding rules yield similar outcomes. We used data from four different sources, all of which were designed to identify protests [ICEWS, GDELT, Gold Standard Reporgrot (GSR), and Social, Political, and Economic Event Database (SPEED)]. GDELT and ICEWS are completely integrated applications that represent the most up-to-date attempts at real-time global event data. Since 2011, the non-profit MITRE Corporation has been hand-coding GSR data from local and international newswires in Latin America. The University of Illinois created SPEED, a semi-automated global event data system that detects events using a mixture of human and automated techniques [44]. It makes a point of bragging about how precise the case coding is. GSR and SPEED were created to give people a sense of fact, but scaling up their approaches would be complicated and expensive. About the fact that these systems have different origins (for example, ICEWS was designed to encode geopolitical exchanges, mostly among nation-states, while GSR was designed to focus on tactical, local issues), we expect that the time series of events produced by these projects would have a high correlation, even if the event counts are not comparable. Finally, most people believe that there is a weak association between case datasets. [Correlation coefficient (r) 0] [Correlation coefficient (r) 0] [correlation coefficient (r) .3] [45].

The average correlation between GDELT and the GSR in for example Asia is 0.222, although the average correlation between ICEWS and the GSR is 0.229. SPEED and GDELT records only match 17.2 percent of the time (i.e., both datasets registered a protest on the same day). SPEED and ICEWS agree on just 10.3 percent of the cases. With an average similarity of 0.317 across Asian countries, ICEWS and GDELT are scarcely in agreement. These ties vary considerably between countries and get stronger as the number of events rises dramatically. For example, both ICEWS and GDELT

capture the massive rise in protests in Venezuela in January 2014. When time scales are roughened, they shift as well (from daily to weekly or monthly). Reliance on English language news media results in greater comparisons with states that gain more exposure in the Western press (e.g., Brazil) [46].

4.4.3. Validity

To assess the authenticity of event-coding projects, we used a special function of the GDELT data set—the degree to which they react to unique real-world events. After its inception on April 1, 2013, GDELT has provided URLs for the majority of its coded activities. We looked at all protest events that occurred between July 2, 2015 and July 2, 2018 (431,549 documents), collected material from records that had a legitimate URL (344,481 records), and iterated them to see if their classification as protest events was right. This yielded 113,932 distinct, non-duplicated instances, all of which were almost probably about protests at the time of publication. Just 49.5 percent of these listed documents are labelled as referring to actual demonstrations, which is comparable to what we found in 1000 human-coded records. After keyword and temporal filtering, de-duplication of events, and machine learning sorting of real events from non-events or predicted events, only 21% of GDELT's legitimate URLs indicate a true protest occurrence. Although the ICEWS scheme was more reliable (approximately 80% of keyword filtered cases were classified as protests), duplicate events remained an issue [47].

4.4.4. Information Gain

The primary goal of developing a decision tree algorithm is to obtain information. The decision tree algorithm will always try to maximise knowledge gain. The highest attributes will be evaluated first by the knowledge gain. Shannon coined the term "entropy" to describe the measurement of impurity in an input set. The entropy of a decision tree determines how it wants to divide the data. Impurity in a series of instances is the subject of information theory. The reduction in entropy is known as knowledge acquisition. The difference between entropy before splitting and average entropy after splitting the dataset based on attribute values is known as the information gain [48].

4.4.5. Gain Ratio

The knowledge gain ratio of decision tree learning is the ratio of information gain to essential information. To minimise a stigma toward multi-valued characteristics, it is recommended that branch size and number be considered when selecting a characteristic. The trait of information acquisition is distorted in some results. It means that attributes with a lot of different qualities are preferred. The splitting attribute has the highest gain ratio [49].

4.4.6. Gini Index

Gini's approach used a decision tree algorithm classification and regression tree to produce break points (CART). The dividing characteristic for the selected subset in the case of a discrete-valued attribute is a subset of the lowest possible Gini index. When dealing with continuous-valued characteristics, the technique is to choose a possible split point for each pair of adjacent values, with the point with the lowest Gini index being selected as the split point. The splitting function selects the lowest Gini index attribute [50].

4.5. Unrest Modelling and Prediction

4.5.1. Feature Selection

Feature selection has been used to increase precision, minimise overfitting, and shorten training time. The random forest algorithm is used to select features in this analysis. The process chooses the data attributes that have the greatest influence on the prediction attribute automatically. The random forest algorithm is used in the study because it only uses a small subset of features rather than an all-features model, and it outperforms other methods. The definition of knowledge theory is used by the random forest method in the mathematical theory of communication to select the most important attribute by looking at a prediction variable. The Gini criterion was also used to reduce the likelihood of misclassification [51]. The following table represents the features and their respective relevance (top 20):

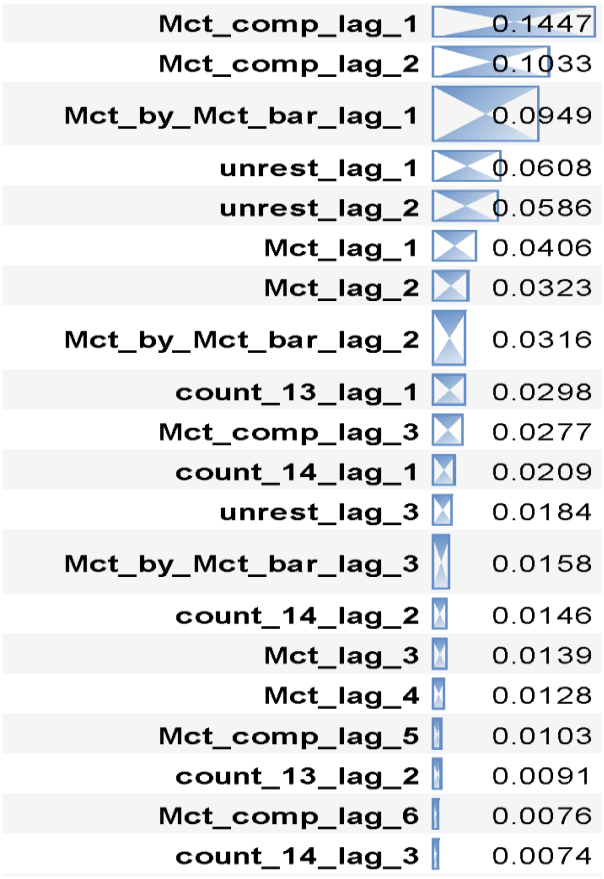

Figure 8. Feature significance

4.5.2. Training & Testing

The pre-processed dataset is split for the training and testing phases after pre-processing. Training has a 65 percent split, while testing has a 35 percent split. A total of 2222 rows is used in the dataset (for an example of Pakistan). The total number of rows used for training is 1458, while the total number of rows used for testing is 764. The pre-processed 7 columns are used as input variables for training and research [52]. Five function hashed columns and two data scaled columns make up the input variable's seven columns. The Random Forest classifier is used for both training and research (although other methods were used for performance baselining). The base model in the stacking technique is a decision tree, and the meta model is a Support vector machine. Once the preparation and research yielded a satisfactory outcome in terms of performance metrics and uncertainty matrix, it was time to put it all together [53].

4.5.3. Performance Metric

The learned model's efficiency is discussed in detail using an output metric. The intrusion's efficiency, behaviour, and actions in the network are evaluated using the Performance metric. According to the performance metric, the precision is 85 percent, with a mean absolute error of 0.15 degrees when looking at the current data and 75% when looking at interpolated data (looking ahead forecasting).

As a result, we can be certain that the algorithm has been sufficiently trained and that the final product is both accurate and effective.

4.5.4. Comparison of the Different Classifier Performance

The proposed method produced excellent results, so we compared the accuracy of the random forest method to that of three other feature extraction classifiers to illustrate the applicability of the random forest method. The following methods were used as comparisons:

- Gaussian
- SVC
- KNN
- Decision tree
- Neural Network

As a result, we can conclude that the random forest approach is the best approach for this analysis. It not only produces stronger and more reliable results, but it also has a fast-training speed, which is critical in practice [54]. The below tables shows a detailed comparison of overall classifier results sorted by accuracy for 3 example countries: Pakistan, Egypt and Sudan:

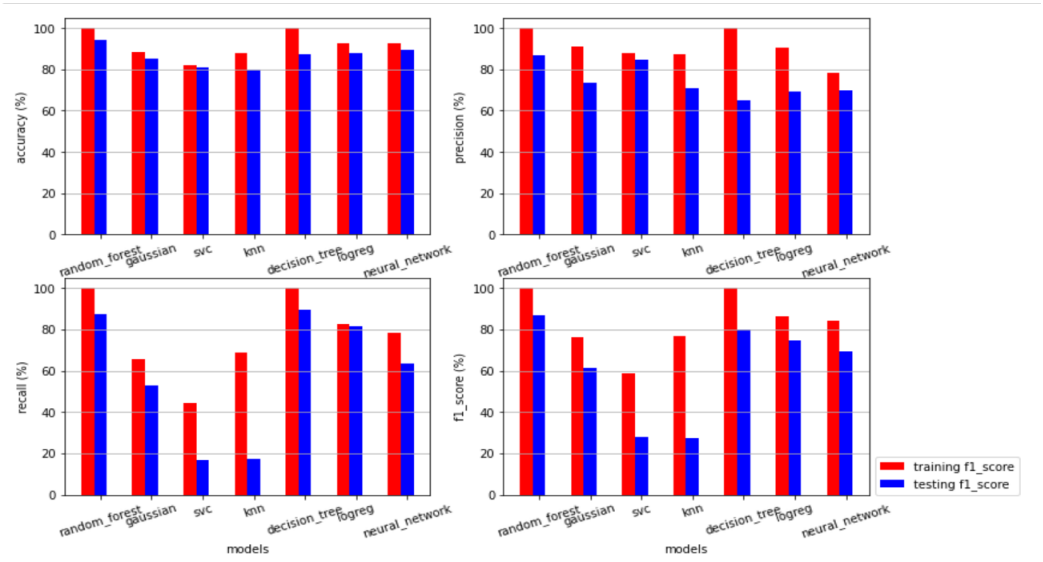

Figure 9. Accuracy comparison for Pakistan

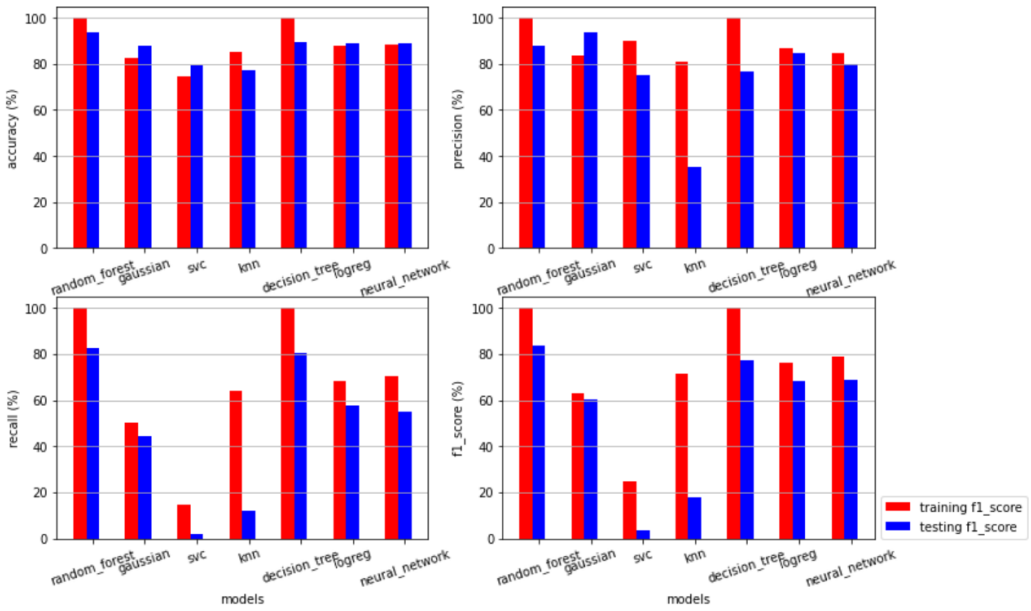

Figure 10. Accuracy comparison for Egypt

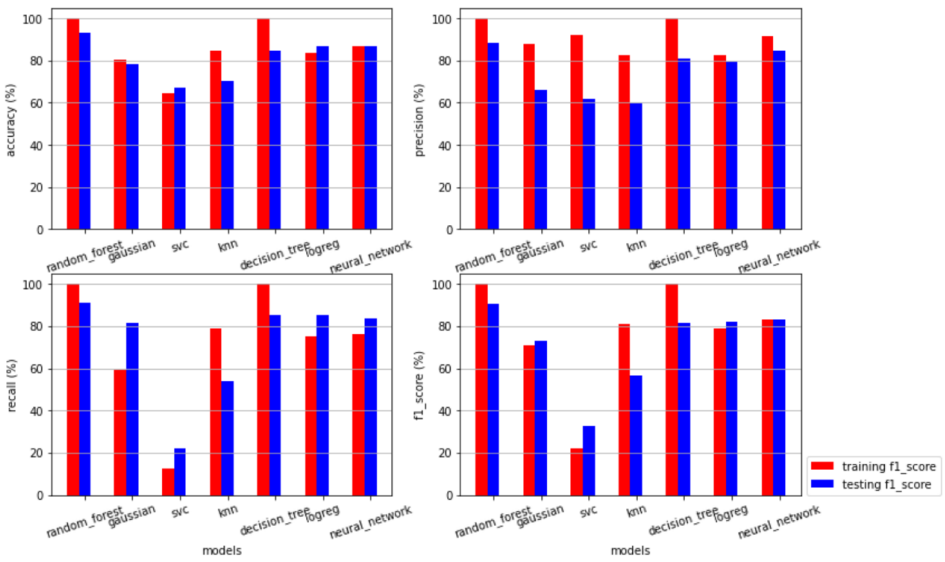

Figure 11. Accuracy comparison for Sudan

When compared to the other classifier models, the bar plots in Figures [9,10,11] clearly shows that my proposed model accuracy is very efficient and reliable. The rest of the models are lagging behind. As a result, we can conclude that the model can effectively define the probability of social unrest given the data and clustering applied.

5. Conclusions

The key contribution of the presented work is the creation of a framework for developing alternative economic and financial metrics that catch investor sentiments and topic popularity using GDELT, the Global Data on Events, Location, and Tone database, a free open portal that contains real-time world's radio, print, and web news. This research is being carried out as part of a project aimed at developing more accurate forecasting approaches for analysing the overall sector of a few countries. We have posted some early results from using this method to predict existing market conditions. This application shows that the technique works well at first, indicating that the method is right. Using the information obtained from the news media contained in GDELT, together with a deep Long Short-Term Memory Network opportunistically educated and tested using a rolling window system, we were able to achieve very good forecasting performance [55].

To summarise, policymakers could now use the experimental conflict case modelling technique applied to the GDELT dataset to map deteriorating or de-escalating circumstances in a country on a weekly or monthly basis. However, event-based frameworks will need further research to compensate for the databases' flaws, such as automatic data confirmation, new classifiers and dictionaries that represent the changing nature of violence, and, most importantly, evidence on the mechanisms between civil instability and violent conflict [58].

Working with GDELT required a considerable amount of acquired informal knowledge about its quirks and shortcomings, which was mostly obtained from the large and involved GDELT user community. Future databases should be even more transparent about how their components are produced, especially for developments that go beyond standard actor and verb dictionaries. Although computer coding cannot absorb all of a specialist's expertise, it can be improved by scholars contributing to dictionaries and coding schema, resulting in potentially more useful datasets for researching political mobilisation. Datasets should preferably include URLs for all events and easily replicable coding schemes to enable end users to understand how the data was generated and the results of improvements in dictionaries, coding algorithms, and sources. To summarise, we conclude that while using vast volumes of machine-coded

data to study phenomena such as civil society, political mobilisation, and government repression should be approached with care, it can be a useful analysis method. Indeed, the current widespread use of machine-coded event data for conflict forecasting can be expanded to include a much wider variety of comparative questions and approaches, which would be made possible by increased clarity and accuracy.

Machine coding of case datasets has the benefit over human coding in that it can process a vast number of files in a limited period of time. In this short essay, we looked at how one of these case datasets—GDELT—can be used for geo-spatial analysis at the subnational level. Machine-coded event data has traditionally been used to research international affairs and has proved to be accurate and effective in this sense. However, there have been few attempts to see whether focusing on machine code produces data that is equally suitable for subnational studies. According to findings, there is a noticeable difference between data that has been human-coded and data that has been machine-coded. We show that this is mostly due to geo-localisation problems. Despite the fact that GDELT tends to monitor the temporal ups and downs of conflict as identified by human-coded datasets, it overcounts incidents in more distant areas by clustering a disproportionately significant number of them near a country's capital.

This may be a point of interest for crime geospatial research. If we cannot be confident that the spatial specificity of events is within reasonable limits, machine-coded case datasets can be difficult to use for fine-grained analyses of the essence of violence on the field. However, we are hopeful that further research will be able to resolve these problems. As a first step toward greater consistency in the machine-coding process, datasets may provide pointers to the initial papers used to code a case. More in-depth validation analyses of both automatic geocoding and event content coding would be possible as a result of this [59]. Despite the fact that GDELT is linked to trends in other datasets, there is always more space for development. Users will refer back to the original articles using the dataset's trace back details to see whether, for example, GDELT's coding for "protest" applies to the kind of "protest" they're interested in. Overall, we agree that GDELT should be used to complement rather than replacing existing event records. Because of the high degree of noise in the GDELT data and the regional accuracy problems we discovered, we believe that using GDELT instead of a more detailed hand-coded dataset to describe spatial complexities of civil war conflict

can lead to distorted or incorrect inferences. This could (and should) be the path event data collection goes with further focus on automatic coding refinement [60].

As a conclusion, it is worth correlating the predicted results to the actual events occurred.

6. References

[1] F. Qiao, X. Zhang, and J. Deng, "Learning Evolutionary Stages with Hidden Semi-Markov Model for Predicting Social Unrest Events," *Discrete Dynamics in Nature and Society*, vol. 2020, pp. 1–16, Oct. 2020, doi: 10.1155/2020/3915036.

[2] H. Wang, B. Zhou, Z. Gu, and Y. Jia, "Contextual Gated Graph Convolutional Networks for Social Unrest Events Prediction," in *2020 IEEE Fifth International Conference on Data Science in Cyberspace (DSC)*, Jul. 2020, pp. 320–325, doi: 10.1109/DSC50466.2020.00056.

[3] S. M. Fast, L. Kim, E. L. Cohn, S. R. Mekaru, J. S. Brownstein, & N. Markuzon, "Predicting social response to infectious disease outbreaks from internet-based news streams," *Annals of Operations Research*, vol. 263, n. 1–2, pp. 551–564, Apr. 2018, doi: 10.1007/s10479-017-2480-9.

[4] A. Okutan, G. Werner, S. J. Yang, and K. McConky, "Forecasting cyberattacks with incomplete, imbalanced, and insignificant data," *Cybersecurity*, vol. 1, no. 1, p. 15, Dec. 2018, doi: 10.1186/s42400-018-0016-5.

[5] K. Buckingham, J. Brandt, W. Anderson, L. F. do Amaral, and R. Singh, "The untapped potential of mining news media events for understanding environmental change," *Current Opinion in Environmental Sustainability*, vol. 45, pp. 92–99, Aug. 2020, doi: 10.1016/j.cosust.2020.08.015.

[6] V. Voukelatou, L. Pappalardo, I. Miliou, L. Gabrielli, and F. Giannotti, "Estimating countries' peace index through the lens of the world news as monitored by GDELT," in *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, Oct. 2020, pp. 216–225, doi: 10.1109/DSAA49011.2020.00034.

[7] L. Huang, H. Yuan, J. Chen, and L. Deng, "Short-term Forecast and Analysis of Mass Incidents Based on Time Series Model," in *2018 14th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, Jul. 2018, pp. 787–791, doi: 10.1109/FSKD.2018.8687248.

[8] J. Ponticelli and H.-J. Voth, "Austerity and anarchy: Budget cuts and social unrest in Europe, 1919–2008," *Journal of Comparative Economics*, vol. 48, no. 1, pp. 1–19, Mar. 2020, doi: 10.1016/j.jce.2019.09.007.

[9] K. Pogorelov, D. T. Schroeder, P. Filkukova, and J. Langguth, "A System for High Performance Mining on GDELT Data," in *2020 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, May 2020, pp. 1101–1111, doi: 10.1109/IPDPSW50202.2020.00182.

[10] O. Schaer, N. Kourentzes, and R. Fildes, "Demand forecasting with user-generated online information," *International Journal of Forecasting*, vol. 35, no. 1, pp. 197–212, Jan. 2019, doi: 10.1016/j.ijforecast.2018.03.005.

[11] N. Levin, S. Ali, and D. Crandall, "Utilizing remote sensing and big data to quantify conflict intensity: The Arab Spring as a case study," *Applied Geography*, vol. 94, pp. 1–17, May 2018, doi: 10.1016/j.apgeog.2018.03.001.

[12] Y. Keneshloo, J. Cadena, G. Korkmaz, and N. Ramakrishnan, "Detecting and forecasting domestic political crises," in *Proceedings of the 2014 ACM conference on Web science - WebSci '14*, 2014, pp. 192–196, doi: 10.1145/2615569.2615698.

[13] L. C. Jacaruso, "A method of trend forecasting for financial and geopolitical data: inferring the effects of unknown exogenous variables," *Journal of Big Data*, vol. 5, no. 1, p. 47, Dec. 2018, doi: 10.1186/s40537-018-0160-5.

[14] C. Searle and J. H. van Vuuren, "Modelling forced migration: A framework for conflict-induced forced migration modelling according to an agent-based approach," *Computers, Environment and Urban Systems*, vol. 85, p. 101568, Jan. 2021, doi: 10.1016/j.compenvurbsys.2020.101568.

[15] P. De Brabanter, "Why Quotation Is Not a Semantic Phenomenon, and Why It Calls for a Pragmatic Theory," 2017, pp. 227–254.

[16] GDELT, "GDELT Project," 2021. https://www.gdeltproject.org/.

[17] N. Levin, S. Ali, D. Crandall, and S. Kark, "World Heritage in danger: Big data and remote sensing can help protect sites in conflict zones," *Global Environmental Change*, vol. 55, pp. 97–104, Mar. 2019, doi: 10.1016/j.gloenvcha.2019.02.001.

[18] V. Voukelatou, L. Pappalardo, I. Miliou, L. Gabrielli, and F. Giannotti, "Estimating countries' peace index through the lens of the world news as monitored by GDELT," in *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, Oct. 2020, pp. 216–225, doi: 10.1109/DSAA49011.2020.00034.

[19] K. Pogorelov, D. T. Schroeder, P. Filkukova, and J. Langguth, "A System for High Performance Mining on GDELT Data," in *2020 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, May 2020, pp. 1101–1111, doi: 10.1109/IPDPSW50202.2020.00182.

[20] F. Qiao, P. Li, X. Zhang, Z. Ding, J. Cheng, and H. Wang, "Predicting Social Unrest Events with Hidden Markov Models Using GDELT," *Discrete Dynamics in Nature and Society*, vol. 2017, pp. 1–13, 2017, doi: 10.1155/2017/8180272.

[21] F. Qiao and K. Chen, "Predicting Protest Events with Hidden Markov Models," in *2016 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)*, Oct. 2016, pp. 109–114, doi: 10.1109/CyberC.2016.30.

[22] H. Wang, B. Zhou, Z. Gu, and Y. Jia, "Contextual Gated Graph Convolutional Networks for Social Unrest Events Prediction," in *2020 IEEE Fifth International Conference on Data Science in Cyberspace (DSC)*, Jul. 2020, pp. 320–325, doi: 10.1109/DSC50466.2020.00056.

[23] F. Qiao and H. Wang, "Computational Approach to Detecting and Predicting Occupy Protest Events," in *2015 International Conference on Identification, Information, and Knowledge in the Internet of Things (IIKI)*, Oct. 2015, pp. 94–97, doi: 10.1109/IIKI.2015.28.

[24] N. Ramakrishnan *et al.*, "'Beating the news' with EMBERS," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, Aug. 2014, pp. 1799–1808, doi: 10.1145/2623330.2623373.

[25] F. Qiao, P. Li, X. Zhang, Z. Ding, J. Cheng, and H. Wang, "Predicting Social Unrest Events with Hidden Markov Models Using GDELT," *Discrete Dynamics in Nature and Society*, vol. 2017, pp. 1–13, 2017, doi: 10.1155/2017/8180272.

[26] F. Qiao, X. Zhang, and J. Deng, "Learning Evolutionary Stages with Hidden Semi-Markov Model for Predicting Social Unrest Events," *Discrete Dynamics in Nature and Society*, vol. 2020, pp. 1–16, Oct. 2020, doi: 10.1155/2020/3915036.

[27] G. Korkmaz, J. Cadena, C. J. Kuhlman, A. Marathe, A. Vullikanti, and N. Ramakrishnan, "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*, Aug. 2015, pp. 258–265, doi: 10.1145/2808797.2808847.

[28] R. A. Blair and N. Sambanis, "Forecasting Civil Wars: Theory and Structure in an Age of 'Big Data' and Machine Learning," *Journal of Conflict Resolution*, vol. 64, no. 10, pp. 1885–1915, Nov. 2020, doi: 10.1177/0022002720918923.

[29] N. N. Daud, S. H. Ab Hamid, M. Saadoon, F. Sahran, and N. B. Anuar, "Applications of link prediction in social networks: A review," *Journal of Network and Computer Applications*, vol. 166, p. 102716, Sep. 2020, doi: 10.1016/j.jnca.2020.102716.

[30] O. Schaer, N. Kourentzes, and R. Fildes, "Demand forecasting with user-generated online information," *International Journal of Forecasting*, vol. 35, no. 1, pp. 197–212, Jan. 2019, doi: 10.1016/j.ijforecast.2018.03.005.

[31] J. Haneczok and J. Piskorski, "Shallow and deep learning for event relatedness classification," *Information Processing & Management*, vol. 57, no. 6, p. 102371, Nov. 2020, doi: 10.1016/j.ipm.2020.102371.

[32] M. Manacorda and A. Tesei, "Liberation Technology: Mobile Phones and Political Mobilization in Africa," *Econometrica*, vol. 88, no. 2, pp. 533–567, 2020, doi: 10.3982/ECTA14392.

[33] V. Voukelatou *et al.*, "Measuring objective and subjective well-being: dimensions and data sources," *International Journal of Data Science and Analytics*, Jun. 2020, doi: 10.1007/s41060-020-00224-2.

[34] D. Jacobson, "Al-Qaeda and Islamist Militant Influences on Tribal Dynamics: The Northern Mali and Northeastern Nigeria Regions," UNIVERSITY OF SOUTH FLORIDA TAMPA United States, 2016.

[35] J. Deng, F. Qiao, H. Li, X. Zhang, and H. Wang, "An Overview of Event Extraction from Twitter," in *2015 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery*, Sep. 2015, pp. 251–256, doi: 10.1109/CyberC.2015.24.

[36] H. Li, "Political Effect of Economic Data Manipulation: Evidence from Chinese Protests." Duke University, 2017.

[37] W. Wang, "Event detection and extraction from news articles." Virginia Tech, 2018.

[38] A. Katagiri & E. Min, "Identifying threats: Using machine learning in international relations," 2015.

[39] S. Kühn, "1 Global employment and social trends," *World Employment and Social Outlook*, vol. 2020, no. 1, pp. 15–38, Feb. 2020, doi: 10.1002/wow3.158.

[40] G. Korkmaz, J. Cadena, C. J. Kuhlman, A. Marathe, A. Vullikanti, and N. Ramakrishnan, "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*, Aug. 2015, pp. 258–265, doi: 10.1145/2808797.2808847.

[41] N. Levin, S. Ali, and D. Crandall, "Utilizing remote sensing and big data to quantify conflict intensity: The Arab Spring as a case study," *Applied Geography*, vol. 94, pp. 1–17, May 2018, doi: 10.1016/j.apgeog.2018.03.001.

[42] J. Wilkerson and A. Casas, "Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges," *Annual Review of Political Science*, vol. 20, no. 1, pp. 529–544, May 2017, doi: 10.1146/annurev-polisci-052615-025542.

[43] C. Searle and J. H. van Vuuren, "Modelling forced migration: A framework for conflict-induced forced migration modelling according to an agent-based approach," *Computers, Environment and Urban Systems*, vol. 85, p. 101568, Jan. 2021, doi: 10.1016/j.compenvurbsys.2020.101568.

[44] B. E. Bedasso and N. Obikili, "A Dream Deferred: The Microfoundations of Direct Political Action in Pre- and Post-democratisation South Africa," *The Journal of Development Studies*, vol. 52, no. 1, pp. 130–146, Jan. 2016, doi: 10.1080/00220388.2015.1036041.

[45] S. Basnet, L.-K. Soh, A. Samal, and D. Joshi, "Analysis of Multifactorial Social Unrest Events with Spatio-Temporal k-Dimensional Tree-based DBSCAN," in *Proceedings of the 2nd ACM SIGSPATIAL Workshop on Analytics for Local Events and News*, Nov. 2018, pp. 1–10, doi: 10.1145/3282866.3282870.

[46] A. V. Rozhnov and I. A. Lobanov, "Investigation of the joint semantic environment for heterogeneous robotics," in *2017 Tenth International Conference Management of Large-Scale System Development (MLSD)*, Oct. 2017, pp. 1–5, doi: 10.1109/MLSD.2017.8109678.

[47] M. Ma, P. Fang, J. Gao, and C. Song, "Does ideology affect the tone of international news coverage?," in *2017 International Conference on Behavioral, Economic, Socio-cultural Computing (BESC)*, Oct. 2017, pp. 1–5, doi: 10.1109/BESC.2017.8256368.

[48] D. Jacobson, "Al-Qaeda and Islamist Militant Influences on Tribal Dynamics: The Northern Mali and Northeastern Nigeria Regions," UNIVERSITY OF SOUTH FLORIDA TAMPA United States, 2016.

[49] H. Li, "Political Effect of Economic Data Manipulation: Evidence from Chinese Protests." Duke University, 2017.

[50] E. L. Loginov and V. E. Loginova, "The stability-saving process of the cooperative behavior of autonomous agentsteams in dynamic problematic spheres of the digital economy," *Market economy problems*, no. 1, pp. 33–38, 2018.

[51] A. Fronzetti Colladon, "Forecasting election results by studying brand importance in online news," *International Journal of Forecasting*, vol. 36, no. 2, pp. 414–427, Apr. 2020, doi: 10.1016/j.ijforecast.2019.05.013.

[52] A. N. Usanov and T. Sweijs, "Models Versus Rankings: Forecasting Political Violence," *SSRN Electronic Journal*, 2017, doi: 10.2139/ssrn.2930104.

[53] M. Nadolski and J. Fairbanks, "Complex systems analysis of hybrid warfare," *Procedia Computer Science*, vol. 153, pp. 210–217, 2019, doi: 10.1016/j.procs.2019.05.072.

[54] N. N. Daud, S. H. Ab Hamid, M. Saadoon, F. Sahran, and N. B. Anuar, "Applications of link prediction in social networks: A review," *Journal of Network and Computer Applications*, vol. 166, p. 102716, Sep. 2020, doi: 10.1016/j.jnca.2020.102716.

[55] Y. Keneshloo, J. Cadena, G. Korkmaz, and N. Ramakrishnan, "Detecting and forecasting domestic political crises," in *Proceedings of the 2014 ACM conference on Web science - WebSci '14*, 2014, pp. 192–196, doi: 10.1145/2615569.2615698.

[56] L. Zhao, F. Chen, C.-T. Lu, and N. Ramakrishnan, "Spatiotemporal Event Forecasting in Social Media," in *Proceedings of the 2015 SIAM International Conference on Data Mining*, Jun. 2015, pp. 963–971, doi: 10.1137/1.9781611974010.108.

[57] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science*, vol. 2, no. 1, pp. 1–8, Mar. 2011, doi: 10.1016/j.jocs.2010.12.007.

[58] N. Kallus, "Predicting crowd behavior with big public data," in *Proceedings of the 23rd International Conference on World Wide Web - WWW '14 Companion*, 2014, pp. 625–630, doi: 10.1145/2567948.2579233.

[59] K. Radinsky and E. Horvitz, "Mining the web to predict future events," in *Proceedings of the sixth ACM international conference on Web search and data mining - WSDM '13*, 2013, p. 255, doi: 10.1145/2433396.2433431.

[60] X. Wang, M. S. Gerber, and D. E. Brown, "Automatic Crime Prediction Using Events Extracted from Twitter Posts," 2012, pp. 231–238.