Near Real-Time Monitoring
and
Forecasting
of
Civil Unrest in South Africa
utilising
Bilingual Sources

by W.S.J. Marais

September 2023

Promotor: Prof. J.E. Kotzé
Co-promotor: Dr B.A. Senekal

Abstract

Several initiatives with the objective of coding event information with regard to civil unrest have been implemented. However, most of these initiatives are global in nature and do not focus on all countries' local events. Other efforts provide only a near real-time picture of current local unrest. This research project focuses on the development of a model with an automated pipeline to collect and extract security event information from Afrikaans and English social media, blogs and digital news sources by applying state-of-the-art machine and deep learning techniques in order to forecast and monitor civil unrest in South Africa in near real-time. The automated pipeline will have two stages with a machine learning / deep learning model for each stage of the pipeline. Stage 1's model will have the capability to extract and classify event information into different categories and monitor current incidents in near real-time. Stage 2's model will have the capability to forecast possible civil unrest in South Africa. An event information vector space will be created by experimenting with different word embedding techniques. Word2vec, Doc2Vec, fastText and GloVe, will be used for static word embeddings, while BART, selected versions of BERT, and selected versions of GPT will be used for contextual word embeddings. Events will be classified by experimenting with machine learning algorithms such as naive Bayes, SVMs, logistic regression, random forest and XGBoost, as well as deep learning algorithms such as CNNs, LSTM RNNs and GRU RNNs. The data that will be used to train the models will be obtained by means of OSINT sources. This project will rely on text data available on online digital platforms. Although information, conveyed through audio, video and images, can be transformed into text data, this falls outside the scope of this research project. Possible online digital sources from which event data can be collected include news websites, social media, blogs, forums and search engines as well as already existing event data repositories. Selecting sources that provide accurate, reliable and local event data will be vital to the forecasting of possible civil unrest in South Africa, as will determining the type of events (features) or combinations thereof that will result in an accurate forecast of civil unrest.

Keywords

Machine Learning, Deep Learning, Near Real-time Open-Source Intelligence, OSINT, Forecast, Civil Unrest, Event Coding, South Africa

Table of Contents

List of Tables

List of Figures

1. Introduction and Background

Armed conflict in Lesotho due to an attempted *coup d'état* during 2014 (Gebremichael et al., 2019; Vhumbunu, 2015). Crime waves in South Africa during the coronavirus lockdown that left dozens of schools burnt to the ground after being broken into and vandalised (Smith, 2020). Mass violence during the *#FeesMustFall* protests in South Africa that started in October 2015 (Mavunga, 2019). Anarchy knocking on the door of Cape Town when the worst drought in the history of South Africa almost forced the city to shut off taps to four million people (Welch, 2018). Riots in KwaZulu-Natal and Gauteng that lasted from 9 to 21 July 2022, sparked by the imprisonment of former South African president Jakob Zuma (*South Africa Zuma Riots: Looting and Unrest Leaves 72 Dead - BBC News*, n.d.). These are all forms of social unrest, caused by political, economic, social or environmental factors (Braha, 2012). The conditions that stem from these factors can give rise to acts of civil disobedience such as strikes and protests which are usually peaceful (*CIVIL DISOBEDIENCE | Meaning in the Cambridge English Dictionary*, n.d.), but can escalate to acts of civil unrest such as riots, looting, vandalism, arson and even armed conflict which are more violent in nature (*Civil Disorder Dictionary Definition | Civil Disorder Defined*, n.d.; Qiao et al., 2017). A growing interest in forecasting acts of civil unrest, such as international conflicts, civil wars, coups and mass killings (Chadefaux, 2017) has sparked collaboration between academia and the military in an effort to avoid past intelligence failures and misprojections.

Several approaches to predict the outbreak of armed conflict have been met with different levels of success (Chadefaux, 2014, 2017). These approaches include:

- Experts: Opinions and predictions provided by experts (Chadefaux, 2017).
- Econometric: Algorithms and models ranging from traditional regression techniques to more intricate random forest or neural network models (Brandt & Freeman, 2006; Hegre et al., 2017; Muchlinski et al., 2016; Rummel, 1969).
- Modelling: Methods such as game theory (Goodwin, 2002) and agent-based models (Bhavnani et al., 2014; Cederman, 2002).
- Wise crowds: Combining different approaches through the aggregation and weighting of various models and opinions to form ensemble forecasts (Montgomery et al., 2012), tournaments (Tetlock & Gardner, 2016) and markets (Chadefaux, 2017).

Approach	Methodology	Advantages	Disadvantages
Experts	Manual	Might contribute to better algorithm performance	Subjective due to vested interest
Econometric	Fully Automated	Large amount of real-time data sources	Possible irrelevant, biased and partial data sources
Modelling	Semi-Automated	Model incorporates all relevant information	Complexity of models limits their applicability, power and inferences
Wise Crowds	Manual, Semi-Automated or Fully Automated	Linked to the combined methods from selected approaches	Linked to the combined methods from selected approaches

Table 1: Methodology, advantages and disadvantages of forecasting approaches

When comparing the different approaches (depicted in Table 1), several advantages and disadvantages regarding each approach are apparent:

- According to Chadefaux (2017), opinions and insight from experts with regard to political events do not provide any better predictive performance than those from novices or simple econometric algorithms (Green & Armstrong, 2007). In fact, Tetlock (2006) states that experts' predictive performance is on average no better than random guesses. This might be because of their vested interest regarding the political event involved. Furthermore, the expert approach is inherently a manual process which also inhibits timely forecasts. However, even though experts have poor predictive power, their opinion and insight might still contribute to better algorithm performance (Chadefaux, 2017).
- Econometric approaches have evolved from relying on structural measures of tensions like regime type, gross domestic product (GDP), ethnicity and terrain (all measured in yearly intervals) to relying on short-term measures of tensions derived from real-time events that take place (a.k.a. event data). Furthermore, advances in natural language processing (NLP) have hugely contributed to econometric approaches with regard to the automation of event coding, therefore, enabling the processing of large amounts of real-time data sources. It is important to note however, that these data sources might contain biased, partial, or even irrelevant information that could influence the predictive performance of this approach (Chadefaux, 2017; Gleditsch & Ward, 2013; O'Brien, 2002).
- Modelling approaches, such as game theory and agent-based models have the advantage of incorporating all relevant information in the model, omitting no detail. Unfortunately, this leads to very complex models that limit their applicability and power, as well as the inferences that can be drawn from the model. This approach has

delivered several successes when it comes to forecasting events, but it requires a lot of human intervention (Chadefaux, 2017).

- Finally, according to Chadefaux (2017), combining the aforementioned approaches produce the most promising forecast performance of all. This is referred to as wise crowds where various methods, within the same or across different approaches, are aggregated according to selected weights. This approach in itself has certain advantages and disadvantages linked to the combination of chosen methods. This approach, therefore, can follow a manual, semi-automated or fully automated methodology.

The aim of this research project is to design, develop and evaluate a state-of-the-art predictive model capable of forecasting civil unrest in South Africa in real-time, or as close to real-time as possible. To have the capability to forecast in near real-time, the predictive model has to be fully automated. As discussed above, econometric and wise crowd approaches that combine several econometric approaches provide forecasting methods that can be fully automated (Chadefaux, 2017; Senekal & Kotzé, 2019) which would therefore be ideal for this research project. Kotzé et al. (2020a) reinforces this point of view by stating that supervised machine learning can be used to automate and speed-up the analysis of data.

Several initiatives have been implemented to classify events that involve matters of civil unrest:

- Political Instability Task Force (PITF) (Ulfelder & Valentino, 2008)
- Armed Conflict Location and Event Data (ACLED) (Raleigh et al., 2010)
- Integrated Crisis Early Warning System (ICEWS) (O'Brien, 2010)
- Social Political and Economic Event Database (SPEED) (Nardulli et al., 2011)
- Uppsala Conflict Data Program (UCDP) (Sundberg et al., 2012)
- Global Data on Events Language and Tone (GDELT) (Leetaru & Schrodt, 2013)
- Conflict Early Warning Signal Index (CEWSI) (Chadefaux, 2014)
- Phoenix (with EL:DIABLO) (Beieler, 2016)
- South African Violent Incident Classifier (SAVIC) (Senekal & Kotzé, 2019)
- Social Conflict Analysis Database (SCAD) (Salehyan et al., 2020)

Although, these initiatives have contributed greatly to the classification of events that involve matters of civil unrest, researchers are still busy searching for techniques to improve the

automation thereof (Chadefaux, 2017; Kotzé et al., 2020b). With the exception of SAVIC, these initiatives focus primarily on international events and circumstances surrounding those events that are not inherently unique to South Africa; therefore, neglecting any deviation there might exist when applied to a South African context (Kotzé et al., 2020b; Senekal & Kotzé, 2019). Also, because these initiatives focus on international events (except for SAVIC), text sources with languages indigenous to South Africa do not form part of their training or testing corpus.

This research project aims to enrich this field of research by providing new insight and knowledge to the classification of domestic events caused by circumstances that are unique to South Africa. By focussing on the inherent political, economic, social and environmental factors in South Africa, this research project will lend itself to the development of a unique state-of-the-art model that is capable of forecasting civil unrest in South Africa in near real-time. Finally, the research project will also assist in the enhancement of current automated methods for the extraction and classification of bilingual (Afrikaans and English) event data and prediction of civil unrest.

2. Rationale for the Research

Most countries in the world have been involved in some or other form of civil unrest (Chadefaux, 2017). The ability to predict possible civil unrest is, therefore, beneficial to different role players for various reasons: (1) It can serve as a warning to civilians to help them prepare for possible disruptions. (2) Policymakers, investors and financial institutions can benefit by proactively implementing countermeasures to avoid potential governing, economic and financial disasters. (3) International aid organisations can use the predictions to increase their response time when their assistance might be needed. (4) It can also minimise the chances of a country's defence and security forces being caught off guard in the event of an attempted *coup d'état* (Chadefaux, 2014). Various methods have been implemented to forecast civil unrest, but unfortunately, according to Chadefaux (2017), several intelligence failures and misestimations have occurred in the past. Therefore, more research with regard to the methods and techniques that are applied to predict possible civil unrest is crucial to increase the accuracy and confidence of these predictions. Furthermore, the recent technological advancements as well as the breakthroughs with regard to machine

learning and deep learning techniques have made research in this field much more appealing and achievable (Hürriyetoğlu et al., 2020).

In South Africa, the political, economic, and social climate is ripe for possible civil unrest. Issues such as expropriation without compensation (Roets et al., 2019), the deterioration of municipal services (Longa, 2018), the country's weakening currency (Euginia, 2017), the high crime and murder rate (Lancaster & Newham, 2020), the low standard of living conditions (Longa, 2018) and lately, the restrictive firearms amendment bill (*REPUBLIC OF SOUTH AFRICA FIREARMS CONTROL AMENDMENT BILL*, 2021), to name a few, all contribute to the possibility of civil unrest in South Africa. In fact, small scale civil unrest has erupted in the past on several occasions but has not yet led to full scale civil war (Alexander et al., 2018; Cook, 2020; Lee, 2018).

According to Kotzé et al. (2020b), there remains a need for research in security related matters with regard to the automatic extraction and classification of event data, especially with regard to domestic events (Senekal & Kotzé, 2019). It, therefore, stands to reason that although main events taking place in South Africa might be included by current event detection systems such as ICEWS and GDELT, fine-grained events specific to South Africa might get overlooked.

Open-source intelligence (OSINT) has become a dominant source of intelligence for event data (Piskorski & Jacquet, 2020; Senekal & Kotzé, 2019; Williams & Blum, 2018). OSINT sources include social media platforms such as X (Twitter), Facebook, Signal, Whatsapp, online news sites, forums and blogs (Chadefaux, 2017; Kotzé et al., 2020b; Vadapalli et al., 2018). The abundance of OSINT sources makes event data easily obtainable, which in turn makes research in this field worthwhile.

3. Preliminary Literature Survey

According to Chadefaux (2014), earlier methods that were used to forecast armed conflict were seldom successful because the measures that were used to identify geopolitical tensions were inadequate. By analysing a comprehensive dataset of historical newspaper articles published during the last century, Chadefaux created a weekly risk index that, when applied to a dataset of all wars within and between countries ranging from 1902 until 2001,

the onset of a war within the following few months could be predicted with a confidence level of up to 85%. Chadefaux (2017), however, also stated that there were limitations with regard to the ability to forecast armed conflict because of imperfect models and data at that point in time.

Later, however, Chadefaux (2017) stated that there existed a growing interest in forecasting conflict situations and that the ability to forecast such events was becoming more successful. The research that followed from this, resulted in a shift from qualitative to quantitative forecasting methods (Senekal & Kotzé, 2019), because qualitative methods lacked the ability to quantify an event or situation. Through quantitative methods, events can be measured and evaluated mathematically or statistically to determine a potential outcome (Goldstone, 2008). Event data is generated through a process called event extraction (Hürriyetoğlu et al., 2020) or event coding (Beieler, 2016; Osorio & Reyes, 2017; Wang & Kennedy, 2016), which involves gathering information about incidents that can be determined from text sources. It follows the form who-did-what-to-whom (Hürriyetoğlu et al., 2020; Olsson et al., 2020) and can also include the date (when) as well as the geographical location (where) of the event.

3.1 Event Coding

Originally, event coding was a manual process where datasets were generated through the use of human coders who identified and recorded events obtained from news articles (Beieler, 2016; Hürriyetoğlu et al., 2020; Qiao et al., 2017; Yörük, 2012). The development of coding software later automated the event coding process. At first, coding software used shallow parsing techniques which consisted of part-of-speech tagging on the words within a sentence. Later developments of coding software implemented deep parsing methods where the entire syntactic structure of a sentence is analysed and comprehended. Early deep parsing methods combined part-of-speech tagging for individual words with noun and verb phrase chunking, as well as syntactic information about the relation between noun and verb phrases (Beieler, 2016).

It is important to note that the correctness of the events identified depends on the accuracy and reliability of the parsing techniques that are implemented to identify the relevant noun and verb phrases (Beieler, 2016). This led to major new developments in NLP and computational linguistics, but unfortunately, even the latest state-of-the-art automated event extraction systems need human intervention to ensure accurate event detection (Hürriyetoğlu

et al., 2020). Ganino et al. (2018) identified the following NLP steps to extract meaningful information from unstructured data sources:

- Tokenization: In this step the raw text is broken up into words or sentences, keeping meaningful punctuation marks in tacked. During tokenization some other preprocessing steps, which Ganino et al. (2018) failed to mention, are also applied. This includes unnecessary punctuation and white spaces that have to be removed, while emoticons are usually kept because they might contribute extra meaning to the text. Stop words also need to be removed. Stop words are words that are most commonly used and do not add much value to the meaning of the text (Raschka & Mirjalili, 2017). To complete the tokenization process, each word needs to be converted to its stem (base form) through a process called stemming or lemmatisation.
- Part-of-speech (POS): During this step each word is labelled with a unique tag which indicates its syntactical role, for example noun, verb, or pronoun.
- Named entity recognition (NER): Specific rules and statistical machine learning techniques are applied to categorise and label elements in the sentence as either a person, location or action. Three approaches can be followed to classify the entities: (1) Entities can be classified by way of precomputed lists of entities, (2) though rule-based methods that analyse the context or key characteristics of the entities or (3) by means of machine learning techniques.
- Semantic role labelling (SRL): This process interprets the meaning of an entire sentence by determining the meaning of the individual words and the relationships between them. During this process the text is transformed into a vector space through vectorisation or embedding techniques, of which several exists that are discussed in Section 3.5.

It stands to reason that the automation of event coding increased the rate at which event information can be processed. This, however, requires a digital form of open-source information. Therefore, it is no surprise that event information is primarily obtained from digital free text OSINT sources (Piskorski & Jacquet, 2020; Senekal & Kotzé, 2019; Williams & Blum, 2018). OSINT provides intelligence based on publicly available sources such as social media, forums, blogs and online news sites (Ganino et al., 2018). Since the inception of the Foreign Broadcast Monitoring Service (FBMS) in 1941, OSINT has played an important role in monitoring and analysing open-source media (Williams & Blum, 2018). The first generation of OSINT faced challenging efforts with regard to the collection of information because it had

to be obtained physically. But the expansion of the Internet into a global presence and the establishment of social media and big data analytics in the early twenty-first century have revolutionised OSINT. According to Williams and Blum (2018), the usefulness of intelligence sources depends on their credibility, reliability and validity, and determines whether an intelligence source is considered single-source or all-source. OSINT should therefore be evaluated for trustworthiness and not just taken at face value to ensure that event datasets contain truthful information.

3.2 Previous Projects

Research into conflict prediction gave rise to several event coding and prediction projects, which produced huge event datasets of incidents that occurred in countries all around the world (Chadefaux, 2014). Table 2 contains a list of such projects, summarising whether the relevant project focuses on event coding, or on forecasting possible outcomes resulting from the event information. How the project's event data is maintained, the sources of the data and the scope of the events that are monitored are also summarised.

Project	Focus	Maintenance	Data Sources	Event Scope
PITF (Ulfelder & Valentino, 2008)	Event Coding	Manually	Local / International	Global
ACLED (Raleigh et al., 2010)	Event Coding	Manually	Local / Regional	Unstable Countries
ICEWS (O'Brien, 2010)	Forecasting	Automated	Local / International	Global
SPEED (Nardulli et al., 2011)	Event Coding	Semi-automated	Local / International	Global
UCDP (Sundberg et al., 2012)	Event Coding	Manually	Local / International	Global
GDELT (Leetaru & Schrodt, 2013)	Event Coding	Automated	International	Global
CEWSI (Chadefaux, 2014)	Forecasting	Manually	International	Global
Phoenix (Beieler, 2016)	Event Coding	Automated	RSS feeds	Global
SAVIC (Senekal & Kotzé, 2019)	Event Coding	Automated	Local Whatsapp	South Africa
SCAD (Salehyan et al., 2020)	Event Coding	Manually	International	Africa / Latin America

Table 2: Summary of event coding and prediction projects

The Political Instability Task Force (PITF) manually maintains a dataset on the deliberate killing of civilians worldwide. The dataset can be used for the development of statistical forecasting models to identify countries where such atrocities are anticipated, or to determine a possible increase in the rate or intensity of atrocities that are already taking place. The dataset can also be used to derive descriptive statistics about atrocities taking place across the whole world. Incidents are coded manually using information obtained from international and local digital commercial news sources retrieved from Factiva. Each event contains the following information: The nature by which the incident was reported, the date, location,

perpetrators, perpetrator characteristics, victim, victim characteristics, casualties, death ambiguities, the severity level of the incident, weapons that were used, the perpetrators' intent, whether there was collateral damage, related tactics about assassinations, and a short description of the incident in English. Also, coded for each event is the primary and secondary sources of information, the initials of the coder, as well as any comments pertaining to the sources and coding (Ulfelder & Valentino, 2008).

The Armed Conflict Location and Event Data (ACLED) project implements manual coding methods to identify the actions of rebels, governments, and militias within unstable countries. Different locations and dates are also included with regard to battle events that occurred, military control that was transferred, headquarter establishment, civilian violence and rioting. Data sources primarily include press articles from local and regional news sources, as well as the Integrated Regional Information Network (IRIN), Relief Web, a global news monitoring and search engine called Factiva, and humanitarian agencies (Raleigh et al., 2010). A baseline artificial neural network (ANN) and a deep neural network (DNN) were later both trained on the ACLED dataset and respectively achieved a 90% and 89% accuracy in conflict prediction (Olaide & Ojo, 2021).

The Integrated Crisis Early Warning System (ICEWS) is an automated system with the capability to monitor, evaluate and forecast possible national, subnational and international crises across the globe. It integrates three forecasting model approaches that include macro-structural factors, event data patterns and leader profile models to form an aggregated Bayesian forecasting model. Macro-structural factors include regime type, gross domestic product (GDP) per capita and the degree of hostility and cooperation between government and civilian society. Event data is generated in the form of who-did-what-to-whom-where-when-how through the use of an automated event coding pipeline. The pipeline implements the Conflict and Mediation Event Observations (CAMEO) coding ontology by means of the Textual Analysis by Augmented Replacement Instructions (TABARI) system. Leader profiles are created from personalities, leadership styles, and operational conduct amongst other leadership trait characteristics. ICEWS collects local and international event information that are available from about three hundred different publishers' commercial news sources (O'Brien, 2010).

The Social, Political, and Economic Event Database (SPEED) is a semi-automated event coding project that uses a combination of human and machine-coding methods to identify

events for basically every country across the globe. It implements the Societal Stability Protocol (SSP) ontology by means of BIN, an automated text classification system and the Event Annotation Tool (EAT). SSP focuses on fine-grained incidents relating to civil unrest, such as protests, strikes, politically motivated attacks, disruptive state acts, and irregular transfers of power. Focusing on fine-grained incidents helps to generate insights into incidents that result in events of greater significance. Event information is obtained from a global archive of local and international news reports by means of a web scraper. Each event contains the following information: The initiators and victims of the event, whether there was international involvement, the type of action carried out, the impact it had, consequences for the initiators, reactions to the events and also if there were subsequent events. Further, each event also includes the weapons used or mode of expression with regard to protests, the location and the date of the event, as well as the societal context and circumstances that led to the event (Nardulli et al., 2011).

The Uppsala Conflict Data Program (UCDP) contains manually coded event data about direct and deliberate killings of civilians in intrastate armed conflicts by organised groups, such as governments or rebels. Event information is retrieved from five international news sources: Reuters News, BBC World Monitoring, Agence France Presse, Dow Jones International News and Xinhua News Agency, as well as EFE News Service (focuses on Latin America), by means of a data search via the VRA software system. Supplementary reports from the United Nations (UN), Human Rights Watch, Amnesty International, and local NGO data were also included. The event data contains information about the actor that commits the killings, whether the violence was one-sided, the number of civilians killed per year, whether there was at least one previous war within the past fifteen years, and if a civil war broke out in the same year of the event. It also contains information about the level of autocracy or democracy in the government, the annual trade as a percentage of GDP, as well as whether the actor is the government or part of a rebel group (Eck & Hultman, 2007; Sundberg et al., 2012). The UCDP dataset was utilised in a competition where fifteen international teams competed by developing models that were capable of forecasting the (de)-escalation in state-based violence. The resulting models included the use of recurrent neural networks (RNN), graph convolutional neural networks (GCNN) and hidden Markov models (HMM) (Hegre et al., 2022).

The initiative with regard to Global Data on Events Language and Tone (GDELT) contains an automated pipeline for coding political event information. Originally, it used the TABARI

system which implements the CAMEO coding ontology. TABARI was later replaced by the Python Engine for Text Resolution and Related Coding Hierarchy (PETRARCH), which later evolved into PETRARCH2. GDELT contains over two hundred million global geo located events specifying the event date, event action, the source actor, the target actor, the event location, each actor's country, organisation (IGO/NGO/rebel), ethnic group and religion they belong to, as well as their societal role. Data is mostly collected from a variety of international news sources, but national sources such as the New York Times, the Associated Press and Google News are also included (Leetaru & Schrodt, 2013).

Chadefaux (2014) created a Conflict Early Warning Signal Index (CEWSI) that could be used to estimate the probability of a war breaking out within and between countries. The resulting risk index reflected a fine-grained weekly measure of geopolitical tensions between the relevant parties that was manually derived from a comprehensive dataset of historical conflict-related newspaper articles between 1 January 1902 and 31 December 2001. The research with regard to CEWSI showed that the number of conflict-related news articles increased dramatically prior to the onset of wars, thus being a good measure for the rise of tensions. CEWSI was capable of predicting the probability of a war erupting within the following couple of months with a confidence interval of 85%. This significantly improved upon existing methods available at the time (Chadefaux, 2014). In more recent research Chadefaux (2022, 2023) focused on predicting the onset of war through pattern recognition techniques by means of machine learning methods. The author used entire event sequences, instead of independent observation sets to uncover hidden patterns in pre-conflict sequences. The time intervals over which these sequences occurred were aligned by applying a dynamic time warping algorithm (Berndt & Clifford, 1994), which then made it possible to identify similar patterns by calculating the Euclidean distance between the sequences (Chadefaux, 2022, 2023).

Phoenix contains an automated pipeline for coding political event information which includes the date of the event, the action that was carried out, the source actor, the target actor, and the location where the action occurred. The main objective of the project is to provide political event information in near real-time. The pipeline utilises the PATRACH2 event coding system which also implements the CAMEO coding ontology. A virtual machine, called EL:DIABLO, was created which contains a script that automatically installs and sets up the Phoenix environment. To obtain data, Phoenix makes use of RSS feeds of a list of predefined news websites (Beieler, 2016).

The South African Violent Incident Classifier (SAVIC), implements an automated pipeline to collect and code event information occurring in South Africa (Senekal & Kotzé, 2019). The main purpose of this project is to monitor the state of mass violence taking place across South Africa via an interactive dashboard, with updates in near real-time. The project originally applied part-of-speech and rule-based NLP techniques by means of Python toolkits to identify and classify incidents either as crime, protests, farm attacks or land expropriation. Event information obtained via twenty-one Whatsapp groups from English and Afrikaans users were used to code the event dataset. Unfortunately, due to the bilingualism of the Whatsapp data, part-of-speech was unable to categorise nouns and verbs accurately. The project, therefore, decided to use a rule-based approach instead. Later the project was extended to implement distributed word and document vectors with a machine learning classifier which outperformed the rule-based classifier (Kotzé et al., 2020a, 2020b).

The Social Conflict Analysis Database (SCAD) (Salehyan et al., 2020), previously known as the Social Conflict in Africa Database (Salehyan et al., 2012), provides a manually coded event dataset of social and political unrest focused on protests, riots, strikes, and small-scale armed attacks in Africa and Latin America. Event data is collected internationally by means of keyword searches from the Associated Press and Agence France Presse on the Lexis-Nexis academic database. The source actor, target actor, action that was carried out, location where the event took place, time frame of the event, the magnitude of the action, the fatalities resulting from the action, and the issues that caused the action are included in each event. In Salehyan et al. (2020), the actors were also categorised into organisations that they belonged to.

3.3 Event Coding Methods

Most of the automated coding projects that were discussed in Section 3.2 (see Table 2), implement a sequence classification tool to extract the different characteristics, used to construct events, from the relevant news sources. The Kansas Event Data System (KEDS) was one of the first attempts at such a tool (Schrodt et al., 1994). By parsing text, it was able to extract words, that were pre-defined in dictionaries, as actors and actions. KEDS was replaced by TABARI which had the ability to recognise passive-voice sentences or disambiguate verbs from nouns. Later PATRARCH replaced TABARI and was capable of extracting event information in the form of who-did-what-to-whom-where-when-how. All these tools use symbolic (rule-based) NLP methods to identify actors and actions from text. However, Phoenix (Beieler, 2016) on the other hand, implements more advanced machine

learning techniques, by means of a character-based convolutional neural network (CNN), to identify the type of event action.

Other authors, namely Adhikari et al. (2019), have also been experimenting with machine and deep learning algorithms, such as bidirectional encoder representations from transformers (BERT), to extract the topic of news articles. Olsson et al. (2020) continued with similar experiments using bag-of-words, embeddings from language models (ELMo), BERT and universal language modelling fine-tuning (ULMFiT). The experiment illustrated the advantages of fine-tuning pre-trained language models on domain-specific data. From these examples it is apparent that deep learning and pre-trained neural language models play a significant role in event coding as well as other NLP and text mining operations such as text classification, machine translation and generative AI (Jurafsky & Martin, 2023). It, therefore, seems fit to briefly discuss the fundamental principles of machine learning, deep learning, and pre-trained neural language models.

3.4 Machine and Deep Learning

Machine learning has played a significant role in the advancement of NLP (Jurafsky & Martin, 2023). Machine learning is a subfield of artificial intelligence (AI) – intelligence demonstrated by machines by imitating human-like behaviour (Goodfellow et al., 2016; Poole et al., 1998). According to Goodfellow et al. (2016), AI had some drawbacks when it came to solving tasks that were easy for people to perform but hard to describe formally, such as recognising spoken words or objects in images, because such AI systems relied on hard-coded knowledge. Machine learning alleviated this drawback by creating AI systems with the ability to acquire their own knowledge by extracting patterns from raw data (Goodfellow et al., 2016). Similarly, deep learning evolved out of machine learning to address certain weaknesses of machine learning (Géron, 2019; Goodfellow et al., 2016). Figure 1 illustrates the relationship between AI, machine learning, deep learning and NLP.

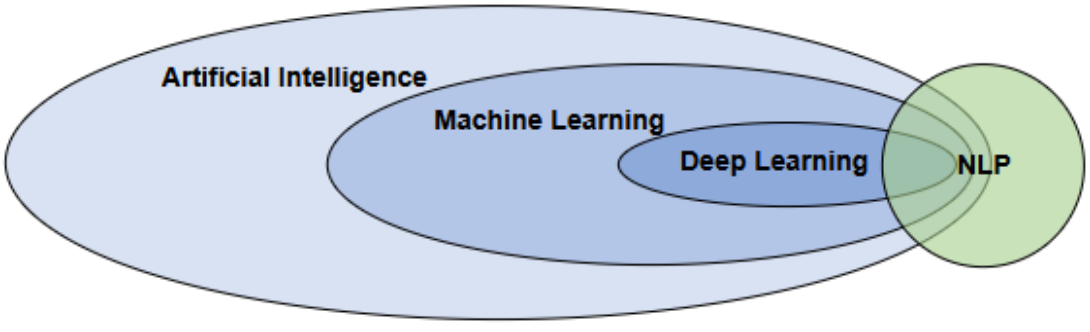

Figure 1: Machine learning, deep learning and NLP relationship within the realm of AI

Although traditional machine learning algorithms have the capability to solve a wide variety of problems, they have certain limitations. The handcrafted feature extraction used in traditional machine learning is not scalable for large-sized data sets (Shrestha & Mahmood, 2019). Also, due to the high-dimensional order of input data involved in speech and object recognition (Goodfellow et al., 2016), as well as text mining and NLP (Jurafsky & Martin, 2023), traditional machine learning algorithms have difficulty generalising when new input data is introduced. In fact, generalisation becomes exponentially more challenging as the dimensions of the data increase, thus rendering the generalisation mechanisms of traditional machine learning insufficient in high-dimensional spaces (Goodfellow et al., 2016).

Deep learning algorithms, such as deep neural networks (DNN) (see Figure 2), CNNs and recurrent neural networks (RNN), to name a few, can overcome this limitation (Bre et al., 2018; Géron, 2019; Goodfellow et al., 2016).

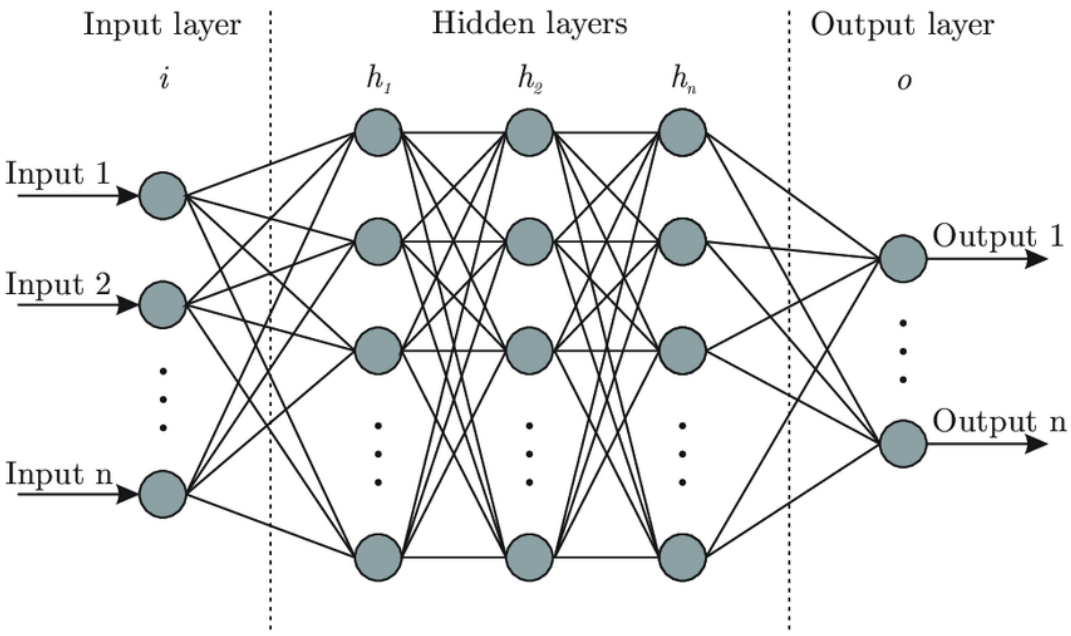

Figure 2: Deep neural network architecture with multiple hidden layers (Bre et al., 2018)

Deep learning algorithms also have the ability to learn features contained in the data automatically, instead of using features that were derived manually as in the case with traditional learning algorithms (Jurafsky & Martin, 2023).

Machine learning can be divided into four major types (see Figure 3): Supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning, each with its own set of appropriate algorithms (Géron, 2019; Goodfellow et al., 2016; Sarker, 2021).

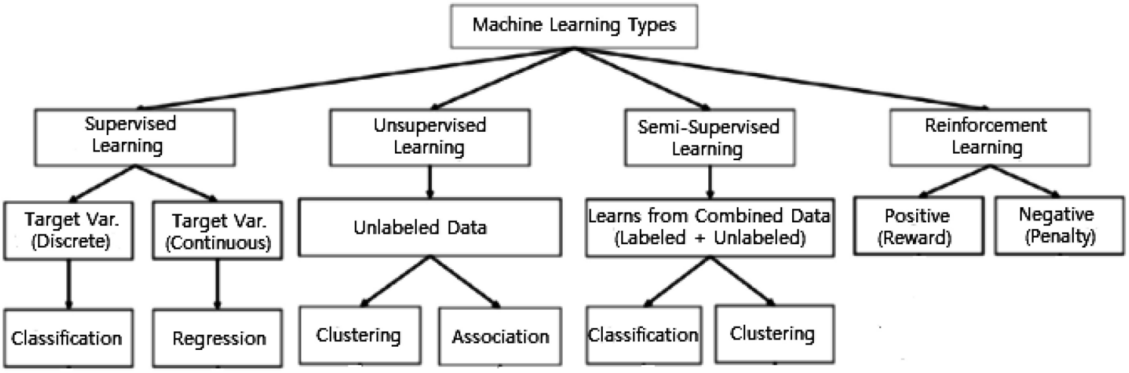

Figure 3: Machine learning types (Sarker, 2021)

3.4.1 Supervised Learning

With supervised learning, labelled input data is used to train a model so that the model will be able to determine the labels (future outcomes) of new input data (*6. Learning to Classify Text*, n.d.; Goodfellow et al., 2016; Jurafsky & Martin, 2023) as illustrated in Figure 4. Supervised machine learning can be used to solve classification problems or regression problems. Algorithms, such as naive Bayes, k-nearest neighbours, decision trees, random forest, support vector machines and logistic regression are better suited for classification problems where the output data contains discrete values, referred to as labels.

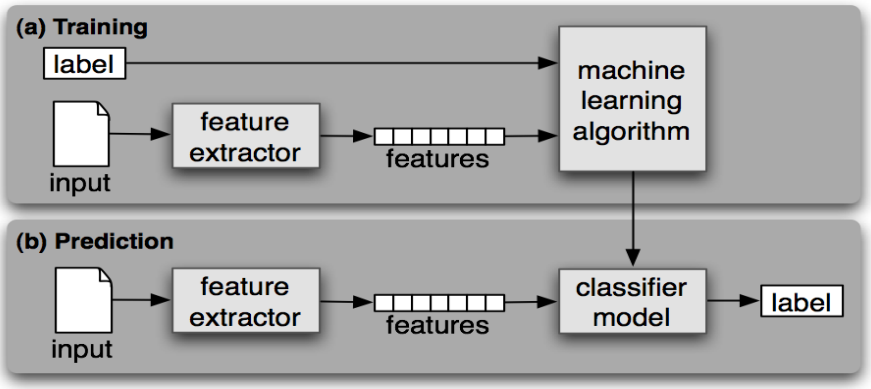

Figure 4: Supervised learning (6. Learning to Classify Text, n.d.)

Other algorithms, like linear regression, multiple regression, polynomial regression, support vector regression and decision tree regression are better suited for regression problems where the output data contains continuous values, referred to as target values (Géron, 2019). Neural networks like ANNs, CNNs and RNNs can be applied to either classification or regression problems (Valliani et al., 2019).

3.4.2 Unsupervised Learning

Unsupervised learning makes use of unlabelled input data and, therefore, implements algorithms that can find similarities, differences, patterns, or relationships within the data to get meaningful insight, segment the data into similar groups or to simplify the data (Géron, 2019; Igual & Seguí, 2017; Sharda et al., 2020). This is then used to automatically assign labels to the input data. These labels are essentially dividing the data into separate clusters, as depicted in Figure 5, when algorithms such as hierarchical clustering, spectral clustering or k-means are implemented. Alternatively, the data can be simplified by means of dimensionality reduction when algorithms such as principal component analysis are used (Géron, 2019; Igual & Seguí, 2017).

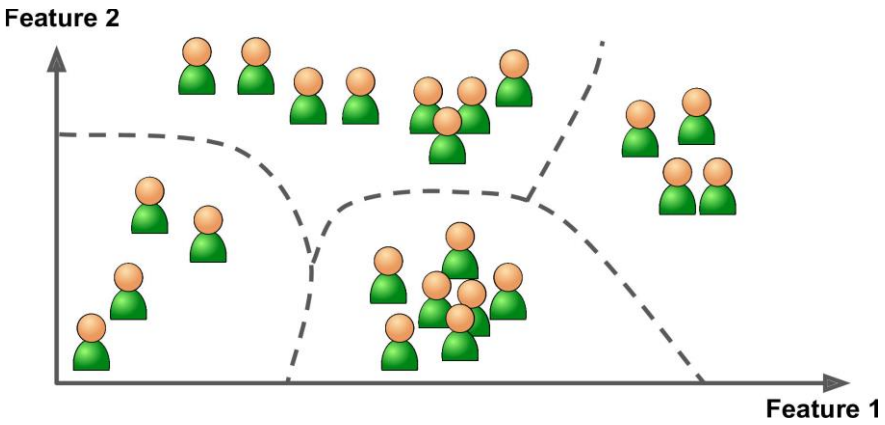

Figure 5: Unsupervised learning (Géron, 2019, p. 11)

3.4.3 Semi-supervised Learning

When only a small number of instances in the training dataset is labelled, semi-supervised learning can be used to propagate the labels to all the other instances (Géron, 2019; Goodfellow et al., 2016). Firstly, a small part of the data is either manually annotated or might already contain labels as depicted by the triangles and squares in Figure 6. This labelled subset is then used to train a base model using supervised learning algorithms. Thereafter, a process known as pseudo-labelling is applied where the base model is used to make predictions for the rest of the dataset, thereby generating labels for the unlabelled portion of the dataset. Finally, the pseudo-labelled data in combination with the original labelled subset is used to train an improved model (Géron, 2019). Semi-supervised learning has certain advantages when compared to supervised and unsupervised learning: (1) Semi-supervised learning can be applied to solve a variety of problems from classification and regression problems to clustering and association problems. (2) It also reduces manual annotation expenses and cuts down on data preparation time.

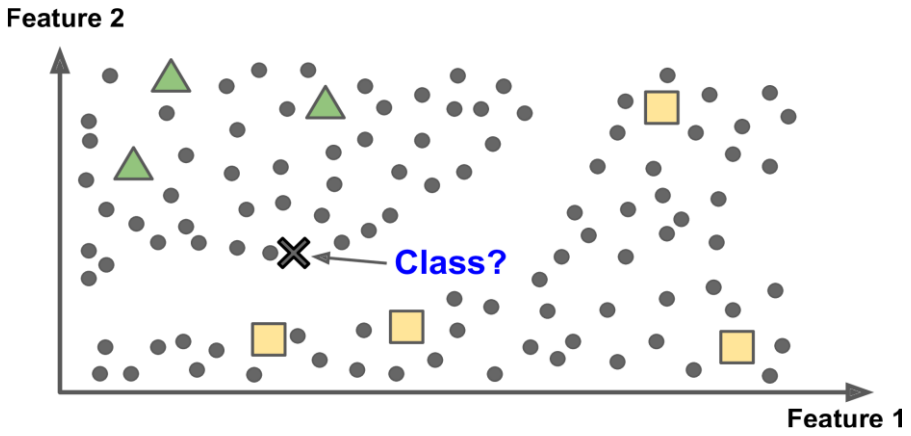

Figure 6: Semi-supervised learning (Géron, 2019, p. 14)

3.4.4 Reinforcement Learning

If the necessary amount of data to train a model is unobtainable, unreliable, or outdated, reinforcement learning can be utilised to overcome this drawback. Figure 7 illustrates how reinforcement learning is applied.

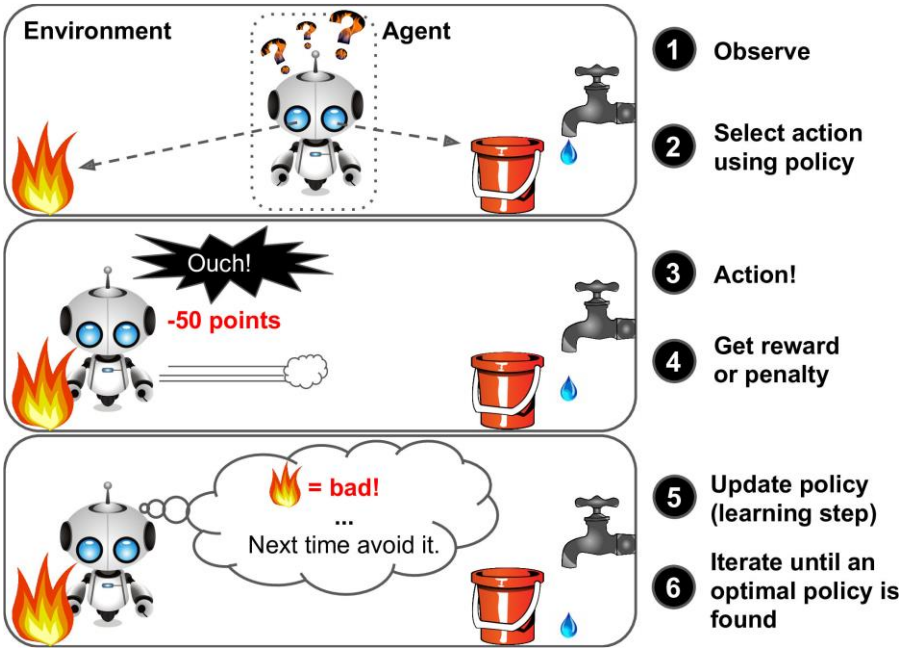

Figure 7: Reinforcement learning (Géron, 2019, p. 15)

Instead of using a dataset to train a model, reinforcement learning algorithms interact with the problem environment to reach solutions on its own. The algorithms are guided in the right direction by either rewarding or penalising the outcome of certain actions performed by the algorithm. The reward or penalty is then taken into account by means of a feedback loop the next time the algorithm performs an action (Géron, 2019; Goodfellow et al., 2016; Szepesvári, 2010).

Although traditional machine learning algorithms can be applied to text classification and other NLP problems (Piskorski & Jacquet, 2020), deep learning algorithms have had more successes, especially when the training and testing domains differ substantially and generalisation is, therefore, a top priority (Büyüköz et al., 2020). This is also evident from several projects that were discussed in Section 3.2. The demand to discover better methods of information extraction from text sources in an effort to advance event coding practices also contributed to the evolution of NLP (Hürriyetoğlu et al., 2020), which has researchers experimenting with transformers and pre-trained language models.

3.5 NLP and Pre-trained Language Models

ELIZA was one of the first NLP systems that implemented symbolic NLP methods in order to emulate natural language understanding by means of a limited conversation with a user (Jurafsky & Martin, 2023). Since then, NLP has evolved from symbolic methods such as regular expressions, to statistical NLP methods like n-grams, naive Bayes and logistic regression language models, to more advanced neural NLP methods such as neural language models, transformers, transfer learning and fine-tuning (Jurafsky & Martin, 2023).

Regular expressions were first defined by Kleene (1951, 1956) and used in text searching by Thompson (1968). N-gram language models started off with the use of Markov chains to predict whether a vowel or a consonant would be the next letter to follow in *Eugene Onegin*, a novel in verse written by Alexander Pushkin (Markov, 1913). It was also later implemented in speech recognition systems (Baker, 1975; Jelinek, 1976). The standard baseline for n-gram language models is based on experiments performed by Chen and Goodman (1996) as well as Goodman (2001) where different discounting algorithms, cache models, class-based models, and other language model parameters were compared.

When it comes to text classification, Maron (1961), and Mosteller and Wallace (1963) led the way; Maron with a naive Bayes classifier that was used to assign subject categories to journal abstracts, and Mosteller and Wallace with a Bayesian probabilistic model that could determine the authors of written essays. Naive Bayes has also been implemented in spam detection systems since 1998 (Sahami et al., 1998) as well as sentiment analysis systems (Liu & Zhang, 2012; Pang & Lee, 2008). In the 1990s logistic regression was commonly used in the field of information retrieval and speech processing (Jurafsky & Martin, 2023) by way

of language modelling, part-of-speech tagging and parsing, as well as text classification (Nigam et al., 1999; Ratnaparkhi, 1996, 1997; Rosenfeld, 1996).

Information retrieval and speech processing rely on the correct representation of words, sentences, and even documents (Jurafsky & Martin, 2023; Salton, 1971) to determine the true context of text sources. Sparse vector semantic models such as term frequency – inverse document frequency (TF-IDF) and positive pointwise mutual information (PPMI) are the original models capable of performing this task. TF-IDF can determine the similarity between words or documents while PPMI can determine the similarity between words only (Jurafsky & Martin, 2023). Dense vector semantic models, also referred to as embeddings, originated from latent semantic analysis (LSA) using singular value decomposition (SVD). LSA was also applied to perform spell checking, language modelling and essay grading to name a few (Coccaro & Jurafsky, 1998; Jones & Martin, 1997; Rehder et al., 1998). Other well-known statistical embedding methods include word2vec, doc2vec, fastText, continuous bag-of-words (CBOW) and Global Vectors (GloVe) (Jurafsky & Martin, 2023).

To alleviate the limitations of symbolic and statistical NLP methods, neural NLP methods were eventually introduced (Géron, 2019; Goodfellow et al., 2016) that could be applied to similar problems such as speech recognition, text classification and language modelling (Jurafsky & Martin, 2023). One example is a neural network model by Jaech et al. (2016) that was able to identify the language of written text. To accomplish this, it implemented a char2vec language model whereby two neural networks, a CNN and a long short term memory (LSTM) RNN, were trained together.

In the early twenty-first century Bengio et al. (2003) discovered that a simple feed-forward neural network (FFNN) could be used to create embeddings for neural language models. This eventually led to the development of RNN-based language models (Mikolov et al., 2010) that can be applied to sequence labelling, sequence classification (text classification), and text generation including the encoder-decoder architecture mostly known for machine translation (Jurafsky & Martin, 2023). Text generation is a fast-growing research field of which the latest advancements involve pre-trained language models such as bidirectional and auto-regressive transformers (BART) (M. Lewis et al., 2020), BERT (Devlin et al., 2019), a-light-BERT (ALBERT) (Lan et al., 2020), a distilled version of BERT (DistilBERT) (Sanh et al., 2019), ULMFiT (Howard & Ruder, 2018), and generative pre-trained transformer (GPT) models (Radford et al., 2018) that make use of bidirectional transformer encoders.

Transformers are non-recurrent networks that implement self-attention and positional embeddings to help represent time and how words relate to each other over long distances (Jurafsky & Martin, 2023). Two factors contributed to the popularity of pre-trained language models: (1) Their ability to be fine-tuned on application-specific parameters (Olsson et al., 2020) and (2) the advantage of contextual embeddings due to words being represented by different vectors each time their context changes (Jurafsky & Martin, 2023). It is, therefore, no surprise that several authors adopted the latest language modelling techniques to automate the extraction of socio-political events from news sources (Hürriyetoğlu et al., 2020), which forms the foundation of this project.

4. Problem Statement and Aim

As can be seen from literature (see Section 3.2), the vast majority of research pertaining to security event extraction and classification fixates on event coding, relegating event forecasting to the margins of investigation. Out of the whole list of projects discussed (see Table 2) only ICEWS and CEWSI implemented forecasting models as well, although the datasets produced from ACLED and UCDP were also utilised in later projects to develop forecasting models (Hegre et al., 2022; Vesco et al., 2022). Furthermore, the move towards machine and deep learning techniques provides a renewed sense of purpose for event coding in taking it a step further (Hürriyetoğlu et al., 2020).

This research project will produce feedback with regard to the accuracy and reliability of advanced machine and deep learning techniques when applied to event coding and forecasting. It will build and improve on the work of Chadefaux (2014, 2017, 2023) and O'Brien (2010) by using fine-grained international and local historical conflict-related news articles combined with current articles of social unrest as well as civil unrest (discussed in Section 1) within South Africa to train a fully automated state-of-the-art predictive model.

Several existing event coding projects rely on manual-coding (see Table 2). However, according to Leetaru and Schrodt (2013), human coders code more or less only six events per hour in comparison to TABARI, which codes about thirty million events per hour. Even PETRARCH, which is slower than TABARI, codes at least three hundred thousand events per hour. PETRARCH is slower because it applies deep parsing instead of shallow parsing methods, and it was programmed in Python instead of C# as TABARI was. Therefore, to

provide near real-time information about events taking place locally or around the world, automated event coding is a necessity, because automation increases the scale of event coding significantly.

Manual coding is also quite cognitively challenging for human coders (Nardulli et al., 2011), and although certain authors are of the opinion that human coders are the best way of ensuring that reliable, consistent, and accurate events are coded (Raleigh, 2020; Raleigh et al., 2010), others have shown that the accuracy of machine coding is similar to human coding (Schrodt & Gerner, 1994; Senekal & Kotzé, 2019). For instance, BIN, an automated text classification tool that screens news reports, is capable of identifying about 96 – 99% of reports relevant to social, political and economic events (Nardulli et al., 2011). According to tests that were conducted by Schrodt and Gerner (1994), the accuracy of human coding is actually quite low. Furthermore, Leetaru and Schrodt (2013, p. 15), found that "*the sustained decision-making required for human coding presents an almost perfect storm for inducing fatigue, inattention, and a tendency to use heuristic shortcuts.*"

Africa, and more specifically South Africa, has been the focus of very little research into event coding and forecasting initiatives. Senekal and Kotzé (2019) and Kotzé et al. (2020a) conducted research that involved the extraction and classification of security events focussing on domestic event data with regard to South Africa. The research by Salehyan et al. (2012, 2020) include event information about South Africa, but focuses on the whole African continent. Furthermore, Afrikaans, one of South Africa's twelve official languages, has also only been included in the research of Kotzé and Senekal (2019) and Kotzé et al. (2020a). Also, research conducted in this field focused mostly on international events, instead of domestic fine-grained events (Senekal & Kotzé, 2019). Lastly, the political, social and economic climate in South Africa warrants the need for the development of a civil unrest forecasting model (Alexander et al., 2018; Cook, 2020; Lee, 2018).

The aim of this research project is to design, develop and evaluate a model capable of forecasting civil unrest in South Africa and monitoring it in near real-time. The model should apply state-of-the-art machine and deep learning techniques as well as pre-trained language models such as BERT and GPT in order to improve on similar existing models and event coding methods. For near real-time updates the model will need an automated pipeline to collect and extract event information for monitoring purposes. Furthermore, the model should

focus on the South African context. Therefore, the model should be able to collect fine-grained event information from social media, blogs and online news sources in Afrikaans and English.

5. Thesis Statement

The research outcome of this thesis is the design, development and evaluation of a model with an automated pipeline to collect and extract security event information from Afrikaans and English social media, blogs and digital news sources by applying advanced machine and deep learning techniques in order to forecast and monitor civil unrest in South Africa in near real-time.

6. Research Questions

The following research questions have been formulated to address the problem statements of this research project:

- Firstly, which neural language modelling techniques can be implemented to create an automated model to extract and classify fine-grained socio-political, criminal and violent events, as well as natural threatening events pertaining to South Africa in an accurate and reliable way, from OSINT sources?
- Secondly, what machine or deep learning methods can be implemented to accurately and reliably monitor and forecast possible civil unrest in near real-time in countries like South Africa, using classified fine-grained socio-political, criminal, violent and natural threatening news events?
- Thirdly, to what extent can an automated pipeline combine the entire process from finding the relevant news sources, to extracting and classifying the event information, to monitoring and forecasting possible civil unrest in near real-time?
- Lastly, how successful can an automated pipeline utilise bilingual textual data sources of fine-grained socio-political, criminal, violent and natural threatening reports pertaining to South Africa?

7. Research Objectives

To answer the research questions pertaining to this project, the theoretical and empirical objectives that will measure the successful outcome of the project need to be defined (Dudovskiy, 2022).

7.1 Theoretical Objectives

The theoretical objectives for this research project include the following:

- To determine which neural language modelling techniques have the capability to extract and classify fine-grained event information, from Afrikaans and English OSINT sources, pertaining to the South African context in an accurate and reliable way.
- To determine which machine or deep learning methods have the capability to accurately and reliably monitor and forecast possible civil unrest in countries like South Africa.
- To design an automated pipeline that incorporates the whole process to provide near real-time feedback.

7.2 Empirical Objectives

Machine and deep learning methods are usually evaluated according to the following performance metrics: (1) Accuracy, (2) precision, (3) recall and (4) f1-score (Géron, 2019; Goodfellow et al., 2016; Jurafsky & Martin, 2023; O'Brien, 2002). Therefore, the empirical objectives for this research project include the following:

- To evaluate the selected neural language modelling techniques, that have the capability to extract and classify fine-grained event information, and compare them with each other according to their accuracy, precision, recall, and f1-scores.
- To evaluate the selected machine or deep learning methods, that have the capability to monitor and forecast possible civil unrest, and compare them with each other according to their accuracy, precision, recall, and f1-scores.
- To evaluate the final automated pipeline that incorporates the whole process and compare it with current models or processes according to effectiveness and efficiency.

8. Hypotheses

To evaluate the performance of the selected language modelling techniques' capability to extract and classify fine-grained event information according to their accuracy, precision, recall and f1-scores, the following null hypotheses were formulated:

- $H_{0,1}$: There is no significant difference in the accuracy scores between the selected neural language modelling techniques, used to extract and classify event information.
- $H_{0,2}$: There is no significant difference in the precision scores between the selected neural language modelling techniques, used to extract and classify event information.
- $H_{0,3}$: There is no significant difference in the recall scores between the selected neural language modelling techniques, used to extract and classify event information.
- $H_{0,4}$: There is no significant difference in the f1-scores between the selected neural language modelling techniques, used to extract and classify event information.

To evaluate the performance of the selected machine or deep learning methods' capability to monitor and forecast possible civil unrest according to their accuracy, precision, recall and f1-scores, the following null hypotheses were formulated:

- $H_{0,5}$: There is no significant difference in the accuracy scores between the selected machine or deep learning methods, used to monitor and forecast possible civil unrest.
- $H_{0,6}$: There is no significant difference in the precision scores between the selected machine or deep learning methods, used to monitor and forecast possible civil unrest.
- $H_{0,7}$: There is no significant difference in the recall scores between the selected machine or deep learning methods, used to monitor and forecast possible civil unrest.
- $H_{0,8}$: There is no significant difference in the f1-scores between the selected machine or deep learning methods, used to monitor and forecast possible civil unrest.

To evaluate the effectiveness and efficiency of the final automated pipeline model that incorporates the whole process, the following null hypotheses were formulated:

- $H_{0,9}$: There is no significant difference in performance (accuracy, precision, recall and f1-scores) between using an automated pipeline and not using an automated pipeline.

- $H_{0,10}$: There is no significant time difference between producing forecasting results using an automated pipeline and producing forecasting results without using an automated pipeline.

9. Research Design and Methodology

To successfully answer the research questions pertaining to the project, it is important to determine which data is needed, as well as the way in which the data will be obtained. However, before data collection techniques and analysis procedures can be decided upon, the underlying aspects need to be determined as illustrated by the research onion of Saunders et al. (2023).

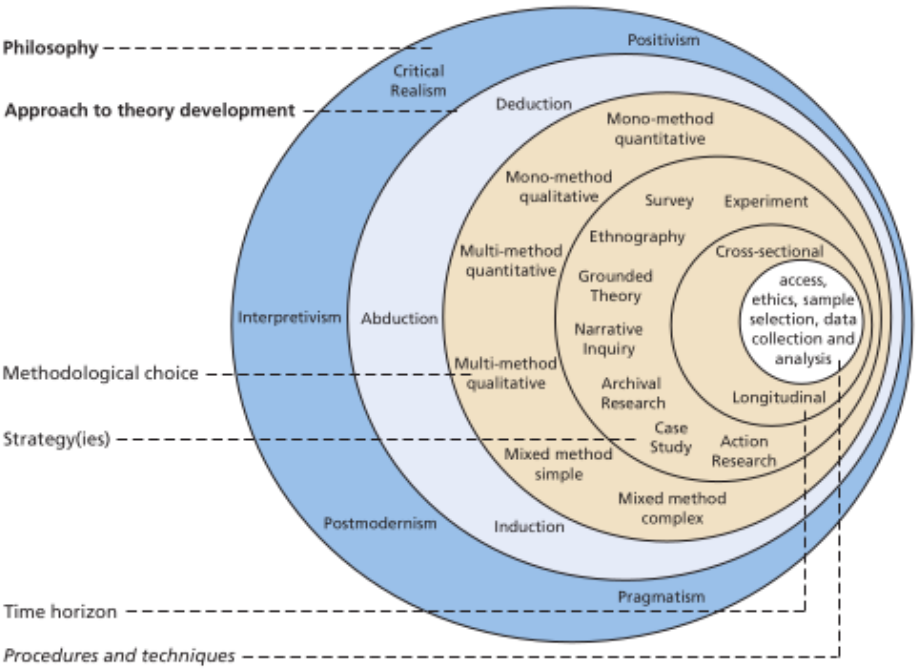

Figure 8: The research onion (Saunders et al., 2023, p. 131)

9.1 Research Philosophy

To justify the research methodology and strategy of this project, the research philosophy needs to be ascertained (Crotty, 1998). New knowledge will be generated while conducting the research. The research philosophy that will be adopted relates to the new knowledge that will be generated as well as the nature thereof. The philosophical stance plays a significant role in understanding the research project and is described by means of ontological, epistemological and axiological perspectives (Saunders et al., 2023). This research project

will be conducted from a positivist perspective. Ontologically, this entails perceiving social entities as real and observable as physical objects while conducting research. Epistemologically, the research focuses on the discovery of observable and measurable facts and patterns to produce credible and meaningful data. Because of the quantifiable data, the axiological perspective entails conducting research in a value-free way as far as possible without influencing the findings by remaining neutral (Saunders et al., 2023).

9.2 Research Approach

From the project's research objectives (see Section 7) and hypotheses (see Section 8) it is clear that it involves the formulation of a theory or theories that will be tested to determine their truthfulness. Therefore, the project will be approached in a deductive manner where the null hypotheses are evaluated and either rejected or not rejected (Saunders et al., 2023). Note that according to Bower and Colton (2003) the null hypothesis can never be "accepted".

9.3 Research Methodology

The research methodology of the study is influenced by the adopted research philosophy and approach, with the research objectives and data collection methods also playing a role (Saunders et al., 2019). Therefore, since the research philosophy (see Section 9.1) and approach (see Section 9.2) of this project involve discovering real facts that are measurable and quantifiable, a quantitative research design will be most appropriate for this project. Also, since the research objectives (see Section 7) involve the determination and evaluation of different language modelling techniques' as well as machine and deep learning methods' capabilities, the study's mode of inquiry will be of an exploratory and evaluative nature. Furthermore, the sources from where data will be collected include online news sites, social media, blogs, and forums (see Section 4), as well as event data repositories from existing projects (see Section 3.2) that will be used for training data. Considering all these aspects, the complete methodology of the research project can be defined as an exploratory and evaluative study, using a multi-method quantitative research design.

9.4 Research Strategy

The South African Civil Unrest Prediction (SACUP) model should be able to monitor acts of civil unrest as close to real-time as possible for the model to be of value. It should also be able to accurately and reliably forecast possible civil unrest. Therefore, this project will implement an experimental research strategy (Saunders et al., 2019) where several machine

and deep learning algorithms will be experimented with, and compared, to determine which one yields the most accurate and reliable forecasts. The experimental strategy will be based on the cross industry standard process for data mining (CRISP-DM), which is viewed as *"the most popular framework for executing data science projects"* (Saltz, 2022).

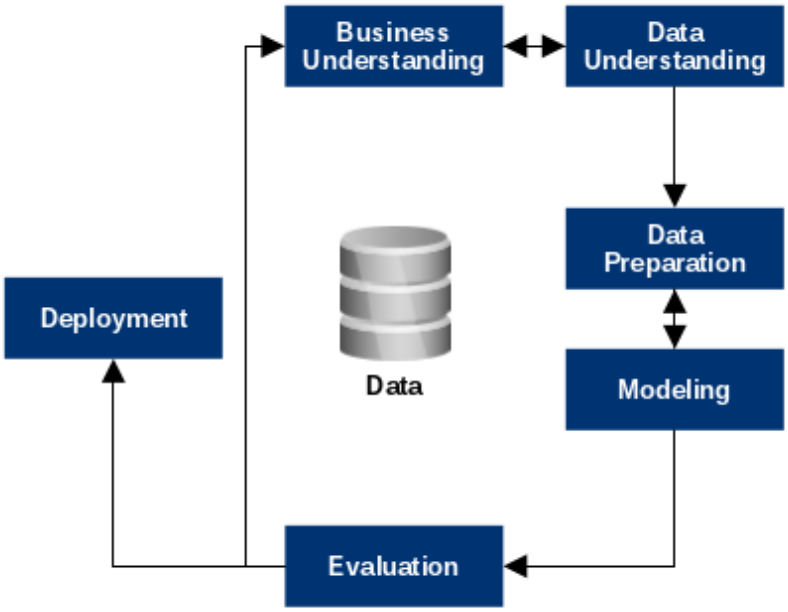

Figure 9: The CRISP-DM process model

The CRISP-DM process model outlines six stages involved in the data science life cycle. These stages are briefly described below:

- **Business understanding:** This stage focuses on determining the objectives and requirements of the project by understanding the customer's needs.
- **Data understanding:** To accomplish the project goals, the necessary data needs to be identified, collected, and analysed.
- **Data preparation:** During this stage the data is pre-processed to ensure that it conforms to the correct format necessary for modelling.
- **Modelling:** Various models, based on different modelling techniques and algorithms, are trained and assessed.
- **Evaluation:** The trained models are then evaluated to determine which model(s) best meets the objectives.
- **Deployment:** The final stage then either entails generating a report regarding the research findings or implementing a system that applies the final model to satisfy the objectives that were identified during the first stage.

9.5 Time Horizon

Event data can be collected from social media platforms such as X (Twitter), Facebook, Signal and Whatsapp, online news sites like News 24 and Netwerk 24, as well as forums and blogs. These sources all generate new content constantly and the time horizon for this project would, therefore, be best described as longitudinal.

9.6 Data Collection and Analysis

The purpose of conducting the experimental and evaluative study is to experiment with multiple language modelling as well as machine and deep learning algorithms and techniques to determine which combination of these algorithms and techniques will produce the most accurate model for each stage of the pipeline with the capability of monitoring acts of civil unrest in near real-time, and also forecast possible civil unrest. For the forecast to be of value and provide meaningful information, a time frame and confidence interval would have to be calculated in which the possible civil unrest might break out.

For the first stage of the pipeline, social media and other digital online text-based OSINT sources, that can provide possible event information most applicable to South Africa, need to be identified. This includes sources that provide event data that would usually be missed by international sources. It can also include OSINT sources from other countries that reference or influence South Africa. An article from Senekal and Kotzé (2019), sums up several such OSINT sources. The more OSINT sources (focussing on text-based sources) can be identified and included in the project, the larger the dataset will be for model training, validation and testing. Automated data collection tools, such as web scrapers and APIs (for example the X (Twitter) API) can be utilised to ensure that the pipeline is fully automated.

From the selected sources needs to be determined which socio-political, criminal, violent and natural threatening incidents (i.e. demonstrations, riots, murders, violence and other forms of lawlessness, as well as economic and political events) are good indicators for possible civil unrest. This will be identified from studies already conducted (see section 3.2 and 3.3), as well as by experimenting with incidents that might be relevant but has not yet been identified in previous studies. All experiments will be conducted using Python libraries such as scikit-learn (Pedregosa et al., 2011), Tensorflow (Abadi et al., 2016), Keras (Chollet, 2017), PyTorch (Paszke et al., 2019), Gensim (Rehurek & Sojka, 2010) and NTLK (Bird et al., 2009).

Once the relevant data sources have been collected, the next step is to perform feature engineering and construct vector spaces. This will include hand-crafted linguistic features (surface, syntactical, lexicon) and word embeddings (semantic). Word2vec (Mikolov et al., 2013), fastText (Bojanowski et al., 2017) and GloVe (Pennington et al., 2014) will be used for the static (non-contextual) word embeddings. Doc2vec (Le & Mikolov, 2014) will be used to learn embeddings to convert paragraphs into a fixed-dimensional vector representation, thus creating static (non-contextual) document embeddings. In addition to these, state-of-the-art contextual embeddings will also be created using BART (M. Lewis et al., 2020), BERT (Devlin et al., 2019), ALBERT (Lan et al., 2020), DistilBERT (Sanh et al., 2019), ULMFiT (Howard & Ruder, 2018) and the latest versions of GPT (Radford et al., 2018) to construct language models that are capable of extracting event information.

Techniques such as Chi^2 and information gain will be applied to determine the most relevant hand-crafted features. Following that, machine learning techniques will be applied to these hand-crafted features and experimentation with different algorithms such as naive Bayes (D. D. Lewis, 1998), SVM (Noble, 2006), logistic regression (Demaris, 1995), random forest (Breiman, 2001) and XGBoost (T. Chen & Guestrin, 2016) will commence to create event classification models. In addition, deep learning algorithms such as CNNs (Kim, 2014), LSTM RNNs (Sherstinsky, 2020) and GRU RNNs (Ravuri & Stolcke, 2016) will also be applied to classify extracted event information.

Before commencing with the second stage of the pipeline, it is important to evaluate and compare all the event classification models and select the best performing model. Performance metrics such as accuracy, precision, recall and f1-scores, provided by the relevant scikit-learn libraries for Python (Pedregosa et al., 2011), as well as learning curves and confusion matrices will be used for this process.

Stage two will use the relevant event classification data as input. Next, stage two would have to create a feature vector space from the classified event information data and other possible variables that were identified from civil unrest research projects conducted in other countries. The feature vector space will be used as training data to experiment with different machine learning algorithms such as naive Bayes (D. D. Lewis, 1998), SVM (Noble, 2006), logistic regression (Demaris, 1995), random forest (Breiman, 2001) and XGBoost (T. Chen & Guestrin, 2016), as well as deep learning algorithms such as CNNs (Kim, 2014), LSTM RNNs

(Sherstinsky, 2020) and GRU RNNs (Ravuri & Stolcke, 2016) to create a forecasting model for possible civil unrest in South Africa. Again, the accuracy, precision, recall and f1-scores, combined with learning curves and confusion matrices will be used to evaluate and compare the different forecasting models to determine which model provides the best forecast.

9.7 The SACUP model

Civil unrest can be defined as activities that arise from a mass act of civil disobedience, such as demonstrations, riots, strikes, murders or other forms of lawlessness, in which the participants become hostile toward authority and the authorities are unable to maintain public safety and order (Braha, 2012). As an initial attempt, acts of civil unrest will be divided into four main categories, which will most likely be revised as the project progresses:

- Military conflict – stemming from political factors and socio-economic factors.
- Crime waves – armed robberies, cash in transit robberies, mass murders, etc.
- Mass violence – stemming from demonstrations, riots, strikes and violent incitement (including hate speech).
- Anarchy – due to natural threats and disasters.

The final product will consist of an automated pipeline that implements language models and machine/deep learning models across two separate stages:

- The first stage's model should automatically extract and classify socio-political, criminal, violent and natural threatening news events, and be used to monitor the current rate of these events that could lead to civil unrest.
- Possible civil unrest will then be forecasted by the model in the second stage by using these events as variable inputs for a machine or deep learning algorithm (that will have to be identified from several possible algorithms) to create a fully-fledged civil unrest prediction model.

The SACUP model (see Figure 10) will form a civil unrest warning system that will monitor all forms of activities (events) that can lead to civil unrest and use the current state of events to determine the possibility of civil unrest in any of the four specified categories with a time frame and confidence interval.

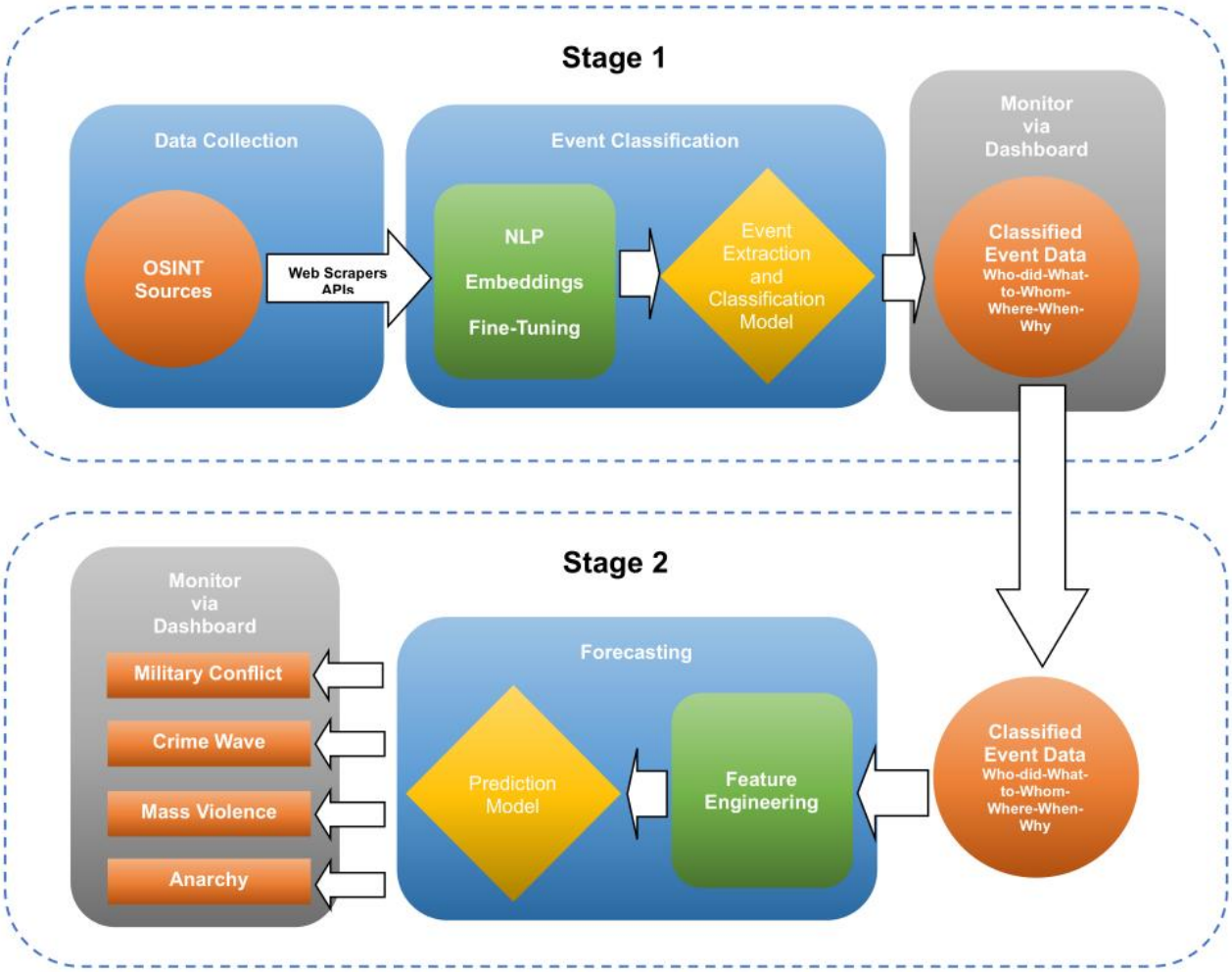

Figure 10: The SACUP model

10. Contribution and Value of the Research

The value that this research project will add to the body of knowledge can be expressed as theoretical and practical contributions.

10.1 Theoretical Contributions

This research project will contribute to the body of knowledge by developing a civil unrest language corpus for Afrikaans and English that is specific to South Africa but could also be applied in other African or third world countries that face similar challenges. The possible identification of new OSINT sources relevant to civil unrest will also add to the body of knowledge. Furthermore, new classification and forecasting models will be produced, which will make a significant contribution to the state-of-the art of the discipline.

10.2 Practical Contributions

The outcome of these models will produce new methods for event coding and also increase the focus on forecasting possible civil unrest, instead of just monitoring the situation. Finally, the automated civil unrest early warning system will provide the citizens of South Africa with peace of mind in times of peace and a sense of preparedness for when civil unrest might erupt.

11. Limitations of the Project

This research project will not use video, audio or image-based data sources, only natural language text-based data sources, because the focus of the study is on NLP, not image, video or audio recognition. Although audio can be converted into text (Basystiuk et al., 2021) from which data can be obtained, text-based data sources provide more than enough information. Also, image recognition and audio to text conversion is totally a different research field and can probably produce a whole thesis on its own. Images can have multiple interpretations (Petrou & Petrou, 2010), which might provide inaccurate information if the image is interpreted incorrectly. However, future research into the possible inclusion of audio, image and video sources might be considered.

12. Ethical Considerations

Possible ethical issues have to be considered, since data will be collected from various OSINT sources. Most OSINT source data is publicly obtainable from social media such as X (Twitter) and Facebook, as well as online news sites, forums and blogs, while data from other sources like Signal and Whatsapp are only obtainable when subscribed to a group. Therefore, to collect data from Signal and Whatsapp groups (Kotzé et al., 2020a), consent would need to be obtained from the group administrators and members. Other OSINT source data, even though publicly obtainable, would still need to be handled in a confidential manner in an effort to minimize possible exposure and harm to individuals' reputations.

13. Estimated Budget

At certain stages of the project's life cycle the processes involved will be subject to expenditure. Table 3 provides a list of possible expenses.

CRISP-DM	Process	Description	Estimated Amount
Data understanding	Data collection	To collect data from online news sources a subscription fee will need to be paid	R4 000
Data preparation	Data annotation	Annotators will need to be paid for annotating the training data	R10 000
Modelling	Event classification	To use the GPT-3.5 Turbo and GPT-4 APIs it is necessary to acquire tokens	R2 000
			R16 000

Table 3: Estimated budget

14. Thesis Layout

The thesis for this research project will use the following layout:

- Abstract
- Table of Contents
- List of Tables
- List of Figures
- Chapter 1 – Introduction
- Chapter 2 – Literature Review
- Chapter 3 – Research Design and Methodology
- Chapter 4 – Model Design
- Chapter 5 – Research Findings
- Chapter 6 – Conclusion
- References
- Appendices

15. Research Schedule

The following schedule is proposed that will allow the completion of the project within approximately two years:

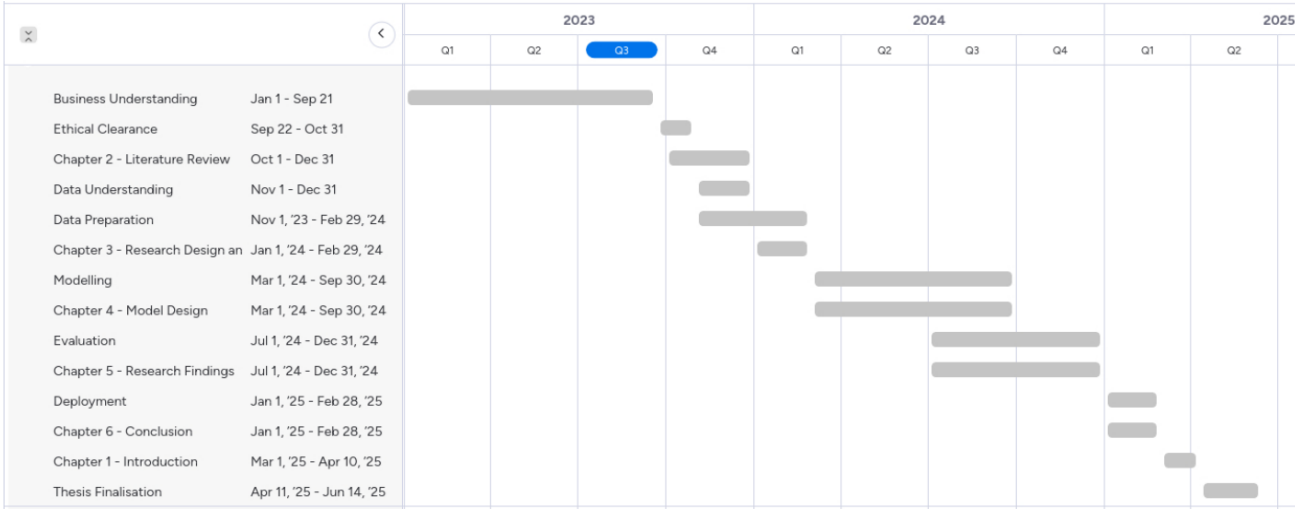

Figure 11: Proposed research schedule

16. References

6. Learning to Classify Text. (n.d.). Retrieved August 27, 2023, from https://www.nltk.org/book/ch06.html

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., & Others, A. (2016). TensorFlow: A System for Large-Scale Machine Learning. *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI '16)*, 265–283. https://www.usenix.org/conference/osdi16/technical-sessions/presentation/abadi

Adhikari, A., Ram, A., Tang, R., & Lin, J. (2019). *DocBERT: BERT for Document Classification*. http://arxiv.org/abs/1904.08398

Alexander, P., Runciman, C., Ngwane, T., Moloto, B., Mokgele, K., & van Staden, N. (2018). Frequency and turmoil. *Sa Crime Quarterly No. 63*, *63*, 27–42.

Baker, J. K. (1975). The DRAGON system--An overview. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *23*(1), 24–29. https://doi.org/10.1109/TASSP.1975.1162650

Basystiuk, O., Shakhovska, N., Bilynska, V., Syvokon, O., Shamuratov, O., & Kuchkovskiy, V. (2021). The Developing of the System for Automatic Audio to Text Conversion. *IT&AS*, 1–8.

Beieler, J. (2016). Creating a Real-Time, Reproducible Event Dataset. *ArXiv Preprint ArXiv:1612.00866*. http://arxiv.org/abs/1612.00866

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A Neural Probabilistic Language Model. *Journal of Machine Learning Research*, *3*, 1137–1155.

Berndt, D. J., & Clifford, J. (1994). Using dynamic time warping to find patterns in time series. *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, 359–370.

Bhavnani, R., Donnay, K., Miodownik, D., Mor, M., & Helbing, D. (2014). Group Segregation and Urban Violence. *American Journal of Political Science*, *58*(1), 226–245. https://doi.org/10.1111/ajps.12045

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc."

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146. https://doi.org/10.1162/tacl_a_00051

Bower, K. M., & Colton, J. A. (2003). Why We Don't "Accept" the Null Hypothesis. *American Society for Quality, Six Sigma Forum*, 16. http://trustminitab.com/uploadedFiles/Shared_Resources/Documents/Articles/not_accepting_null_hypothesis.pdf

Braha, D. (2012). Global Civil Unrest: Contagion, Self-Organization, and Prediction. *PLoS ONE*, *7*(10), 1–9. https://doi.org/10.1371/journal.pone.0048596

Brandt, P. T., & Freeman, J. R. (2006). Advances in Bayesian time series modeling and the study of politics: Theory testing, forecasting, and policy analysis. *Political Analysis*, *14*(1), 1–36. https://doi.org/10.1093/pan/mpi035

Bre, F., Gimenez, J. M., & Fachinotti, V. D. (2018). Prediction of wind pressure coefficients on building surfaces using artificial neural networks. *Energy and Buildings*, *158*(April 2018), 1429–1441. https://doi.org/10.1016/j.enbuild.2017.11.045

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32.

Büyüköz, B., Hürriyetoğlu, A., & Özgür, A. (2020). Analyzing ELMo and DistilBERT on Socio-political News Classification. *Proceedings of the Workshop on Automated Extraction of Socio-Political Events from News 2020*, *May*, 9–18. https://www.aclweb.org/anthology/2020.aespen-1.4

Cederman, L. E. (2002). Endogenizing geopolitical boundaries with agent-based modeling. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(SUPPL. 3), 7296–7303. https://doi.org/10.1073/pnas.082081099

REPUBLIC OF SOUTH AFRICA FIREARMS CONTROL AMENDMENT BILL, (2021) (testimony of Bheki Cele).

Chadefaux, T. (2014). Early warning signals for war in the news. *Journal of Peace Research*, *51*(1), 5–18. https://doi.org/10.1177/0022343313507302

Chadefaux, T. (2017). Conflict forecasting and its limits. *Data Science*, *1*(1–2), 7–17. https://doi.org/10.3233/ds-170002

Chadefaux, T. (2022). A shape-based approach to conflict forecasting. *International Interactions*, *48*(4), 633–648. https://doi.org/10.1080/03050629.2022.2009821

Chadefaux, T. (2023). An automated pattern recognition system for conflict. *Journal of Computational Science*, *72*(May). https://doi.org/10.1016/j.jocs.2023.102074

Chen, S. F., & Goodman, J. (1996). An empirical study of smoothing techniques for language modeling. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, *1996-June*(June), 310–318. https://doi.org/10.3115/981863.981904

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 785–794.

Chollet, F. (2017). Introduction to Keras. In *Deep Learning with Python*.

CIVIL DISOBEDIENCE | meaning in the Cambridge English Dictionary. (n.d.). Retrieved August 11, 2020, from https://dictionary.cambridge.org/dictionary/english/civil-disobedience

Civil disorder dictionary definition | civil disorder defined. (n.d.). Retrieved August 11, 2020, from https://www.yourdictionary.com/civil-disorder

Coccaro, N., & Jurafsky, D. (1998). Towards Better Integration of Semantic Predictors in Statistical Language Modeling. *5th International Conference on Spoken Language Processing (ICSLP 1998)*. https://doi.org/10.21437/icslp.1998-642

Cook, N. (2020). *South Africa: Current Issues, Economy, and U.S. Relations*. https://crsreports.congress.gov

Crotty, M. J. (1998). *The Foundations of Social Research : Meaning and Perspective in the Research Process* (p. 256). SAGE Publications Ltd. http://digital.casalini.it/9781446283134

Demaris, A. (1995). A Tutorial in Logistic Regression. *Journal of Marriage and Family*, *57*(4), 956–968.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*(Mlm), 4171–4186.

Dudovskiy, J. (2022). *An ultimate guide to writing a dissertation in business studies: A step-by-step assistance*.

Eck, K., & Hultman, L. (2007). One-sided violence against civilians in war: Insights from new fatality data. *Journal of Peace Research*, *44*(2), 233–246. https://doi.org/10.1177/0022343307075124

Euginia, K. (2017). *The effects of sentiments on the dollar rand (USD/ZAR) exchange rate*. University of the Witwatersrand.

Ganino, G., Lembo, D., Mecella, M., & Scafoglieri, F. (2018). Ontology population for open-source intelligence: A GATE-based solution. *Software - Practice and Experience*, *48*(12), 2302–2330. https://doi.org/10.1002/spe.2640

Gebremichael, M., Feyissa, T. K., Kidane, A., Mesfin, E., & Belay, T. (2019). *KINGDOM OF CONFLICT INSIGHT* (Vol. 1, Issue January).

Géron, A. (2019). Hands-on Machine Learning whith Scikit-Learing, Keras and Tensorfow. In *O'Reilly Media, Inc*.

Gleditsch, K. S., & Ward, M. D. (2013). Forecasting is difficult, especially about the future: Using contentious issues to forecast interstate disputes. *Journal of Peace Research*, *50*(1), 17–31. https://doi.org/10.1177/0022343312449033

Goldstone, J. A. (2008). *Using Quantitative and Qualitative Models to Forecast Instability* (Vol. 204, Issue March). www.usip.org

Goldstone, J. A., Bates, R. H., Epstein, D. L., Gurr, T. R., Marshall, M. G., Ulfelder, J., & Woodward, M. (2010). A Global Model for Forecasting Political Instability. *American Journal OfPolitical Science*, *54*(1), 190–208.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Goodman, J. T. (2001). A bit of progress in language modeling. *Computer Speech & Language*, *15*(4), 403–434.

Goodwin, P. (2002). Forecasting games: Can game theory win? *International Journal of Forecasting*, *18*(3), 369–374. https://doi.org/10.1016/S0169-2070(02)00022-5

Green, K. C., & Armstrong, J. S. (2007). The Ombudsman: Value of expertise for forecasting decisions in conflicts. *Interfaces*, *37*(3), 287–293. https://doi.org/10.1287/inte.1060.0262

Hegre, H., Metternich, N. W., Nygård, H. M., & Wucherpfennig, J. (2017). Introduction: Forecasting in peace research. *Journal of Peace Research*, *54*(2), 113–124. https://doi.org/10.1177/0022343317691330

Hegre, H., Vesco, P., & Colaresi, M. (2022). Lessons from an escalation prediction competition. *International Interactions*, *48*(4), 521–554. https://doi.org/10.1080/03050629.2022.2070745

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, *1*, 328–339. https://doi.org/10.18653/v1/p18-1031

Hürriyetoğlu, A., Zavarella, V., Tanev, H., Yörük, E., Safaya, A., & Mutlu, O. (2020). Automated Extraction of Socio-political Events from News (AESPEN): Workshop and Shared Task Report. *Proceedings of the Workshop on Automated Extraction of Socio-Political Events from News 2020*, *May*, 1–6. http://arxiv.org/abs/2005.06070

Igual, L., & Seguí, S. (2017). Introduction to Data Science: A Python Approach to Concepts, Techniques and Applications. In *Springer International Publishing*.

Jaech, A., Mulcaire, G., Hathi, S., Ostendorf, M., & Smith, N. A. (2016). Hierarchical Character-Word Models for Language Identification. *Proceedings of the Fourth International Workshop on Natural Language Processing for Social Media*, 84–93. https://doi.org/10.18653/v1/w16-6212

Jelinek, F. (1976). Continuous speech recognition by statistical methods. *Proceedings of the IEEE*, *64*(4), 532–556. https://doi.org/10.1109/72.286885

Jones, M. P., & Martin, J. H. (1997). Contextual Spelling Correction Using Latent Semantic Analysis. *5th Conference on Applied Natural Language Processing*, 166–173. https://doi.org/10.3115/974557.974582

Jurafsky, D., & Martin, J. H. (2023). *Speech and Language Processing, (January 7, 2023 draft)*. Prentice Hall.

Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. ArXiv. *Preprint*.

Kleene, S. C. (1951). *Representation of Events in Nerve Nets and Finite Automata*.

Kleene, S. C. (1956). REPRESENTATION OF EVENTS IN NERVE NETS AND FINITE AUTOMATA^1. *Automata Studies: Annals of Mathematics Studies. Number 34*, *34*, 3.

Kotzé, E., & Senekal, B. (2019). Oopbronintelligensie (OSINT) vir veiligheidsdoeleindes: Die ontwikkeling van ’n dataontledingspyplyn om relevante WhatsAppboodskappe te ontleed. *Suid-Afrikaanse Tydskrif Vir Natuurwetenskap En Tegnologie*, *38*(1), 1–10. https://doi.org/10.36303/satnt.2019.38.17.717

Kotzé, E., Senekal, B., & Daelemans, W. (2020a). Automatic classification of social media reports on violent incidents in South Africa using machine learning. In *South African Journal of Science* (Vol. 116, Issues 3–4). Academy of Science of South Africa. https://doi.org/10.17159/sajs.2020/6557

Kotzé, E., Senekal, B., & Daelemans, W. (2020b). Exploring the Classification of Security Events using Sparse and Dense Representation of Text. *2020 International SAUPEC/RobMech/PRASA Conference*, 1–6. https://doi.org/10.1109/SAUPEC/RobMech/PRASA48453.2020.9041092

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). Albert: A Lite Bert for Self-Supervised Learning of Language Representations. *8th International Conference on Learning Representations, ICLR 2020*, 1–17.

Lancaster, L., & Newham, G. (2020). *The State of Crime and Safety in SA Cities*. www.saferspaces.org.za.

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. *International Conference on Machine Learning*, 1188–1196.

Lee, M. (2018). *Civil Unrest in South Africa : Insights from Cognitive Linguistics and Critical Discourse Analysis By*. *March*.

Leetaru, K., & Schrodt, P. A. (2013). GDELT: Global Data on Events, Location and Tone, 1979-2012. *ISA Annual Convention*, *2*(4), 1–49.

http://data.gdeltproject.org/documentation/ISA.2013.GDELT.pdf

Lewis, D. D. (1998). Naive (Bayes) at forty: The independence assumption in information retrieval. *European Conference on Machine Learning*, 4–15.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020). *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension*. 7871–7880. https://doi.org/10.18653/v1/2020.acl-main.703

Liu, B., & Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In *Mining text data* (pp. 415–463). Springer.

Longa, S. (2018). *ANALYSIS OF FACTORS INFLUENCING PROVISION OF MUNICIPAL SERVICES IN THE RURAL DISTRICTS: THE CASE STUDY OF LUWINGU DISTRICT COUNCIL OF ZAMBIA*.

Markov, A. A. (1913). An example of statistical analysis of the text of eugene onegin illustrating the association of trials into a Chain. *Bulletin de LAcadamie Imperiale Des Sciences de St. Petersburg, Ser*, *6*, 153162.

Maron, M. E. (1961). Automatic Indexing: An Experimental Inquiry. *Journal of the ACM (JACM)*, *8*(3), 404–417. https://doi.org/10.1145/321075.321084

Mavunga, G. (2019). #FeesMustFall Protests in South Africa: A Critical Realist Analysis of Selected Newspaper Articles. *Journal of Student Affairs in Africa*, *7*(1), 81–99. https://doi.org/10.24085/jsaa.v7i1.3694

Mikolov, T., Karafiát, M., Burget, L., Černocký, J., & Khudanpur, S. (2010). Recurrent neural network based language model. *Interspeech*, *2*(3), 1045–1048.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 3111–3119. https://doi.org/10.18653/v1/d16-1146

Montgomery, J. M., Hollenbach, F. M., & Ward, M. D. (2012). Improving predictions using ensemble bayesian model averaging. *Political Analysis*, *20*(3), 271–291. https://doi.org/10.1093/pan/mps002

Mosteller, F., & Wallace, D. L. (1963). Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed Federalist Papers. *Journal of the American Statistical Association*, *58*(302), 275–309.

Muchlinski, D., Siroky, D., He, J., & Kocher, M. (2016). Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data. *Political Analysis*, *24*(1), 87–103. https://doi.org/10.1093/pan/mpv024

Nardulli, P. F., Leetaru, K. H., & Hayes, M. (2011). Event data, civil unrest and the social, political and economic event database (SPEED) project: post World War II trends in political protests and violence. *Annual Meeting of the International Studies Association*.

Nigam, K., Lafferty, J., & Mccallum, A. (1999). Using Maximum Entropy for Text Classification. *IJCAI-99 Workshop on Machine Learning for Information Filtering*, *1*(1), 61–67.

Noble, W. S. (2006). What is a support vector machine? *Nature Biotechnology*, *24*(12), 1565–1567. https://doi.org/10.1038/nbt1206-1565

O'Brien, S. P. (2002). Anticipating the good, the bad, and the ugly: An early warning approach to conflict and instability analysis. *Journal of Conflict Resolution*, *46*(6), 791–811.

https://doi.org/10.1177/002200202237929

O'Brien, S. P. (2010). Crisis early warning and decision support: Contemporary approaches and thoughts on future research. *International Studies Review*, *12*(1), 87–104. https://doi.org/10.1111/j.1468-2486.2009.00914.x

Olaide, O. B., & Ojo, A. K. (2021). A Model for Conflicts' Prediction using Deep Neural Network. *International Journal of Computer Applications*, *183*(29), 8–12. https://doi.org/10.5120/ijca2021921667

Olsson, F., Sahlgren, M., ben Abdesslem, F., Ekgren, A., & Eck, K. (2020). Text Categorization for Conflict Event Annotation. *Proceedings of the Workshop on Automated Extraction of Socio-Political Events from News 2020*, *May*, 19–25. https://www.aclweb.org/anthology/2020.aespen-1.5

Osorio, J., & Reyes, A. (2017). Supervised Event Coding From Text Written in Spanish: Introducing Eventus ID. *Social Science Computer Review*, *35*(3), 406–416. https://doi.org/10.1177/0894439315625475

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*(1–2), 1–135.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury Google, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Xamla, A. K., Yang, E., Devito, Z., Raison Nabla, M., Tejani, A., Chilamkurthy, S., Ai, Q., Steiner, B., … Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, *32*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., & Others, A. (2011). Scikit-learn: Machine Learning in Python. *Journal OfMachine Learning Research*, *12*(2011), 2825–2830.

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation Jeffrey. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, 1532–1543.

Petrou, M. M., & Petrou, C. (2010). *Image Processing: The Fundamentals*. John Wiley & Sons.

Piskorski, J., & Jacquet, G. (2020). TF-IDF Character N-grams versus Word Embedding-based Models for Fine-grained Event Classification: A Preliminary Study. *Proceedings of the Workshop on Automated Extraction of Socio-Political Events from News 2020*, *May*, 26–34. https://www.acleddata.com/%0Ahttps://www.aclweb.org/anthology/2020.aespen-1.6

Poole, D. L., Mackworth, A., & Goebel, R. G. (1998). Computational Intelligence and Knowledge. *Computational Intelligence: A Logical Approach*, *Ci*, 1–22. https://www.cs.ubc.ca/~poole/ci.html

Qiao, F., Li, P., Zhang, X., Ding, Z., Cheng, J., & Wang, H. (2017). Predicting Social Unrest Events with Hidden Markov Models Using GDELT. *Discrete Dynamics in Nature and Society*, *2017*. https://doi.org/10.1155/2017/8180272

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training*. https://doi.org/10.4310/HHA.2007.v9.n1.a16

Raleigh, C. (2020). Keynote Abstract: Too soon? The limitations of {AI} for event data. *Proceedings of the Workshop on Automated Extraction of Socio-Political Events from*

News 2020, *May*, 7. https://www.aclweb.org/anthology/2020.aespen-1.2

Raleigh, C., Linke, A., Hegre, H., & Karlsen, J. (2010). Introducing ACLED: An Armed Conflict Location and Event Dataset. *Journal of Peace Research*, *47*(5), 651–660. https://doi.org/10.1177/0022343310378914

Raschka, S., & Mirjalili, V. (2017). *Python Machine Learning* (2nd Editio). Packt Publishing.

Ratnaparkhi, A. (1996). A Maximum Entropy Model for Part-Of-Speech Tagging. *Conference on Empirical Methods in Natural Language Processing*, 133–142.

Ratnaparkhi, A. (1997). A Linear Observed Time Statistical Parser Based on Maximum Entropy Models. *ArXiv Preprint Cmp-Lg/9706014*. http://arxiv.org/abs/cmp-lg/9706014

Ravuri, S., & Stolcke, A. (2016). A comparative study of recurrent neural network models for lexical domain classification. *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6075–6079. https://doi.org/10.1109/ICASSP.2016.7472844

Rehder, B., Schreiner, M. E., Wolfe, M. B. W., Laham, D., Landauer, T. K., & Kintsch, W. (1998). Using latent semantic analysis to assess knowledge: Some technical considerations. *Discourse Processes*, *25*(2–3), 337–354. https://doi.org/10.1080/01638539809545031

Rehurek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 46–50.

Roets, E., Lamberti, R., Oberholzer, L.-M., & Gildenhuys, B. (2019). *Expropriation Without Compensation: A Disaster in Waiting*.

Rosenfeld, R. (1996). A maximum entropy approach to adaptive statistical language modelling. *Computer Speech and Language*, *10*(3), 1–37.

Rummel, R. J. (1969). Forecasting international relations: A proposed investigation of three-mode factor analysis. *Technological Forecasting*, *1*(2), 197–216.

Sahami, M., Dumais, S., Heckerman, D., & Horvitz, E. (1998). A Bayesian approach to filtering junk e-mail. *Learning for Text Categorization: Papers from the AAAI Workshop*, *WS-98-05*(Cohen), 55–62. http://research.microsoft.com/en-us/um/people/horvitz/junkfilter.htm

Salehyan, I., Feinberg, A., & Naughton, K. A. (2020). Merging actors with events: introducing the social conflict analysis dataset - organizational properties (SCAD-OPs). *International Interactions*, *46*(1), 133–149. https://doi.org/10.1080/03050629.2019.1687466

Salehyan, I., Hendrix, C. S., Hamner, J., Case, C., Linebarger, C., Stull, E., & Williams, J. (2012). Social Conflict in Africa: A New Database. *International Interactions*, *38*(4), 503–511. https://doi.org/10.1080/03050629.2012.697426

Salton, G. (1971). *The SMART retrieval system—experiments in automatic document processing*. Prentice-Hall, Inc.

Saltz, J. (2022). *CRISP-DM is Still the Most Popular Framework for Executing Data Science Projects*. https://www.datascience-pm.com/crisp-dm-still-most-popular/

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv Preprint ArXiv:1910.01108*, 2–6. http://arxiv.org/abs/1910.01108

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, *2*(3), 1–21. https://doi.org/10.1007/s42979-021-00592-x

Saunders, M., Lewis, P., & Thornhill, A. (2019). *Research Methods for Business Students* (Eighth). Pearson Education International.

Saunders, M., Lewis, P., & Thornhill, A. (2023). Understanding research philosophy and approaches to theory development. In *Research Methods for Business Students* (Ninth, pp. 128–174). Pearson Education International.

Schrodt, P. A., Davis, S. G., & Weddle, J. L. (1994). Political Science: KEDS-A Program for the Machine Coding of Event Data. *Social Science Computer Review*, *12*(4), 561–587. https://doi.org/10.1177/089443939401200408

Schrodt, P. A., & Gerner, D. J. (1994). Validity Assessment of a Machine-Coded Event Data Set for the Middle East , 1982-92. *American Journal of Political Science*, *38*(3), 825–854.

Senekal, B., & Kotzé, E. (2019). Open source intelligence (OSINT) for conflict monitoring in contemporary South Africa: Challenges and opportunities in a big data context. *African Security Review*, *28*(1), 19–37. https://doi.org/10.1080/10246029.2019.1644357

Sharda, R., Delen, D., & Turban, E. (2020). *ANALYTICS, DATA SCIENCE, & ARTIFICIAL INTELLIGENCE: SYSTEMS FOR DECISION SUPPORT* (11th ed.). Pearson.

Sherstinsky, A. (2020). Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network. *Physica D: Nonlinear Phenomena*, *404*, 132306. https://doi.org/10.1016/j.physd.2019.132306

Shrestha, A., & Mahmood, A. (2019). Review of deep learning algorithms and architectures. *IEEE Access*, *7*, 53040–53065. https://doi.org/10.1109/ACCESS.2019.2912200

Smith, E. (2020). *South Africa's Ramaphosa blasts "despicable" crime wave during coronavirus lockdown*. https://www.cnbc.com/2020/04/13/south-africas-ramaphosa-blasts-despicable-crime-wave-during-coronavirus-lockdown.html

South Africa Zuma riots: Looting and unrest leaves 72 dead - BBC News. (n.d.). Retrieved August 26, 2022, from https://www.bbc.com/news/world-africa-57818215

Sundberg, R., Eck, K., & Kreutz, J. (2012). Introducing the UCDP Non-State Conflict Dataset. *Journal of Peace Research*, *49*(2), 351–362. https://doi.org/10.1177/0022343311431598

Szepesvári, C. (2010). Algorithms for reinforcement learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, *9*, 1–89. https://doi.org/10.2200/S00268ED1V01Y201005AIM009

Tetlock, P. E. (2006). *Expert Political Judgment: How Good Is It? How Can We Know?* (Course Boo). Princeton University Press.

Tetlock, P. E., & Gardner, D. (2016). *Superforecasting: The Art And Science Of Prediction*. Broadway Books.

Thompson, K. (1968). Programming Techniques: Regular expression search algorithm. *Communications of the ACM*, *11*(6), 419–422. https://doi.org/10.1145/363347.363387

Ulfelder, J., & Valentino, B. (2008). Assessing Risks of State-Sponsored Mass Killing. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1703426

Vadapalli, S. R., Hsieh, G., & S. Nauer, K. (2018). TwitterOSINT. *Proceedings of International Conference Security and Management (SAM)*, 220–226.

https://csce.ucmss.com/cr/books/2018/LFS/CSREA2018/SAM9750.pdf

Valliani, A. A. A., Ranti, D., & Oermann, E. K. (2019). Deep Learning and Neurology: A Systematic Review. *Neurology and Therapy*, *8*(2), 351–365. https://doi.org/10.1007/s40120-019-00153-8

Vesco, P., Hegre, H., Colaresi, M., Jansen, R. B., Lo, A., Reisch, G., & Weidmann, N. B. (2022). United they stand: Findings from an escalation prediction competition. *International Interactions*, *48*(4), 860–896. https://doi.org/10.1080/03050629.2022.2029856

Vhumbunu, C. H. (2015). *Appraising the efficacy of SADC in resolving the 2014 Lesotho conflict – ACCORD*. https://www.accord.org.za/conflict-trends/appraising-the-efficacy-of-sadc-in-resolving-the-2014-lesotho-conflict/

Wang, B. W., & Kennedy, R. (2016). *Automated event coding raises promise and concerns*. *353*(6307), 1502–1504.

Welch, C. (2018). *How Cape Town Is Coping With Its Worst Drought on Record*. https://www.nationalgeographic.com/news/2018/02/cape-town-running-out-of-water-drought-taps-shutoff-other-cities/

Williams, H., & Blum, I. (2018). Defining Second Generation Open Source Intelligence (OSINT) for the Defense Enterprise. In *Defining Second Generation Open Source Intelligence (OSINT) for the Defense Enterprise*. https://doi.org/10.7249/rr1964

Yörük, E. (2012). *THE POLITICS OF THE TURKISH WELFARE SYSTEM TRANSFORMATION IN THE NEOLIBERAL ERA: WELFARE AS MOBILIZATION AND CONTAINMENT*. Johns Hopkins University.