An investigation of the effectiveness of using Twitter data for predicting South African protests with Graph Neural Networks

Derwin **Ngomane**

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA
Denkleiers • Leading Minds • Dikgopolo tša Dihlalefi

Faculty of Engineering, Built Environment & IT,
Department of Computer Science, University of Pretoria, Pretoria.

A mini-Dissertation submitted to the Faculty of Science in fulfilment of the requirements for the Master degree in Big Data Science.

Supervised by
1st Supervisor - Vukosi **Marivate**
2nd Supervisor - Maxamed **Ahmed**

May 21, 2024

Declaration

I, Derwin Thulani Ngomane, hereby declare the content of this dissertation to be my own work unless otherwise explicitly referenced. This dissertation is submitted in partial satisfaction of the requirements for Master degree in Big Data Science at the University of Pretoria, Pretoria. This work has not been submitted to any other university, nor for any other degree.

Signed: _____

Date: _____

"Education is the most powerful weapon which you can use to change the world."

Nelson Mandela

"If people did not do silly things, nothing intelligent would ever get done."

Ludwig Wittgenstein

Abstract

Social media creates an echo chamber effect that is closely related to social movement theory, which aims to mobilise people to change society. In South Africa, there has been an increase in protests that appear to have started on social media. For example, consider the riots that occurred in July 2021 following the arrest of former President Jacob Zuma. Protests in South Africa, on the other hand, have culminated in violent incidents, such as the July 2021 protest. In that situation, the South African Human Rights Commission found that social media sites such as WhatsApp, Facebook, and Twitter aided the violence by sharing protest information. This study investigates whether social media can be utilised to signal upcoming South African protests.

This research investigates the effectiveness of nose reduction techniques on Twitter data for predicting protest-related events in South Africa using Graph Neural Networks. It addresses research gaps by addressing the need for graph-based methodologies in the South African context, addressing the lack of noise reduction research for Twitter data, and using an automated method to extract relevant keywords in the word networks. The work aims to provide a new avenue for noise reduction in real-world scenarios where future events have not occurred.

This study examines a three-year data window between 2019 and 2021 using the Global Dataset of Events, Location, and Tone (GDELT) and Twitter data. GDELT focuses on CAMEO codes related to protests and conflict, while Twitter extracts social media text related to protest-related posts. A sliding window approach is used to combine the data, with noise-reduction filtration techniques guiding the filtration. This work explores the potential of processing Twitter data to reveal signals for improved predictive capability. Derivative metrics, from hashtags, links, and mentions, are used to reveal such signals.

The study compares different machine learning methods, including Logistic Regression, Graph Convolutional Networks, and Graph Isomorphism Networks, to model the data. It is discovered that the geometric deep learning methods struggle with overfitting in hold-out testing data but are stable and have better cross-validation scores. The GIN model exhibits higher accuracy and isomorphism detection, making it suitable for the task. However, graph neural networks struggle with limited data and hence overfit the training data, as well as isomorphism and isolated nodes due to message-passing paradigm.

The intricacy of Twitter interactions and conversations is highlighted in this work, emphasising the need for future research in data processing and model building. The study excluded other data features to add more information about the data space's complexity, such as user interactions. Keyword selection was done independently, but node eigenvector centrality could be used for informed decision-making. The graph neural network paradigm of message passing has limited capability in the existence of isolated nodes, and isomorphism is crucial for network performance. Further research should investigate dynamic capabilities and edge weights in GIN networks.

Acknowledgements

- I want to express my gratitude to my family [Idah Ngwenya, Jeanette Ngwenya, SipheSihle Mthunywa, Matshie Maponya, and Thembi Ngomane] for all of their help and encouragement during my educational journey.
- I would like to thank my Supervisors, Prof V Marivate and Mr M Ahmed for all the guidance, support and encouragement; including all the members of the Data Science for Social Impact Research group. This work would have not been possible without all of them.
- I would like to thank my friends [Busisiwe Jumbe, Ronald Ndlovu, Otumiseng Kgarume, Zethu Tlhako] for the motivation and being a caring ear when I needed one
- I would like to thank my biggest supporter, Rhulani Maluleke, for being there throughout this journey and always being my biggest cheerleader.
- Finally, I would like to thank my cat, Beauty, for keeping me company throughout the writing of this work and being there when I felt alone.
- This work is dedicated to all the young people from the streets of Umjindi, Barberton that dare to dream.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Purpose of Study

Social Media is a technology where people can communicate and share opinions with like-minded individuals. Social media users could either be from the same social circles or even from a broader community of subscribers. Hence, social media has a tendency to create environments where users opinions are reinforced by their peer groups. This is the echo chamber effect, where the environment reinforces the beliefs, opinions, and political leanings of individuals [Cinelli et al., 2021]. Therefore, social media has the capability of reinforcing echo chambers in society.

The echo chamber effect has close ties to social movement theory. Individuals' shared collective values serve as the driving force behind social movements with the aim of bringing about societal change [Killian et al., 2020]. Social movements are about mobilising individuals in order to create a collective of like-minded individuals pursuing the same purpose. The speed at which information spreads through social media creates an opportunity for the mobilisation of individuals for the purpose of social movements, such as protests. The motivation behind a protest is sometimes actually for the benefit of a social movement [Loya and McLeod]. Hence, the echo chamber effect as a result of social media usage can benefit protest action.

Recently, there has been a spark of protests that have been mobilised through social media. Examples of such protests have occurred in the United States, where, for example, crowds mobilised after the election of former President Donald Trump [Mele and Correal, 2016]. However, some protests that have been mobilised using social media have had violent outcomes, whereby citizens have been injured and some have lost their lives. Examples of these violent protests include the incitement tweet that the former President Donald Trump released, inciting violence at the U.S. Capitol [Holland et al., 2021], and the murder of George Floyd, which led to a violent confrontation between citizens and the police [Hu, 2020]. These are a few examples of social media-mobilised protests in the United States. South Africa has not been immune to social media-driven protests.

South Africa recently also experienced a social media-mobilised protest that turned violent. The protest was motivated by the arrest of former President Jacob Zuma in July 2021. The

protest resulted in billions of Rands being lost due to damaged property and the loss of the lives of more than 350 citizens [Mokoena, 2021]. The South African Human Rights Commission conducted a national investigative hearing around the protest, and according to the Institute of Security Studies (ISS), social media played an "instrumental" role in the protest that turned into an unrest [Pillay and Mtshali, 2021].

The purpose of the investigative hearing was to bring together experts and affected parties in order to testify about the human rights violations that occurred due to the July protest. The opinion of the ISS was also echoed during the hearing by one of the residents of an area called Phoenix in Kwa-Zulu Natal, where racial tensions started during the protests. The Phoenix resident believes that "half-truths" were being spread on social media and resulted in the racial tensions [Mfundo, 2021]. Additionally, a panel of experts asserted that social media platforms such as WhatsApp, Facebook, and Twitter facilitated the violence during the protest by disseminating information about what was occurring or was about to occur [Africa et al., 2021]. These assertions serve as the basis for this study, which aims to determine whether social media can be altered to provide a high signal for the future occurrence of protests in South Africa.

1.2 Problem and Thesis Statement

Social media can be a powerful weapon for those who use it. It has been discussed how social media can assist in the occurrence of protests. However, protests in South Africa have become devastating and damaging to commercial properties, causing physical harm to citizens [Bonga, 2021]. This has resulted in the loss of lives and has negative consequences for the economic well-being of the country. The July unrest resulted in the loss of lives, property, and loss of income for a number of South Africans. In addition, social media can also be used by security services in order to place operational measures to respond to escalating protests.

Despite the abundance of opportunities and benefits of social media that have been highlighted, very little has been done for the purpose of preventing escalating protests in the South African context. This research aims to address this shortcoming by using Twitter data in order to anticipate the occurrence of a protest in South Africa. While the ability to forecast protest-related events using word relation networks derived from Twitter data has demonstrated potential, there is currently very little research on how automatic noise removal can increase this method's accuracy in a South African context.

Our research aims to answer the following question: What is the effectiveness of nose reduction techniques on Twitter data in order to predict protest-related events in South Africa using Graph Neural Networks?

In order to answer the research question, the following sub-questions need to be addressed:

- How can Twitter data be processed in order to reveal a signal that will improve the predictive capability of machine learning models?
- How are word relation networks derived from Twitter data?

- Can graph neural networks on Twitters' word relation networks be used to determine the occurrence of a future protest?

Utilising social media data generates a rich stream of huge, high-velocity data that can be used for social good. Using graph networks, this research will seek to identify word patterns in South African Twitter posts to predict the risk of a protest. This will reveal the themes in social media keywords that contribute to potential protests.

1.3 Thesis Structure

This section provides a summary of the chapters presented in this study.

- **Chapter 2** - provides the current literature and research focus on event prediction for protest. Additionally, this section also provides the literature gap that this study will attempt to address. Finally, the section provides the technical background of the methods and entities that will be used in this study.
- **Chapter 3** - provides a thorough overview of the research methodology. In this section, the research design, the research data, the limitations of the work, and the ethical considerations are discussed.
- **Chapter 4** - provides an analysis of the research data that is used in an attempt to answer the research questions.
- **Chapter 5** - provides an overview of the model results.
- **Chapter 6** - provides an discussion of the model results, the implications, and possible future possible research directions.
- **Chapter 7** - provides concluding remarks on the research and a summary of the research findings. Additionally, the direction for future possible research is also outlined.

Chapter 2

Background & Literature Review

2.1 Introduction

This section provides a full background on the research problem. Initially, the definition of a protest is described, as is the motivation behind protests in South Africa. Thereafter, a brief background of protests in South Africa according to literature and the motivation behind them is discussed, as is the way social media has been used in this context. Additionally, a theoretical review of social protest using social capital theory under the view of the fifth estate and how social media is used for efficient mobilisation under the theory of resource management is discussed. Finally, the value and benefit of predicting protests in South Africa are discussed.

In addition to the review on protests, this section delves deeper into the methods that are used to predict protests, according to the literature. This will provide an overview of statistical methods, machine learning methods, and deep learning methods currently being used for protest-related event prediction.

2.2 Literature Review

Participants in protests use a form of collective action to express their dissatisfaction with a system and influence change. People use political protests to shape public policy and advance democracy [Passarelli and Tabellini, 2017]. Accordingly, Chan [2017] highlights that the use of protest action can be characterised by three (3) factors: injustice towards an in-group, the belief that a protest action will result in immediate results, and the in-group identity of protest participants either through race, gender, or political party association [Chan, 2017]. Additionally, Barbera and Jackson [2019] add that the homophily of learning, which involves learning through shared attributes between individuals, can influence potential participation in a collective action such as protest [Barbera and Jackson, 2019]. Therefore, in-group identity and associated ties between individuals are two of the biggest factors that influence a person's decision to partake in a protest and attempt to influence change.

These factors have been observed in a wide variety of prominent protests around the world. The Arab Spring is considered a protest that originated in Tunisia in December 2010 due to political frustration and instability [Steinert-Threlkeld, 2017]. The protests then spread across the Arab nations of North Africa and the Middle East. Hence, this can be deemed a phenomenon of learning through homophily. Additionally, this is an example of a protest that was driven by injustice towards an in-group and the belief that a protest action will result in change.

Similarly, there have been a lot of protests motivated by the fight for democracy and inequality. Examples of such protests include the "occupy movement" protests. The first occupy movement protest was in the United States, named the Occupy Wall Street protest of 2011 [Calhoun, 2013]. The Occupy movement continued to spread to Nigeria in 2012, due to the fuel subsidy removal by the government in Nigeria [Uwalaka and Watkins, 2018] and the Occupy Hong Kong protest in 2014 [Qiao and Wang, 2015]. More recently, the Black Lives Matter protest of 2020 in the United States was due to police brutality towards minority races [Mundt et al., 2018]. These examples all demonstrate the association between protests and perceived in-group injustice, and the learning from successful protests by shared-attribute participants.

South Africa has not been immune to protest actions that are due to dissatisfaction with the current or past political system. The protests have been sparked by socio-economic issues related to poverty, an increasing unemployment rate, and increased corruption [Bonga, 2021]. These protests have been going on since 1994, when the country became a democracy [Matebesi and Botes, 2017]. However, the origin of the protest actions of citizens in South Africa can be tracked back to pre-democratic years [Mottiar, 2013]. It is a democratic right in South Africa to partake in protest action under the constitution. However, the protests in South Africa have become more violent over the years. For the period of 2011/12, 1 091 protests out of a total of 11 033 protest incidents were classified as unrest [Mottiar, 2013].

2.2.1 Protests in South Africa

There has been a rise in protests in South Africa. The rise in the number of protests has also been associated with a rise in the chances of the protest turning violent [Bonga, 2021]. Alexander et al. [2018] believe that the rise in community protests that are disorderly, disruptive, and violent has been evident since 2006 [Alexander et al., 2018]. However, it has also been argued that this trend has been evident since 1994, after the first democratic elections in the country [Matebesi and Botes, 2017]. According to Lancaster [2018], South African citizens believe that disruptive and violent protest lead to successful outcomes [Lancaster, 2018]. Therefore, South Africans believe that through violent protests, there will be immediate results. However, as per the factors remarked by Chan [2017], there is also an injustice towards South Africans. Therefore, the primary theme that has been highlighted is all related to socio-economical issues.

There are several socio-economic-related protests aimed at the state, but a significant number of events against businesses are significant [Lancaster, 2018]. Additionally, economic and political conditions are central to the emergence of protests in the South African context [Mare, 2014; Bonga, 2021]. Specifically, violent protests have been due to the lack of transformation in South Africa [Bonga, 2021]. Bedasso and Obikili [2016] conducted a cohort analysis using

GDELT data and World Values Survey data in order to reveal the predictors of social unrests in South Africa post-apartheid between 1990 and 2001. They concluded that the issues related to the rise of protests were due to economic issues, unfulfilled expectations by the post-apartheid government and was more pronounced with the youth demographic [Bedasso and Obikili, 2016]. This work sheds light on the high likelihood of young individuals involvement in protest actions. However, due to the use of survey data in the study, the method cannot be used in real-time. The majority of young people use social media, and the data is constantly streaming in near real-time.

According to Mare [2014] during a crisis, social media platforms are used by activists as instruments for early warning, bypassing mainstream media, and passing solidarity messages [Mare, 2014]. This was evident in the Rhodes Must Fall protest, where social media was used as a medium of communication to generate discussion for those that were not part of the protest [Bosch, 2017]. South Africans use social media as a form of citizen journalism [Mare, 2014] and also makes it easier to organise protests [Bonga, 2021]. The internet has boosted young political engagement in the country, and it has been established that countries with a large youth population are more prone to instability [Bosch, 2017]. It is evident that social media is used as an important tool before or during a protest, especially among the youth population that has an active online presence.

The Rhodes Must Fall protest is considered one of the most prominent protests in South Africa. The protest was driven by the students of the University of Cape Town, South Africa. The demonstration triggered a surge of what is referred to as "meme events" [Frassinelli, 2018]. The protest continued the wave of protests that were due to public dissatisfaction with service delivery, crime, corruption, and growing levels of unemployment [Bosch, 2017]. These protests include the #FeesMustFall protest of 2016 at the University of the Witwatersrand, South Africa, due to fee increases at the university; and the #ZumaMustfall protest of 2017 due to the corruption allegations of the former president of South Africa [Frassinelli, 2018]. Although some of the demonstrations were unrelated, they were all labelled with the #mustfall hashtag on social media.

The organic spread of information relating to these protests was driven by what is referred to as a hashtag [Saxton et al., 2015]. A hashtag is used by Twitter in order to collectively gather similar information and make it easier to find related information [Frassinelli, 2018].The use of the hashtag can be considered a new method to increase the mobilisation of in-group identity protest participants.

2.2.2 Mobilisation using Social Media

Mainstream media has long been considered the fourth arm of legislation after the executive, the legislature, and the judiciary [Amodu et al., 2014]. The purpose of the fourth estate is to act as an overseer that makes sure the government is accountable to the people and that the people can be a part of the process of governance. The media should provide a platform for the exchange of information, public commentary, and criticism. However, the media has been found to be biased in its reporting [Chan, 2017].The South African media has been criticised for biased

reporting on social movements in the country, such as the ongoing student protests and the 2012 Marikana Massacre [Rodny-Gumede, 2017a].The fourth estate falls short by ignoring views and misrepresenting citizens [Chan, 2017].Therefore, alternative media are then used to disseminate anti-establishment views.

Mainstream media creates an environment where citizens are consumers of news. However, social media enables citizens to become citizen journalists [Mare, 2014]. The potential benefits of social media mobilisations include removing the barrier to information and communication across the world and through networks. Social media is simple to use, elevates citizens from content consumers to content producers, and enables real-time event documentation [Dunu Ifeona and Uzochukwu, 2015].

Social capital theory argues that society acts in such a way that relationships are assets that may be used to create and accumulate human capital. In the age of social media technology and from the standpoint of social capital theory, social media might be regarded as the fifth estate where important relationships are created [Uwalaka and Watkins, 2018].A state that values collaborative engagements, open debate, and social issue discourses. Social media users have a high degree of social capital, which is correlated to their usage of the platform [Hwang and Kim, 2015].The platform has the ability to allow citizens to be directly involved in the governance of a country and forming a fifth estate.

The impact of the fifth estate can be seen from multiple examples where the fourth estate is considered to have failed. Dunu Ifeona and Uzochukwu [2015] demonstrate this using the failed 2013/14 election by INEC (Independent Electoral Commission) in Nigeria [Dunu Ifeona and Uzochukwu, 2015]. INEC failed to get Nigerian citizens to participate in the registration process due to the fact that it only used mainstream media for communication and no social media. Even though there were successful initiatives in Nigeria when it comes to social media usage, e.g., parties in Anmbra State and South Eastern Nigeria during the 2013/14 elections used social media extensively in their campaigning, as did the child immunisation for bird flu and the eradication of the Ebola virus [Dunu Ifeona and Uzochukwu, 2015]. Similarly, social media was also paramount to the execution of the Occupy Nigeria protest of 2012 in terms of planning, coordination, and mobilising the protest [Uwalaka and Watkins, 2018]. Additionally, it has been argued that there was an interrelationship between social media and mainstream media during the Arab Spring protests [Aday et al., 2012]. These are a few examples of the impact of social media on influencing behaviour.

Social media can encourage behaviour change in a dynamic, personalised, and participatory manner. Additionally, social media can facilitate the amplification of a narrative through shares and reposts [Mundt et al., 2018]. The platform can be used by activists in order to shape the narrative about their cause [Zeitzoff, 2017]. The exchange of information plays a role in organising a protest by lowering the cost, increasing the speed of information transmission, and bypassing mainstream media [Zeitzoff, 2017].Social media has the edge over mainstream media due to its collaborative nature and increasing shared values and beliefs [Uwalaka and Watkins, 2018]. Social media is changing the way in which like-minded citizens can communicate in order to conduct social movements.

According to the theory of resource management, resources such as time, money, organisational skills, and social opportunities are essential to the success of a social movement. Nahed Eltantawy and Wiest [2011] claim that the 2011 Egyptian revolution can be explained using resource mobilisation theory in an attempt to understand the mobilisation inspired by social media [Nahed Eltantawy and Wiest, 2011]. Similarly, it has been found that there is a link between social media and mobilisation, especially in relation to the scaling up of a social movement [Mundt et al., 2018]. Therefore, in the context of social movements, social media can be considered a resourcing tool.

One of the drivers of protest is the in-group identity of participants, which is correlated to the formation of weak and strong ties amongst social media users [Mundt et al., 2018].These ties are also formed among participants in a social movement. The strong ties' theory comes into effect when groups facing similar marginalisation come together in order to work against any societal injustice, as seen in the Black Lives Matter movement [Mundt et al., 2018].Activists use social media to find like-minded individuals or groups with whom they may form links in order to boost the chance of political participation [Zeitzoff, 2017; Chan, 2017].

It is evident that social media plays an important role in offline mobilisation. The speed and ease of transmission of a narrative allow participants to pursue like-minded individuals to get involved in a protest. The use of social media also allows the organisers of protests to avoid surveillance and suppression from authoritarian governments that will prevent social movement mobilisation.

2.2.3 The importance of predicting protests in South Africa

Protests in South Africa have become devastating and are causing damage to commercial property and physical harm to citizens. This has resulted in the loss of lives and has negative consequences for the economic well-being of the country [Bonga, 2021].There are a number of protests that have had negative consequences for the economy due to the disruption of infrastructure during the protests. Examples of such protests include the 2021 July unrest that resulted in jobs affected [Africa et al., 2021], the burning of manufacturing firms in the 2016 Mandeni protest that resulted in more than 2 000 lost jobs [Khambule et al., 2019], and the burning of school infrastructure in the 2016 Vumani protest [Khambule et al., 2019]. There have been negative effects for citizens as a result of violent protests.

The consequences of such violent protests have a negative impact on infrastructure and job security. The destruction of public infrastructure results in an impact on future capability opportunities for citizens [Khambule et al., 2019]. For example, schools that need to be rebuilt due to their demolition and the loss of jobs when employment infrastructure is destroyed. Additionally, El-Mallakh et al. [2018] argue that protest activity led to a reduction in intra-household gender disparity in labour force participation in regions with a greater number of fatalities due to men's increased income uncertainty [El-Mallakh et al., 2018]. Hence, the government will need to focus on ex ante measures to prevent the destruction of important infrastructure that prevents communities from moving forward [Khambule et al., 2019].

There is a need to deploy appropriate responses to curb the escalation of protests. It has been found that the inclusion of security forces during an escalated protest results in the protest tending to be more violent [Bonga, 2021]. This can be observed from the number of student protests that have turned violent and the Marikana protest [Mottiar, 2013].There is a requirement to monitor not only the frequency of protests but also the nature of the grievances being expressed and the places where protests are located [Lancaster, 2018]. The retrospective view does not prevent the events that have occurred or the results after an escalation to violence. There is a need to be proactive, predict whether a protest event will occur, and prepare accordingly in order to prevent any unlawful activities.

2.2.4 Event Prediction

The study of events is the analysis of a single occurrence or a series of occurrences. In addition to studying the event itself, it is necessary to study the event's location, time, and topic, as events are characterised by these three (3) aspects [Zhao, 2021].For instance, the assignment may consist of analysing the likelihood of a Malaria outbreak in South Africa over the span of a few years. Therefore, event prediction is a problem with multiple dimensions. However, event analyses can be categorised into three distinct research areas: event summarization [Chakrabarti and Punera, 2011], event detection [Aldhaheri and Lee, 2017], and event prediction [Zhao, 2021].

The problem of event detection and event summarisation seek to look at an event in a retrospective view [Zhao, 2021]. Whereas, the problem of event prediction seeks to predict the occurrence of future events that satisfy a certain criteria. The problem of event prediction can be defined to jointly predict the three (3) facets of an event or defined to predict one facet. Therefore, we can have four sub-problems in event prediction i.e. time prediction, location prediction, semantic prediction and joint prediction [Zhao, 2021]. In this review, the sole focus is on the event prediction task, in particular time prediction. This focus aligns with the research question of this study. The benefit of studying event prediction is that we are able to put measures in place to avoid the losses associated with the event.

The time prediction task solely focuses on the prediction of the time and occurrence of an event. Different approaches have been applied to the event occurrence prediction task. The techniques span threshold-based methods, traditional statistical methods and deep learning approaches.

In order to account for uncertainty, Bayesian methods have been used in a study conducted for the detection of protests in Australia using Twitter data and a custom dataset that recorded historical events in the region [Tuke et al., 2020]. Similarly, Qiao et al. [2017b] used a Hidden Markov Model approach on the Global Database of Events, Language and Tone (GDELT) data in order to predict the occurrence of protests in Southeast Asia and use the Bayes method to classify the generated sequence. Probabilistic methods have the benefit of accounting for uncertainty; however, they are unable to use large and complex data that can capture the dynamics of protests.

Researchers have started using deep learning approaches for event detection in order to capture complex relationships in the data. Smith et al. [2017] use a Long-Short-Term Memory (LSTM)

neural network on the numerical features of the GDELT data in order to predict the monthly count of unrests in Afghanistan. Similarly, Halkia et al. [2020] also use an LSTM model in order to predict the monthly occurrence of an unrest as an early warning system. Additionally, Zambezi [2021] also uses an LSTM in order to predict counts of protests in South Africa using the GDELT data. Parrish et al. [2018] compares multiple machine learning models on a dataset that includes the GDELT data and social and economic data for 158 countries in order to predict disruptive events. It is established that the Gated Recurrent Unit (GRU) obtains the best Area Under the Curve (AUC) value across all the categories. Authors use the Recurrent Neural Network (RNN) based neural network because of its ability to capture long term dependencies in sequential data [Zaremba et al., 2014]. The recurrent deep learning methods have proven to work when it comes to counts data but have not been used in data of other forms.

The literature on event occurrence prediction also includes the use of methods based on traditional statistical methods and using text based data sources. Radinsky and Horvitz [2013] use 22 years of news archives from the New York Times and enrichment data from 90 sources in order to develop a probabilistic model that predicts the likelihood of an event of interest [Radinsky and Horvitz, 2013]. Similarly, Chakraborty et al. [2014] examine ensemble models to predict Influenza-like illness counts for 15 Latin American nations using seven distinct data sources, including Twitter data [Chakraborty et al., 2014]. Additionally, Kallus [2014] also uses multiple sources of public data, including Twitter data, in order to train a random forest classifier in order to predict protests and crowd behaviour in seven (7) languages in 18 countries [Kallus, 2014]. However, as much as the combination of multiple sources of data proves to be beneficial, the work does not consider the noise and dynamics that originates from using social media data.

2.2.5 Event Prediction using Social Media

Social media, as previously mentioned, has been used in the context of event prediction. Social media has been used in order to predict movie revenues from Twitter data [Asur and Huberman, 2010], find an association between voting polls and sentiment over social media [O'Connor et al., 2010], and predict the stock market using sentiment analyses and mood analysers [Bollen et al., 2011; Arias et al., 2014]. Research has focused on improving signal detection in social media due to the vast number of applications of event prediction using social media.

Several researchers have used different variants of logistic regression to predict events in social media data. Korkmaz et al. [2015] used a LASSO-based logistic regression due to its feature selection capability in order to predict civil unrest in Latin American countries and by combining multiple data sources on top of Twitter, such as news sources and Tor [Korkmaz et al., 2015].Their work uses a pre-compiled list of protest-related keywords in order to create a keyword volume from the extracted tweets [Korkmaz et al., 2015]. Similarly, Early Model Based Event Recognition using Surrogates (EMBERS), a computerised system, also combines many data sources, including Tweets, blogs, news media, etc., in order to forecast social unrest [Ramakrishnan et al., 2014]. Many different models are used in this system, including the LASSO-based logistic regression.

In a similar manner, Wu and Gerber [2017] developed a logistic regression model on social media data, they add protest participation theory features derived from tweets focusing on the 2011 Egyptian revolution. Similarly, Zhao et al. [2015] used a multi-task approach by combining keywords extracted using LASSO and a multi-task learning framework in order to constrain features that are similar in different location. The benefit of using logistic regression is clearly demonstrated; however, the method requires the design of features that may not capture all the dynamics of a social unrest.

Approaches based on deep learning, such as RNNs, have been discussed as beneficial for event prediction, but they suffer from a lack of explainability in the predictions generated [Deng and Ning, 2021]. Hence, researchers have been looking at approaches that can give insights into the prediction. One such methodology used an attention-based neural network that has been extensively used in Natural Language Processing [Vaswani et al., 2017]. The attention network has been used for traffic flow prediction [Do et al., 2019]. Similarly, [Ertugrul et al., 2019] leverage the same idea of having spatial and temporal attention in order to develop a unified approach to forecasting unrest for the Black Lives Matter protest and the Charlottesville white supremacy rally [Ertugrul et al., 2019].

In addition to the deep-learning methods, researchers have started using graph theory in order to predict events in social media data so as to capture the social dynamics of social media and the hypothesis that Twitter tends to indicate offline user behaviour [Steinert-Threlkeld, 2017].Jin et al. [2014] developed a bispace model using Geometric Brownian motion in order to model the propagation of information on Twitter and the Poisson distribution to model the propagation of information outside of Twitter [Jin et al., 2014]. Similarly, Cadena et al. [2015] use Twitter data from Brazil, Mexico, and Venezuela in order to build information cascades on graphs and use Lasso logistic regression to predict the probability of an unrest on a particular day. However, due to the amount of information that a social network graph may capture, the authors Chen and Neill [2014] created a "sensor" network that connects various types of features in social media and applies a non-parametric graph scan method.

It has been found that in Social media, such as Twitter, the distribution of followers follows a power-law distribution [Kwak et al., 2010a], users connect to people who are like them [Al Zamal et al., 2012], friends follow each other on Twitter [Xie et al., 2012], interaction decreases with geographic distance [Kulshrestha et al., 2012]. Due to these properties, researchers have started using graph theory methods in order to predict social unrests. In order to capture the temporal-based evolution of an unrest and the semantic information in social media data, Deng et al. [2019] built a context-aware, dynamic graph network on social media data in order to predict social unrests in India, Egypt, Thailand, and Russia. Similarly, the work by Wang et al. [2020] improves the contextual ability by adding a squeeze and excitation module in order to increase the weight of meaningful words for event prediction and suppress weaker ones. Deng [2021] also attempts to improve the context awareness of the network by using entity interaction in the dynamic graph. Consequently, combining the temporal idea with the Graph Convolution Network with GRU is investigated in order to combine Twitter data and news data from GDELT [Jiang et al., 2021]. Qiao and Wang [2015] also use the GDELT data to build an event interaction graph on the Occupy Protests in America and Hong Kong. The authors use a logistic regression

model to estimate the probability of an event occurring on a particular day. There is enormous versatility with graph neural network, and hence research has progressed in that area.

2.2.6 Literature limitations and Gaps

The work described thus far validates the use of social media in order to predict protests. The following research gaps have been identified:

- The literature has not considered using graph-based methodologies in the South African context. Hence, this research is an extension of the work initiated by previous research using the GDELT data in South Africa, such as the work by Zambezi [2021]. However, this work looks into using Twitter in order to predict protests in South Africa and using geometric deep learning methods to fulfil the task.
- The literature that has used social media for the purposes of protest prediction has not presented the noise-reduction process in the Twitter data. The work that has been presented either assumes that the original Twitter data is good enough to be used for machine learning methods or does not explicitly describe the cleansing process. This work addresses these research gaps and provides one avenue for noise reduction.
- The work that studies a significant protest in South Africa by Chinta et al. [2021] reveals that there are no signals on Twitter prior to the event but that they only occur while the protest is occurring. However, this work only considers English tweets, and South Africa is a multilingual society, so considering only a limited language may be a disadvantage to the prediction task.
- The work by Wang et al. [2020] exclusively uses news articles that are related to the protest being predicted. Which is a shortcoming in reality, whereby the event has not occurred. Hence, this work looks at using an automated method in order to extract the keywords used in the network, which is applicable in a real-world scenario where the future event has not occurred.

The aforementioned support the research conducted to complete the research gaps identified.

2.3 Background

This section of the study provides background knowledge necessary for the rest of the work that is detailed in the study. This section also summarises the feature engineering techniques that have been used in this study. Thereafter, a detailed background on the geometric deep learning methods used in this study are discussed. Additionally, this section will also explain the evaluation metrics that are used for model evaluation in this work.

2.3.1 Feature engineering

This section of the study presents the methodological data processing background that is used to answer the research questions. Initially, a discussion on the preliminaries of graphs is discussed, and thereafter, distributional semantics in NLP is discussed. Thereafter, this section will build the theoretical framework for the use of geometric deep learning neural networks in the prediction of protests.

2.3.1.1 Preliminaries on Graphs

A graph is a mathematical object that is used to represent the relationship between two or more elements. A graph that is typically represented with $\mathcal{G}$ is made of edges and nodes (vertices). The nodes of a graph is the set $\mathcal{V}$ of elements a and the edges $\mathcal{E}$ is a set of ordered pairs of elements of the nodes set $\mathcal{V}$ of the graph, i.e $\mathcal{E} = \{(a, b)|a, b \in \mathcal{V}\}$, which represents the relationship between the nodes of the graph. Additionally, a graph can contain additional feature vectors containing additional information about the nodes. Hence, a graph $\mathcal{G}$ is said to be a tuple containing a set of nodes $\mathcal{V}$, a set of edges $\mathcal{E}$, and a feature vector u, i.e., $\mathcal{G} = (\mathcal{V}, \mathcal{E}, u)$.

Points (such as circles) and arcs connecting nodes can graphically represent a graph. There are two categories of graphs: directed graphs and undirected graphs. A directed graph is an asymmetric graph with arcs indicating the directional relationship between nodes, whereas an undirected graph is a symmetric graph without arrows indicating the existence of a relationship between the nodes. Figure 2.1 depicts an example of a directed and an undirected graph. In addition, there are both heterogeneous and homogeneous graphs within directed or undirected graphs. A homogeneous network consists of nodes of the same type, whereas a heterogeneous graph may contain nodes of various varieties. Similarly, both static and dynamic graphs exist. A dynamic graph has an evolving topology, whereas a static graph does not.

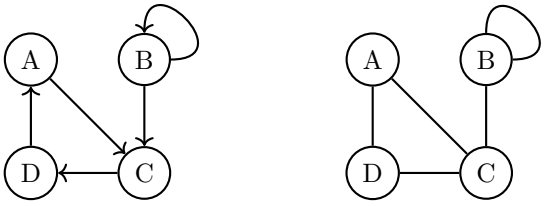


FIGURE 2.1: Left: Directed graph. Right: Undirected graph

A graph can also be represented as a matrix, which is called an adjacency matrix. The adjacency matrix is a square matrix that represents in matrix form the relationship between nodes and edges. Figure 2.1 depicts graphs that can also be represented as an adjacency matrix. It is said that two nodes a and b in $\mathcal{V}$ are adjacent to each other if there is an edge $e \in \mathcal{E}$ connecting them such such that $e = (a, b)$. The set of edges for both the directed graph and undirected graph in figure 2.1 is $\mathcal{E} = \{(A, C), (B, C), (B, B), (C, D), (D, A)\}$ for the set of nodes $\mathcal{V} = \{A, B, C, D\}$.

The edge set $\mathcal{E}$ of the directed graph is represented as an adjacency matrix $\mathbf{A}_1$, as depicted in Equation 2.1. The edge set of the undirected graph is represented as an adjacency matrix, as depicted by $\mathbf{A}_2$. It can be seen that the adjacency matrix of the undirected matrix is symmetric since $\mathbf{A}_2 = \mathbf{A}_2^T$. As a note, the graphs both have a self connection on node B.

$$\mathbf{A}_1 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}, \mathbf{A}_2 = \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix} \tag{2.1}$$

There are metrics that can be extracted from a graph in order to understand the properties of the graph. These metrics assist in understanding the interaction between the nodes and edges of the graph. In order to understand the number of connections a node has, we can use the degree of the the node. The degree of a node $d(a)$ is defined by Equation 2.3 and Equation 2.2 for an directed graph and undirected graph, respectively.

$$d(a) = |\{a \in \mathcal{V} \, : \, (a,b) \in \mathcal{E} \text{ or } (b,a) \in \mathcal{E}\}| \tag{2.2}$$

$$d(a) = |\{a \in \mathcal{V} \, : \, (a,b) \in \mathcal{E}\}| \tag{2.3}$$

The degree of a graph $\mathcal{G}$ can be represented as a diagonal matrix, $\mathbf{D}$, with the number of edges for each node. The degree matrices of the directed and undirected graphs is depicted in Equation 2.5 by $\mathbf{D}_1$ and $\mathbf{D}_2$, respectively. As it can be observed that node B has a high degree on the directed graph; and node A and C have the highest degree on the undirected graph since they are connected to more nodes. Additionally, the average degree of a graph defined by Equation 2.4 can be derived from the node degree values.

$$\overline{d} = \frac{\sum_{a \in \mathcal{V}} d(a)}{|\mathcal{V}|} \tag{2.4}$$

$$\mathbf{D}_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \mathbf{D}_2 = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix} \tag{2.5}$$

From the adjacency matrix and degree matrix, the unnormalised Laplacian matrix can be derived, which is defined as $L = D - A$.

The degree of a node only captures the number of connections of a node and does not capture the influence of a node in the graph. The eigenvector centrality measures whether a node is connected to other influential nodes and hence measures the nodes influence in the graph. The eigenvector centrality is calculated using a recursive function defined by

$$\lambda \mathbf{a} = \mathbf{A}\mathbf{a} \tag{2.6}$$

In addition to the defined connectivity metrics above, clustering coefficient is also considered. The local clustering coefficient, defined by Equation 2.7, quantifies the proportion of closed triangles in a graph, i.e., nodes that are connected to one another, and thus the triangle density of a graph.

$$c(a) = \frac{|(a_1, a_2) \in \mathcal{E} \ : a_1, a_2 \in \mathcal{N}(a)|}{\binom{d_a}{2}} \tag{2.7}$$

The clustering coefficient is able to detect nodes that have strong ties to one another in a graph and also assist in detecting structural phenomena of a graph. Additionally, the global (average) clustering coefficient of a graph defined by Equation 2.8 can be derived from the local clustering coefficient values.

$$\overline{c} = \frac{\sum_{|\mathcal{V}|} c}{|\mathcal{V}|} \tag{2.8}$$

The relationship between two or more graphs can also be defined. There is a concept of a subgraph, which can be considered as the building blocks of networks. A subgraph is a graph $\mathcal{H} = (\mathcal{V}', \mathcal{E}', u)$, where $\mathcal{V}' \subseteq \mathcal{V}$ and $\mathcal{E}' \subseteq \mathcal{E}$ i.e. the nodes and edges of $\mathcal{G}$ are contained in $\mathcal{H}$. Additionally, the structural similarity between graphs is defined as isomorphic. Two graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ are isomorphic, $\mathcal{G}_1 \cong \mathcal{G}_2$, if and only if two nodes in $\mathcal{G}_1$ are connected by an edge and equivalent nodes in $\mathcal{G}_2$ are also connected by an edge.

Graphs can be used to illustrate several naturally occurring processes. Examples of fields where graphs are used include genetics, computer science, economics, and sociology. In this research graphs are used to portray word relation networks on Twitter as the primary focus of our research. In this research the relationship between words is modelled using graphs.

2.3.1.2 Distributional semantics

Typically, words have meanings that they contain. In languages such as English, the meaning of words can vary depending on the context in which they are used. According to the distributional hypothesis, the meanings of terms that appear in the same context are similar. This theory gives rise to the computational methods of word vector embeddings.

Word vector embeddings methods are a set of methods that are used in order to represent words as continuous vectors. Additionally, the representation is meant to preserve the meaning of the words and the capture similarity between words in a corpus. Examples of such methods include the Term Frequency - Inverse Document Frequency (TF-IDF) [Ramos et al., 2003], Pointwise Mutual Information (PMI) Church and Hanks [1990], the word2Vec model [Mikolov et al., 2013] and the Bidirectional Encoder Representations from Transformers (BERT) Devlin et al. [2018]. The methods are briefly discussed below:

Term Frequency - Inverse Document Frequency - The TF-IDF is a combination between two statistics i.e. the Term Frequency (TF) and the Inverse Document Frequency (IDF). The term frequency is the frequency of a word w in the entire document d:

$$tf_{wd} = \log_{10}(f_d(w) + 1) \tag{2.9}$$

where: $f_d(w)$ - frequency of a word w in document d

The Inverse Document Frequency (IDF) which is defined as:

$$idf_w = \log_{10}\left(\frac{N}{df_w}\right) \tag{2.10}$$

where:

N - Number of documents in the corpus

df_w - Number of documents containing word w

The final TF-IDF is a combination of the equation 2.9 and equation 2.10. The TF-IDF is meant to down-weight words in a corpus using the Inverse Document Frequency. The TF-IDF is presented below:

$$tf{-}idf_{wd} = tf_{wd} \times idf_w \tag{2.11}$$

The TF-IDF can be used in order to identify important words in a corpus [Qaiser and Ali, 2018]. The words that appear less frequently across documents are considered important and have a high TF-IDF value.

Positive Pointwise Mutual Information - The Positive Pointwise Mutual Information is a statistic that measures the association between two words in a corpus. The statistic is based of the Pointwise mutual information to define mutual information between two events x and y [Church and Hanks, 1990]. Pointwise Mutual Information is defined as in Equation 2.12, that quantifies the joint probability of the events together with observing the events independently. The statistic should capture whether the occurrence is genuine or by chance.

$$I(x, y) = \log_2 \frac{P(x, y)}{P(x)P(y)} \tag{2.12}$$

The same equation can be extended to looking at words in a corpus instead of events. Hence, the Pointwise Mutual Information (PMI) between word w_1 and context word w_2 is depicted by Equation 2.14.

$$PMI(w_1, w_2) = \log \frac{N\hat{P}(w_1, w_2)}{\hat{P}(w_1)\hat{P}(w_2)} \tag{2.13}$$

where:

N - The number of documents in the corpus
$\hat{P}(w_1, w_2)$ - Number of documents where both w_1 and w_2 both appear.

$\hat{P}(w_1)$ - Number of documents containing the occurrence of w_1
$\hat{P}(w_2)$ - Number of documents containing the occurrence of w_2.

The PMI can attain any real number. There are instances due to the size of a corpus where the PMI value can negative due to the word pairs not being observed in the training data [Salle and Villavicencio, 2019]. This is avoided by only considering the positive values of PMI, which leads us to the Positive Pointwise Mutual Information (PPMI) as defined by Equation

$$PPMI(w_1, w_2) = \begin{cases} 0, & \text{if } PMI < 0 \\ \log \frac{N\hat{P}(w_1,w_2)}{\hat{P}(w_1)\hat{P}(w_2)}, & \text{otherwise} \end{cases} \tag{2.14}$$

The $PPMI$ solves the problem of negative values. However, it was discovered that both $PPMI$ and PMI have bias towards infrequent events/words [Levy et al., 2015]. Hence, Levy et al. [2015] have suggested a smoothing version of $PPMI$ that lowers the weight of the rare context word w_2 by α as depicted by Equation 2.15.

$$PPMI(w_1, w_2) = \begin{cases} 0, & \text{if } PMI < 0 \\ \log \frac{N\hat{P}(w_1,w_2)}{\hat{P}(w_1)\hat{P}_\alpha(w_2)}, & \text{otherwise} \end{cases} \tag{2.15}$$

with:

$$\hat{P}_\alpha(w_2) = \frac{f_d(w_2)^\alpha}{\sum_{w_2} f_d(w_2)^\alpha} \tag{2.16}$$

where:

$f_d(w_2)$ - frequency of w_2 in document d

Word2Vec - Word2Vec is a machine learning algorithm used to generate dense vector representations of words in a continuous vector space from a large data set that contains a huge vocabulary [Mikolov et al., 2013]. These representations are known as word embeddings. These embeddings are intended to capture semantic relationships and similarities among words.

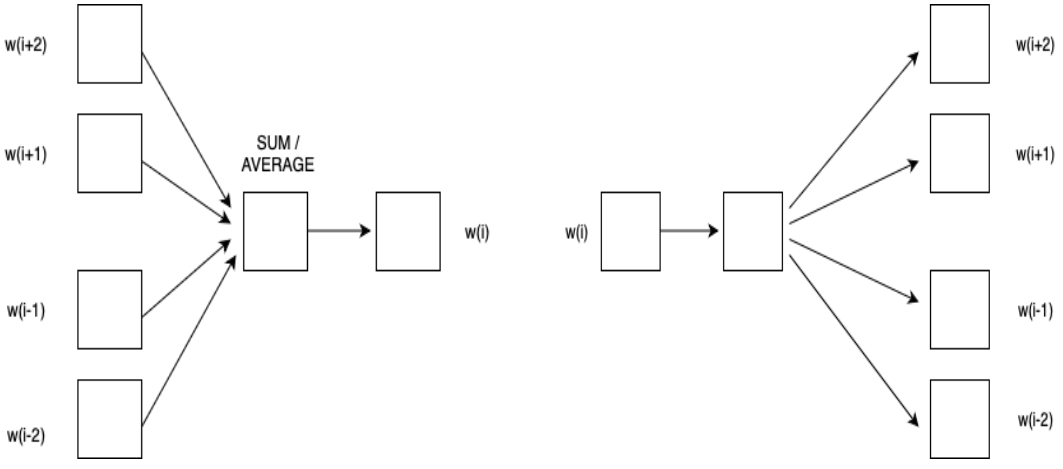


FIGURE 2.2: Word2Vec architectures

Word2Vec employs the concept of distributional theory, which claims that words that appear in similar contexts share semantic ties [Goldberg and Levy, 2014]. Using the theory, Word2Vec proposes two (2) architectures, i.e., Continuous Bag of Words (CBOW) and Skip-gram. The architectures predict either context words given a target word or the target word given context words.

The CBOW architecture predicts a target word based on the surrounding context words of a specified window size, as depicted by Figure 2.2 (left) [Mikolov et al., 2013]. The objective of the model is to maximise the probability of the target word given the context words. When training on large datasets, this architecture is efficient.

The Skip-gram architecture predicts the context terms for a target word given its context words, as depicted by Figure 2.2 (right) [Mikolov et al., 2013]. The objective of the model is to treat the context words as classes that ought to be predicted by the model. This approach tends to perform better with a small dataset and provides good word embeddings for uncommon words. In this work, the Skip-gram architecture is used due to the size of the data set and the word frequency dynamics of Twitter data.

2.3.2 Geometric Deep Learning

Geometric deep learning is a subfield of machine learning concerned with the creation of deep learning algorithms that can operate on non-Euclidean geometric data structures such as graphs and manifolds [Bronstein et al., 2017]. In contrast to the success of traditional machine learning methods and deep learning methods, which make the assumption that the data is represented in a Euclidean space, such as a grid or vector space, this method assumes data is represented in a non-Euclidean space [Cao et al., 2020]. Due to its prospective applications in a wide range of domains, including computer vision, natural language processing, social network analysis, and medical imaging, geometric deep learning has received considerable attention in recent years. There are a wide range of applications for geometric deep learning methods such as the use in recommendation systems [Cao et al., 2020], detecting fake news in social media [Monti et al., 2019] and for drug discovery [Atz et al., 2021].

This chapter explains the concept of geometric deep learning as well as the algorithms that are used in this study. This section examines the application of geometric deep learning to NLP. The chapter also examines three (3) geometric deep learning methods that are used in subsequent chapters.

2.3.3 Graph Neural Networks for Natural Language Processing

Graph Neural Networks (GNNs) are a set of deep learning methods used to process graph data that form part of Geometric deep learning [Gori et al., 2005]. GNNs can be used for a variety of purposes in Natural Language Processing (NLP) by learning global text representation through the aggregation of information from node neighbours [Zhang and Zhang, 2020]. GNNs can capture semantics information words in text [Liu et al., 2021], making them very effective for

NLP. The tasks that can be tackled using GNNs are text classification, sequence labelling, machine translation, event detection, and question answering [Liu et al., 2021].

The learning can be divided into two levels: node focused and graph focused. The goal of representation learning is to learn representative properties of the graph. The various tasks that can be completed with GNNs can be divided into three categories: graph level tasks, node level tasks, and edge level tasks. A graph level task is the task of predicting a single outcome for an entire graph, a node level task is a task to predict an outcome on a node level and an edge level task is a task to predict the connectivity between two or more nodes in a graph [Zhang et al., 2021]. This work solely focuses on graph focused representation learning.

Graph representation learning uses the message passing framework. The framework seeks to propagate features between adjacent nodes in a graph. The message-passing framework consists of 3 steps, as presented by Equation 2.17:

1. Initialisation of the representation of the node
2. Aggregation - calculate the influence between nodes
3. Update - update the representation of the node

$$n_v^{(k+1)} = UPDATE^{(k)}\left(n_v^{(k)}, AGGREGATE\left(\{n_u^{(k)} : \forall u \in \mathcal{N}(v)\}\right)\right) \tag{2.17}$$

All the message-passing paradigm GNNs use Equation 2.17 as a standard formula. The differences between the algorithms is the manner by which the algorithms perform the UPDATE and AGGREGATION step.

2.3.3.1 Graph Convolutional Neural Networks

The graph Convolutional Neural Network (GCN) is a neural network architecture that is designed to perform representation learning on graphs. The architecture was proposed by Kipf and Welling [2016]. This section explains the forward operations of the graph convolutional network.

The GCN operates on the input graph $\mathcal{G}$ that is represented by adjacency matrix $A \in \mathbb{R}^{N \times N}$, where N is the number of nodes in the graph. The nodes have feature vectors represented by the initial word vector embedding matrix $H^{(0)} \in \mathbb{R}^{N \times D}$, where D represents the dimension of the word embedding space. The one layer GCN message passing forward operation is represented by Equation 2.18.

$$H^{(1)} = g\left(\hat{A}H^{(0)}W^{(0)} + b^{(0)}\right) \tag{2.18}$$

where:
$\hat{A} \in \mathbb{R}^{N \times N}$ - Normalised symmetric adjacency matrix
$H^{(0)} \in \mathbb{R}^{N \times D}$ - initial word vector embedding

$W^{(0)} \in \mathbb{R}^{D \times D'}$ - Learnable weight matrix
$b^{(0)} \in \mathbb{R}^{D'}$- Learnable bias matrix
g - nonlinear activation function

The purpose of the operation $AH^{(0)}$ is to allow a node to pass features to neighbouring nodes. However, that operation that does not include the feature to the original node, hence a self-connecting mechanism is included where an identity matrix is added i.e. $\tilde{A} = A + I_N$.

However, in order to introduce instability in the network due to the multiplication of $\hat{A}$ with the feature matrix, $H^{(0)}$, and the weight matrix, $W^{(0)}$. The instability causes exploding gradients during backprobagation. The self-looping adjacency matrix is normalised using a self-connecting Degree matrix $\tilde{D}$ as depicted in Equation 2.19, resulting in $\hat{A}$.

$$\hat{A} = \tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} \tag{2.19}$$

Multiple GCN layers can be stacked together using Equation 2.18 which results in passing of information from further neighbours i.e. going from local to global message passing. Adding L layers in the networks results in Equation 2.20 for the l^{th} layer.

$$H^{(l+1)} = g\left(\hat{A} H^{(l)} W^{(l)} + b^{(l)}\right) \tag{2.20}$$

2.3.3.2 Dynamic Graph Convolutional Neural Network

The data can also be processed sequentially. In order to achieve sequential processing of data, Deng et al. [2019] propose an updated formulation of the GCN layer that is able to account for temporal dynamics. To account for the temporal inputs, Equation 2.20 is transformed to Equation 2.21.

$$H_{t+1}^{(l+1)} = g\left(\hat{A}_t H_t^{(l)} W^{(t)} + b^{(t)}\right) \tag{2.21}$$

where:
$\hat{A}_t \in \mathbb{R}^{N \times N}$ - Normalised symmetric adjacency matrix at time t
$H_t^{(l)} \in \mathbb{R}^{N \times D^{(t)}}$ - embedding at time t
$W^{(t)} \in \mathbb{R}^{D^{(t)} \times D^{(t+1)}}$ - Learnable weight matrix
$b^{(t)} \in \mathbb{R}^{D^{(t+1)}}$- Learnable bias matrix
g - nonlinear activation function

The Temporal Encoding layer is based on the work by Deng et al. [2019]. The purpose of the layer is to preserve and carry the encoding of the keywords from the initial semantic word embedding, $H^{(0)} \in \mathbb{R}^{N \times D}$ at $t = 0$ throughout each time-step $t > 0$ of the dynamic GCN network. The temporal encoding units acts in a similar manner as a residual connection in the Microsoft ResNet Convolutional Neural Network architecture [He et al., 2016]. The benefit of the residual connection is to reduce information loss during training.

In the unit two parameters are learned: $H_p^{(t)}$ and $H_e^{(t)}$. The two parameters are then concatenated to form the new $\hat{H}_t$ to be passed onto the next time-step t. The encoding process is defined as follows [Deng et al., 2019]:

$$H_p^{(t)} = H_t W_p^{(t)} + b_p^{(t)} \tag{2.22}$$

$$H_e^{(t)} = H_0 W_e^{(t)} + b_e^{(t)} \tag{2.23}$$

$$\hat{H}_t = \tanh\left(H_p^{(t)} || H_e^{(t)}\right) \tag{2.24}$$

where:
$W_p^{(t)} \in \mathbb{R}^{D^{(t)} \times \alpha}$
$b_p^{(t)} \in \mathbb{R}^{\alpha}$
$W_e^{(t)} \in \mathbb{R}^{D \times \left(D^{(t)} - \alpha\right)}$
$b_e^{(t)} \in \mathbb{R}^{\left(D^{(t)} - \alpha\right)}$
$0 \leq \alpha \leq D^{(t)}$

The α is a hyper-parameter that controls the dimension of the resulting $\hat{H}_t$ encoding matrix. This value is dependant on the number of dimensions of the initial initial semantic word embedding, $H^{(0)}$. Deng et al. [2019] report a high evaluation score with $\alpha = 60$ for a embedding size of 100. Hence, based on the dimension size we use on the BERT model of 512 dimensions, we will optimise for the optimal α.

The $\hat{H}_t$ replaces the $H_t^{(l)}$ in Equation 2.21, hence the dynamic GCN layers becomes:

$$H_{t+1} = g\left(\hat{A}_t \hat{H}_t W^{(t)} + b^{(t)}\right) \tag{2.25}$$

The main purpose of using keywords in a graph to predict a social unrest is to extract the most meaningful and important words that can increase the accuracy of the model.

2.3.3.3 Graph Isomorphism Network

The concept of graph isomorphism in graph theory explains a phenomena whereby two graphs are topologically similar. The graph isomorphism problem is whereby there in an attempt to distinguish isomorphic graphs [Xu et al., 2019]. The Weisfeiler-Lehman (WL) algorithm [Weisfeiler and Leman, 1968] for graph isomorphism test has been the one classical solution that has been proposed to solve the problem in a computationally efficient manner on a number of graphs by checking whether two graphs are non-isomorphic. The Graph Isomorphism Network (GIN) [Xu et al., 2019] is a network that attempts to solve this problem for representation learning on graphs.

The GIN achieves the same purpose as the WL algorithm by introducing two injective functions that applied on message passing framework for the $AGGREGATE$ and $UPDATE$ components in Equation (2.17). Using the universal approximation theorem of multi-layer perceptron (MLP) which states that at least one unit of an MLP can approximate any continuous function with a finite state [Hornik et al., 1989]. Hence, the injective functions are learned using Equation (2.26).

$$h_v^{(k+1)} = MLP^{(k+1)}\left(\left(1+\epsilon^{(k+1)}\right)\cdot h_v^k + \sum_{u\in\mathcal{N}(v)} h_u^k\right) \tag{2.26}$$

2.3.3.4 Output Layer

The model needs to predict whether a protest will occur or not i.e. the output layer needs to return one output giving the probability of an unrest. In to order achieve this, the final $\tilde{H}$ matrix is passed into a fully connected layer that encodes the weights from each node as depicted in Equation 2.27. The final layer output is passed through a sigmoid σ in order to return the $\hat{y}$, probability of a protest.

$$\hat{y} = \sigma\left(W_1\tilde{H}_T + b_1\right) \tag{2.27}$$

2.3.4 Learning Procedure

The model will need to update and train the weights in order to learn the correct encoding of the data. The method backprobagation that is used as the learning procedure using Adam [Kingma and Ba, 2017]. The method tries to optimise the binary cross-entropy as depicted by Equation 2.28.

$$\mathcal{L} = -\frac{1}{N}\sum_{i=0}^{N} y_i \log\left(p_i\right) \tag{2.28}$$

2.3.5 Model Evaluation

The objective of this work is to use the model to predict a binary outcome for data not used to train the model. As such a confusion matrix is used as depicted by Figure 2.3. To evaluate the performance of the model on data that was not observed during training, confusion matrix-defined metrics are used [Hossin and Sulaiman, 2015].

		Actual Class	
		Positive	Negative
Predicted Class	Positive	True Positive (tp)	False Negative (fn)
	Negative	False Positive (fp)	True Negative (tn)

FIGURE 2.3: Confusion Matrix

The metrics will make use of are the Sensitivity/Recall (2.29), Specificity (2.30) and Balanced Accuracy (2.31).

$$Sensitivity = \frac{tp}{tp + tn} \tag{2.29}$$

$$Specificity = \frac{tn}{tn + fp} \tag{2.30}$$

$$Balanced \quad Accuracy = \frac{1}{2}\left(Sensitivity + Specificity\right) \tag{2.31}$$

The metrics capture different aspects about the model's predictions, i.e. Sensitivity is a metric that indicates a model's ability to recognise only positive samples. Specificity indicates the model's ability to recognise negative samples. The balanced accuracy is used in order to combine the two measures. The balanced accuracy results in an unbiased performance measure for a classification task with imbalanced data [Brodersen et al., 2010].

2.4 Conclusion

This literature review and background has investigated the current theories behind protest and the justification of using social media as an avenue to use to predict social unrests. However, as far as this work is concerned, social media use in South Africa has not been cited as input data in this prediction task. This work extends the research of previous researchers by focusing on South Africa, in a multilingual setting for keyword extraction rather than monolingual setting. Additionally, this work investigates the automatic extraction of keywords to be used in the creation of word relation networks that can be used for geometric deep learning methods

Chapter 3

Methodology

3.1 Introduction

The rise in violent protests in South Africa is the driving force behind interest in this work. Hence, the primary focus of this work is to understand whether deep learning techniques can be used to anticipate the occurrence of a protest. In order to achieve the objective of this work, the methodology section is separated based on the previously mentioned research questions. The objective of the study is to investigate the effectiveness of nose reduction techniques on Twitter data in order to predict protest-related events in South Africa using Graph Neural networks.

This section of the study details the technique that lead to the conclusion of the thesis statement. The section begin by describing the research design for the project, allowing a grouping of the efforts into a specific design. Following this, this section outlines the instruments and methodology for gathering data. Additionally, the section outlines the modelling process conducted with the data. Finally, the limitations and ethical considerations of this work are further discussed.

3.2 Research Design

The goal of this research is to find a relationship between social media data and the prevalence of protests in South Africa. The research examines protests in South Africa for the period between 2019 and 2021, i.e., a 3-year window. The work looks at using graph neural networks for predicting protest-related events in the South African context using word relation networks that are derived from the text data on Twitter. The premise is that if a model can accurately predict the occurrence of a protest, then deductions can be made from the dynamics that led to the predicted discontent.

There are challenges to using text from social media. Social media data is prone to being noisy and containing incorrect information [Barbier and Liu, 2011]. South Africa is also unique due to its multilingual society. There is also the argument that data from social media is biased

and not representative of the population [Schwartz and Ungar, 2015]. However, the adoption of social media has grown since its inception, with an estimated number of users of over 3.8 billion in 2020 [Jayaram et al., 2020]. There are also some privacy concerns with the use of social media data; however, since this work is only concerned with aggregated text content of publicly available social media text, we deem this work not to be a violation of the privacy of the users of the platform. The issues that are present in this type of research design can be circumvented by the use of exploratory data analysis and appropriate data cleaning techniques.

The use of social media data provides the advantage of acquiring a large amount of data on the opinions of a diverse collection of people [Chen et al., 2014]. Social media platforms are also considered avenues for marginalised groups in society to participate and co-create in social discourse [Rodny-Gumede, 2017b]. This study's purpose would benefit from this data because we would be collecting qualitative data from South Africans without resorting to research questionnaires that are usually limited in scale and variety. The analysis of text based on social media is data-driven and hence not driven by a priori assumptions contained in questionnaires.

3.3 Research Instruments & Datasets

In this section of the study, the instruments that have been used to collect the data for the purposes of our study are described. Additionally, the datasets obtained are also described. The section details the two major data sources that have been used in the study: the Global Dataset of Events, Location, and Tone (GDELT) [Leetaru and Schrodt, 2013] and Twitter data [Kwak et al., 2010b].The background of the data source, its design, purpose, and reliability are discussed.

3.3.1 Global Dataset of Events, Location, and Tone (GDELT)

Kalev H. Leetaru founded the GDELT Project in 1979 with the intention of keeping an eye on global societal issues affecting people. The GDELT Project currently comprises over a billion data points from 1979 to the present. The dataset is an automated event data collection system [Ward et al., 2013] using machine coding. The GDELT data relies purely on an automated event encoding system [Deng and Ning, 2021]. The underlying event coding uses the Conflict and Mediation Event Observation (CAMEO) actor coding scheme, a framework used for the purposes of coding events related to political disputes between parties. The GDELT Project has been separated into two (2) versions, i.e., GDELT 2.0 and GDELT 1.0. The GDELT 2.0 format has data starting from February 2015, and the GDELT 1.0 format was from January 1979 to the beginning of February 2015. Considering the duration of this study, the focus of the discussion is on the GDELT 2.0 version.

The GDELT project collects data from numerous data sources around the globe every 15 minutes and makes it publicly available through raw Comma-Separated Values (CSV) files and Google BigQuery. These include online and print media, blogs, and videos. These sources include AfricaMedia, News24, BBC Monitoring, and the Washington Post [Rogers, 2013].The data

collected covers over 65 languages globally that are translated into English. The database contains multiple event-related records. An example of a South African event captured in the project is a peaceful march in Durban which was captured from here: https://www.roadsafety.co.za/2015-12/march-in-durban-against-police-killings-xenophobia-and-all-forms-of-crime/.

The GDELT project also collects data related to 2 300 emotions and themes using sentiment analysis. The initiative also monitors the geographical source of the data and the progress of the event. In addition to collecting data related to events, the GDELT Project also provides a global knowledge graph that extracts the entities, themes, locations, and tone contained in the news articles.

The data is initially extracted from the data sources previously mentioned. Thereafter, a translation algorithm is used for the purposes of translating text from non-English sources into English using the GDELT Translingual system. Thereafter, with the translated English text, NLP techniques are applied in order to extract themes and entities from the text.

The major technology used for data collection is the Google Cloud Platform. The services that monitor the sources and extract data are based on the Google Cloud compute engine. Google BigQuery and Google Cloud Storage are used to store the data that has been collected. The events data is collected every 15 minutes. The data is collected in two tables, i.e., the events table and the mentions table. The tables are partitioned by day using BigQuery's functionality due to the scale of the data. The events table contains all the information as extracted from the data source, including but not limited to the date of the event, the CAMEO code attributes of the event, the attributes of the event, and the geographical location of the event. The mentions table is a table containing all the mentions of the event in the events table. The mentions has a one-to-many association with the events table, hence assessing the trajectory of a particular event.

The GDELT Project does have some limitations. It has been found that the GDELT Project has no procedures for eliminating false positive reports in their sources [Ward et al., 2013]. In order to overcome this challenge, the project provides a confidence score on the mentions table. The score provides a measure of confidence for the extraction of the event in an article. It is also noted, according to Bunte and Vinson, that GDELT does not capture inter-religious conflicts due to the coverage bias of international newspapers. However, this will not impact our study since the focus is on national conflicts.

There are other projects similar to the GDELT Project. A similar project is the Armed Conflict Location and Event Data (ACLED) Project [Raleigh et al., 2010]. In comparison, the ACLED data is manually updated according to news reports. According to Manacorda and Tesei, ACLED suffers from Type I error in comparison to the GDELT Project, which suffers from Type II error. However, there is a high correlation between the events reported by GDELT and ACLED [Manacorda and Tesei, 2020]. As a result, using the GDELT Project in this work does not make the outcomes data-dependent.

3.3.2 Twitter Data

The secondary data requirement of the study requires data from social media. The study have opted to use Twitter as the main social media data source. The research instrument that was used for the collection is Twarc[1]. Twarc is a Python library for obtaining data on Twitter. This section will briefly discuss the module.

Twarc is a Python wrapper for the Twitter Application Programming Interface (API) [Weber et al., 2020]. Twarc was developed in 2006 to enable the collection of Twitter data. Twarc is part of the "Document the Now" project [Galarza, 2018]. The library has been used for social media research such as analysing COVID-19 [Dashtian and Murthy, 2021] and crisis research [Haq et al., 2022]. Twarc requires a user to apply for the Twitter developer account in order to get access. Additionally, Twarc has been updated in order to be able to work with the Twitter API v2, including academic access.

There are limitations with the use of twarc that are inherent in the Twitter API v2. The Twitter API limits searches to 10 million tweets per month for academic access and 2 million tweets for developer access. In addition to the monthly limit, tweets that have been deleted, tweets from accounts that are private or suspended, and tweets from users who are private at the time of extraction cannot be retrieved through twarc. The benefit of twarc, or the Twitter API v2, is that it provides tweet annotations that are based on Twitter's own semantic analysis in order to avoid the use of a large set of search keywords.

3.3.3 Research Data

This section of the study describes the actual data that is used for the research. As briefly discussed, the GDELT database is used as the main database for data related to protests and Twitter for extracting the textual information needed for our model. The research data is separated in the discussion into two types: training data and evaluation data. The training data is the data that is used for modelling purposes. The evaluation data is to be used in order to evaluate the performance of the model.

The ground truth data that is used for the events data is initially created. This data serves as the training data. The objective of the training data is to learn the association between social media text data and output events provided by the GDELT database. For the purposes of this study, data from 2019 through 2021 will be considered. For the purposes of temporal generalisability, the training data will cover the first two years, 2019 and 2020, while the evaluation data will be for 2021.

Data from the GDELT Project is maintained in a public BigQuery database and updated every 15 minutes. The data contains records with 58 fields in the events table that contains information on an event occurring around the world. The database also contains a mentions table with 61 fields containing all the sources of the events in the events table. Among the 61 fields in the GDELT Project events database, we are only interested in GLOBALEVENTID, SQLDATE,

[1] https://twarc-project.readthedocs.io/en/latest/

Actor1CountryCode, ActionGeo_CountryCode,EventCode, EventRootCode, and SOURCEURL for this study. We are only interested in the GLOBALEVENTID field on the mentions table in order to use it to merge the table with the events table and the Confidence field. A description of the fields is provided below:

- GLOBALEVENTID - a unique ID of each record on the events table
- SQLDATE - The date of the event as per the source
- Actor1CountryCode - The CAMEO code for the country of Actor1.
- ActionGeo_CountryCode - The location of the event, using the 2-character FIPS10-4 country code
- EventCode - The raw CAMEO action code that describes the action done by Actor1 on Actor2
- EventRootCode - The level two of the CAMEO action code describing the action done by Actor1 on Actor2
- SOURCEURL - The url of the source where the event was extracted
- Confidence - The confidence percentage of the extraction from the source.

CAMEO codes are developed in a three-level taxonomy. The three-level taxonomy is used in order to demonstrate different degrees of specificity for the event. For example, the root level could be '14' for 'protests', and the second level would be '141' for 'demonstrate or rally', and the third level is '1411' for 'demonstrate for leadership change'. As can be observed, the event classification becomes more specific as we move down the taxonomy. For the purposes of our research, we will only use the root-level CAMEO code.

The primary focus of the research is on protest-related conflict, and hence only a select subset of the CAMEO codes that are related to protests and conflict are extracted. As a result, only the CAMEO codes '14' (PROTEST), '15' (EXHIBIT FORCE POSTURE), '17' (COERCE), '18' (ASSAULT), '19' (FIGHT), and '20' (USE UNCONVENTIONAL MASS VIOLENCE) are examined[2]. Additionally, non-protest-related events are considered in order to create a binary classification problem. The CAMEO codes that are chosen to reflect this phenomenon are '01' (MAKE PUBLIC STATEMENT), '02' (APPEAL) and '03' (EXPRESS INTENT TO COOPERATE) [3] . Due to the length of the investigation, the SQLDATE field is restricted to dates between 2019 and 2021. Additionally, the ActionGeo CountryCode field is restricted to 'SF' since the investigation focuses on social protests in South Africa. Finally, the focus is on events initiated by the South African population; hence, ActionGeo_CountryCode is restricted to 'SF'. Additionally, our interest is on events initiated by the South African population, hence, we set Actor1CountryCode to 'ZAF'. As stated above, because the GDELT Project is machine-coded, it is susceptible to Type II errors. To reduce the error, we only consider sources with a confidence level of 90% or higher.

[2]Table A.1
[3]Table A.2

In November 2022, there were 237.8 million monetizable active daily users on Twitter [Dixon, 2022].In order to obtain valid and relevant data for the investigation, there are limits that are applied to the data extracted from Twitter. Similar to the GDELT data, the Twitter data is constrained to only South African Twitter users that were active during the window of investigation, i.e., 2019 to 2021.

The twarc API used to collect data from Twitter requires keyword matching. The keywords are the brief phrases that the API uses to retrieve the tweets of interest. The research focuses on protest-related events; therefore, the keywords used must correspond to this purpose. A number of social protests have occurred in the country as a result of government corruption and a lack of reform [Mare, 2014; Bonga, 2021]. In addition to the keywords 'protest', 'unrest', and 'strike', which are all words related to protests in South Africa, since it has also been highlighted that the main drivers of protests in South Africa are a lack of transformation and corruption in government [Mare, 2014; Bonga, 2021], government-related keywords are included: 'corruption', 'ANC', 'Jacob Zuma', and 'transformation'.

There is also evidence that hashtags are extensively used before a demonstration. The most popular hashtag used during the Rhodes University protest was'mustfall' [Frassinelli, 2018]. In addition to the 'mustfall' hashtag, we include the 'mustfall' keyword that is closely related to public discourse [Daniels, 2016] and the globally popular keyword 'occupy' [Mottiar, 2013]. Finally, location-based keywords were used as a means to highlight areas where protest participants would convene. These are all the major cities in South Africa.

The twarc API makes use of query search operators that allow the user to modify the returned results from the API. These are in addition to the keywords provided to the API. In this work, three (3) operators are used, i.e.,

- 'is:retweet' - a Boolean on whether to extract retweet tweets
- 'is:reply' - a Boolean operation on whether to return reply tweets
- 'has:media' - a Boolean operation on whether to return tweets containing media

In this work all the above mentioned operators are set to 'False'. The final list of fields that are extracted are described below:

- author_id - the unique user ID for the tweet
- conversation_id - a unique conversation ID for the tweet
- created_at - the create date of the tweet in the format 'day-month-year H:M:S'
- text - the original text of the tweet
- public_metrics.like_count - total number of likes for the tweet
- public_metrics.reply_count - total number of replies for the tweet
- public_metrics.retweet_count - total number of retweets for the tweet

- author.location - Geographic location of the user
- author.country_code - Country code for the Geographic location of the user

The Twitter data and the GDELT data are merged together in order to provide the complete set of the research data required for the study. The research question for this work is to derive Twitter data dynamics that signal the likelihood of a protest event. Therefore, the requirement is that the Twitter data and GDELT data are synchronised in time, i.e., tweets are ordered in sequence of time before an event indicated by the GDELT data. To achieve this purpose, a sliding windowing technique comparable to that presented by Qiao et al. is used. The method only has an observation period and a prediction period, as depicted in 3.1. The observation period will contain tweets for five (5) historical days before the prediction period. It has been uncovered that a lead time of greater than four (4) provides good model performance since protests take a while to plan [Deng et al., 2019].

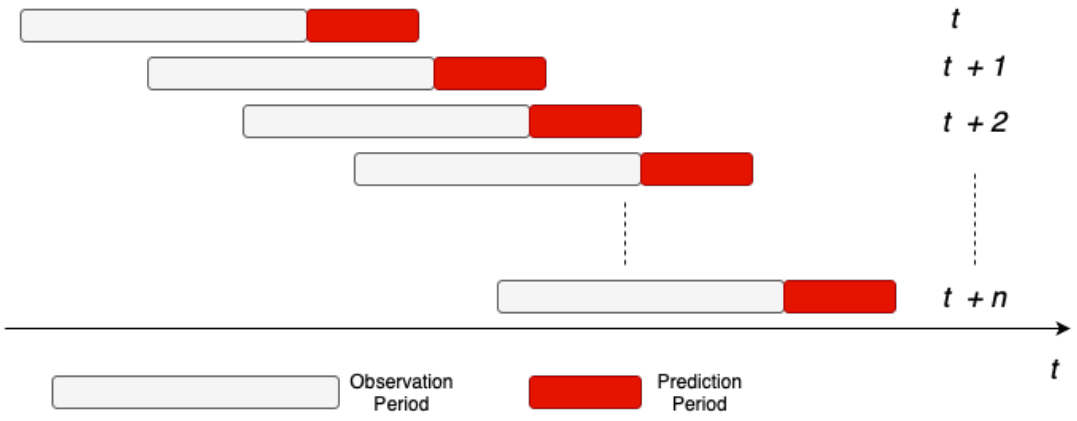


FIGURE 3.1: Sliding window

3.3.4 Modelling

In this section of the study, the modelling procedure used for the proposed data in the study is described. The techniques that are described below are for the complete data, i.e., the combined version of the GDELT data and the Twitter data.

The data analysis for this work is chosen in such a way that it assists in answering the research questions. Initially, an exploratory data analysis is conducted on the data using bar charts and other summary statistics visualisation techniques. This is used in order to identify trends and patterns in the data. This type of analysis will allow the identification of erroneous entries in the data. Thereafter, textual analysis methods are used on the Twitter text data to pre-process the data.

Twitter text data is known to contain noise that might impact any downstream modelling that is performed on the data. Different noise-reduction techniques are investigated in order to remove the noise from the text. The noise reduction is driven by using metrics calculated from the text. The metrics are: the average token length (3.2), the ratio of hashtags used in a tweet (3.3), the

ratio of mentions in a tweet (3.4), and the ratio of links contained in a tweet (3.5).These metrics are used to guide the filtering process of noise reduction.

$$\text{Number of tokens} = \sum_{i=0}^{n} \mathbf{1}_{w_i} \tag{3.1}$$

$$\text{Average token length} = \frac{\sum_{i=0}^{n} |w_i|}{n} \tag{3.2}$$

$$\text{Ratio of hashtags} = \frac{\sum_{i=0}^{n} \mathbf{1}_{\#\in w_i}}{\sum_{i=0}^{n} \mathbf{1}_{w_i}} \tag{3.3}$$

$$\text{Ratio of mentions} = \frac{\sum_{i=0}^{n} \mathbf{1}_{mentions\in w_i}}{\sum_{i=0}^{n} \mathbf{1}_{w_i}} \tag{3.4}$$

$$\text{Ratio of links} = \frac{\sum_{i=0}^{n} \mathbf{1}_{links\in w_i}}{\sum_{i=0}^{n} \mathbf{1}_{w_i}} \tag{3.5}$$

where:

$$w_i - \text{token in a tweet}$$
$$n - \text{total number of tokens in a tweet}$$

It has been found that for sentiment analysis, a pre-compiled stopwords list has a major impact on classification performance on Twitter data. In addition to the noise-reduction metrics, we also compile a stopwords list using the TF1 method [Saif et al., 2014]. The benefits of using the TF1 method are that it reduces the feature space and the sparsity degree of a dataset [Saif et al., 2014]. In addition to the above filtration methods, the following are applied to the Twitter data to enhance the signal:

- Remove URL's, punctuation and usernames
- All the text is converted to lowercase
- Tokenization of the text

The pre-processed data is further applied to different machine learning techniques to assess their classification performance. The methods that will be compared in this work are

the previously described methods: the Logistic Regression, Graph Convolutional Neural Network (GCNN), Dynamic Graph Convolutional Neural Network (DGCNN), and the Graph Isormophism Network (GIN). In order to apply machine learning methods to the data, the data needs to be converted to a suitable format for the model being applied.

In order to use the Twitter data for the prediction of a protest-related event using logistic regression, a feature matrix is generated from the Twitter text corpus. The feature matrix is generated using the TF-IDF values of the tokens in the corpus [Ramos et al., 2003]. However, for the purposes of the graph networks GCN, DGCN, and GIN, the data is transformed into a graph representation. The advantage of using graph representation for this task is that graphs are non-euclidean and can capture complex relationships that euclidean-based representations cannot. Two versions of the graph generation process are described: dynamic graph generation and static graph generation.

The Twitter data is initially ordered sequentially using the date of the post. Thereafter, for each event as given by the GDELT data, the prior 5 days of tweets are extracted. However, Twitter data contains many keywords that make it infeasible to use them all due to scalability. Hence, there is a selection process for relevant keywords. A relevancy score is defined by Equation 3.6using the TF-IDF score of each token in the corpus. The relevancy score is defined as the mean of each TF-IDF vector of a word w_i over all documents k in the corpus that contain w_i. The benefits of using the TF-IDF in the relevancy score are preserved from the original formulation of the TF-IDF. The relevancy score is similar to that proposed by Horn et al. [2017], but in this scenario, the class is not considered, and this formulation accounts for the number of documents that have the word. The benefit of the normalisation is to have values within the same range as the original TF-IDF values.

$$r(w_i) = \frac{1}{k_{w_i}} \sum_{k_{w_i}} \mathbf{x}_k \tag{3.6}$$

For the static generation, the TF-IDF matrix is calculated for a 5-day period, whereas for the dynamic generation, the TF-IDF matrix is calculated for each day. Thereafter, 100 keywords are chosen based on their relevancy score. The extracted 100 keywords are used to construct a word relation network with nodes as the keywords and edges based on the smoothed positive pointwise mutual information, as shown in Equation 2.9.A Word2Vec word representation that has been pre-trained on the entire corpus represents each node's feature matrix. The procedure is illustrated in Figure 3.2.

3.4 Limitations

In this section of the study, the limitations of the methodology are described. The section describes two (2) types of limitations, i.e., the data limitations and the model limitations.

The methodology described makes assumptions around the dynamics that result in a protest in South Africa and assumes that these dynamics can be captured through the use of Twitter

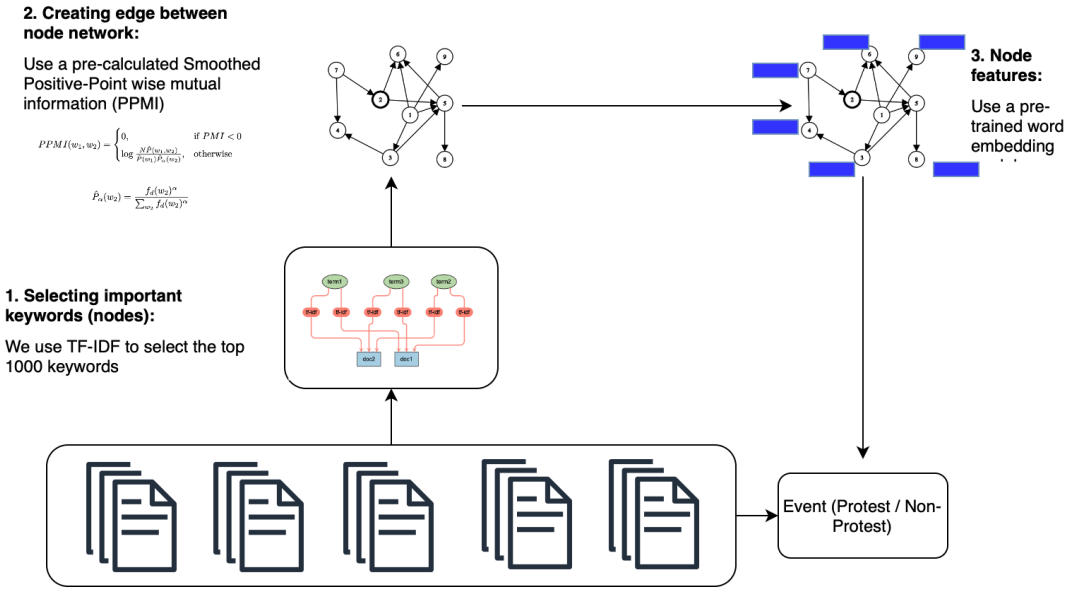


FIGURE 3.2: Graph generation process

social media data. The data that is used to infer the occurrence of a protest is primarily based on Twitter data. However, this is only a fraction of the information diffusion prior to a protest event. A most recent study indicates that Facebook and WhatsApp were the most popular social media sites in South Africa in 2021 [Lord, 2021]. Consequently, the study is susceptible to sampling and prejudice biases. However, due to privacy restrictions on these platforms, they cannot be used in this work. Apart from the online media, there is still a large population that communicates through word-of-mouth and has no digital footprint. This study makes use of data that is publicly available and has been widely used in research for event monitoring and prediction.

In this study, the network is limited to lexical information on the tweets. This limits the study to building contextual features and graphs in order to predict a protest. Data related to other salient information related to users, such as the number of followers, the geographical location of the user, the influence of the user in a network, etc., are not considered. The benefit of including such information would result in descriptive features that are able to capture the influence of user types on the occurrence of a protest. This study is therefore susceptible to exclusion bias. The possibility that user accounts are private or that Twitter has removed them as a result of community violations is the drawback of such information. There are also privacy violations that arise when we consider user-related data.

The model types that have been chosen for this study are also limited in the amount of information they may capture. Hence, the work is also susceptible to model selection bias. The models only use a select number of keywords and are unable to process larger networks. In this work, the size has been chosen to be 100 nodes. Additionally, graph-based neural networks leverage the concept of homophily in the modelling process, i.e., nodes that are similar to one another will be connected. This could be an incorrect presumption regarding protest-related occurrences.

To the best of our knowledge, all significant limitations of our methodology have been considered. In light of these limitations and the supplied justifications for their exclusion from the study, it is believed that the findings are valid for the research questions in this work.

3.5 Ethical Considerations

In this section of the study, a description of all the ethical considerations that have been considered and circumvented in our study is discussed. The ethical issues that originate in this study are related to the data that is used for our study and the environmental implications of large processing computations.

The data that is used is from publicly available profiles on Twitter at the time of extraction and has no direct identification with the user that made the social media post. The data is limited to only the keywords of the tweets and not the direct tweets from the user. None of the users data is stored in the study in order to refer back to the original author. Masking and removal of any usernames at the pre-processing stage of the data are applied so that no usernames form part of the outcome of the study.

The computational processing of large neural networks has a direct impact on carbon emissions. All the computations in this study make use of cloud infrastructure, which has been proven to be more computationally efficient and resilient. The use of cloud resources has also been found to reduce carbon emissions by close to 80% [Beardmore, 2020]. In this study, every effort has been made to avoid violating the privacy of Twitter users and to have no negative impact on the environment.

3.6 Summary

This chapter describes the research methodology for the purposes of answering the research questions. The research design for this work seeks to use social media data in order to find a relationship between posts and the occurrence of a protest.

This study focuses on a 3-year window data between 2019 and 2021. The data that is used for this work is the Global Dataset of Events, Location, and Tone (GDELT) and Twitter data for the period of investigation. The data from GDELT particularly focuses on the CAMEO codes that are related to protests and conflict. The codes that are used are codes for 'protest', 'exhibit force posture', 'coerce', 'assault', 'use conventional mass violence', and 'fight'. In order to create non-protest related instance, the CAMEO codes that are used are 'make public statement', 'appeal', and 'express intent to cooperate'. The data extracted is primarily for South African events.

Twitter is used for the social media text data. The keywords that are used for the text extracted are all related to protest-related keywords that have been found to be related to South African protests. Such keywords include, for example, 'protest', 'strike', and 'corruption'. In order to combine the Twitter data and the GDELT data, a sliding window approach is taken.

Where the observation window is Twitter text for a period of 5 days and the output variable is the GDELT indicator variable.

Twitter data is known to contain a lot of noisy information. A noise-reduction filtration technique is used. The metrics that are used to guide the filtration are the average token length, the ratio of hashtags in a tweet, the ratio of hashtags in a tweet, and the ratio of links. In addition to these normalisation steps, steps are taken to enhance the signal of the data.

For the purposes of modelling, the work compares different machine learning methods. The proposed methods that are used are Logistic Regression, Graph Convolutional Networks and Graph Isomorphism Networks. In order to use the data for logistic regression, a feature matrix is generated using the TF-IDF values of the tokens in the corpus. For the purposes of the graph networks, a relevancy score is defined, which is defined using the TF-IDF. Word relation networks are constructed using an adjacency matrix made of the Point-wise Mutual Information between the keywords and the node features given by the Word2Vec word representation.

The limitations of this work are also highlighted. Firstly, the work only uses Twitter data in order to model the dynamics of the origination of a protest in South Africa. However, there are other social media platforms that are used in the country that may be used but cannot be included in this work due to privacy concerns. Secondly, the work limits the dynamic to the content in Twitter posts, and there could be other driving factors besides the content in a post.

This section also considers all the ethical considerations of this study. The first consideration is related to the data; however, due to the fact that no personal identifiable information is used, the study does not violate any user privacy. Secondly, the computational infrastructure for this work does not have implications for carbon emissions due to the usage of cloud-based platforms.

Chapter 4

Data

4.1 Introduction

This section of the report investigates the data used for modelling and answering the research questions. The GDELT data are analysed and interpreted to generate the output variable for the modelling procedure. Both the GDELT data and the Twitter data undergo an exploratory analysis utilising graphs and summary statistics.

In this section, the noise reduction filtration procedure is explained using various metrics calculated from Twitter text data. This section then examines the process of generating features for the models that will be used to answer the research questions.

4.2 Exploratory Data Analysis

In this section, exploratory analysis and preprocessing of the data are conducted. The purpose of this section is to gain insights and patterns from the data and also identify any erroneous data. The data with which the exploratory analysis is done is the GDELT protest data extracted from the GDELT database and the Twitter data that is extracted using Twarc. Summary statistics and visualisation are used in order to perform the exploratory analysis.

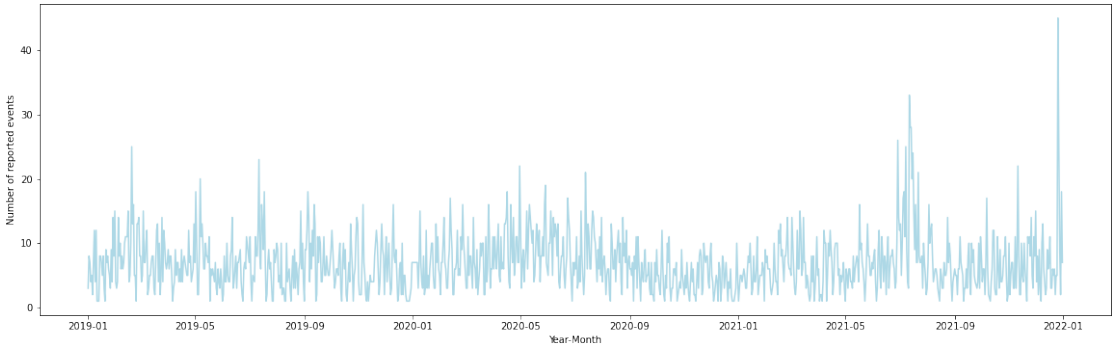


FIGURE 4.1: Number of reported events

The GDELT protest data has a total of 7 417 events for the period between 01 January 2019 and 31 December 2021, as depicted in Figure 4.1. On average, there are more events reported for the year 2021. It is noted that multiple types of events may be reported for one day; hence, the total does not represent the number of days with events. The unique number of days that have events is 1 067 days for the data.

As per the methodology of the GDELT data collection, the events are extracted from different online media sources. Figure 4.2 depicts the top 20 sources in the data. The top three (3) sources of events data are Eye Witness News (EWN), Independent Online (IOL), and News24.

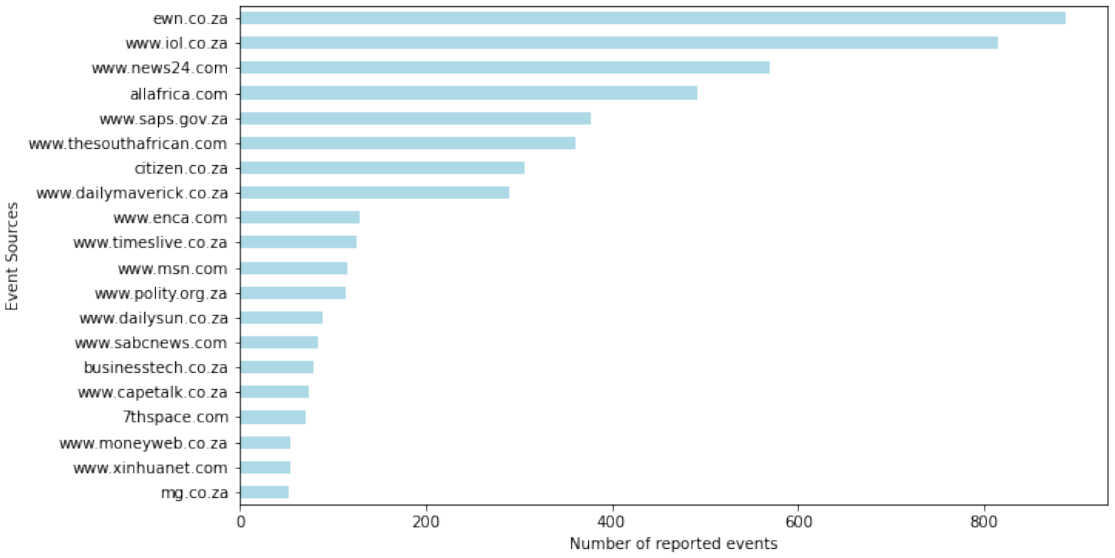


FIGURE 4.2: Number of reported events by source

Reported events are separated by event codes. Pre-selected event codes have been chosen so that they are suitable for this investigation. In total, there are 80 events codes in the data. Figure 4.3 depicts the top 20 event codes. It is observed that event code 10 (Make statement, not specified) and 20 (Appeal, not specified) are the highest event codes contained in the data.These two event codes were chosen to represent non-protest-related events forming part of event root level 01 and 02, respectively. The highest occurrence of protest-related event codes is 173 (Arrest, detain, or charge with legal action) and 190 (Use conventional military force, not specified).

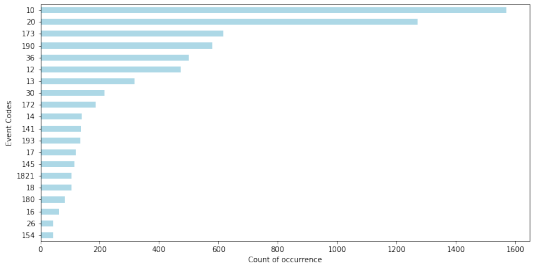


FIGURE 4.3: Count of top 20 event codes

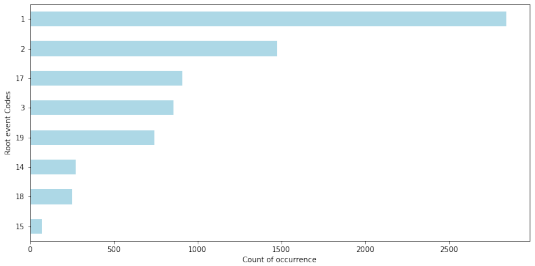


FIGURE 4.4: Count of root event codes

Root event codes, which stand for the higher order of the hierarchy, group the event codes. Figure 4.4 depicts the count of the root event codes in the data. Similar to Figure 4.3, it can be observed that the root codes with the highest number of instances are 1 (Make public statement)

and 2 (Appeal). The definitions of the rest of the root codes and event codes are contained in Appendix A.

Figure 4.5 depicts the root code count per day. The study observed that the root codes 14 (Protest), 15 (Exhibit force posture) and 17 (coerce) increased on average in mid-2021. This period coincides with the July 2021 unrest from 9 July 2021 to 18 July 2021.

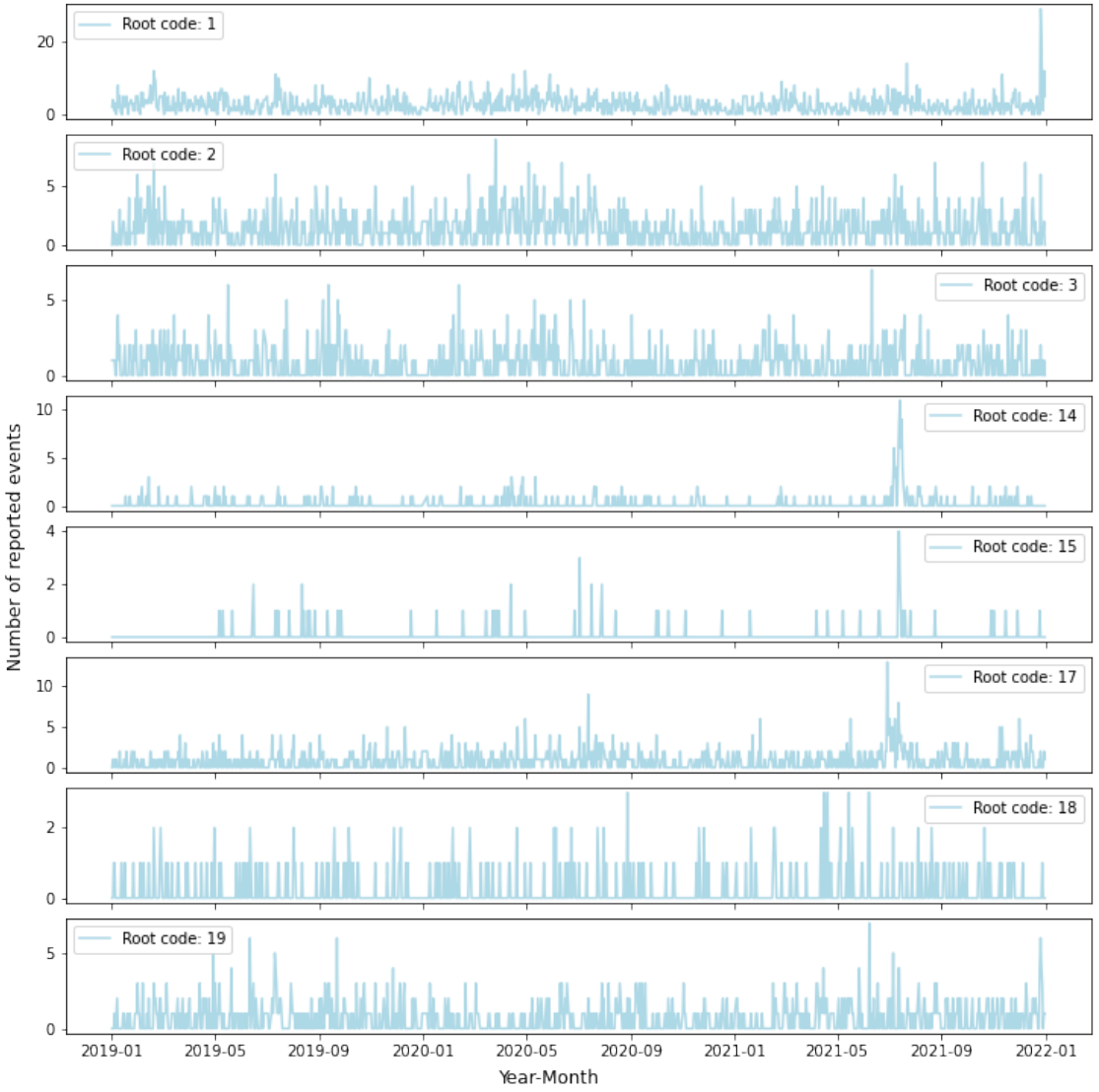


FIGURE 4.5: Count of root event codes per day

In order to use the GDELT data for modelling, the root event is transformed into a binary indicator. The binary indicator is a grouping of non-protest-related root codes and protest-related root codes, as depicted in Figure 4.6. However, note that there may be multiple conflicting event codes reported in one day. In order to get one indicator for the day, we default to a protest indicator if there was a protest-related event on the day; otherwise, the day is classified as non-protest. The final distribution of the output variable is depicted in Figure 4.7.

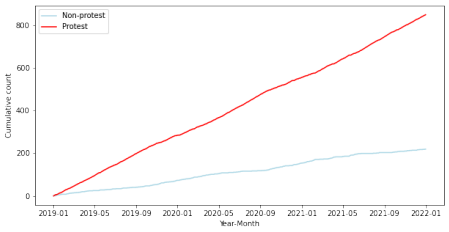


FIGURE 4.6: Cumulative count per day

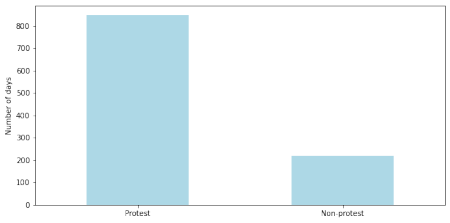


FIGURE 4.7: Number of days per class

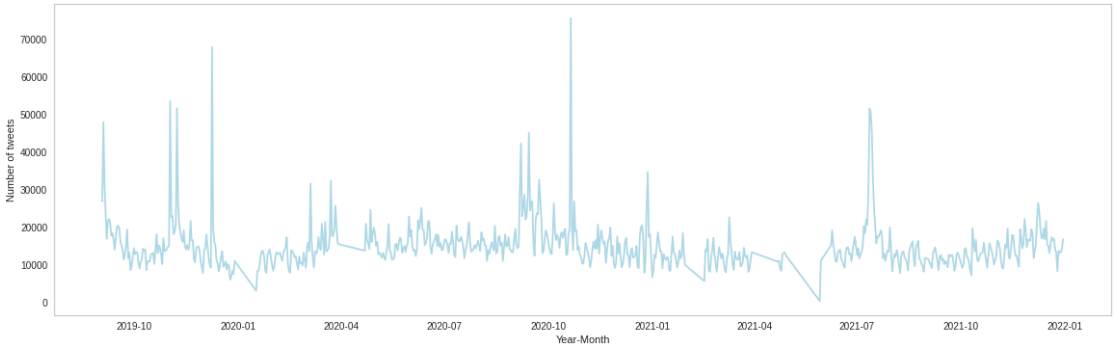


FIGURE 4.8: Number of daily events

The total number of observations for the Twitter data is 11 000 665 between 01 October 2019 and 31 December 2021. Figure 4.8 depicts the daily number of tweets from our data. It is observed that the trend is not consistent. There are days that have spikes in usage. One such day is the period we have noted as the July 2021 unrest.

A Twitter user is able to indicate a location on their Twitter profile. It can observed from Figure 4.9 that a large number of users indicate "South Africa" as their location instead of the actual town or city. However, it is also noted that users from the major South African cities are amongst the high number of users that indicate a town or city.

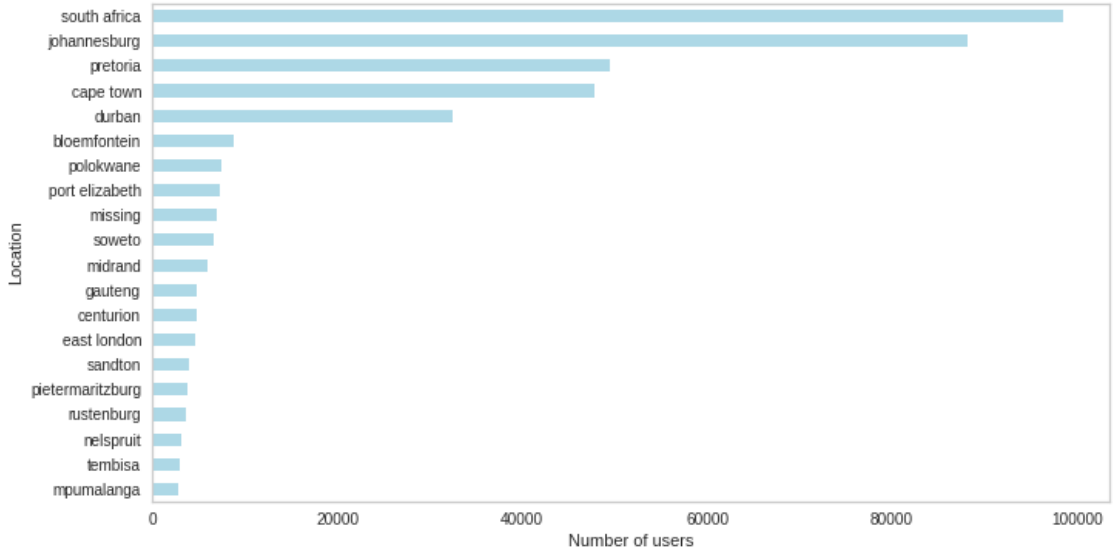


FIGURE 4.9: Number of users by location

In order to use the Twitter data for modelling purposes, the contents of the text data are initially analysed. However, due to the fact that in 2021 part of the data will be used for evaluating the temporal generalisability of the models, the analysis will only be conducted on the data for 2019 and 2020.

The individual tweets are initially tokenized using the TweetTokenizer module from the Natural Language Toolkit (NLTK) [Bird and Loper, 2004] in Python. Tokenization is the process of decomposing text into its basic units [Webster and Kit, 1992], which are usually words for common corpora. In the case of social media text, there are special cases that are unique to Twitter, such as emojis and hashtags. Hence, the use of a dedicated tokenizer for Twitter data is vital. Thereafter, the text data is processed using the following process:

- Convert all text to lower case
- Remove all numeric values from the text
- Remove any HTML tags from the text
- Remove punctuation from text
- Filter tokens that have less than 2 characters

Using the data that has been processed above, we use Equation 3.1 in order to calculate the number of tokens in a tweet. Figure 4.10 shows the distribution of the tokens in a tweets. There are an average of 20 tokens in a tweet for the data. We also show the 99^{th} percentile of the distribution.

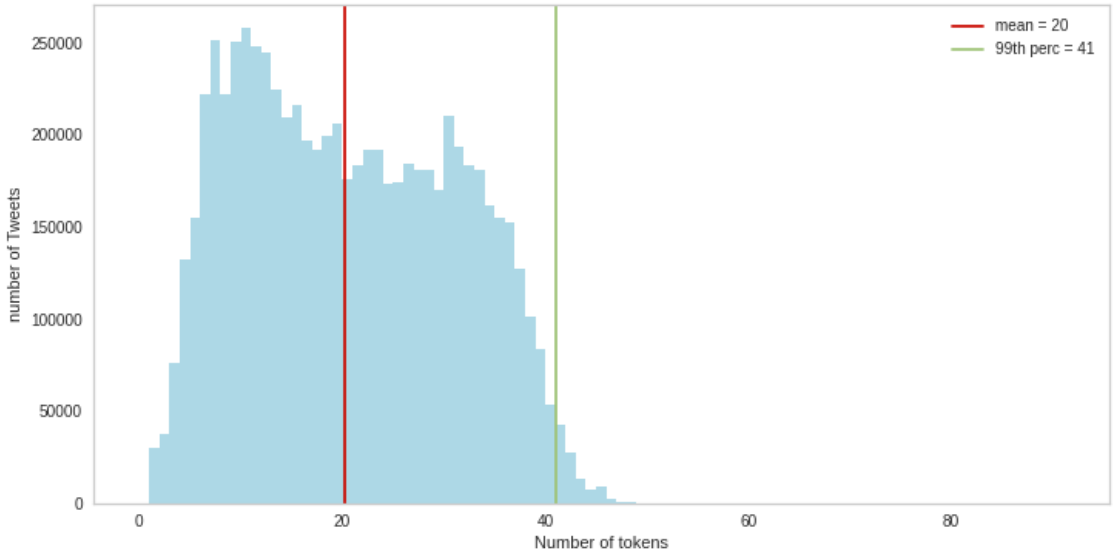


FIGURE 4.10: Distribution of tokens per tweet

Twitter is synonymous with users attempting to increase engagement and views in their tweets. However, other users may be posting spam for malicious intent [Santos et al., 2014]. These spam posts serve no other purpose than to increase views or redirect users to external web-pages [Santos et al., 2014]. In order to remove these non-spam tweets, filtration methods are used that look at

different content characteristics of tweets. The content characteristics are the average length of tokens in a tweet, the hashtags of a tweet, links in a tweet, and mentions in a tweet [Aggarwal et al., 2012].

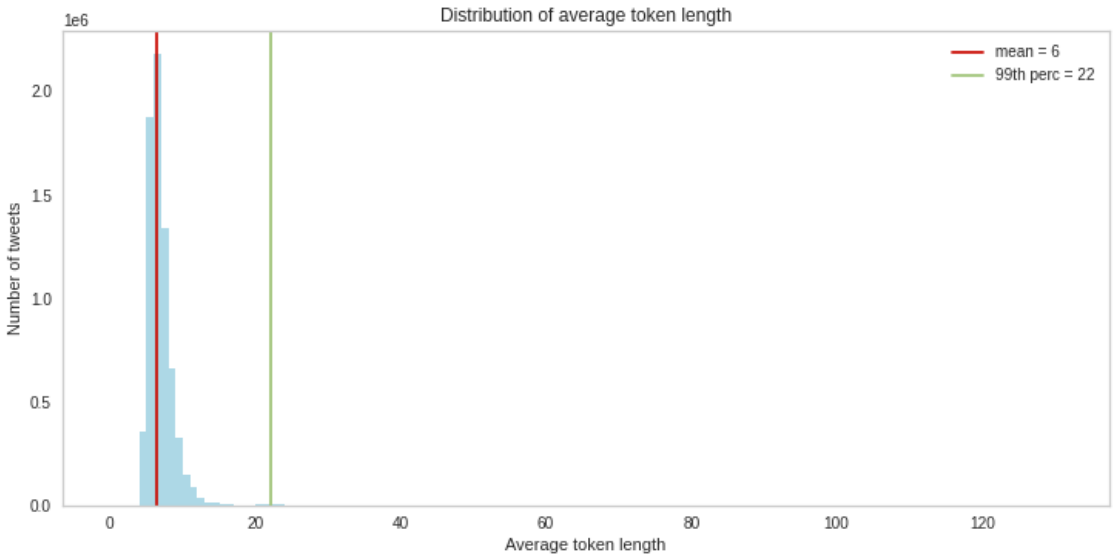


FIGURE 4.11: Distribution of average token length per tweet

The average token length in a tweet indicates the type of content in the tweet. We may also be able to identify misspellings and grammatically incorrect words. The average token length in the corpus is 6 characters, as depicted in Figure 4.11. Our corpus also contains tokens that are greater than 22 characters. It has been noticed that these are tweets that only contain external URLs. Hence, a threshold is set for the average number of tokens in a tweet to be the 99^{th} percentile.

Hashtags are frequently used with tweets. Hashtags serve as the topic of a tweet [Aggarwal et al., 2012]. Figure 4.12 depicts the most discussed topics for 2019 and 2020, respectively. These are the different topical discussions that were prevalent in the years 2019 and 2020. It is noteworthy that for the year 2020, the coronavirus became the most widely mentioned hashtag since South Africa was at the peak of the pandemic in 2020.

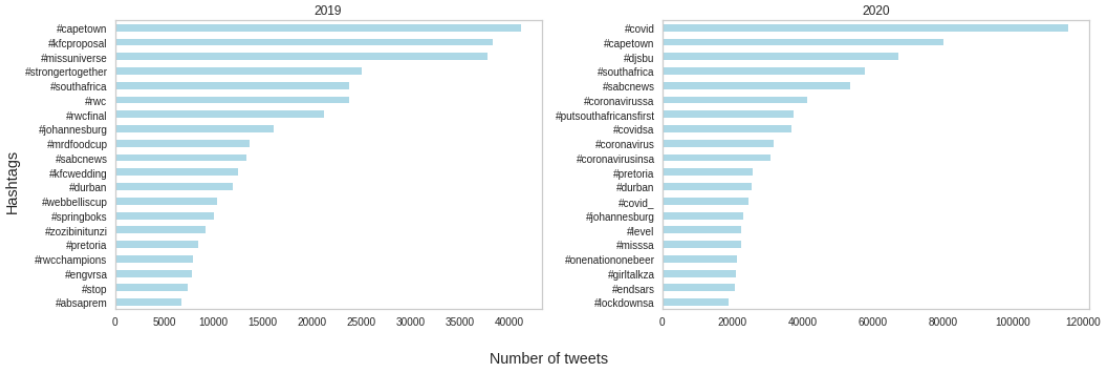


FIGURE 4.12: Top 20 hashtags per year

Figure 4.13 shows the hashtags that are similar to some of the keywords that were used in the Twitter search query:

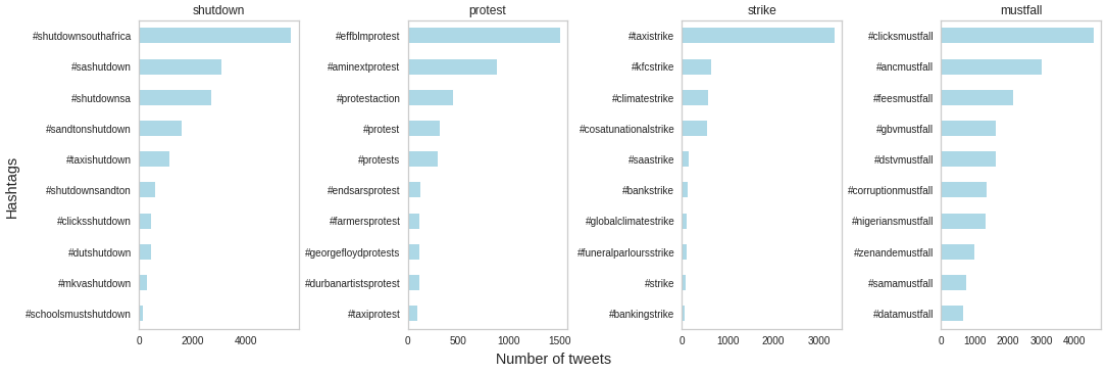


FIGURE 4.13: Top 10 hashtags per search keyword

Hashtags are used to increase the exposure and engagement of a Twitter post. When a hashtag is popular, it is considered a trending topic [Aggarwal et al., 2012]. These trending topics can be used by malicious users in order to increase the reach of their posts [Aggarwal et al., 2012]. Malicious users can use a combination of the trending hashtags in their posts, irrespective of their relevance to the post. However, users post content of different lengths. Due to the difference in the number of characters in a post, raw counts of the number of hashtags contained in a tweet cannot be considered on their own. For example, a user may post a tweet that contains only hashtags but would not be identified as potential spam if the fraction of words that are hashtags in the post were not considered. Hence, the relative frequency of hashtags per tweet is used in order to account for the differences in characters in a tweet.

The frequency of hashtags per tweet in our data ranges from one (1) hashtag to thirty (30) hashtags. Figure 4.14 depicts the percentage of hashtags per tweet. It is observed that a majority of tweets contain close to zero hashtags. However, it is important to note that some tweets have more than 50% of their content as hashtags. This figure is used as the percentile trimming threshold.

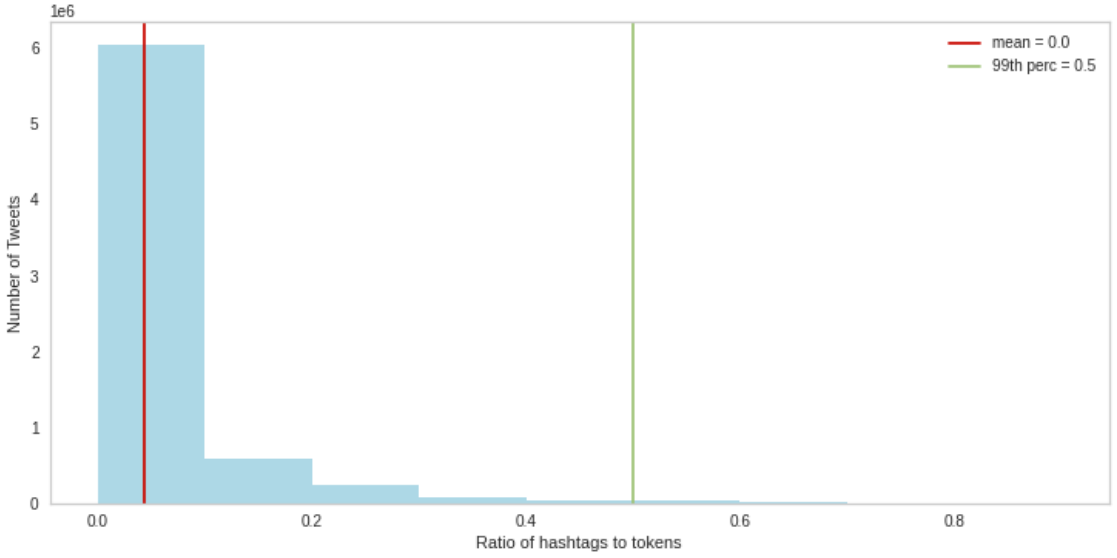


FIGURE 4.14: Percentage of hashtags per tweet

Twitter also uses the mention functionality as a way to tag a particular user in a tweet. Similar to the hijacking of hashtags, malicious users may also use mentions in order to push the relevance of a post [Santos et al., 2014] and to also directly push content to the timeline of a user mentioned in the post [Aggarwal et al., 2012]. Similar to hashtags, the percentage of mentions in a tweet relative to the number of tokens is used. Figure 4.15 depicts the percentage of mentions in a tweet. It is observed that the fraction can be as high as 80% of the tokens in a tweet post, which indicates suspicious behaviour since re-tweets have been removed in our data extraction. Hence, all the tweets that have 40% of mentions in a post are removed.

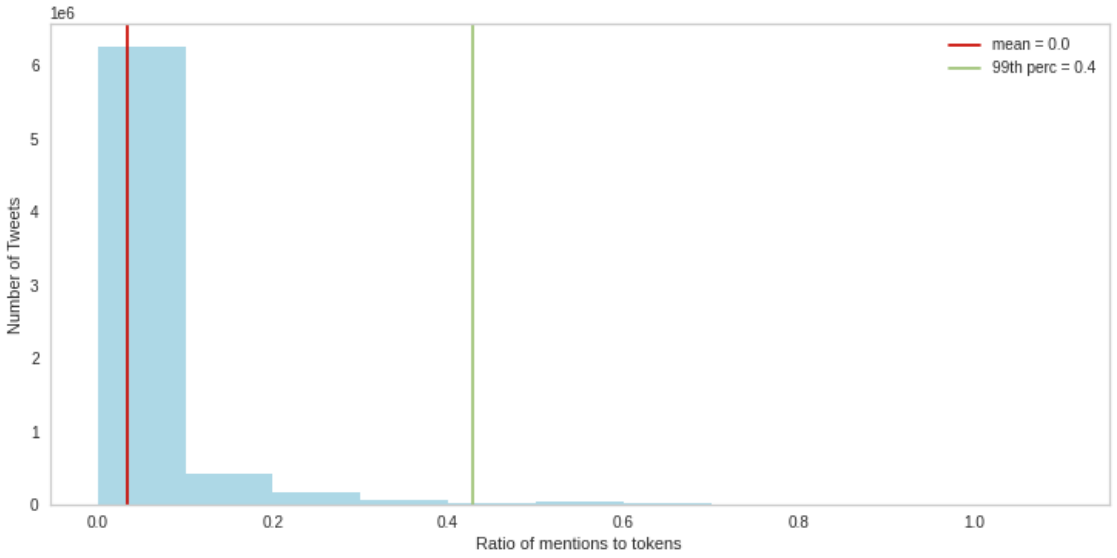


FIGURE 4.15: Percentage of mentions per tweet

In addition to the visibility and increase in reach of posts, malicious Twitter users may include external links in posts in order to redirect users to malicious external websites [Aggarwal et al., 2012]. By using URL obfuscation to shorten URLs, this is also possible. Figure 4.16 depicts the percentage of links in a post. The average number of links is 10% of the number of tokens in a tweet. A threshold for the percentage of links in a tweet is set at 30% of all tweets.

Stopwords are words that are considered to have non-discriminative capability in a classification task, such as prepositions and conjunctions [Silva and Ribeiro, 2003].In addition to prepositions and conjunctions, there are others that are considered non-informative for a classification and hence added to the stopwords list. There are pre-existing stopwords, such as the list used in the gensim and Spacy Python packages [Stone et al., 2011]. These are primarily used in English-language classification tasks.

However, it has been found that stopwords are language and context-dependent [Silva and Ribeiro, 2003; Saif et al., 2014]. Additionally, the pre-compiled list of stopwords from Python modules is incompatible with certain tokenizers [Nothman et al., 2018]. The shortcomings are also exacerbated when dealing with Twitter data that is known to have shorthand texts, misspellings, and differing languages [Saif et al., 2014].Hence, in this work, a stopword list is created by using words that appear only once (TF1 method) [Saif et al., 2014]. Figure 4.17 depicts the rank ordering of each keyword in the Twitter corpus and the cut-off point for the creation

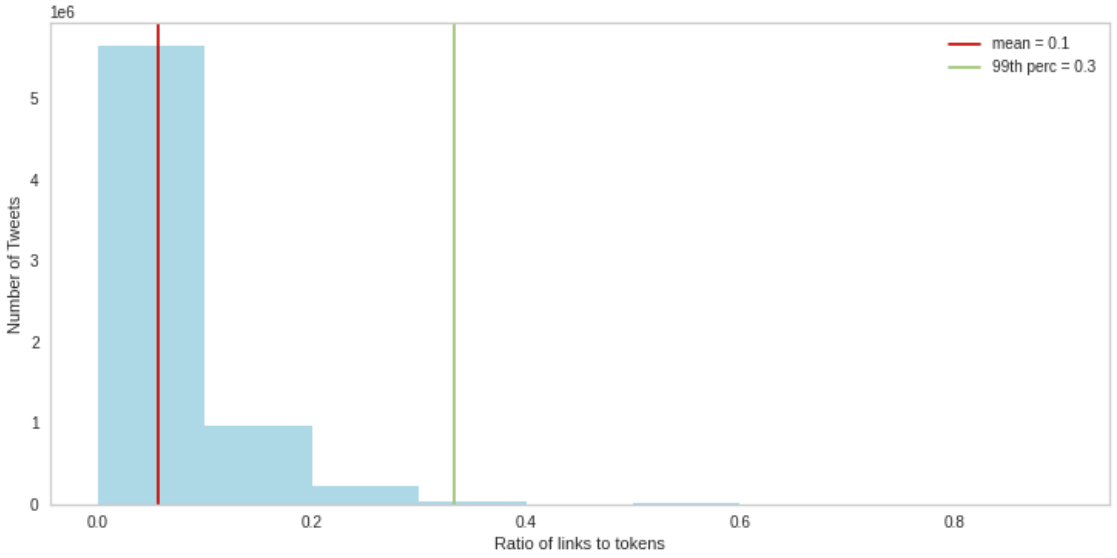


FIGURE 4.16: Percentage of links per tweet

of the stopwords, i.e., low-rank keywords.The stopwords list contains a total of 209 712 keywords, which contain many misspelt and irregular terms such as “unsocilited”, “trafik” and “laaaaaaaaannnnnnnddddddd”.

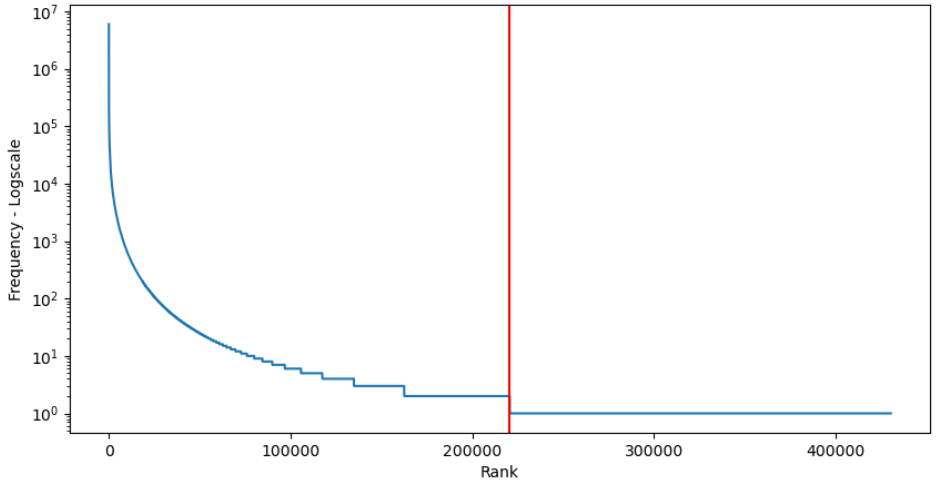


FIGURE 4.17: Rank-Frequency distribution

In addition to the above filtration steps, the following pre-processing is also performed:

- Remove URL’s
- Remove emoji’s
- Remove hashtags
- Remove numbers, and
- Remove reserved words

4.3 Feature Generation

This section covers the feature generation process that is used for modelling purposes. The modelling is separated between the baseline model, logistic regression, and the geometric deep learning methods. The features needed for both methods are different. However, the pre-processing of the data and the process of combining the GDELT data with the Twitter data in order to create the research data remain the same between the methodologies.

The research data is separated into training and testing data. Table 4.1 depicts the distribution of the output class variable. It is observed that there are more protest-related days than non-protest-related days. This distribution of the output variable indicates that the classification problem is an imbalanced classification problem.

Class	**Training data**	**Testing data**
Protest	335	248
Non-protest	123	57
Total	**458**	**305**

TABLE 4.1: Output class distribution of research data

4.3.1 Logistic regression

This section outlines the results of the features that are generated for logistic regression. In order to implement the logistic regression learning procedure, the Twitter text data must be represented numerically. To accomplish this, the TF-IDF technique [Ramos et al., 2003] is used as described in Equation 2.11. As seen in Figure 3.1, a document in this problem space consists of all the extracted tweets from the observation window. The TF-IDF is used because it has the capacity to incorporate weights into a token. There are a total of 199,604 feature tokens for the model to use to learn the output variable.

The frequencies of the classes in the output variable are highly imbalanced, as seen in Table 4.1. The Synthetic Minority Oversampling Technique (SMOTE) technique is used in order to handle the data imbalance. The purpose of using SMOTE is to create synthesised data that improves the learning capacity of the learning algorithm. The SMOTE algorithm creates instances for the minority class, i.e., non-protest, in order to balance the frequencies of the output variable. Hence, there were an additional 212 instances that are created in order to have a total of 335 instances for the non-protest class. Thereafter, the synthesised data serves as an input to the logistic regression model.

4.3.2 Geometric Deep Learning

This section describes the feature generation process for the geometric deep learning methods. The geometric deep learning methods all operate on word relation network data. Word relation network data consists of different parts, such as the adjacency matrix and the node feature representation. This section describes the individual components of the feature generation as depicted in Figure 3.2.

The initial process of the feature generation process is the selection of the keywords that form part of the nodes of the word relation network. As previously mentioned, the relevant keywords are extracted using Equation 3.6. The relevancy score results in keywords that have a higher TF-IDF score on average for each observation window. Figure 4.18 depicts the distribution differences between the keywords that are selected in comparison to all the keywords in the observation windows. There is a clear indication that the selected scores have a higher TF-IDF than all the other keywords.

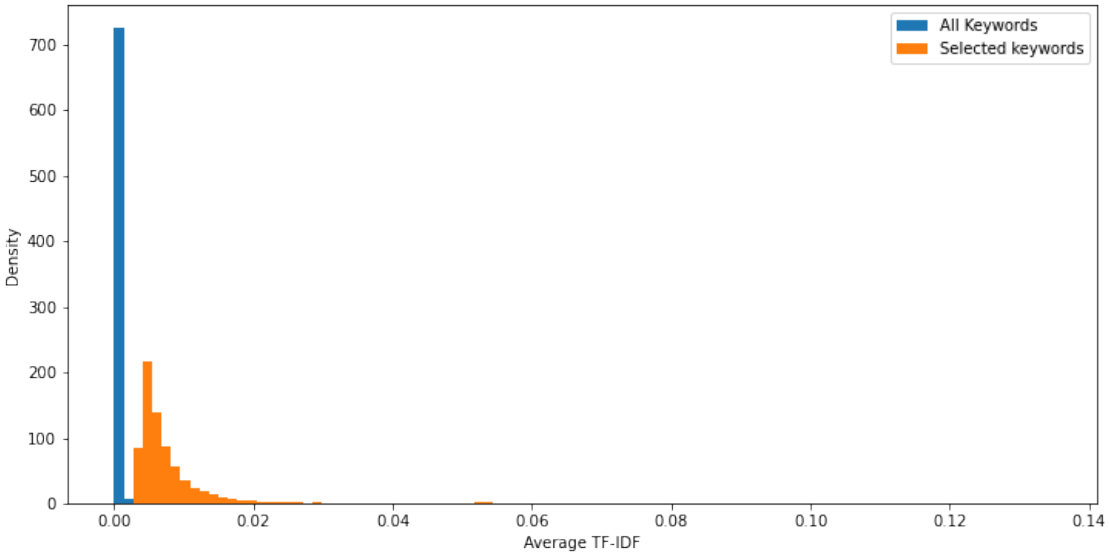


FIGURE 4.18: Frequency of TF-IDF keywords

The second requirement of the word relation network is the adjacency matrix. The adjacency matrix is an indication of the relationship between the keywords and the strength of the relationship. The adjacency matrix is calculated using the smoothed Pointwise mutual information from Equation 2.15. In this work, the value of α is set to 0.75, as per the finding from Levy et al. [2015], that it improves the embedding on a range of tasks. Since this work uses the PPMI, the range of values that we have for the association between keywords is $[0, \infty)$, where a zero (0) value of the PPMI indicates that there is no association between the keywords, and a value greater than zero provides the strength of the association.

The final requirement of the word relation network is the initial feature representation of the nodes, i.e., keywords. The feature representation of the nodes is chosen to be the word2Vec representation of the keyword. Each keyword is represented by a 100-dimensional vector representation. The representation is trained using the training dataset in order to avoid information leakage. A context window of 8 dimensions is used to generate 100-dimensional vectors for each token in a corpus. It has been shown that a larger context window encompasses more topical representation [Levy and Goldberg, 2014]. Figure 4.19 depicts the word embedding representation decomposed to two dimensions using Manifold Approximation and Projection (UMAP) algorithm [McInnes et al., 2018]. In Figure 4.19, there is a a zoomed in portion of the 10 closely associated words to the following tokens: “protest”, “strike” and “unrest”, which are the focus of this study. The visual depicts that in the embedding space, the chosen keywords are similar

to one another, indicating the topical similarity of the chosen keywords. Other notable keywords that are in the cluster are “rioting”, “looting”, “violent”, and “picket”, which are common words associated with protests. Conclusively, the embeddings are able to capture good word representations.

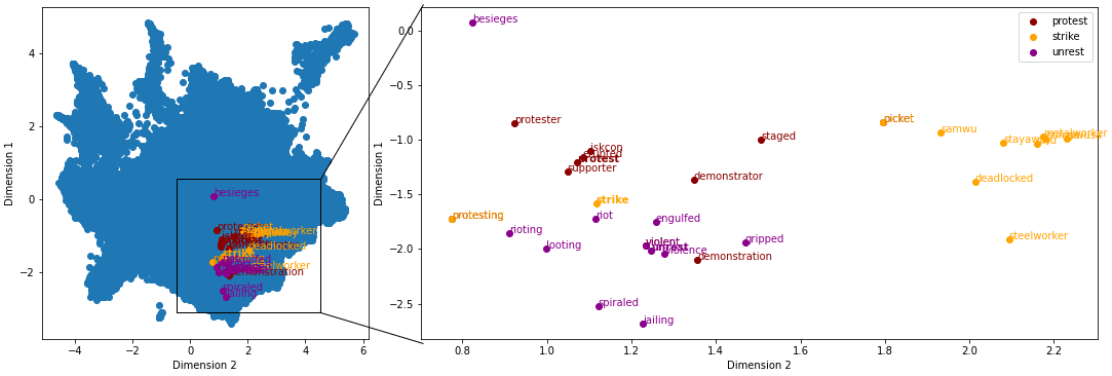
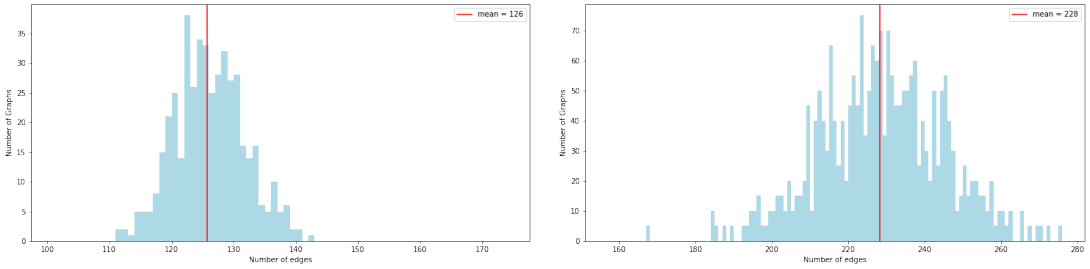


FIGURE 4.19: UMAP representation of the embeddings

With the above components, the word relation network graph can be constructed for the purposes of geometric deep learning methods. The work is separated into static and dynamic network graphs. Using the above process, the static process generates a network graph for the entire observation window, i.e., 5 days. The dynamic process generates a network graph for each day of the observation window, which implies there will be on average 5 graphs per observation window. The static process shares similarities with the logistic regression feature matrix; hence, there are a total of 458 network graphs for the training data and 305 network graphs for the testing data. However,the dynamic process has a total of 2 205 network graphs.



Static Network Graphs Dynamic Network graphs

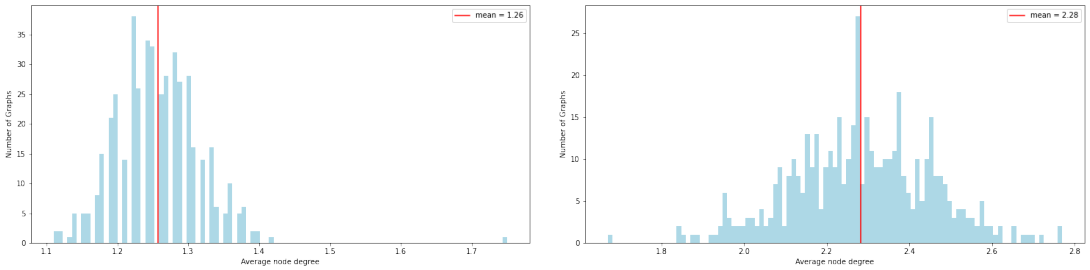
FIGURE 4.20: Distribution of the number of edges

There are metrics that can be extracted from a graph in order to understand its properties. The metrics are derived from the adjacency matrix of the graph. These metrics help in measuring how nodes and edges in the graph interact with one another. The metrics that can be derived are the total number of edges, the average degree of the nodes, the eigenvector centrality, and the clustering coefficient.

The number of edges in a network graph is the number of connections that are in the adjacency matrix. These are the total connections that are present in the network topology. Figure 4.20 depicts the distribution of the number of edges of the static network graph (left) and the dynamic network graph (right). The average number of edges in the static network graph is 126 edges,

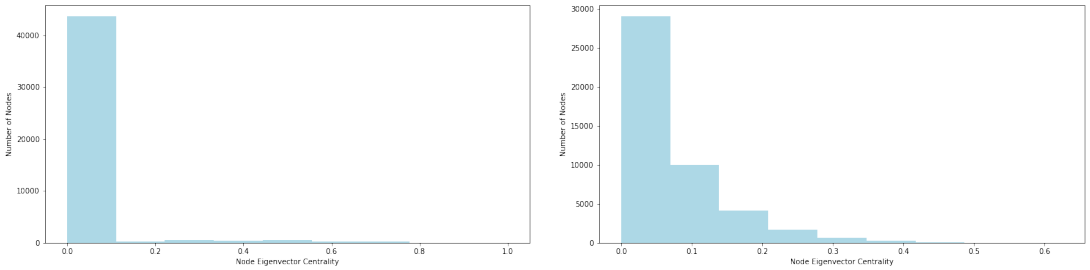
and for the dynamic network graph is 228. However, this total number contains self-connections in the network graphs. Hence, the deduction is that for a network graph of 100 nodes with self-connections, there are on average 26 connecting edges between the nodes for the static network graph and 128 connecting edges for the dynamic network graph. This makes sense since a day would have more co-incident words than a period of 5 days.

Using the number of edges in a network graph, the degree of a network can be derived. The degree of a network graph is defined by the average number of connections from a node to other nodes. This measures the neighbours of a node. This measure is crucial for the current training scheme of the geometric deep learning methods since every node's embedding is based on its neighbourhood. Figure 4.21 demonstrates the average degree of the static graphs (left) and the dynamic graphs (right). Similar to the number of edges, the dynamic graph has nodes with more links (an average of 2.28 connections per node) as compared to the static graph, where on average a node has 1.26 links. The degree is also an indication of sparsity in the adjacency matrix. Hence, from the average degree of the different network graphs, a deduction can be established that the static network graph is more sparse and nodes are more isolated than in its dynamic counterpart.



Static Network Graphs Dynamic Network graphs

FIGURE 4.21: Distribution of the Average Graph degree



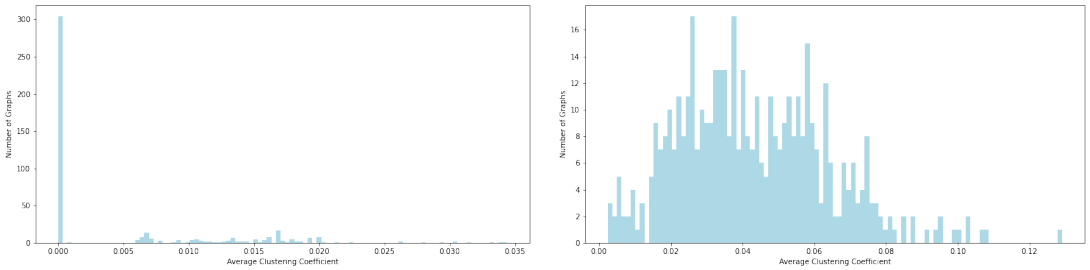
Static Network Graphs Dynamic Network graphs

FIGURE 4.22: Distribution of the Node Eigenvector Centrality

The average degree calculates how many links or connections a node has in the graph network. However, it does not indicate influence in the graph network. In order to understand the influence a node has on the network, the measure of eigenvector centrality can be calculated. The eigenvector centrality measure measures whether a node is connected to other influential nodes in the network. Figure 4.22 depicts the eigenvector centrality of the different nodes in the two graph types. Evidently, there are more nodes in the static network graphs that have no influence in the network topology. Similarly, the same behaviour is seen with the dynamic network graph.

However, the dynamic graph networks have more nodes that have influence in their respective networks. Even though the static network has fewer influential nodes, the eigenvector centrality is as high as 0.99 for the few influential nodes compared to the dynamic network graph, which has a maximum eigenvector centrality of 0.66. This is due to the degree of the nodes, where the static network graph nodes are connected to a few but influential nodes and the dynamic graph nodes are connected to many but less influential nodes.

In addition to the connectivity metrics, there is a clustering coefficient. The clustering coefficient is a measure of the fraction of triangles in a graph network, i.e., nodes that are connected to one another, and hence the density of triangles in a network. In the context of the word relation network, these triangular configurations depict the words that tend to be used together in a document. Figure 4.23 depicts the distribution of the average graph clustering coefficient of the static network graph and the dynamic network graph, respectively. As expected from the metrics above, such as the degree of the network, the dynamic graph network has more transitivity than the static graph network.



Static Network Graphs Dynamic Network graphs

FIGURE 4.23: Distribution of the Average Graph Clustering Coefficient

4.4 Summary

In summary, this section has explored the data that has been used in order to answer the research questions. The section initially evaluates and analyses the events data from GDELT. The events are shown to be correlated with well-publicised media protests in the country. Thereafter, explaining the process of converting the events data into an output variable for the models that are evaluated in this work. It is shown that, due to the conversion process, the output variable is highly imbalanced. Due to the imbalance nature of the problem, there would need to be methods in order to deal with the imbalanced data while using the machine learning methods.

The section also analyses the Twitter data that also forms part of the research data used in this work. The steps that were taken in order to remove noise from the data are explained. Additionally, other standard normalisation techniques are applied to the Twitter data. The steps that are taken result in training data that has 458 instances and testing data that has 305 instances.

Finally, the feature generation process is explained in detail. The feature generation process is separated between logistic regression and geometric deep learning methods. The class imbalance

solution is presented using the SMOTE algorithm, which results in a slight modification of the data distribution of the input data. The generated network graphs that are used for the geometric deep learning methods are analysed in terms of the average degree of the networks, the eigenvector centrality, and the clustering coefficient.

Chapter 5

Results

5.1 Introduction

This section of the study discusses the results of the machine learning models applied to the research data. Initially, the logistic regression model is discussed. The training data for the logistic regression is presented, as is the resulting performance of the model on the hold-out testing data. Thereafter, the geometric deep learning methods are explored, the training process is described, and the performance of the models is presented on the hold-out testing data. Finally, the stability of the models is assessed using cross-validation on the entire research data.

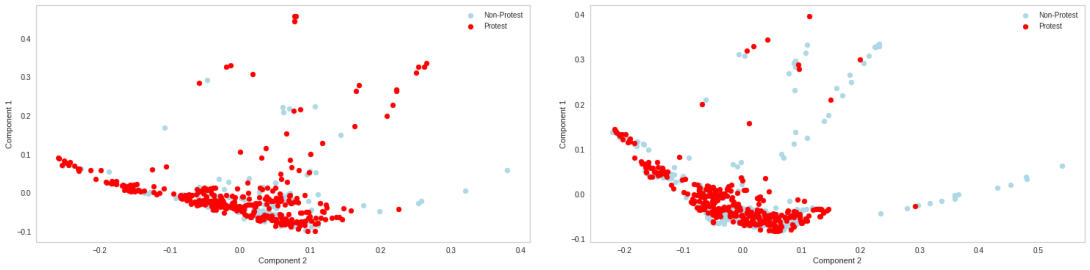
5.2 Logistic Regression

In this section, the logistic regression model results are described in detail. The section will delve deeper into the data that has been used for training the model. Thereafter, the hyper-parameter optimisation process will be studied, and the resulting optimal hyper-parameters will be presented. Finally, the results of the optimal test on the testing data will be presented.

5.2.1 Training data

The model is applied to the TF-IDF feature matrix that has already been described. The feature matrix contains calculated weights using Equation 2.11 for each token that forms part of an instance in the research data. Thereafter, due to the class imbalance of the problem space, SMOTE is applied to the feature matrix in order to generate synthesised data that will be used for the training of the logistic regression model.

Figure 5.1 depicts the Principal Component Analysis (PCA) decomposition of the original feature matrix (left) and the synthesised feature matrix (right). Evidently, in the original feature matrix, we have a substantial number of protest instances that overlap with the non-protest instances. Due to the synthesised minority class data instances that SMOTE creates, it is



Original feature matrix Synthesised feature matrix

FIGURE 5.1: PCA decomposition of the TF-IDF feature matrix

noticeable that there is a distinct separation between the classes in the synthesised feature matrix. However, as it can be observed, the methodology alters the original data distribution that was present in order to make the minority class more distinguishable from the majority class.

In order to use logistic regression on the synthesised data, the data distribution needs to be rescaled. The re-scaling of the data is done by transforming the data to have a zero (0) mean and a standard deviation of one (1). The re-scaling is important in order to assist in the convergence of the model and the interpretation of the resulting feature coefficients. Figure 5.2 depicts the PCA decomposition of the rescaled feature matrix of the synthesised data. It can be clearly observed that the data is zero (0) mean-centred after the re-scaling.

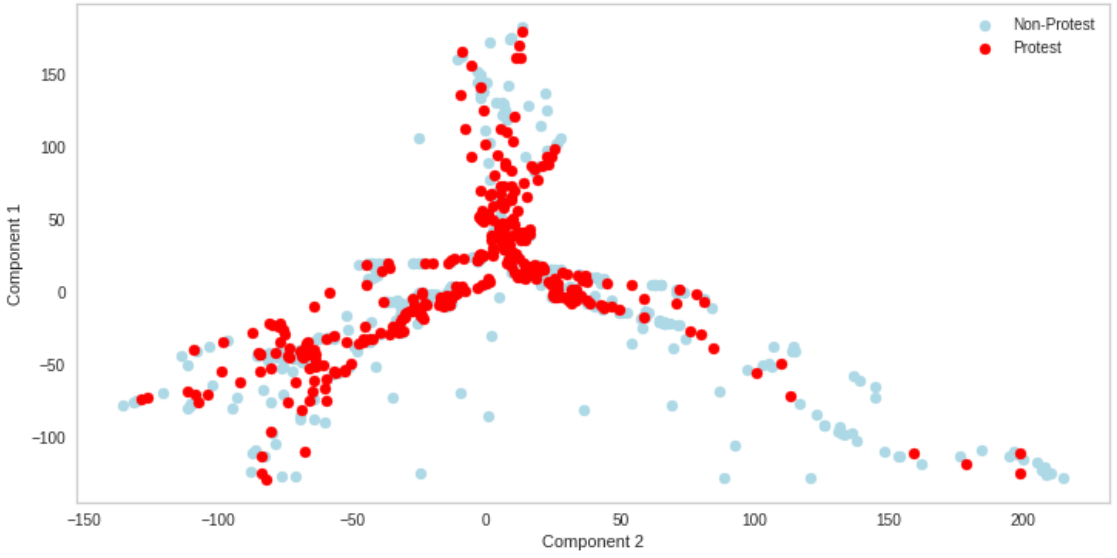


FIGURE 5.2: PCA decomposition of the TF-IDF feature matrix re-scaled

5.2.2 Model Architecture

The data is then used for training the logistic regression model. However, in order to best model the data at hand, there are hyperparameters that need to be optimised for the data in order to obtain high accuracy values. The original training data is used in order to search for the optimal hyperparameters. In order to remain within the same scheme as the training and testing split

for the research problem, the time-aware five-fold cross-validation strategy is used for the hyper-parameter searching. The cross-validation strategy is presented in Figure 5.3, where the testing set is always ahead of the training data in terms of time.

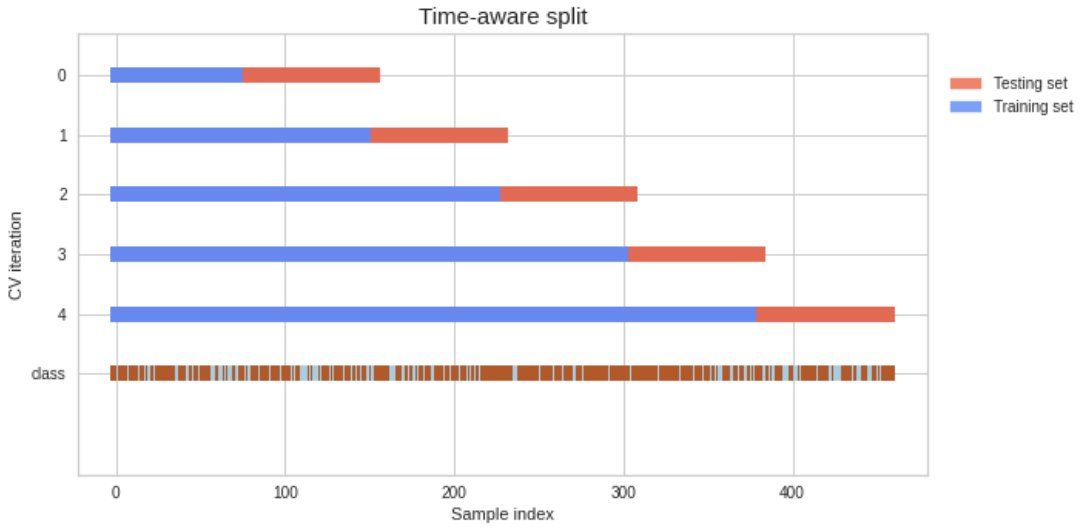


FIGURE 5.3: Time-aware cross-validation split

The hyper-parameter optimisation is done using the Python module Optuna [Akiba et al., 2019]. The hyper-parameters that are optimised are the regularisation weight C, the regularisation penalty, and the logistic regression optimisation solving scheme. In addition to the logistic regression hyper-parameters, dimension reduction is done on the feature matrix since there are way more features than there are observations. The feature dimension reduction is done using PCA, and the number of optimal dimensions forms part of the hyper-parameter optimisation.

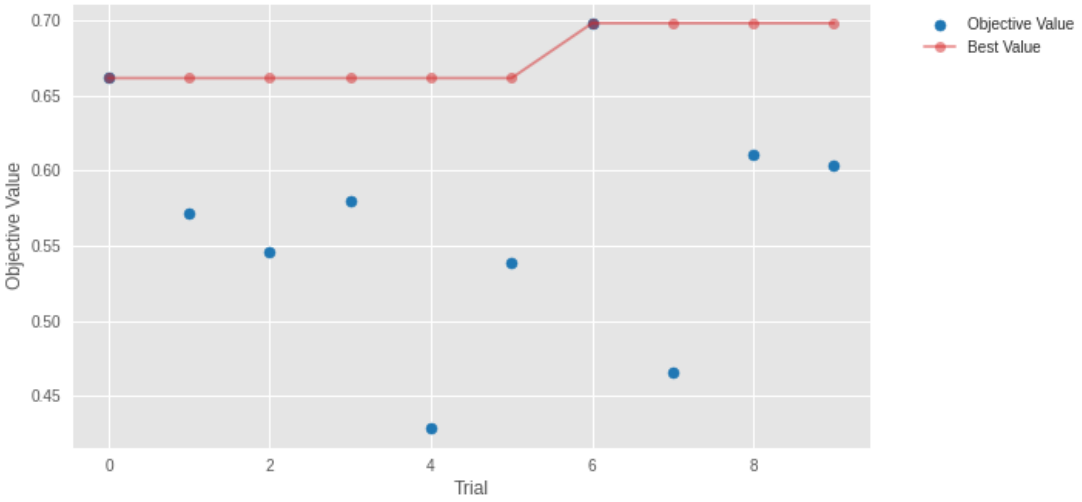


FIGURE 5.4: Optimisation history

The optimisation is done over 8 trials for the model. Figure 5.4 depicts the optimisation history of the algorithm. The optimal value obtained during the search is an average balanced

accuracy of 0.69. The optimal hyper-parameters that are selected from the process are $C = 2.91$, solver = saga, penalty = L1, and feature dimension = 14.

The hyper-parameter importance is presented in Figure 5.5. Due to the existence of redundant features that the dimension reduction eliminates, it is obvious that the number of features chosen for PCA are the most crucial hyper-parameters. Hence, the penalty is of the least importance since the purpose of the penalty would also be to play a feature reduction role in the model.

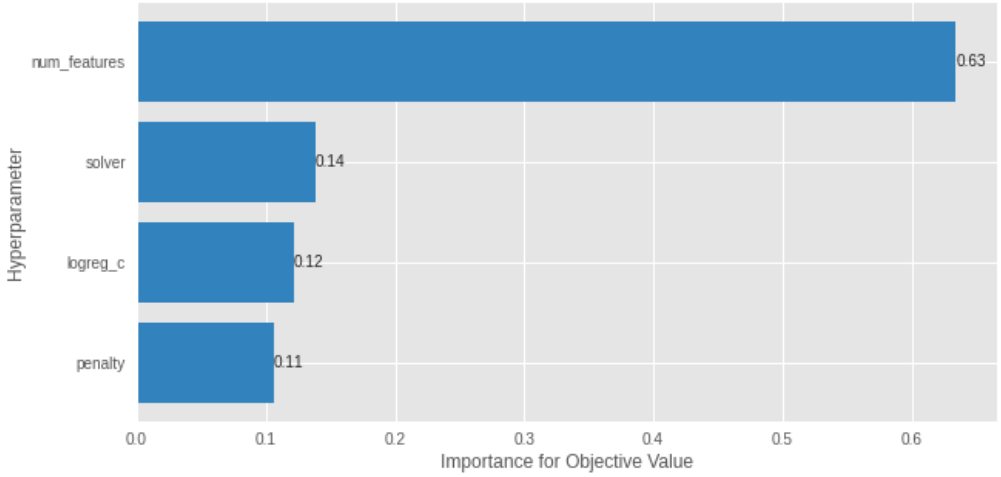


FIGURE 5.5: Hyper-parameter importance

5.2.3 Model Evaluation

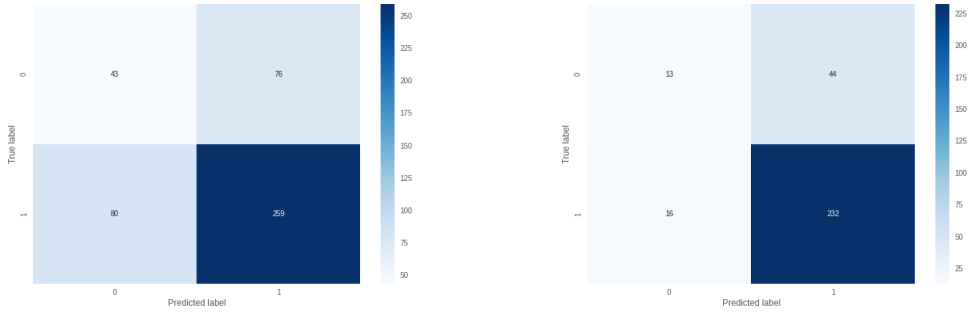
The optimal model that is selected using the hyper-parameter optimisation scheme thereafter gets used for evaluating its performance on the hold-out set. Figure 5.6 depicts the performance of the model on the hold-out set. The balanced accuracy of the holdout testing data is 0.58, as presented in Table 5.1. It can be seen that the model has a higher sensitivity value than a higher specificity value. This implies that the model is better at identifying days that have protest actions than it is at identifying non-protest days. However, note that these values are based on a prediction threshold of 0.5.

	Specificity	**Sensitivity**	**Balanced Accuracy**
Training data	0.34	0.77	0.56
Test data	0.22	0.935	0.58

TABLE 5.1: Classification report for Logistic Regression

5.3 Geometric Deep Learning

In this section, the results of the geometric deep learning method are described in detail. This section will study the training data used for the model and the individual models in detail. Thereafter, the hyper-parameter configuration and model architecture will be detailed and explained.



Training data Testing data

FIGURE 5.6: Confusion matrices of research data

Finally, the results of the chosen model architecture on the testing data will be presented.

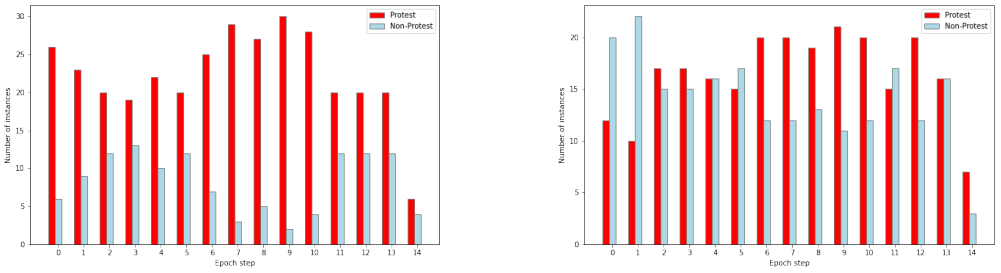
5.3.1 Training data

In this section of the study, the training data is discussed in detail. The purpose of this section is to explain the process of handling the data imbalance during the training process of the geometric deep learning methods.

The geometric deep learning methods are trained on network data. The data contains the adjacency matrix, the weight matrix, and the node feature matrix. Additionally, the output variable also forms part of the data that is used for training the models. As previously discussed, the data that is being modelled is highly imbalanced, where there are more instances of protest-related incidents than there are non-protest-related incidents. In order to update the model's weights, the input is transmitted through the neural network in batches of 32. The difficulty with such a large class imbalance is that the model will not generalise well to the minority class, and the data that flows through the model during the training phase will only contain examples of the majority class.

During the training process, two gradient descent methods are used for the different geometric deep learning models. The DynamicGCN will be using the stochastic gradient descent method, and the GCN and GIN will be using the mini-batch stochastic gradient descent. In order to handle the class imbalance in the context of the GCN and GIN models, the weighted random sampling method is implemented. The purpose of the weighted random sampling method is to increase the weight of the minority class instances, which will therefore result in an increase in the likelihood of selecting an instance from the minority class and also prevent the chance of overfitting. The class weight is chosen to be the inverse of the count of instances in the training data. Hence, protest-related and non-protest instances will have a weight of $2.99e^{-03}$ and $8.13e^{-03}$, respectively.

Figure 5.7 depicts the distribution of the class instance per epoch step before re-weighting (left) and after re-weighting (right). It is observed that after the re-weighting is applied, the



Before re-weighting After re-weighting

FIGURE 5.7: Epoch step data distribution

non-protest instances are sampled at a much higher rate. This re-weighting scheme is only applied to the training data, and the validation and test data still have the original data distribution.

5.3.2 Model Architecture

In this section of the report, the model architectures of the different geometric deep learning methods are presented. In order to use the different geometric deep learning methods on the research data, the most optimal architecture needs to be obtained. In order to obtain the optimal architecture for the model, a manual search was conducted for each of the models. The architecture that presented the highest validation accuracy was chosen as the optimal model.

Table 5.2 depicts the layers of the optimal Graph Convolutional Network. The network contains 5 layers, with an input layer (`conv0`, which takes an input dimension $(100, 100)$ and outputs a shape of $(100, 4)$, followed by a batch normalisation layer. Thereafter, there have 4 layers, `conv1`, of GCN layers, each with dimension $(4, 4)$ and each with an accompanying batch normalisation layer. Finally, the last layer has the dimensions $(4, 1)$. Additionally, there is a dropout layer with a dropout rate of 0.025. The model is trained using using the Adam optimiser with a learning rate of $1.52e-04$ and a weight decay of $1.27e-05$.

Layer	Input Shape	Output Shape	#Param
GCN	[3200, 3200], [3200]	[32, 1]	529
├─(conv1)ModuleList	--	--	80
│ └─(0)GCNConv	[3200, 4], [2, 5747], [5747]	[3200, 4]	20
│ └─(1)GCNConv	[3200, 4], [2, 5747], [5747]	[3200, 4]	20
│ └─(2)GCNConv	[3200, 4], [2, 5747], [5747]	[3200, 4]	20
│ └─(3)GCNConv	[3200, 4], [2, 5747], [5747]	[3200, 4]	20
├─(conv1_bns)ModuleList	--	--	32
│ └─(0)BatchNorm1d	[3200, 4]	[3200, 4]	8
│ └─(1)BatchNorm1d	[3200, 4]	[3200, 4]	8
│ └─(2)BatchNorm1d	[3200, 4]	[3200, 4]	8
│ └─(3)BatchNorm1d	[3200, 4]	[3200, 4]	8
├─(conv0)GCNConv	[3200, 100], [2, 5747], [5747]	[3200, 4]	404
├─(conv0_bn)BatchNorm1d	[3200, 4]	[3200, 4]	8
├─(lin)Linear	[32, 4]	[32, 1]	5

TABLE 5.2: GCN Architecture

Table 5.3 depicts the layers of the optimal Graph Isomorphism Network. The network consists of four (4) layers, with two (2) GIN layers each that have one (1) multi-layer unit of dimensions $(100, 4)$ and two linear layers. In addition to the model architecture, the dropout rate is chosen to be 0.24, the learning rate is $2.96e-03$ and the weight decay is $1.55e-05$.

Layer	Input Shape	Output Shape	#Param
GIN	[30500, 30500], [30500]	[305, 4], [305]	525
├─(pre_mp)Linear	[30500, 100]	[30500, 4]	404
├─(convs)ModuleList	--	--	80
│ └─(0)GINConv	[30500, 4], [2, 38372]	[30500, 4]	40
│ └─(1)GINConv	[30500, 4], [2, 38372]	[30500, 4]	40
├─(bns)ModuleList	--	--	16
│ └─(0)BatchNorm1d	[30500, 4]	[30500, 4]	8
│ └─(1)BatchNorm1d	--	--	8
├─(post_mp)Sequential	[305, 4]	[305, 1]	25
│ └─(0)Linear	[305, 4]	[305, 4]	20
│ └─(1)LeakyReLU	[305, 4]	[305, 4]	--
│ └─(2)Dropout	[305, 4]	[305, 4]	--
│ └─(3)Linear	[305, 4]	[305, 1]	5

TABLE 5.3: GIN Architecture

Finally, the architecture of the optimal architecture for the Dynamic Graph Convolutional Network is presented in Table 5.4. The network contains three (3) convolutional layers with two (2) temporal encoding units, each with a batch normalisation. The optimal model does not use the contextual module. Additionally, the model uses a dropout rate of 0.45 and a learning rate of $1.59e-05$ with a weight decay of $8.068e-04$.

Layer	Input Shape	Output Shape	#Param
DynamicGCN	[1, 100, 100], [1, 2, 119], [1, 119], [1, 100]	[1]	40,901
├─(layer_stack)ModuleList	--	--	20,301
│ └─(0)GCNConv	[100, 100], [2, 119], [119]	[100, 100]	10,100
│ └─(1)GCNConv	[100, 100], [2, 119], [119]	[100, 100]	10,100
│ └─(2)GCNConv	[100, 100], [2, 119], [119]	[100, 1]	101
├─(temporal_cells)ModuleList	--	--	20,200
│ └─(0)TemporalEncoding	[100, 100], [100, 100]	[100, 100]	10,100
│ └─(1)TemporalEncoding	[100, 100], [100, 100]	[100, 100]	10,100
├─(bn_stack)ModuleList	--	--	400
│ └─(0)BatchNorm1d	--	--	200
│ └─(1)BatchNorm1d	--	--	200
├─(contextual_cells)ModuleList	--	--	--
├─(mask)MaskLinear	[100, 1]	[1]	--

TABLE 5.4: DynamicGCN Architecture

The aforementioned architectures and related hyper-parameters for geometric deep learning algorithms are utilised for training on training data. The completed model is then used to evaluate the testing results. The final model findings are reported in the following section.

5.3.3 Model Evaluation

In this section of the report, the results from the geometric deep learning methods are discussed. The optimal model architectures are used on the hold-out set in order to evaluate the balanced accuracy of the models. Initially, the GCN is discussed, and thereafter, the results of the DynamicGCN are discussed, and finally, the results of the GIN are evaluated.

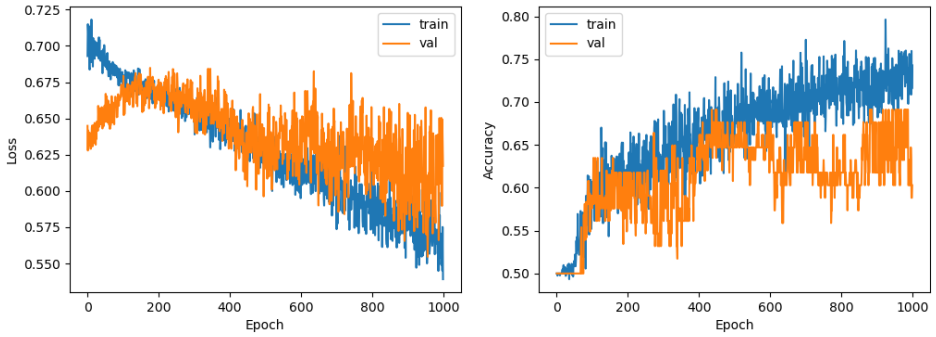
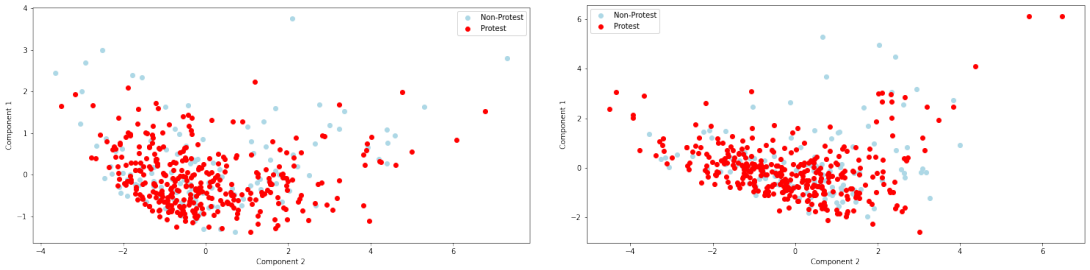


FIGURE 5.8: GCN learning curves

Figure 5.8 depicts the learning curves of the GCN model. It can be observed that the accuracy for the training data reaches 0.743 and the validation accuracy reaches an average of 0.603. However, the accuracy of the training data is based on the re-weighting scheme applied during the model training process.

The geometric deep learning methods are meant to learn a graph embedding on the input training data such that the same embeddings can be used in the hold-out set. The models are all initialised with random weights, and the process of backpropagation allows the models to learn proper embeddings for the input data. In Figure 5.9, there is a depiction of the graph representation embeddings pre-training (left) and and the learned embeddings post-training (right), and this is not of the balanced training version. As it can be seen, the separation between protest and non-protest instances is overlapping the majority of the time, and this is due to the high number of instances that are present in the original data distribution.



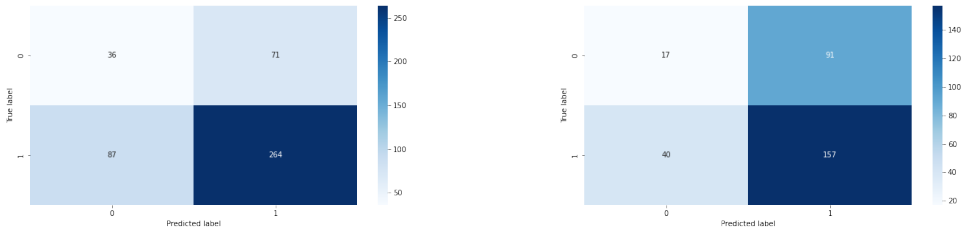
Pretraining Post-training

FIGURE 5.9: PCA representation of the GCN graph embeddings

The trained model is then used on the unseen testing data. Figure 5.10 depicts the results of the GCN model on the testing data. The overall accuracy score on the testing data is 0.46, with a specificity value of 0.28 and a sensitivity value of 0.63, as seen in Table 5.5. A random day is more likely to be predicted as a protest-related day than a non-protest day by the algorithm. However, the protest-related days could potentially be misidentified.

	Specificity	**Sensitivity**	**Balanced Accuracy**
Training data	0.29	0.78	0.54
Test data	0.28	0.63	0.46

TABLE 5.5: Classification report for GCN



Training data Testing data

FIGURE 5.10: GCN confusion matrices

Similarly, the GIN architecture is used to train on the training data of 458 instances. In order to prevent overfitting during the training process, the same re-weighting scheme is applied during the training process. Figure 5.11 depicts the learning curves of the model during training. It can be observed that the model starts to overfit the training data at about epoch 150, where the training loss starts to increase while the validation is increasing. This similar behaviour is seen with the accuracy of the model, where the accuracy of the training data increases, whereas the validation data starts to decrease. Due to the overfitting behaviour of the model, an early stopping mechanism with a patience of 50 epochs is implemented, and the training of the model stops at 200 epochs.

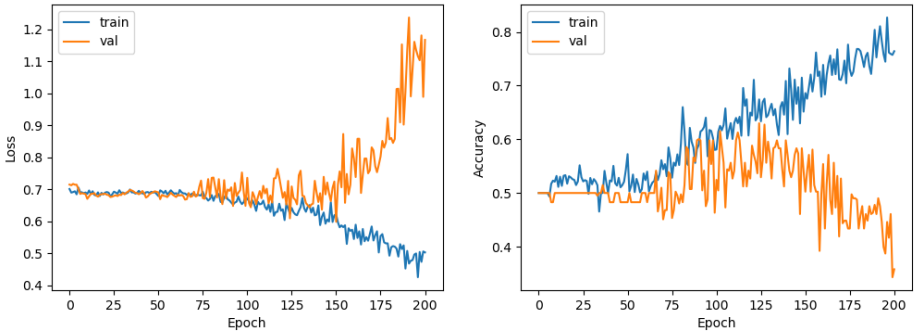


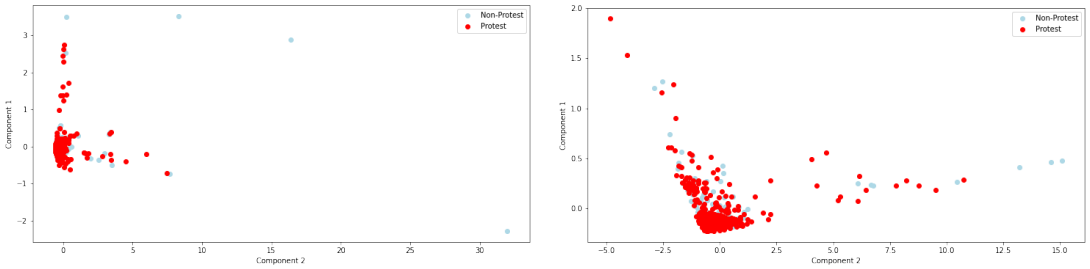
FIGURE 5.11: GIN learning curves

In order to understand the learned representation for the original training graphs, Figure 5.12 depicts the graph embeddings of the training data pre-train (left) and post-training (right). The overlapping mechanism still presents itself in the learned embeddings. Hence, there was no distinct separation between the protest-related instances and the non-protest instances.

In Table 5.6 and Figure 5.13, the accuracy scores for the training data and the testing data show that the two classes can't be separated in the graph embedding space. Both datasets exhibit a high sensitivity but a low specificity. This means that the model is more likely to be able to forecast protests than non-protests.

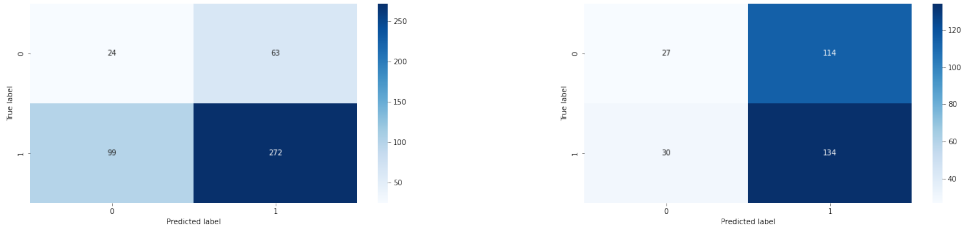
	Specificity	**Sensitivity**	**Balanced Accuracy**
Training data	0.19	0.81	0.50
Test data	0.45	0.60	0.53

TABLE 5.6: Classification report for GIN



Pretraining Post-training

FIGURE 5.12: PCA representation of the GIN graph embeddings



Training data Testing data

FIGURE 5.13: GIN confusion matrices

Finally, the DynamicGCN is applied to the training data in order to optimise the model parameters. The learning curves in Figure 5.14 are a representation of the DynamicGCN model's training process. Similarly to what happens with the GIN model, the DynamicGCN model starts overfitting the training data early in the training process. It can also be seen that the accuracy increases for the training portion of the data but remains on average the same for the validation data.

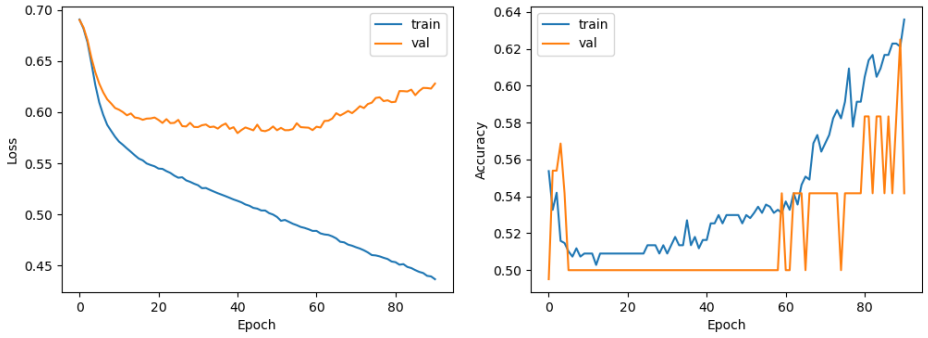
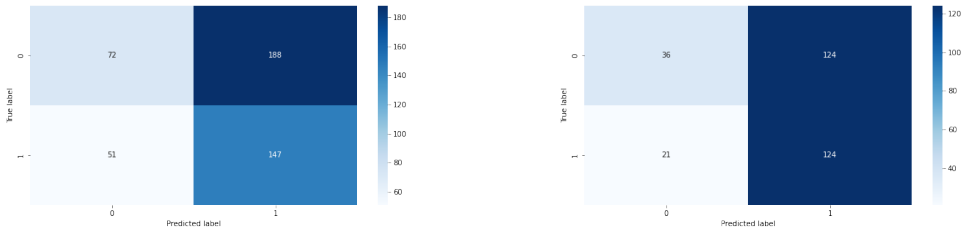


FIGURE 5.14: DynamicGCN learning curves

The DynamicGCN is thereafter evaluated on the entire training data and the hold-out testing data. Table 5.7 and Figure 5.15 depict the results of the model on the training data and the testing data. In comparison to the other geometric deep learning methods investigated, the DynamicGCN model has on average higher specificity values and lower sensitivity values. This implies that the model has a higher rate of detecting instances that are non-protest-related than instances that are protest-related.

	Specificity	**Sensitivity**	**Balanced Accuracy**
Training data	0.58	0.44	0.512
Test data	0.63	0.5	0.57

TABLE 5.7: Classification report for DynamicGCN



Training data Testing data

FIGURE 5.15: DynamicGCN confusion matrices

5.4 Model Stability

This section of the report assesses the stability of the different models presented above. The results reported have been reported for one data instance and may be noisy. Hence, in order to understand how the models would perform with different variations and obtain an estimate of the true mean average performance of the model, it needs to be measured. The way in which model stability is assessed is through cross-validation. The method of cross-validation that is used is the same as that presented in Figure 5.3.

Model	**Accuracy**	**Sensitivity**	**Specificity**
Logistic Regression	0.499 ± 0.042	0.762 ± 0.08	$\mathbf{0.238 \pm 0.022}$
GCN	0.491 ± 0.018	0.973 ± 0.03	0.009 ± 0.020
GIN	$\mathbf{0.516 \pm 0.019}$	0.915 ± 0.085	0.116 ± 0.100
DynamicGCN	0.506 ± 0.020	$\mathbf{0.990 \pm 0.007}$	0.023 ± 0.041

TABLE 5.8: Cross-validation performance on research data

Table 5.8 presents the cross-validation results of the models investigated. It can be observed that all the geometric deep learning methods have a higher average sensitivity value than average specificity value. Additionally, the GIN model has the highest average specificity value of all the geometric deep learning methods, resulting in the GIN model having the highest average accuracy. This behaviour is seen with the hold-out test data in Table 5.6. By design, the GIN model is more powerful at distinguishing isomorphic graphs than the GCN models and hence performs better than all the GCN models. However, the logistic regression model has the highest average specificity of all the models but the lowest average sensitivity value. Hence, the logistic regression model does not have a good accuracy score.

5.5 Summary

This chapter presents the results of the different research methods applied to the research data presented in this work. The chapter explores the optimal model architectures and the results

of the hold-out test data. The logistic regression performs better on the hold-out testing data. However, the logistic regression model is not stable and does not perform as well using cross-validation. The logistic regression model in general has a higher specificity value than a higher sensitivity value.

The geometric deep learning methods struggle to perform equally well as logistic regression on the hold-out testing data due to overfitting that results from the limited training examples. However, the models are more stable and have a better cross-validation score on average. The GIN model presents higher accuracy on hold-out testing data and also during cross-validation. The isomorphism detection capability of the model makes it suited for the task presented, and it can be observed that the GIN model has a higher specificity value amongst all the geometric deep learning methods.

Chapter 6

Discussion

6.1 Introduction

This section of the study discusses the interpretation of the results obtained from the machine learning models used in this study. The purpose of this work is to investigate the effectiveness of using noise reduction techniques on Twitter data in order to predict protests in South Africa when applying graph neural networks. There are two (2) classes of graph neural network methods that are compared, i.e., convolutional graph neural networks and isomorphic graph neural networks. In order to create a baseline model, a TF-IDF logistic regression model is used.

The findings indicate that the geometric deep learning approaches have challenges with overfitting when applied to hold-out testing data. However, these methods demonstrate stability and yield superior cross-validation scores. The baseline TF-IDF logistic regression method performs better for the hold-out set. However, TF-IDF logistic regression struggles with stability and yields suboptimal performance on cross-validation. The Graph Isomorphism Network (GIN) model demonstrates superior accuracy over the Graph Convolutional Network (GCN) models due to its isomorphism recognition capabilities, rendering it well-suited for the given task. Graph neural networks face challenges when dealing with insufficient data that impacts a lot of deep learning methods, leading to overfitting of the training data. Additionally, they encounter difficulties in handling isomorphism and isolated nodes due to the message-passing paradigm.

The remainder of this section focuses on the discussion of the interpretation of the results. This section dedicated to interpretation will additionally examine the impact of the findings on the research questions of this study. Furthermore, this discussion will address potential areas for future recommendations, specifically focusing on two distinct avenues: data processing and modelling. These areas will be identified as requiring further investigation in order to enhance the predictive performance of the model. Finally, this section concludes with a discussion on final remarks.

6.2 Results Interpretation

In this particular section of the study, an in-depth analysis and explanation of the obtained findings are presented and examined. This study demonstrates the effectiveness of the TF-IDF logistic regression approach in achieving favourable results on the hold-out test dataset. The number of features is a crucial factor in hyper-parameter optimisation, mostly influenced by the number of samples present in the study [Maalouf, 2011].

The decrease in the quantity of features benefits the model in mitigating the potential for overfitting. Nevertheless, it should be acknowledged that the model exhibits suboptimal performance in the cross-validation scenario. Due to the dynamic nature of social media and the varying keywords associated with protest-related events, the static model is inadequate in capturing the evolving signals inherent in social media data.

The circumvention of this issue is achieved by the dynamic behaviour demonstrated by geometric deep learning models. The static GCN and GIN models employ distinct sets of relevant keywords for each sliding window. Conversely, the Dynamic Graph Convolutional Network (DynamicGCN) selects various relevant keywords on a daily basis. The selection of appropriate modelling techniques is crucial due to the dynamic nature of keywords utilised on social media platforms. Nevertheless, opting for such a model selection results in graphs that exhibit structural similarities while being distinct in terms of their context, sometimes referred to as isomorphic. Therefore, the GIN model demonstrates proficiency in managing the isomorphic characteristics of research data, but the GCN models encounter difficulties in addressing this behaviour [Xu et al., 2019].

This work has investigated and addressed the literature limitation and gaps. The work has used geometric deep learning methods in order to predict future protests. The outcomes have not yielded highly favourable findings; nevertheless, the research has effectively showcased the capabilities and possibilities of employing graph networks for the purpose of forecasting forthcoming protests. This is due to the complex relationships that are captured by the word relational networks. Additionally, the work presented has shown the importance of noise reduction and dynamic stopwords on social media data. This is demonstrated by the performance of the all the models. This study also examines the prediction task for a society characterised by multilingualism. This study highlights the significance of carefully selecting pertinent keywords to incorporate them into the word-relation network.

The research has limitation with regards to the data. The research was conducted with a restricted dataset due to the inherent limitations of data collection through the Twitter API, which imposes a monthly constraint of 10 million tweets. However, with the amount of traffic in social media and daily tweets it becomes unfeasible to amass the entirety of the data necessary for the research endeavour. Similar work has data spanning over a 5 year period between 2014 and 2019 [Chinta et al., 2021]. However, it has been demonstrated with cross-validation that the results in this work are valid.

6.3 Recommendations

The work in this study has resulted in a few recommendations for improvements to the predictive capability of this problem. The recommendations are separated between data processing research and modelling aspects.

6.3.1 Data Processing

This section of the study examines the data processing processes that could potentially improve the performance of the graph neural network algorithms. The study utilises social media data to identify indicators of upcoming protest-related actions. Nevertheless, the conversations that occur on social media platforms encompass a wide range of issues that evolve over time and lack specificity to any single discussion. This study use a relevancy score to extract significant keywords, which are further utilised to construct word relation networks. The utilisation of the defined relevancy score encompasses the entire corpus, perhaps yielding terms that hold significance across Twitter conversations but lack specificity to protests.

The work by Deng et al. uses news articles in order to predict the occurrence of social unrest. In their work, the authors mention that the keyword selection is done using the TF-IDF; however, the process is not explicitly demonstrated. Similarly, the work by Wang et al. uses Twitter data that is selected based on its relation to the event that is being forecast. This does not work well in the real-world context since the event has not occurred. Therefore, as suggested by Chinta et al., there is a need for a data filtration step in order to select relevant documents that are related to the study. The work uses a pre-trained prediction model that predicts whether a tweet is related to a protest or not [Chinta et al., 2021]. However, the model that is used only works for English tweets, so a multi-lingual model could be beneficial to the work of event prediction using online text data. Evidence of a pre-selection step of documents is seen in the performance of the logistic regression model that uses a bag-of-words method containing all the documents.

In addition to the selection of relevant documents, there is also a requirement to select the keywords that are to be used in the word-relation network. In this work, a relevancy score that is based on the TF-IDF has been used. Nevertheless, other techniques have been investigated to facilitate the automatic extraction of keywords within a given text corpus. Examples of methodologies encompass the robust Twitter Keyword Graph technique, which is employed to extract keywords from tweets by employing graph centrality measures [Abilhoa and De Castro, 2014]. Another method is the LexRank stochastic graph-based approach, which is insensitive to data noise and utilises eigenvector centrality and intra-sentence cosine similarity to extract significant keywords [Erkan and Radev, 2004]. These methods can be used for the type of text data that is used in this study, as they exhibit robustness and insensitivity to noise.

6.3.2 Modelling

This section of the study examines modelling processes that could potentially improve the performance of the graph neural networks. The work makes use of graph neural network methods in

order to predict upcoming protest-related incidents in South Africa. In this work two (2) classes graph neural network methods are compared; the static and dynamic graph neural network. For the static graph neural network, the GCN and GIN models are used; and for the dynamic graph neural network, the DynamicGCN is used.

The phenomenon under consideration is a widely researched occurrence that arises as a result of the message-passing paradigm. This paradigm can be considered as a specific instance of Laplacian smoothing, which in turn leads to the issue of over-smoothing [Li et al., 2018]. Over-smoothing is when the representation of the graph nodes from different classes becomes indistinguishable due to the increase in model complexity. This has a detrimental impact on the performance of the model, particularly when there is is limited labelled data [Li et al., 2018]. A phenomena that can be observed in Figures 5.9 and 5.12. Proposed modelling solutions to overcome over-smoothing include a self-training with a GCN network [Li et al., 2018], another strategy involves incorporating a regularisation term into the training loss based on the Mean Average Distance (MAD) [Chen et al., 2020].

In addition to overcoming the over-smoothing problem inherent to graph neural networks [Chen et al., 2020] other modelling choices can be made in order to potentially improve the performance of the classification task. Such choices include pre-training approaches that provide solutions to limited data environments and out-of-distribution issues arising from real-world graphs [Hu et al., 2020]. Methods of pre-training have been shown to improve the predictive performance of graph neural networks on downstream tasks [Lu et al., 2021; Veličković et al., 2018]

6.4 Conclusion

This section of the study is a discussion of the results that were obtained from the model. This study presents an analysis of the TF-IDF logistic regression approach for predicting future protests. The model's effectiveness is attributed to the decrease in the feature space, which helps mitigate overfitting. However, the static model struggles with cross-validation due to the dynamic nature of social media and the varying keywords associated with protest-related events. Geometric deep learning models, such as the DynamicGCN, overcome this issue. The study also highlights the importance of noise reduction and dynamic stopwords in social media data. The study also examines the prediction task for multilingual societies and the significance of selecting relevant keywords. However, the study has limitations due to data limitations from the Twitter API, which imposes a monthly constraint of 10 million tweets.

Additionally, two avenues of potential recommendations are provided, i.e., data processing and modelling. Data processing techniques include data filtration steps to select relevant documents and using graph based methods to select relevant keywords. The study also discusses the use of different graph neural network models, such as static and dynamic graph neural network models, and the issue of over-smoothing in these models. Proposed solutions to overcome over-smoothing include self-training with a GCN network and incorporating a regularisation term based on MAD. Additionally, pre-training approaches are suggested to improve the performance of graph neural networks on downstream tasks.

Chapter 7

Conclusion

7.1 Introduction

This section of the study concludes the work by summarising all the work that has been done in the research. This section discusses the main research findings and the possible future direction of the work presented in this study.

The purpose of this study was to investigate the effectiveness of using graph neural networks on Twitter data in order to predict protest in the South African context. The work initially starts by explaining the social impact of protests in South Africa, the rise of protests in the country since the post-apartheid era, and the impact that protests have had, limited to loss of lives and damage to properties. Thereafter, the current literature that exists around protest is analysed, and the different theories that attempt to explain the use of social media as a mobilisation and communication tool and hence as a very effective tool in organising a protest are discussed. In this work, the problem of event prediction in the context of protest is explored, and the current methods that have been used to tackle the task are described. Finally, the current limitations that exist with the current methods are outlined, and hence the purpose of this work is to cover some of the existing research gaps, more specifically the lack of research on the use of multi-lingual Twitter data in predicting the occurrence of a protest.

In this work, the research data was used in order to answer the research question. The data from GDELT and Twitter ranging between the years 2019 and 2021 is described. The final data used in the modelling process is created using the sliding window methodology. Thereafter, this data is explored using exploratory analysis methods, where it is shown that the data is highly imbalanced, with more instances of protest than non-protest-related instances. Finally, the noise reduction of Twitter data is performed using different metrics derived from the tweets, such as the number of hashtags in a post, the percentage of links in a tweet, and the percentage of mentions. Additionally, the TF1 method is used in order to improve the signal in the data by removing keywords that appear only once in the corpus. This is shown to be important for this work since keyword selection is done using the TF-IDF score.

The final clean data is thereafter used to create features for the models that were investigated. The logistic regression method uses the TF-IDF feature matrix, and the geometric deep learning methods use the keywords with the highest average TF-IDF score to create word relation networks. It is shown that word relation networks have on average two (2) neighbouring nodes, and there are many isolated nodes in the network.

The model-building process is presented for all the models under investigation. It is revealed that the logistic regression method has the highest balanced accuracy score. The most important hyperparameter of the model is the number of features, and this is due to the number of training examples that were in the research data. Hence, a smaller set of tokens was necessary to improve the model's performance. However, the logistic regression model presents a less stable model because of the dynamic keywords that are present in Twitter data. Hence, it may not be possible to model the dynamics of protests with static keywords.

It is shown that the geometric deep learning methods have a tendency to overfit the training data and hence perform poorly on the hold-out testing data. The overfitting of these models is primarily due to the limited dataset that was available to train them. However, the geometric deep learning methods are more stable and present a higher sensitivity score. The stability in these models is due to the complexity captured in the graph embedding, where not only the individual keywords are preserved but their relationships are also stored. Which implies that the models are more likely to predict protest-related instances. However, the study may not be able to predict non-protest instances.

7.2 Summary of Research Findings

The purpose of this study was to investigate whether there are signals in Twitter data that may be indicative of an upcoming protest. These signals were to be detected using machine learning methods. We have found in this work that noise-reduction techniques using derivative metrics from social media text can be used to extract signals from South African Twitter data. Additionally in this study we were able to show that graph neural networks are capable of handling a multilingual keywords for event prediction. However, it is noted that the further improvement in the models performance is needed due to the dynamic nature of Twitter data. This work has the following answers to the research questions:

- How can Twitter data be processed in order to reveal a signal to that will improve the predictive capability of machine learning models?
  - The text that users are using contains the original noise in Twitter data. In order to reveal signals in Twitter data, derived metrics of the tweets can be used, such as the number of hashtags in a post, the percentage of links in a tweet, and the percentage of mentions. In addition to the derived metrics, since Twitter users tend to use shorthand text, the removal of infrequent words can also improve machine learning models and reduce the sparsity of the modelling data.
- How are word relation networks derived from Twitter data?

  - Word relation networks can be derived from a cleaner version of the Twitter data. Whereby all the noisy tweets have been removed. The use of the TF-IDF score per word can also provide reasonable results in the modelling process.

- Can graph neural networks on Twitters' word relation networks be used to determine the occurrence of a future protest?

  - It has been discovered that graph neural networks struggle to work with limited data and tend to overfit the training data. These models are capable of capturing the signal. However, there is a limit to the capability of the models in a limited data space. Additionally, the models struggle with isomorphism and a high number of isolated nodes due to the mechanism of message passing inherent in the nature of the networks.

7.3 Possible Future Work

The work in this study has resulted in a few directions for future research on this problem. The directions are separated between data processing research and model research. The research directions are described in this section.

7.3.1 Data processing

In this work, it has been seen that the dynamic and stochastic nature of Twitter interactions and conversations makes the problem more complex. However, there are other data features that were purposefully excluded that may add more information about the complexity of the data space. Such information includes data related to users and their interactions on Twitter. This would result in heterogeneous networks with more than one type of edge and node entity. Additionally, the keyword selection in this work was done independently from the keywords influence in the larger network. In order to make an informed decision about keyword selection, the node eigenvector centrality may be used to select influential keywords in the network.

7.3.2 Modelling

This work primarily used the graph neural network paradigm of message passing. However, it has been noticed that message passing has limited capability in a context where the network has many isolated nodes. Further research needs to be considered beyond the message-passing paradigm. Additionally, it was observed that isomorphism was vital to the performance of the network. However, the GIN network is a static network and also does not consider edge weight in the modelling process. Hence, further model development needs to be investigated for the GIN network to allow it to have dynamic capability and also have edge weights as inputs.

Bibliography

[Abilhoa and De Castro, 2014] Willyan D Abilhoa and Leandro N De Castro. A keyword extraction method from twitter messages represented as graphs. *Applied Mathematics and Computation*, 240:308–325, 2014.

[Aday et al., 2012] Sean Aday, Henry Farrell, Marc Lynch, John Sides, and Deen Freelon. New media and conflict after the arab spring. *United States Institute of Peace*, 80:1–24, 2012.

[Africa et al., 2021] Sandy Africa, Silumko Sokupa, and Mojankunyane Gumbi. Report of the expert panel into the july 2021 civil unrest. *The Presidency Republic of South Africa*, 2021.

[Aggarwal et al., 2012] Anupama Aggarwal, Ashwin Rajadesingan, and Ponnurangam Kumaraguru. Phishari: Automatic realtime phishing detection on twitter. In *2012 eCrime Researchers Summit*, pages 1–12. IEEE, 2012.

[Akiba et al., 2019] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

[Al Zamal et al., 2012] Faiyaz Al Zamal, Wendy Liu, and Derek Ruths. Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 6, pages 387–390, 2012.

[Aldhaheri and Lee, 2017] Abdulrahman Aldhaheri and Jeongkyu Lee. Event detection on large social media using temporal analysis. In *2017 IEEE 7th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 1–6. IEEE, 2017.

[Alexander et al., 2018] Peter Alexander, Carin Runciman, Trevor Ngwane, Boikanyo Moloto, Kgothatso Mokgele, and Nicole Van Staden. Frequency and turmoil: South africa's community protests 2005–2017. *South African Crime Quarterly*, 63:27–42, 2018. ISBN: 1991-3877.

[Amodu et al., 2014] Lanre O. Amodu, Suleimanu Usaini, and Oyinkansola Ige. The media as fourth estate of the realm. *ResearchGate*, 2014.

[Arias et al., 2014] Marta Arias, Argimiro Arratia, and Ramon Xuriguera. Forecasting with twitter data. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(1):1–24, 2014.

[Asur and Huberman, 2010] Sitaram Asur and Bernardo A Huberman. Predicting the future with social media. In *2010 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology*, volume 1, pages 492–499. IEEE, 2010.

[Atz et al., 2021] Kenneth Atz, Francesca Grisoni, and Gisbert Schneider. Geometric deep learning on molecular representations. *Nature Machine Intelligence*, 3(12):1023–1032, 2021.

[Barbera and Jackson, 2019] Salvador Barbera and Matthew O Jackson. A model of protests, revolution, and information. *Revolution, and Information (October 2019)*, 2019.

[Barbier and Liu, 2011] Geoffrey Barbier and Huan Liu. Data mining in social media. In *Social network data analytics*, pages 327–352. Springer, 2011.

[Beardmore, 2020] Adele Beardmore. Uncovering the environmental impact of cloud computing. https://earth.org/environmental-impact-of-cloud-computing/, 2020.

[Bedasso and Obikili, 2016] Biniam E Bedasso and Nonso Obikili. A dream deferred: The microfoundations of direct political action in pre-and post-democratisation south africa. *The Journal of Development Studies*, 52(1):130–146, 2016.

[Bird and Loper, 2004] Steven Bird and Edward Loper. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/P04-3031.

[Bollen et al., 2011] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of computational science*, 2(1):1–8, 2011.

[Bonga, 2021] Wellington Garikai Bonga. Impact of repetitive protests on economic development: A case of south africa. *Quest Journals' Journal of Research in Humanities and Social Science*, 9(8):34–39, 2021.

[Bosch, 2017] Tanja Bosch. Twitter activism and youth in south africa: The case of# RhodesMustFall. *Information, communication & society*, 20(2):221–232, 2017. Number: 2 ISBN: 1369-118X Publisher: Taylor & Francis.

[Brodersen et al., 2010] Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M. Buhmann. The balanced accuracy and its posterior distribution. In *2010 20th International Conference on Pattern Recognition*, pages 3121–3124, 2010. doi: 10.1109/ICPR.2010.764.

[Bronstein et al., 2017] Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017.

[Bunte and Vinson, 2016] Jonas B Bunte and Laura Thaut Vinson. Local power-sharing institutions and interreligious violence in nigeria. *Journal of peace research*, 53(1):49–65, 2016.

[Cadena et al., 2015] Jose Cadena, Gizem Korkmaz, Chris J. Kuhlman, Achla Marathe, Naren Ramakrishnan, and Anil Vullikanti. Forecasting social unrest using activity cascades. *PloS one*, 10(6):e0128879, 2015. ISBN: 1932-6203 Publisher: Public Library of Science San Francisco, CA USA.

[Calhoun, 2013] Craig Calhoun. Occupy wall street in perspective. *British journal of sociology*, 64 (1):26–38, 2013. ISBN: 0007-1315 Publisher: The London School of Economics and Political Science.

[Cao et al., 2020] Wenming Cao, Zhiyue Yan, Zhiquan He, and Zhihai He. A comprehensive survey on geometric deep learning. *IEEE Access*, 8:35929–35949, 2020. doi: 10.1109/ACCESS.2020.2975067.

[Chakrabarti and Punera, 2011] Deepayan Chakrabarti and Kunal Punera. Event summarization using tweets. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5, pages 66–73, 2011.

[Chakraborty et al., 2014] Prithwish Chakraborty, Pejman Khadivi, Bryan Lewis, Aravindan Mahendiran, Jiangzhuo Chen, Patrick Butler, Elaine O Nsoesie, Sumiko R Mekaru, John S Brownstein, Madhav V Marathe, et al. Forecasting a moving target: Ensemble models for ili case count predictions. In *Proceedings of the 2014 SIAM international conference on data mining*, pages 262–270. SIAM, 2014.

[Chan, 2017] Michael Chan. Media use and the social identity model of collective action: Examining the roles of online alternative news and social media news. *Journalism & Mass Communication Quarterly*, 94(3):663–681, 2017. Number: 3 ISBN: 1077-6990 Publisher: SAGE Publications Sage CA: Los Angeles, CA.

[Chen et al., 2020] Deli Chen, Yankai Lin, Wei Li, Peng Li, Jie Zhou, and Xu Sun. Measuring and relieving the over-smoothing problem for graph neural networks from the topological view. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 3438–3445, 2020.

[Chen and Neill, 2014] Feng Chen and Daniel B. Neill. Non-parametric scan statistics for event detection and forecasting in heterogeneous social media graphs. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1166–1175, 2014.

[Chen et al., 2014] Xin Chen, Mihaela Vorvoreanu, and Krishna Madhavan. Mining social media data for understanding students' learning experiences. *IEEE Transactions on learning technologies*, 7(3):246–259, 2014.

[Chinta et al., 2021] Abhinav Chinta, Jingyu Zhang, Alexandra DeLucia, Mark Dredze, and Anna L Buczak. Study of manifestation of civil unrest on twitter. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 396–409, 2021.

[Church and Hanks, 1990] Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, 1990. URL https://aclanthology.org/J90-1003.

[Cinelli et al., 2021] Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9), 2021.

[Daniels, 2016] Glenda Daniels. Scrutinizing hashtag activism in the# mustfall protests in south africa in 2015: What role did media play in hashtag activism during the# rhodesmustfall and# feesmustfall protests in south africa in 2015? *Digital Activism in the Social Media Era: Critical Reflections on Emerging Trends in Sub-Saharan Africa*, pages 175–193, 2016.

[Dashtian and Murthy, 2021] Hassan Dashtian and Dhiraj Murthy. Cml-covid: A large-scale covid-19 twitter dataset with latent topics, sentiment and location information. *arXiv preprint arXiv:2101.12202*, 2021.

[Deng, 2021] Na Deng. Predicting social events using entity interaction graph sequences. In *2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)*, pages 1025–1029. IEEE, 2021. ISBN 1-66541-540-1.

[Deng and Ning, 2021] Songgaojun Deng and Yue Ning. A survey on societal event forecasting with deep learning. *arXiv preprint arXiv:2112.06345*, 2021.

[Deng et al., 2019] Songgaojun Deng, Huzefa Rangwala, and Yue Ning. Learning dynamic context graphs for predicting social events. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1007–1016, 2019.

[Devlin et al., 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv*, 2018.

[Dixon, 2022] S Dixon. Number of monetizable daily active twitter users (mdau) worldwide from 1st quarter 2017 to 1st quarter 2022. statista, 2022. URL https://www.statista.com/statistics/970920/monetizable-daily-active-twitter-users-worldwide/.

[Do et al., 2019] Loan NN Do, Hai L Vu, Bao Q Vo, Zhiyuan Liu, and Dinh Phung. An effective spatial-temporal attention based neural network for traffic flow prediction. *Transportation research part C: emerging technologies*, 108:12–28, 2019.

[Dunu Ifeona and Uzochukwu, 2015] Vivian Dunu Ifeona and Uzochukwu Uzochukwu. IOSR journal of humanities and social science. *Social Media: An Effective Tool for Social Mobilization in Nigeria*, 20(4):10–21, 2015. ISSN 2279-0837.

[El-Mallakh et al., 2018] Nelly El-Mallakh, Mathilde Maurel, and Biagio Speciale. Arab spring protests and women's labor market outcomes: Evidence from the egyptian revolution. *Journal of Comparative Economics*, 46(2):656–682, 2018.

[Erkan and Radev, 2004] Günes Erkan and Dragomir R Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479, 2004.

[Ertugrul et al., 2019] Ali Mert Ertugrul, Yu-Ru Lin, Wen-Ting Chung, Muheng Yan, and Ang Li. Activism via attention: interpretable spatiotemporal learning to forecast protest activities. *EPJ Data Science*, 8(1):5, 2019. ISBN: 2193-1127 Publisher: Springer Berlin Heidelberg.

[Frassinelli, 2018] Pier Paolo Frassinelli. Hashtags:# rhodesmustfall,# feesmustfall and the temporalities of a meme event. In *Perspectives on political communication in Africa*, pages 61–76. Springer, 2018.

[Galarza, 2018] Alex Galarza. Documenting the now, 2018.

[Goldberg and Levy, 2014] Yoav Goldberg and Omer Levy. word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method, 2014.

[Gori et al., 2005] M. Gori, G. Monfardini, and F. Scarselli. A new model for learning in graph domains. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 2, pages 729–734 vol. 2, 2005. doi: 10.1109/IJCNN.2005.1555942.

[Halkia et al., 2020] Matina Halkia, Stefano Ferri, Michail Papazoglou, Marie-Sophie Van Damme, and Dimitrios Thomakos. Conflict event modelling: Research experiment and event data limitations. In *Proceedings of the Workshop on Automated Extraction of Socio-Political Events from News 2020*, pages 42–48, 2020.

[Haq et al., 2022] Ehsan-Ul Haq, Gareth Tyson, Lik-Hang Lee, Tristan Braud, and Pan Hui. Twitter dataset for 2022 russo-ukrainian crisis. *arXiv preprint arXiv:2203.02955*, 2022.

[He et al., 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[Holland et al., 2021] Steve Holland, Jeff Mason, and Jonathan Landay. Trump summoned supporters to "wild" protest, and told them to fight. they did. Reuters, 2021.

[Horn et al., 2017] Franziska Horn, Leila Arras, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. Exploring text datasets by visualizing relevant words, 2017.

[Hornik et al., 1989] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

[Hossin and Sulaiman, 2015] Mohammad Hossin and Md Nasir Sulaiman. A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process*, 5(2):1, 2015.

[Hu, 2020] Jane Hu. The second act of social-media activism. The New Yorker, 2020.

[Hu et al., 2020] Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for pre-training graph neural networks, 2020.

[Hwang and Kim, 2015] Hyesun Hwang and Kee-Ok Kim. Social media as a tool for social movements: The effect of social media use and social capital on intention to participate in social movements. *International Journal of Consumer Studies*, 39(5):478–488, 2015. ISBN: 1470-6423 Publisher: Wiley Online Library.

[Jayaram et al., 2020] MA Jayaram, Gayitri Jayatheertha, and Ritu Rajpurohit. Time series predictive models for social networking media usage data: The pragmatics and projections. *Asian J. Res. Comput. Sci*, 2020.

[Jiang et al., 2021] Yijie Jiang, Bin Zhou, Hongkui Tu, and Liqun Gao. A temporal dual graph convolutional network for social unrest prediction. In *Journal of Physics: Conference Series*, volume 1757, page 012005. IOP Publishing, 2021. ISBN 1742-6596. Issue: 1.

[Jin et al., 2014] Fang Jin, Rupinder Paul Khandpur, Nathan Self, Edward Dougherty, Sheng Guo, Feng Chen, B. Aditya Prakash, and Naren Ramakrishnan. Modeling mass protest adoption in social network communities using geometric brownian motion. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1660–1669, 2014.

[Kallus, 2014] Nathan Kallus. Predicting crowd behavior with big public data. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 625–630, 2014.

[Khambule et al., 2019] Isaac Khambule, Amarone Nomdo, and Babalwa Siswana. Burning capabilities: the social cost of violent and destructive service delivery protests in south africa. *African Journal of Peace and Conflict Studies*, 8(1):51, 2019.

[Killian et al., 2020] Lewis M. Killian, Ralph H. Turner, and Neil J. Smelser. "social movement". Encyclopedia Britannica, 2020. URL https://www.britannica.com/topic/social-movement.

[Kingma and Ba, 2017] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[Kipf and Welling, 2016] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[Korkmaz et al., 2015] Gizem Korkmaz, Jose Cadena, Chris J. Kuhlman, Achla Marathe, Anil Vullikanti, and Naren Ramakrishnan. Combining heterogeneous data sources for civil unrest forecasting. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, pages 258–265, 2015.

[Kulshrestha et al., 2012] Juhi Kulshrestha, Farshad Kooti, Ashkan Nikravesh, and Krishna Gummadi. Geographic dissection of the twitter network. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 6, pages 202–209, 2012.

[Kwak et al., 2010a] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600, 2010a.

[Kwak et al., 2010b] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? pages 591–600, 2010b.

[Lancaster, 2018] Lizette Lancaster. Unpacking discontent: Where and why protest happens in south africa. *South African Crime Quarterly*, 64:29–43, 2018. ISBN: 1991-3877.

[Leetaru and Schrodt, 2013] Kalev Leetaru and Philip A Schrodt. Gdelt: Global data on events, location, and tone, 1979–2012. In *ISA annual convention*, volume 2, pages 1–49. Citeseer, 2013.

[Levy and Goldberg, 2014] Omer Levy and Yoav Goldberg. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 302–308, Baltimore, Maryland, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-2050. URL https://aclanthology.org/P14-2050.

[Levy et al., 2015] Omer Levy, Yoav Goldberg, and Ido Dagan. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the association for computational linguistics*, 3:211–225, 2015.

[Li et al., 2018] Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[Liu et al., 2021] Xiaochen Liu, Yang Su, and Bingjie Xu. The application of graph neural network in natural language processing and computer vision. In *2021 3rd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI)*, pages 708–714, 2021. doi: 10.1109/MLBDBI54094.2021.00140.

[Lord, 2021] Richard Lord. The state of social media in south africa. https://themediaonline.co.za/2021/07/the-state-of-social-media-in-south-africa/, 2021.

[Loya and McLeod] Luis Loya and Doug McLeod. Social protest. Oxford Bibliographies. URL https://www.oxfordbibliographies.com/view/document/obo-9780199756841/obo-9780199756841-0005.xml.

[Lu et al., 2021] Yuanfu Lu, Xunqiang Jiang, Yuan Fang, and Chuan Shi. Learning to pre-train graph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 4276–4284, 2021.

[Maalouf, 2011] Maher Maalouf. Logistic regression in data analysis: an overview. *International Journal of Data Analysis Techniques and Strategies*, 3(3):281–299, 2011.

[Manacorda and Tesei, 2020] Marco Manacorda and Andrea Tesei. Liberation technology: Mobile phones and political mobilization in africa. *Econometrica*, 88(2):533–567, 2020.

[Mare, 2014] Admire Mare. Social media: The new protest drums in southern africa? In *Social Media in Politics*, pages 315–335. Springer, 2014.

[Matebesi and Botes, 2017] Sethulego Matebesi and Lucius Botes. Party identification and service delivery protests in the eastern cape and northern cape, south africa. *African Sociological Review/Revue Africaine de Sociologie*, 21(2):81–99, 2017. Number: 2 ISBN: 1027-4332.

[McInnes et al., 2018] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29): 861, 2018.

[Mele and Correal, 2016] Christopher Mele and Annie Correal. 'not our president': Protests spread after donald trump's election. The New York Times, 2016. URL https://www.nytimes.com/2016/11/10/us/trump-election-protests.html.

[Mfundo, 2021] Mkhize Mfundo. Phoenix residents tell SAHRC hearing of racial tension, abuse during july unrest. Times Live, 2021.

[Mikolov et al., 2013] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[Mokoena, 2021] Sakhile Mokoena. Ncop welcomes efforts to reduce effects of july unrest on economy. Parliament of Republic of South Africa, 2021.

[Monti et al., 2019] Federico Monti, Fabrizio Frasca, Davide Eynard, Damon Mannion, and Michael M Bronstein. Fake news detection on social media using geometric deep learning. *arXiv preprint arXiv:1902.06673*, 2019.

[Mottiar, 2013] Shauna Mottiar. From 'popcorn'to 'occupy': protest in durban, south africa. 44(3): 603–619, 2013. ISBN: 0012-155X Publisher: Wiley Online Library.

[Mundt et al., 2018] Marcia Mundt, Karen Ross, and Charla M. Burnett. Scaling social movements through social media: The case of black lives matter. *Social Media+ Society*, 4(4): 2056305118807911, 2018. ISBN: 2056-3051 Publisher: SAGE Publications Sage UK: London, England.

[Nahed Eltantawy and Wiest, 2011] Nahed Eltantawy and Julie Wiest. *International Journal of Communication*, 5:1207 – 1224, 2011.

[Nothman et al., 2018] Joel Nothman, Hanmin Qin, and Roman Yurchak. Stop word lists in free open-source software packages. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 7–12, 2018.

[O'Connor et al., 2010] Brendan O'Connor, Ramnath Balasubramanyan, Bryan Routledge, and Noah Smith. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the international AAAI conference on web and social media*, volume 4, pages 122–129, 2010.

[Parrish et al., 2018] Nathan H. Parrish, Anna L. Buczak, Jared T. Zook, James P. Howard, Brian J. Ellison, and Benjamin D. Baugher. Crystal cube: Multidisciplinary approach to disruptive events prediction. In *International Conference on Applied Human Factors and Ergonomics*, pages 571–581. Springer, 2018.

[Passarelli and Tabellini, 2017] Francesco Passarelli and Guido Tabellini. Emotions and political unrest. *Journal of Political Economy*, 125(3):903–946, 2017. Number: 3 ISBN: 0022-3808 Publisher: University of Chicago Press Chicago, IL.

[Pillay and Mtshali, 2021] Yogashen Pillay and Samkelo Mtshali. Social media was 'instrumental' in july unrest expert tells sa human rights commission. IOL, 2021.

[Qaiser and Ali, 2018] Shahzad Qaiser and Ramsha Ali. Text mining: use of tf-idf to examine the relevance of words to documents. *International Journal of Computer Applications*, 181(1): 25–29, 2018.

[Qiao and Wang, 2015] Fengcai Qiao and Hui Wang. Computational approach to detecting and predicting occupy protest events. In *2015 International Conference on Identification, Information, and Knowledge in the Internet of Things (IIKI)*, pages 94–97. IEEE, 2015.

[Qiao et al., 2017a] Fengcai Qiao, Pei Li, Xin Zhang, Zhaoyun Ding, Jiajun Cheng, and Hui Wang. Predicting social unrest events with hidden markov models using gdelt. *Discrete Dynamics in Nature and Society*, 2017, 2017a.

[Qiao et al., 2017b] Fengcai Qiao, Pei Li, Xin Zhang, Zhaoyun Ding, Jiajun Cheng, and Hui Wang. Predicting social unrest events with hidden markov models using GDELT. *Discrete Dynamics in Nature and Society*, 2017, 2017b. ISBN: 1026-0226 Publisher: Hindawi.

[Radinsky and Horvitz, 2013] Kira Radinsky and Eric Horvitz. Mining the web to predict future events. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 255–264, 2013.

[Raleigh et al., 2010] Clionadh Raleigh, Andrew Linke, Håvard Hegre, and Joakim Karlsen. Introducing acled: an armed conflict location and event dataset: special data feature. *Journal of peace research*, 47(5):651–660, 2010.

[Ramakrishnan et al., 2014] Naren Ramakrishnan, Patrick Butler, Sathappan Muthiah, Nathan Self, Rupinder Khandpur, Parang Saraf, Wei Wang, Jose Cadena, Anil Vullikanti, and Gizem Korkmaz. 'beating the news' with EMBERS: forecasting civil unrest using open source indicators. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1799–1808, 2014.

[Ramos et al., 2003] Juan Ramos et al. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 29–48. Citeseer, 2003.

[Rodny-Gumede, 2017a] Ylva Rodny-Gumede. Questioning the media and democracy relationship: the case of south africa. *Communication*, 43(2):10–22, 2017a. Number: 2 ISBN: 0250-0167 Publisher: Taylor & Francis.

[Rodny-Gumede, 2017b] Ylva Rodny-Gumede. Questioning the media and democracy relationship: the case of south africa. *Communicatio*, 43(2):10–22, 2017b.

[Rogers, 2013] Simon Rogers. Gdelt: A big data history of life, the universe and everything, Apr 2013. URL https://www.theguardian.com/news/datablog/2013/apr/12/gdelt-global-database-events-location.

[Saif et al., 2014] Hassan Saif, Miriam Fernandez, Yulan He, and Harith Alani. On stopwords, filtering and data sparsity for sentiment analysis of Twitter. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 810–817, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/292_Paper.pdf.

[Salle and Villavicencio, 2019] Alexandre Salle and Aline Villavicencio. Why so down? the role of negative (and positive) pointwise mutual information in distributional semantics. *ArXiv*, abs/1908.06941, 2019.

[Santos et al., 2014] Igor Santos, Igor Miñambres-Marcos, Carlos Laorden, Patxi Galán-García, Aitor Santamaría-Ibirika, and Pablo García Bringas. Twitter content-based spam filtering. In *International Joint Conference SOCO'13-CISIS'13-ICEUTE'13: Salamanca, Spain, September 11th-13th, 2013 Proceedings*, pages 449–458. Springer, 2014.

[Saxton et al., 2015] Gregory D Saxton, Jerome Niyirora, Chao Guo, and Richard Waters. # advocatingforchange: The strategic use of hashtags in social media advocacy. *Advances in Social Work*, 16(1):154–169, 2015.

[Schwartz and Ungar, 2015] H Andrew Schwartz and Lyle H Ungar. Data-driven content analysis of social media: a systematic overview of automated methods. *The ANNALS of the American Academy of Political and Social Science*, 659(1):78–94, 2015.

[Silva and Ribeiro, 2003] C. Silva and B. Ribeiro. The importance of stop word removal on recall values in text categorization. In *Proceedings of the International Joint Conference on Neural Networks, 2003.*, volume 3, pages 1661–1666 vol.3, 2003. doi: 10.1109/IJCNN.2003.1223656.

[Smith et al., 2017] Emmanuel M. Smith, Jim Smith, Phil Legg, and Simon Francis. Predicting the occurrence of world news events using recurrent neural networks and auto-regressive moving average models. In *UK Workshop on Computational Intelligence*, pages 191–202. Springer, 2017.

[Steinert-Threlkeld, 2017] Zachary C. Steinert-Threlkeld. Spontaneous collective action: Peripheral mobilization during the arab spring. *American Political Science Review*, 111(2):379–403, 2017. ISBN: 0003-0554 Publisher: Cambridge University Press.

[Stone et al., 2011] Benjamin Stone, Simon Dennis, and Peter J Kwantes. Comparing methods for single paragraph similarity analysis. *Topics in Cognitive Science*, 3(1):92–122, 2011.

[Tuke et al., 2020] Jonathan Tuke, Andrew Nguyen, Mehwish Nasim, Drew Mellor, Asanga Wickramasinghe, Nigel Bean, and Lewis Mitchell. Pachinko prediction: A bayesian method for event prediction from social media data. *Information Processing & Management*, 57(2):102147, 2020.

[Uwalaka and Watkins, 2018] Temple Uwalaka and Jerry Watkins. Social media as the fifth estate in nigeria: An analysis of the 2012 occupy nigeria protest. *African Journalism Studies*, 39(4): 22–41, 2018. ISBN: 2374-3670 Publisher: Taylor & Francis.

[Vaswani et al., 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[Veličković et al., 2018] Petar Veličković, William Fedus, William L. Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm. Deep graph infomax, 2018.

[Wang et al., 2020] Haiyang Wang, Bin Zhou, Zhipin Gu, and Yan Jia. Contextual gated graph convolutional networks for social unrest events prediction. In *2020 IEEE Fifth International Conference on Data Science in Cyberspace (DSC)*, pages 320–325. IEEE, 2020.

[Ward et al., 2013] Michael D Ward, Andreas Beger, Josh Cutler, Matthew Dickenson, Cassy Dorff, and Ben Radford. Comparing gdelt and icews event data. *Analysis*, 21(1):267–297, 2013.

[Weber et al., 2020] Derek Weber, Mehwish Nasim, Lewis Mitchell, and Lucia Falzon. A method to evaluate the reliability of social media data for social network analysis. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 317–321. IEEE, 2020.

[Webster and Kit, 1992] Jonathan Webster and Chunyu Kit. Tokenization as the initial phase in nlp. pages 1106–1110, 01 1992. doi: 10.3115/992424.992434.

[Weisfeiler and Leman, 1968] Boris Weisfeiler and Andrei Leman. The reduction of a graph to canonical form and the algebra which appears therein. *nti, Series*, 2(9):12–16, 1968.

[Wu and Gerber, 2017] Congyu Wu and Matthew S. Gerber. Forecasting civil unrest using social media and protest participation theory. *IEEE Transactions on Computational Social Systems*, 5(1):82–94, 2017. ISBN: 2329-924X Publisher: IEEE.

[Xie et al., 2012] Wei Xie, Cheng Li, Feida Zhu, Ee-Peng Lim, and Xueqing Gong. When a friend in twitter is a friend in life. In *Proceedings of the 4th Annual ACM Web Science Conference*, pages 344–347, 2012.

[Xu et al., 2019] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks?, 2019.

[Zambezi, 2021] Samantha Zambezi. Predicting social unrest events in south africa using lstm neural networks. Master's thesis, Faculty of Science, 2021.

[Zaremba et al., 2014] Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*, 2014.

[Zeitzoff, 2017] Thomas Zeitzoff. How social media is changing conflict. *Journal of Conflict Resolution*, 61(9):1970–1991, 2017. ISBN: 0022-0027 Publisher: SAGE Publications Sage CA: Los Angeles, CA.

[Zhang and Zhang, 2020] Haopeng Zhang and Jiawei Zhang. Text graph transformer for document classification. In *Conference on empirical methods in natural language processing (EMNLP)*, 2020.

[Zhang et al., 2021] Xiao-Meng Zhang, Li Liang, Lin Liu, and Ming-Jing Tang. Graph neural networks and their current applications in bioinformatics. *Frontiers in Genetics*, 12, 2021. doi: 10.3389/fgene.2021.690049.

[Zhao, 2021] Liang Zhao. Event prediction in the big data era: A systematic survey. *ACM Computing Surveys (CSUR)*, 54(5):1–37, 2021.

[Zhao et al., 2015] Liang Zhao, Qian Sun, Jieping Ye, Feng Chen, Chang-Tien Lu, and Naren Ramakrishnan. Multi-task learning for spatio-temporal event forecasting. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1503–1512, 2015.

Appendix A

CAMEO codes

Root level	Second level	Third level
14 PROTEST		
	140 Engage in political dissent, not specified below	
	141 Demonstrate or rally	
		1411 Demonstrate for leadership change
		1412 Demonstrate for policy change
		1413 Demonstrate for rights
		1414 Demonstrate for change in institutions, regime
	142 Conduct hunger strike, not specified below	
		1421 Conduct hunger strike for leadership change
		1422 Conduct hunger strike for policy change
		1423 Conduct hunger strike for rights
		1424 Conduct hunger strike for change in institutions, regime
	143 Conduct strike or boycott, not specified below	
		1431 Conduct strike or boycott for leadership change
		1432 Conduct strike or boycott for policy change
		1433 Conduct strike or boycott for rights
		1434 Conduct strike or boycott for change in institutions, regime
	144 Obstruct passage, block	
		1441 Obstruct passage to demand leadership change
		1442 Obstruct passage to demand policy change
		1443 Obstruct passage to demand rights
		1444 Obstruct passage to demand change in institutions, regime
	145 Protest violently, riot	
		1451 Engage in violent protest for leadership change
		1452 Engage in violent protest for policy change
		1453 Engage in violent protest for rights
		1454 Engage in violent protest for change in institutions, regime
15 EXHIBIT FORCE POSTURE		
	150 Demonstrate military or police power, not specified below	
	151 Increase police alert status	
	152 Increase military alert status	
	153 Mobilize or increase police power	
	154 Mobilize or increase armed forces	
17 COERCE		
	170 Coerce, not specified below	
	171 Seize or damage property, not specified below	
		1711 Confiscate property
		1712 Destroy property
	172 Impose administrative sanctions, not specified below	
		1721 Impose restrictions on political freedoms
		1722 Ban political parties or politicians
		1723 Impose curfew
		1724 Impose state of emergency or martial law
	173 Arrest, detain, or charge with legal action	
	174 Expel or deport individuals	
	175 Use tactics of violent repression	
18 ASSAULT		
	180 Use unconventional violence, not specified below	
	181 Abduct, hijack, or take hostage	
	182 Physically assault, not specified below	
		1821 Sexually assault
		1822 Torture
		1823 Kill by physical assault
	183 Conduct suicide, car, or other non-military bombing, not spec below	
		1831 Carry out suicide bombing
		1832 Carry out car bombing
		1833 Carry out roadside bombing
	184 Use as human shield	
	185 Attempt to assassinate	
	186 Assassinate	
19 FIGHT		
	190 Use conventional military force, not specified below	
	191 Impose blockade, restrict movement	
	192 Occupy territory	
	193 Fight with small arms and light weapons	
	194 Fight with artillery and tanks	
	195 Employ aerial weapons	
	196 Violate ceasefire	
20 USE UNCONVENTIONAL MASS VIOLENCE		
	200 Use unconventional mass violence, not specified below	
	201 Engage in mass expulsion	
	202 Engage in mass killings	
	203 Engage in ethnic cleansing	
	204 Use weapons of mass destruction, not specified below	
	2041 Use chemical, biological, or radiological weapons	
	2042 Detonate nuclear weapons	

Table A.1: Three-level unrest related CAMEO codes event descriptions

Root level	Second level	Third level
01 MAKE PUBLIC STATEMENT		
	010 Make statement, not specified below	
	011 Decline comment	
	012 Make pessimistic comment	
	013 Make optimistic comment	
	014 Consider policy option	
	015 Acknowledge or claim responsibility	
	016 Deny responsibility	
	017 Engage in symbolic act	
	018 Make empathetic comment	
	019 Express accord	
02 APPEAL		
	020 Appeal, not specified below	
	021 Appeal for material cooperation, not specified below	
		0211 Appeal for economic cooperation
		0212 Appeal for military cooperation
		0213 Appeal for judicial cooperation
		0214 Appeal for intelligence
	022 Appeal for diplomatic cooperation, such as policy support	
	023 Appeal for aid, not specified below	
		0231 Appeal for economic aid
		0232 Appeal for military aid
		0233 Appeal for humanitarian aid
		0234 Appeal for military protection or peacekeeping
	024 Appeal for political reform, not specified below	
		0241 Appeal for change in leadership
		0242 Appeal for policy change
		0243 Appeal for rights
		0244 Appeal for change in institutions, regime
	025 Appeal to yield	
		0251 Appeal for easing of administrative sanctions
		0252 Appeal for easing of popular dissent
		0253 Appeal for release of persons or property
		0254 Appeal for easing of economic sanctions, boycott, or embargo
		0255 Appeal for target to allow international involvement (non-mediation)
		0256 Appeal for de-escalation of military engagement
	026 Appeal to others to meet or negotiate	
	027 Appeal to others to settle dispute	
	028 Appeal to others to engage in or accept mediation	
03 EXPRESS INTENT TO COOPERATE		
	030 Express intent to cooperate, not specified below	
	031 Express intent to engage in material cooperation, not specified below	
		0311 Express intent to cooperate economically
		0312 Express intent to cooperate militarily
		0313 Express intent to cooperate on judicial matters
		0314 Express intent to cooperate on intelligence
	032 Express intent to provide diplomatic cooperation such as policy support	
	033 Express intent to provide matyerial aid, not specified below	
		0331 Express intent to provide economic aid
		0332 Express intent to provide military aid
		0333 Express intent to provide humanitarian aid
		0334 Express intent to provide military protection or peacekeeping
	034 Express intent to institute political reform, not specified below	
		0341 Express intent to change leadership
		0342 Express intent to change policy
		0343 Express intent to provide rights
		0344 Express intent to change institutions, regime
	035 Express intent to yield, not specified below	
		0351 Express intent to ease administrative sanctions
		0352 Express intent to ease popular dissent
		0353 Express intent to release persons or property
		0354 Express intent to ease economic sanctions, boycott, or embargo
		0355 Express intent allow international involvement (not mediation)
		0356 Express intent to de-escalate military engagement
	036 Express intent to meet or negotiate	
	037 Express intent to settle dispute	
	038 Express intent to accept mediation	
	039 Express intent to mediate	

Table A.2: Three-level non-unrest related CAMEO codes event descriptions