Tracing the Precursors and Amplifiers of Conflict in the Information Age: An NLP Inquiry of Tensions, Political Communication, and Misinformation

by

Philipp Zimmer

B.S. Management
Kühne Logistics University, 2020

Submitted to the Institute of Data, Systems, and Society and the
Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degrees of

MASTER OF SCIENCE IN TECHNOLOGY AND POLICY

and

MASTER OF SCIENCE IN ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2024

Authored by:	Philipp Zimmer Institute of Data, Systems, and Society January 26, 2024
Certified by:	Fotini Christia Ford International Professor of the Social Sciences, Thesis Supervisor
Certified by:	Manish Raghavan Assistant Professor of Information Technology, Thesis Supervisor
Accepted by:	Frank R. Field III Senior Research Engineer, Sociotechnical Systems Research Interim Director, Technology and Policy Program
Accepted by:	Leslie A. Kolodziejski Professor of Electrical Engineering and Computer Science Chair, Department Committee on Graduate Students

Tracing the Precursors and Amplifiers of Conflict in the Information Age: An NLP Inquiry of Tensions, Political Communication, and Misinformation

by

Philipp Zimmer

Submitted to the Institute of Data, Systems, and Society and the
Department of Electrical Engineering and Computer Science
on January 26, 2024 in partial fulfillment of the requirements for the degrees of

MASTER OF SCIENCE IN TECHNOLOGY AND POLICY

and

MASTER OF SCIENCE IN ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

ABSTRACT

Violent conflicts, in their varied and complex forms, have long been a subject of research and political discourse. Despite increased attention for the field, various nuances and dynamics are yet to be explored. This thesis seeks to study three aspects of the multifaceted nature of conflicts through the lens of natural language processing (NLP), thereby not only offering new insights but also advancing the field's methodological landscape.

First, the study delves into the identification of causal predictors of conflicts. By showcasing the potential of a frame-semantic parser, I am able to quantify the precursors that contribute to conflict and examine the potential for enhancing prediction models with greater qualitative depth. This chapter utilizes a rich but under-examined data source, news articles, which can aid closing the data gap in conflict studies.

In the second chapter, the communication strategies of political leaders during crises are scrutinized to understand the rationale behind their messaging and the impact thereof. I argue that leaders' engagement frequency and style with their citizens is dependent on the political systems' characteristics and that it matters for societal conceptions.

The final chapter addresses the spread of misinformation, such as in times of crisis, investigating which themes are prone to the widespread propagation on social media and presenting a novel ensemble method for the detection of misleading and false content.

By integrating computational techniques with political theory, this work contributes to a nuanced understanding of conflict dynamics and offers rich potential for anticipatory actions of policymakers.

Thesis supervisor: Fotini Christia
Title: Ford International Professor of the Social Sciences

Thesis supervisor: Manish Raghavan
Title: Assistant Professor of Information Technology

Acknowledgments

This thesis is not only a reflection of my academic journey, but also a testament to the incredible support and inspiration I have received from many. I am deeply grateful to each individual who has contributed to my growth, joy, and success during my time at MIT.

First and foremost, I dedicate this work to my family, especially my parents. They instilled in me the curiosity to explore the unknown, encouraging me to see the world and follow my passions. Their example of living life with compassion and a commitment to making the world a better place through knowledge has been my guiding star.

To my friends, especially Malte, Paul and Alexa, who have always had my back through the highs and lows in the past years, thank you for creating a second home filled with love, tea, and music. Your unwavering support has been invaluable.

I am also grateful to the many remarkable individuals I have met along my academic and professional journey. To Tobi, Alexander, Maria, and Kersten – thank you for being more than mentors; I would not want to miss your guidance, friendship, and time.

A special thanks to my advisors, Fotini and Manish, for their continued guidance, trust in my research endeavors, and insightful comments that have greatly enriched my work. Also, a heartfelt thank you to everyone in the lab for many interesting thoughts and conversations.

I extend my gratitude to my collaborators, particularly Sarah, for all her advice and mentorship, teaching me how to think like a political scientist and building theories. My work on political leadership communication is a collaborative project and I look forward to jointly publishing a more elaborate study on the topic.

My experience at the World Bank has been exceptionally rewarding, thanks to colleagues like Sam and Manuel. Working with you, including on the projects informing my chapter on misinformation, has not only pushed my technical knowledge but also shown how research and practice can complement each other effectively.

Finally, I would like to express my strongest appreciation to everyone at IDSS, especially Barb and Frank, for crafting a beautiful graduate program and fostering a positive, supportive community. I must admit I rolled my eyes hearing about TPP being like a family back in 2021, but now I can proudly say how happy I am to remain part of it as an alumnus.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The dawn of the 21st century has witnessed a persistent and troubling proliferation of conflicts across the globe. From the devastating impacts of Russia's invasion of Ukraine, the enduring war in Sudan, to the sectarian strife in Nigeria and the protracted Syrian civil war, the list of armed engagements is lengthy and their conclusion remains elusive. These conflicts, compounded by the exacerbating effects of climate change, are a bitter taste of an alarming predicted five-fold increase in crises in the coming decades [1].

Amidst this unsettling backdrop, quantitative methods like machine learning (ML) have emerged as a beacon of hope in conflict studies, promising a new paradigm in forecasting and understanding the complexities of violent conflicts. Scholars such as Mueller and Rauh [2] have made strides in discerning the countries at risk of conflict, while Solaimani et al. [3] posit the potential of ML and natural language processing (NLP) as tools for early warning systems. Hegre et al. [4] emphasize that the pivotal questions surrounding the future of machine learning methods in peace research concern its form rather than its necessity, underscoring the burgeoning role of technology in this field.

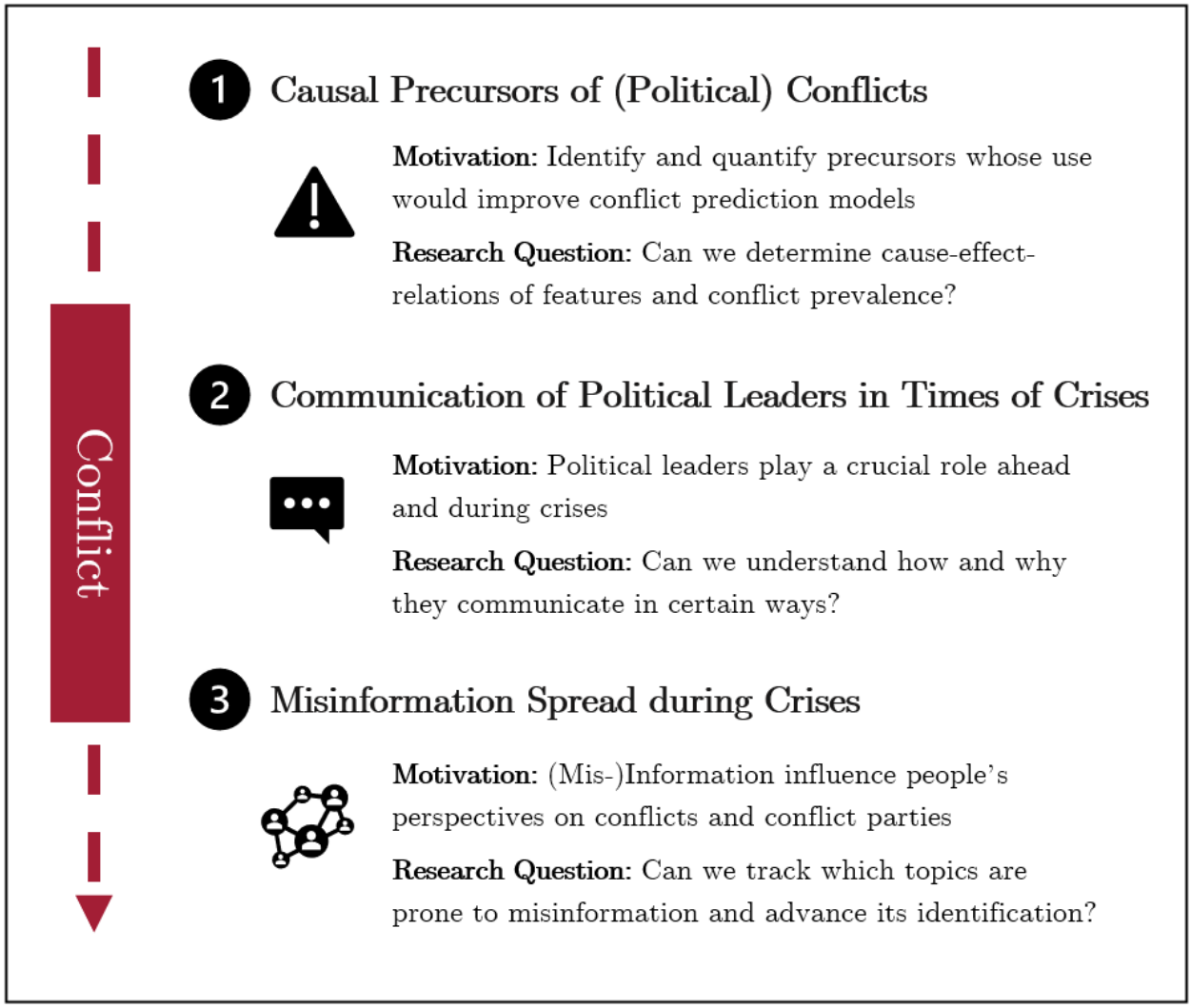
Cederman and Weidmann [5] remind the research community of the unique nature of each conflict, challenging the notion of universal solutions. In this intricate landscape, it seems unlikely that machine learning and artificial intelligence will supplant the nuanced human decision-making process, often based on intuition. Yet, in an era marked by rapid information flow, it is incumbent upon policymakers to leverage technological advancements for enhanced insights to inform their strategic choices.

This thesis is an exploration of that very premise: not to replace human discretion, but to judiciously augment it with computational tools. Specifically, it focuses on the application of natural language processing. The subtleties and embedded values in language, as discussed by KhudaBukhsh et al. [6] and Ahmed et al. [7], advocate for the significance of text data and natural language processing in studying such a nuanced field. Fairclough [8] articulates

the profound connections between language use and power dynamics, highlighting the oft-overlooked role of language in shaping social relations and influencing the balance of power.

I approach the study of conflicts not as a discrete event with a clear beginning and end, but rather as a continuum of interactions that evolve over time. In adopting a comprehensive definition of conflict, I align with scholarly perspectives that recognize the diverse nature of strife, such as Raleigh et al.'s [9] examination of civil war research paradigms, and the extensive literature addressing various manifestations of conflict, including interstate wars, civil unrest, and episodes of political instability [e.g. 10–12]. This inclusive approach underscores the thesis's commitment to understanding conflict in all its forms, highlighting the potential of computational methods, and NLP in particular.

Figure 1.1: Structure of Studying Three Facets of Conflicts with NLP.



To demonstrate the potential of NLP within the domain of conflict studies, this thesis

examines three underexplored facets of conflict, each representing a piece of the larger puzzle of conflicts and each potentially promising venues to use quantitative methods.

As shown in Figure 1.1, Chapter 2 attempts to identify causal precursors to political conflicts by implementing a frame-semantic parser to analyze textual data from news articles. This method enables the quantitative evaluation of conflict predictors within cause-and-effect frames, providing a novel set of indicators for conflict forecasting models. The outcome is a refined approach to addressing the need for better indicators in conflict predictions [13, 14], enhancing the explanatory power of traditional conflict analysis methods.

Chapter 3 examines what happens when "it has happened," that is, when a conflict or crises occurs. In particular, I inquire about the communication strategies of political leaders in times of tensions, focusing on the rhetoric employed during periods of heightened political tension, such as elections and crises. The analysis differentiates communicative behaviors by regime type and individual characteristics of leaders, offering insight into the relationship between leadership communication and the political environment. Additionally, the chapter explores the incidence of toxic messaging, providing a comparative perspective on the discourse during times of crisis.

Next, Chapter 4 addresses the proliferation of misinformation in online spaces, particularly during crises, which has significant implications for societal stability [15, 16]. Utilizing a semantic similarity model, I analyze social media content to detect and track the dissemination of misinformation. The model is calibrated against a corpus of fact-checked false claims to establish a metric for identifying similar deceptive content. The findings contribute to an understanding of the geographical and temporal distribution of misinformation and enhance the methodology for its identification in social media networks.

Finally, in Chapter 5, I present concluding remarks on the identified benefits and limitations of utilizing natural language processing in the context of conflict studies.

Chapter 2

The Seeds of Conflict: Quantifying Causal Precursors of Political Tensions

2.1 Theory

2.1.1 Characteristics and Causes of (Violent) Conflicts

Conflicts, in their varied and complex forms, have long been a subject of both intrigue and concern in global discourse. The eruption of violence in regions like the Middle East has reignited debates about the predictability of conflicts. While some argue that certain conflicts are foreseeable, others maintain that predicting the outbreak and timing of such conflicts is an immensely challenging task. This divergence in views raises fundamental questions: What exactly constitutes a conflict, and is it possible to predict them?

The study of conflicts, particularly within the realm of civil war research, has seen diverse definitions and approaches. Raleigh et al. [9] highlight that while researchers have focused on the causes of war, the definition of war itself varies significantly. This discrepancy often leads to certain types of violence being overlooked in data collection, resulting in a misrepresentation of the nature of internal conflicts. The research has expanded over time to encompass a variety of conflicts, including interstate wars [10], civil wars [11], and other forms of turmoil as state failures, genocides, human rights violations, and ethnic clashes [12].

In this study, I regard conflicts as encompassing all types of violent conflicts, irrespective of the parties involved. The focus is not primarily on predicting conflicts but on how computational methods, particularly those from natural language processing, can inform and enhance the prediction of such conflicts by identifying and quantifying precursors to conflicts.

Recent years have witnessed an escalation in mass atrocities, terrorism, and political

unrest, leading to significant human suffering [3]. The variety of conflicts, both internal and external, is influenced by a variety of factors like climate change. Climate change affects living conditions, deteriorates livelihoods, increases needs, and thus fuels desperation and conflict. Often, natural and violent crises reinforce each other, creating a cycle of deterioration and unrest. On the mission of understanding these patterns, the field of conflict and peace studies has evolved significantly, influenced by historical events like the Cold War, technological developments, and globalization [17]. Though, an overarching notion persists that war can be conceptualized as a manifestation of disorder [17].

In the study of conflict triggers, Beck at al. [10] suggest that certain factors can assume significant importance in some contexts, while they are of relatively minor relevance in the majority of cases. This variance suggests that context-specific analysis is crucial in conflict studies [10]. The psychological profile of political leaders is found to be a significant factor regarding the risk of conflict escalation. Leaders who exhibit a strong need for power, coupled with a tendency towards simplistic thinking and distrustfulness, are often more prone to engage in aggressive political actions. Their nationalistic views and belief in personal control over events further exacerbate conflict risks [18]. The transition towards democracy in a nation reveals a relationship with conflict. Evidence suggests a gradual decline in conflict likelihood as democratic processes stabilize, a relationship that is nuanced and influenced by various socio-political factors [2]. Recent studies have also shed light on the role of social media in civil conflicts. Enhanced social media engagement, especially among political elites, correlates with increased levels of civil unrest. While not a direct cause, social media can amplify societal tensions and political polarization, contributing to conflict escalation [19].

The scholarly landscape on conflict causation is broad, with studies exploring ethnic cleavages, climate change, natural resources, and a mix of political and economic indicators as potential drivers of conflict [20]. With regards to the economy, economic stability plays a critical role in conflict emergence. Research by Collier [21] points out how variations in economic growth rates can significantly impact the likelihood of civil war in a country – noting that a small increase in economic growth can reduce the risk of civil war. Other factors, such as arms races, territorial rivalries, shifts in power, and geographic terrain, have also been associated with increased conflict risks, influencing both international and domestic disputes. Of these traditional conflict risk factors, particularly geographical and ethnic elements, are relatively static, presenting challenges in their predictive value for conflict timing. Likewise, the causal links of these factors to conflict onset are a subject of ongoing research [2].

This diversity in identified causes underscores the complexity of conflict prediction and the need for multifaceted approaches to understand and mitigate the risks of conflict. It also highlights the importance of paving a way to identifying the true causal factors among those

correlated with conflict occurrences in line with a common notion in the machine learning community: correlation versus causality.

2.1.2 The Gravity of the Situation and Current Trends

The study of conflicts reveals both their historical prevalence and the evolving nature of their occurrence. Since the beginning of the 20th century, conflicts have had a significant impact, with over 200 wars leading to approximately 35 million battle deaths [22]. This staggering number is a testament to the enduring nature of conflict in human history. In more recent times, Grady et al. [23] observe that since 1989, violent inter-group conflicts have caused around 2.8 million deaths and displaced over 100 million people. These conflicts have broader implications, threatening food supplies and inflicting psychological trauma on both participants and victims. A poignant example is the conflict in Nigeria's Middle Belt, which parallels other local-level ethnic or tribal conflicts over resources seen globally, such as in Kenya, Uganda, Mali, Yemen, and Afghanistan.

As one delves deeper into overarching temporal trends regarding the prevalence of conflicts, it becomes apparent that the roles of individual leaders and political systems have played a role over time. Krcmaric et al. [24] remark on the recent and expected changes in the influence of individual leaders on geo-politics and confrontations. Guriev and Treisman [25] note a decrease in state-internal violence with the rise of informational autocrats in recent years, leading to less overt forms of violence, such as political killings.

Meanwhile, the call for anticipatory action in conflict research has been growing louder. Newman and van Selm [26] emphasize the need for integrative work that addresses root causes. This is further motivated by findings that over time the accuracy of conflict predictions has varied geographically, with East Africa showing lower accuracy compared to West and Southern Africa [27]. This variance highlights the complex interplay of factors influencing conflicts. In the quest to understand and predict conflicts, researchers grapple with the challenge of human bias in political assessments. Little and Meng [28] argue that perceived democratic declines are more a reflection of changes in human perceptions rather than actual political institutions. This highlights the difficulty in separating subjective interpretation from objective data in conflict studies.

Recent advancements in technology have opened new avenues for conflict prediction. Solaimani et al. [3] speak to the potential of machine learning and natural language processing tools in early warning systems. The geographical and spatial aspects of conflicts are equally crucial. Solaimani et al. [3] note that political conflicts often begin in specific regions and spread through complex inter-dependencies. Likewise, the duration and nature of conflicts

are influenced by various geographical and resource-related factors. Research reveals patterns such as longer conflicts in peripheral areas and the role of local resources like diamonds in prolonging conflicts [29, 30]. The geographical terrain also plays a role in conflict dynamics as mountainous regions are more prone to conflict due to the advantage they provide to non-state forces like rebel groups [31]. The changing composition of the actors in conflicts globally has thus changed both the dynamic and the duration of violence. Raleigh et al. [9] further emphasize the complexities of assessing militias, which can be transient and not fully representative of the ethnic groups they are associated with. Not only does the evolving nature of many militant groups affect the ability to study them, but also does it increase government forces' ability to combat them. Cases all over the globe, including Afghanistan, have contributed to the worrying trend of conflicts "not really ending."

Therefore, in evaluating conflicts, the focus is increasingly shifting towards understanding the broader impact on populations rather than just on fatalities. Raleigh et al. [9] argue that fatality-based intensity assessments can be misleading, advocating for a broader approach to understanding conflict impact. Despite numerous efforts to predict this, particularly establishing causality remains a significant challenge [32]. This old notion is echoed in more recent studies, which revisits and supports Gurr's post-Cold War forecast about the decline of ethnic civil wars due to accommodative politics [5].

Looking towards the future, the potential long-term effects of conflicts are a subject of intense study. A study on the conflict trap shows that the onset of a new conflict substantially increases the long-term incidence of conflict, while successful de-escalation have shown positive effects [4]. All in all, the worrisome developments encourage the need to advancing conflict research.

2.1.3 What are Potential Problems with Conflict Predictions? What is the Value of Knowing Early?

The state of conflict predictions research is far from perfect. The field of conflict prediction encounters numerous challenges, including issues with data accessibility and the complexity inherent in political dynamics. Practical hurdles such as gaining access to open data and interpreting it once obtained are significant barriers [13]. The predictive aspect in conflict studies is fraught with controversy within the academic environment, with ongoing debates about the feasibility of accurate forecasting [5].

Past approaches to forecasting have often been criticized for their poor performance, highlighting a need for methods that enable causal reasoning, a crucial element in political science [10]. The prediction of ongoing conflicts appears more feasible, particularly in

politically unstable countries that frequently fall into a pattern of successive coups, known as the "coup-trap" [33]. The complexity and non-linearity of conflict data challenge traditional linear models in political science, which often fail to capture the nuances of these dynamics [10].

Quantitative studies in international conflict often lack robustness due to their generalizing approach and failure to account for variations across different dyads. Such studies also commonly exhibit selection biases, including the exclusion of both non-war and war cases [10]. The challenge extends beyond identifying potential conflicts to predicting their exact timing, which is notably more difficult [2]. The notion that forecasting new civil wars might have reached its limits has been discussed, with arguments suggesting that the unpredictable nature of violent conflicts poses a significant barrier to prediction [5, 22]. The access to granular geo-referenced data, while reducing aggregation bias, increases the risk of outcome misclassification [14].

Determining the appropriate unit of analysis for conflict studies, especially for events like violence, is a complex matter due to the flexibility in unit of analysis and challenges in achieving the right level of analytical resolution [14]. This is compounded by the often imprecise or inaccurate location information in event data, as exemplified in the UCDP Geo-referenced Event Dataset for Afghanistan [34]. Systematic underreporting and inaccuracies in event characteristics, such as timing and location, present further complications. These issues are exacerbated by inadequate data quality on correlates of events, like economic indicators [35]. Traditional methods of encoding events from news stories are slow, expensive, and ambiguous, underscoring the limitations of human-based approaches [3]. The unprecedented amounts of disaggregated data available today pose their own challenges, with disparate event typologies and units of analysis making it difficult to draw comparable conclusions across different studies [36]. The field has seen limited progress partly due to a focus on using consistently significant variables across studies, often relying on the same databases [37].

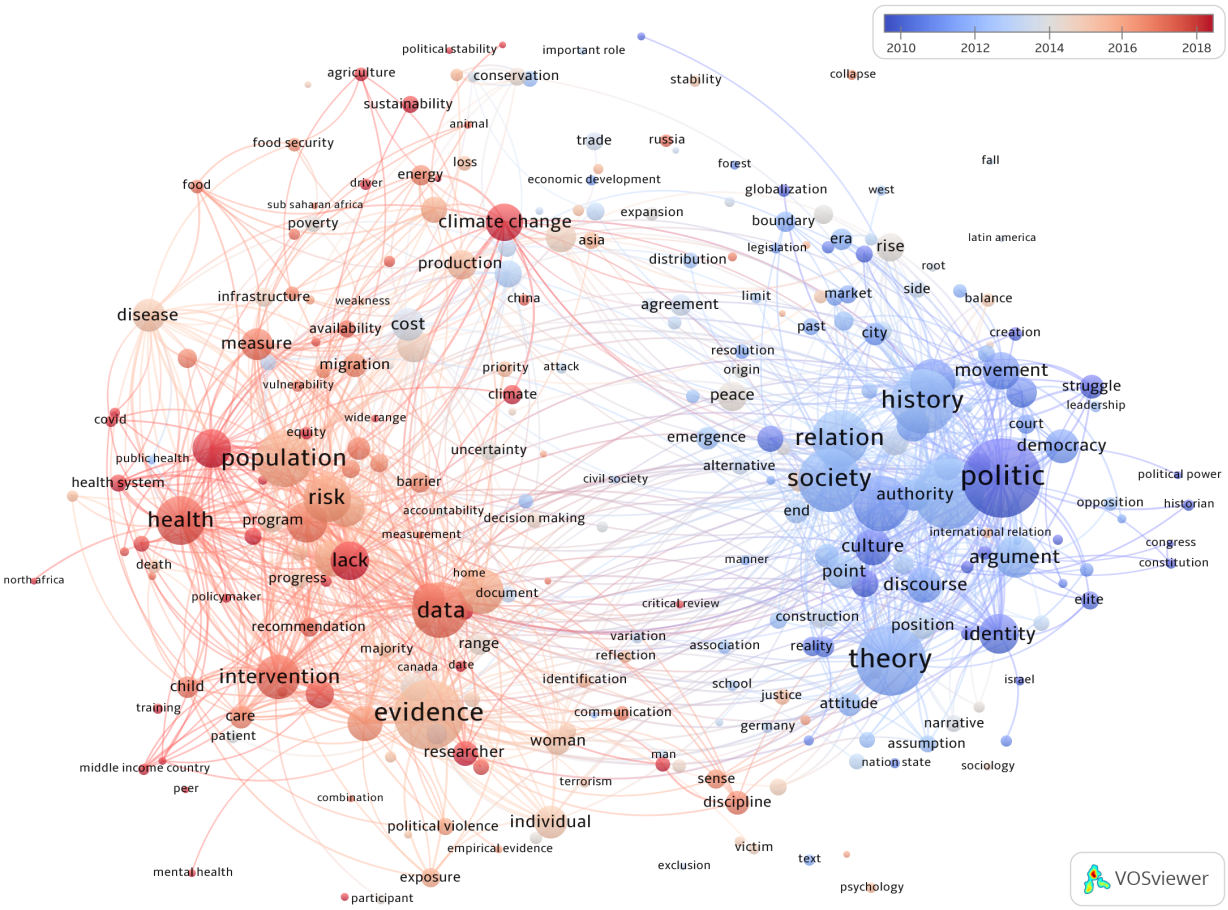
In summary, the pursuit of accurate conflict prediction is challenged by methodological limitations, data complexity, and the inherent unpredictability of conflict dynamics. These issues underscore the need for innovative approaches and the development of more nuanced models to enhance the field's predictive capabilities. Nonetheless, the **relevance of predicting conflicts and the advantages of early awareness** in this domain cannot be overstated. For over two decades, the development of robust conflict early warning systems has been a significant focus in conflict prevention efforts [13]. Despite ongoing debates about the predictive capacity in political studies [38], the potential to save lives through big data-driven forecasts of conflict outbreaks remains a crucial goal. The value of early

conflict knowledge extends far beyond academic inquiry. It plays a critical role in mitigating conflicts, finding diplomatic solutions, and deploying humanitarian aid effectively. An illustrative example is the grain export corridors from Russia to Africa, which underscore the international nature of such efforts. Singer [39] emphasizes that forecasting peace and conflict is a fundamental motivation for peace research.

Reliable forecasting systems could significantly enhance preparedness and intervention strategies [40]. These early warning systems can have varied perspectives and capabilities, ranging from risk assessment to action support, highlighting their importance in reducing humanitarian, economic, and political harms [41]. The historical instance of World War I, where tensions were underestimated until the conflict's outbreak, serves as a poignant reminder of the critical need for early awareness [22]. The efficacy of early predictions in other fields, such as food insecurity, further supports this endeavor. Balashankar et al. [42] and Backer and Billing [27] demonstrate the potential of accurate forecasting in reducing humanitarian costs, with the latter achieving an 84% accuracy in projections across African countries. The former study outlines how novel data sources like news articles can help setup a timely and cost-effective approach to early warning systems. In the same vein, recent advancements in data granularity have been pivotal in enhancing conflict prediction. The shift to more detailed sub-national data enables better understanding of local phenomena [14] and the ability to give more localized predictions. These predictions should not be considered in isolation or as merely a technical exercise. Early warning systems act as direct references for decision makers in the field, feeding data-driven recommendations. With the introduction of more granular data, conflict predictions will be able to link specific historical events to forecasts of future ones [5]. A related development in the field are open-source data repositories, like the xSub data platform, which facilitates comparative sub-national research [36].

The question remains: how do these early warning systems help when widespread conflicts affect populations? The practical implications of early conflict knowledge are substantial. Jahre et al. [43] and Kovács and Spens [44] argue that early localization of needs and swift aid delivery are crucial for effective humanitarian operations. A study on the return on investment of emergency preparedness actions observed significant time and cost savings, highlighting the practical benefits of being prepared [45]. As for governments, early knowledge of potential conflict carpets enables them to deploy resources to stabilize the to-be-affect regions. For the aforementioned reasons, the development and refinement of conflict prediction and early warning systems are not only academically relevant but also carry immense practical and humanitarian significance. These systems have the potential to transform how conflicts are managed and mitigated, ultimately saving lives and reducing the severity of humanitarian crises.

Figure 2.1: Bibliometric Network of Most Relevant Terms in Literature on (Political) Conflicts over Time.



Note: The graph was created with VOSviewer. The approach is described in Appendix A.

2.1.4 Data, Models, and Metrics of Conflict Prediction Research with New Opportunities

The selection and utilization of **data** play a crucial role in conflict prediction research. The field has seen a variety of data sources employed, each contributing unique insights into conflict dynamics. Country-level analyses, typically with a yearly resolution, still form the bulk of conflict prediction research [5]. Historically, efforts have been directed towards improving data and measures of international conflict, adapting statistical models to accommodate conflict data's special features, and deriving models from rational choice theories [10].

The Armed Conflict Location & Event Data Project (ACLED) has emerged as a primary conflict event data source, now supported by a UN-led multinational trust fund, signifying its importance and reliability in the field [13]. While this coding framework is increasingly

adopted in the field, other approaches exist. For instance, the Conflict and Mediation Event Observations (CAMEO) coding scheme, employed by Brandt et al. [46], offers another approach to categorizing conflict events. One of the oldest, continuously available data sources on conflicts is the conflict-cooperation scale for the World Events Interaction Survey (WEIS) events data, a method that dates back to work in the 1970s [47]. Moreover, researchers have utilized a combination of event data based on the Integrated Crisis Early Warning System (ICEWS), achieving notable accuracy in predicting civil war occurrences [48]. Others employ a unique method by relying on keyword counts as data to construct a tension index, used to predict conflict onset weeks in advance [22]. Another approach uses newspaper data, covering several decades and reporting events in all countries, thus capturing even rare occurrences [2].

The quality of sub-national economic activity data, such as that provided by Nordhaus [49], varies significantly within and across countries, posing challenges for researchers seeking consistent and reliable data sources [14]. Population data, such as the Landscan dataset [50], offers high-resolution spatial information, crucial for detailed analyses. Cook and Weidmann [14] highlight the accessibility and value of granular geo-referenced data on social and political phenomena, though they caution that smaller spatial and temporal scales require careful scrutiny and don't always lead to improved inferences.

Most quantitative studies of civil war typically use the country-year as the unit of observation, with common controls like GDP at the country-year level [9]. Other notable data sets in conflict prediction research include UCDP/PRIO [51] and the Correlates of War [52], each contributing different perspectives and data types to the study of conflicts.

In summary, the data employed in conflict prediction research are diverse, ranging from event-specific datasets to broader economic and population data. The evolution of these data sources reflects the field's ongoing efforts to enhance the accuracy and depth of conflict analysis and prediction – with room for novel ideas.

The **methodologies** employing the above data are diverse and continually evolving. As shown in Figure 2.1, until earlier in this century, research on conflicts focused predominantly on theory and history. In recent years, evidence and data-driven studies have become more common. From linear to neural network models, the field adapts to the intricate nature of conflicts, striving to improve accuracy and depth in forecasts. These developments reflect the ongoing commitment to enhancing the predictive capabilities to the complex nature of conflicts.

Forecasting in conflict research is defined as making predictions about unrealized conflict occurrences based on model estimates from realized data, essentially assigning probability distributions to both realized and unrealized outcomes [4]. Initially, conflict prediction relied

heavily on linear models, which gradually proved inadequate in capturing the complexities and varying effects of conflicts. This led to a shift towards more sophisticated neural network models, although this transition sometimes resulted in a loss of interpretability [5]. Neural network models, first introduced to political science in the 1990s [53, 54], have been particularly beneficial for conflict research. These models are capable of allowing the effects of each explanatory variable to differ significantly over dyads, matching the requirements of international conflict analysis [10]. However, the author cautions against exclusively relying on neural network models, suggesting they should complement rather than replace logit models in quantitative studies of international conflict. Other studies have aimed to employ machine learning to distill vast quantities of information into interpretable topics, which are then used in panel regression to predict the onset of conflict [2].

While predictions at the country-year level are still common, modeling approaches increasingly aim to predict on daily, weekly, or monthly resolutions. This shift has been instrumental in pushing the discipline forward, providing more granular insights into conflict dynamics [4, 38]. Another significant area of discussion around modeling choices has been the trade-off between local and regional models [13, 14]. Through simulations, Cook and Weidmann [14] explore how the variance of observations at these different levels influences the choice of an appropriate model. Notwithstanding the individual efforts, the field has also experimented with new collaboration approaches. As such, prediction competitions organized by the ViEWS research team provide a platform for comparing different modeling approaches in forecasting changes in state-based violence. These competitions assess how various models perform over time and at different levels of analysis, and whether collective insights, or the 'wisdom of the crowd,' can offer new perspectives [55].

The evaluation of these models is critical to understanding their effectiveness and improving their accuracy. Various **performance metrics** have been developed and utilized to assess the performance of these models.

One of the key challenges is identifying the most suitable set of performance metrics to effectively capture the multiple dimensions of model success, especially in complex processes like changes in violence [55]. Distinct scores that reward precision, accuracy, innovation, or diversity of contributions are essential in this context. Out-of-sample forecast accuracy is considered the gold standard for model assessment in conflict prediction research [10]. This approach emphasizes the importance of a model's ability to predict new, unseen data accurately, rather than just fitting the model to existing data. Cederman and Weidmann [5] note that focusing on out-of-sample prediction is beneficial as it helps prevent the inclusion of too many explanatory factors, which can potentially degrade a model's predictive performance.

However, several other metrics are currently also being considered to score models in

conflict prediction research. The Mean Squared Error (MSE) is a common metric; however, it is significantly influenced by the proportion of zeros in highly imbalanced datasets, a common issue in conflict data. While MSE provides an indication of how far predictions are from the observed values, it does not necessarily reflect the model's ability to predict specific events like (de-)escalation accurately [55]. Targeted Absolute Distance with Direction Augmentation (TADDA) is a newly developed metric, specifically tailored for changes in fatalities, an outcome of interest in conflict prediction. TADDA could be applied across various research domains, offering a more targeted approach to evaluating model performance. Another innovative metric is the pseudo-Earth Mover Divergence (pEMDiv), which is network-based and context-sensitive. It incorporates Earth Mover Distance calculations while allowing for mis-calibrated predictions and asymmetric distance calculations. This metric addresses the challenge of incorporating asymmetric movement costs and rewards models that produce temporally and spatially approximate predictions, even if they do not precisely hit the target [55].

It becomes apparent that the field continues to grapple with unresolved challenges, in terms of modeling [55], data, and metrics. Moreover, there is a pressing need for improved indicators that can more timely signal impending conflicts, a gap that novel modeling techniques hold promise in addressing [2]. This confluence of challenges and opportunities leads to the chapter's core query.

Research Question of Chapter 2. *How, if so, can natural language processing methods be effectively utilized to develop more timely indicators of crises?*

2.2 Methodology

2.2.1 Data

This project utilizes two primary data sources: news articles and FrameNet data.

News Articles

This section capitalizes on the premise that risk factors triggering a conflict, such as food crises, are frequently mentioned in on-the-ground news reports before being reflected in traditional risk indicators, which can often be incomplete, delayed, or outdated [22, 42]. By harnessing newspaper articles as a key data source, this initiative aims to identify these causal precursors more timely and accurately than conventional methods.

News articles represent a valuable data source, particularly in research domains where timely and detailed information is crucial. In contrast to another "live" data source that

currently revels in popularity amongst researchers – social media data – news articles are arguably less prone to unverified narratives. While news articles typically undergo editorial checks and balances, ensuring a certain level of reliability and credibility, they certainly do not withstand all potential biases and are to be handled with caution. To counteract potential biases of individual news outputs, accessing a diverse range of news sources is essential. Rather than having to scrape or otherwise collect data on news articles, there is a set of resources available:

- NewsAPI[1]: This platform provides convenient access to a daily limit of 100 articles, offering diverse query options. Its integration with a Python library streamlines the process of data retrieval. However, the limitation lies in the relatively small number of data points it offers, potentially restricting the scope of analysis.

- GDELT Database[2]: Renowned for its vast repository of historical information spanning several decades, GDELT stands as a comprehensive data source. Its extensive database is a significant asset, but similar to NewsAPI, it predominantly features article summaries or initial sentences rather than complete texts, which may limit the depth of analysis.

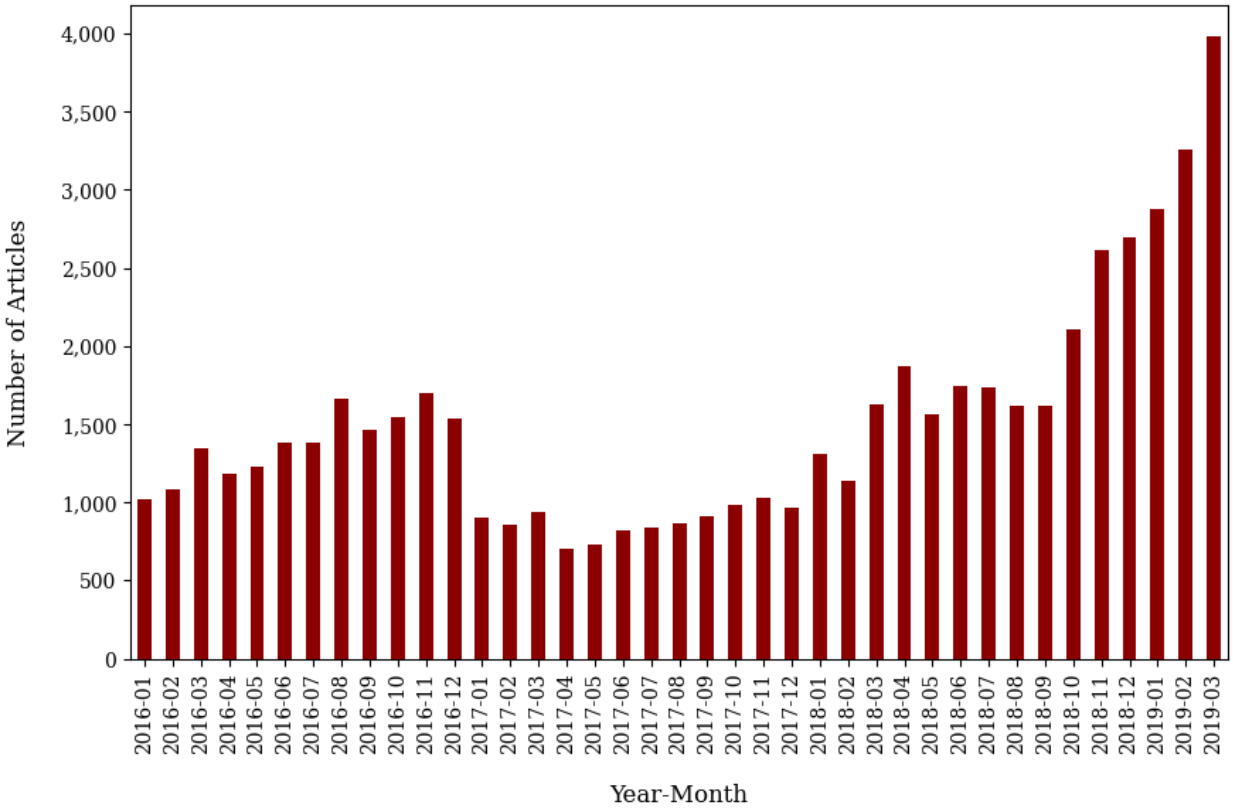
- Factiva[3]: A premium service that grants access to the complete bodies of articles from a plethora of global news sources in multiple languages. While offering an exhaustive depth of data, this resource comes with associated costs, which may be a consideration for budget-constrained projects.

- RealNews[4]: As a cost-free alternative, this dataset encompasses entire newspaper articles collated between 2016 and 2019. Selected for this project due to its unrestricted accessibility and comprehensive nature, it provides a substantial set of articles, making it a valuable resource for in-depth analysis.

Given the outlined conditions, I choose the open-source RealNews dataset, as presented by Zellers et al. [56]. As this chapter outlines the proof-of-concept of utilizing an NLP approach to detecting and monitoring conflict precursors, the modeling is done on a randomly selected subset of 120,821 news articles. Each article entry in this subset provides a rich array of information, including *url*, *url_used*, *title*, *text*, *summary*, *authors*, *publish_date*, *domain*, *warc_date*, and *status*.

[1]NewsAPI: https://newsapi.org/, last accessed January 4th, 2024.
[2]GDELT Database: https://www.gdeltproject.org/, last accessed January 4th, 2024.
[3]Factiva: https://www.dowjones.com/professional/factiva/, last accessed January 4th, 2024.
[4]RealNews: https://paperswithcode.com/dataset/realnews, last accessed January 4th, 2024.

Figure 2.2: News Articles over Time.

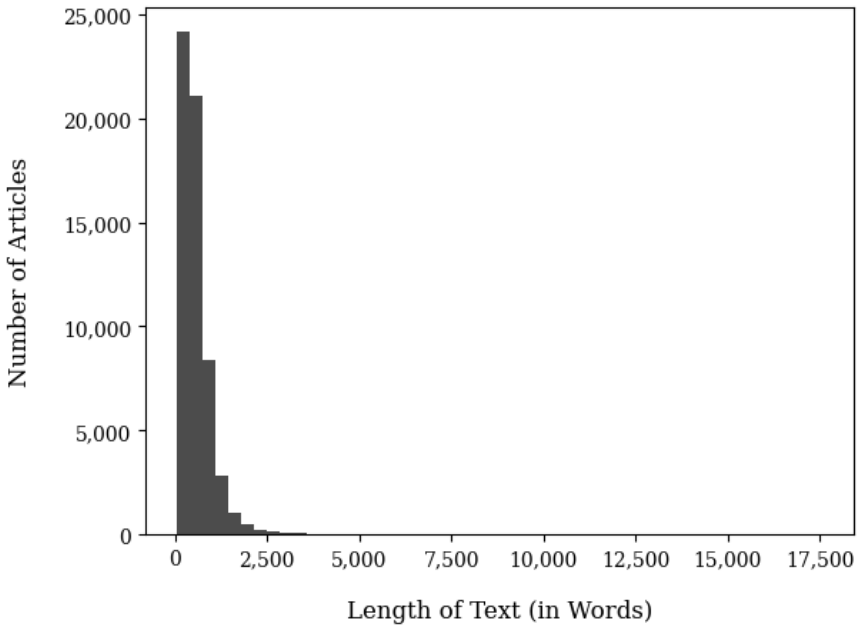


The range of articles spans from 1869 to 2019, but for focused analysis, I narrow the scope to articles from January 2016 through March 2019, which is the end of the currently available version of the RealNews dataset [56]. This temporal delimitation results in a dataset comprising 58,867 articles. These articles originated from an expansive pool of 493 distinct news outlets, offering a broad perspective on global events and narratives. The distribution of these articles across the specified time frame provides the expected observation of increasing news reporting, as visualized in Figure 2.2. The longest (shortest) news article of the dataset has 17,531 (67) words, with a mean of 614 words per news article (see Figure 2.3).

FrameNet Data

As I describe in further detail in Section 2.2.2, I leverage the FrameNet data, an extensive linguistic resource developed by the International Computer Science Institute at the University of California, Berkeley [57]. FrameNet stands as a cornerstone in computational linguistics, offering a richly annotated corpus grounded in frame semantics theory. The core concept of frame semantics is that the meanings of most words can be best understood based on

Figure 2.3: Distribution of News Articles Length.



the semantic frames or conceptual structures they evoke. The FrameNet data is structured around these semantic frames, which are essentially schematic representations of situations involving various participants, objects, and events. Each frame in the database encapsulates a specific type of event, relation, or entity, along with its associated elements (roles played by participants in the event) and lexical units (words or phrases evocative of the frame).

At its core, the FrameNet project provides two key components: (1) the FrameNet lexicon, which details the relations between lexical units and frames, and (2) the annotated corpus, where instances of these lexical units in running text are linked to their respective frames and roles. This annotation is meticulously carried out by linguistic experts, ensuring a high degree of accuracy and reliability [42, 57, 58].

The FrameNet dataset is invaluable for numerous NLP applications, including but not limited to semantic role labeling, information extraction, and text understanding. Its structured approach to semantics makes it an ideal resource for tasks requiring a deep understanding of context, thematic roles, and conceptual relationships in language. In my case, FrameNet serves as an essential tool, providing a nuanced and structured framework for analyzing and interpreting the semantic dimensions of text data with various documented cause-effect relations which I am attempting to identify with regards to conflicts.

2.2.2 Modeling

In order to detect causal precursors of conflicts, I broadly perform three modeling steps. First, I identify semantically proximate words to conflict as potential seed words for the later analysis. Second, I train and utilize a frame-semantic parser to detect cause-effect relations in the news article data presented in Section 2.2.1. Lastly, I utilize the seed words proximate to the term conflict (like war or battle) to validate which cause-effect relations are relevant to the study of conflicts.

Identifying Seed Words via Semantic Similarity

The preparatory step of identifying semantically similar words to the term *conflict* addresses the concern that – given linguistic diversity – there are various words and ways to speak about (violent) conflicts. I begin with a set of initial key phrases *conflict, war, and battle*, ensuring they capture the core essence of the thematic domain. To obtain a list of candidate words and phrases, I collect all unique uni-grams, bi-grams, and tri-grams in the news article dataset, resulting in 565,202 candidates.

Then, I leverage pre-trained word embeddings from the *Gensim* library, *glove-wiki-gigaword-100*, to map the three initial key phrases and all candidate seeds into a high-dimensional semantic vector space. I also experiment with more sophisticated embedding approaches (like transformer-based models) to compute the semantic similarity and thus obtain the seeds. When trading off complexity and time with performance, the simpler pre-trained *Gensim* embeddings perseveres.

I compute the cosine similarity of the candidates' embeddings to measure the semantic proximity of these terms to the initial seed phrases. This process involves averaging the distances or similarities across the seed phrases for each candidate, as illustrated in Equation 2.1. $\mathbf{A}$ and $\mathbf{B}$ are n-dimensional embedding vectors generated with the *Gensim* library. A_i and B_i are the i-th components of the embedding vectors $\mathbf{A}$ and $\mathbf{B}$ respectively.

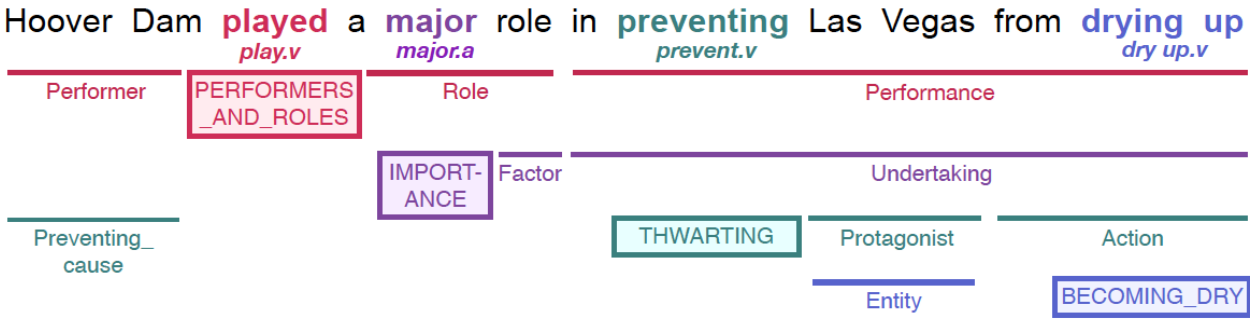
$$\text{cosine similarity } = S_C(\mathbf{A}, \mathbf{B}) := \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|_2 \|\mathbf{B}\|_2} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \cdot \sqrt{\sum_{i=1}^{n} B_i^2}} \tag{2.1}$$

All candidate seeds are then ranked based on their cosine similarity with the key phrases. The top candidates are manually validated to ensure they are not only semantically similar but also synonymous in meaning. By identifying the set of seed phrases, I am able to broaden the analysis to include a diverse array of expressions and viewpoints related to the term *conflict*.

The Frame-Semantic Parser

In the pursuit of bridging the gap between the robust theoretical understanding of conflict dynamics and the practical challenges in data availability, the frame-semantic parser emerges as a promising methodological tool. In a recent study [42], a team of researchers established a proof-of-concept via its successful application of a frame-semantic parser for the study of food insecurity - a field with similar challenges surrounding data access and quality. While this study relied on what can now be considered the "old state-of-the-art," my proposed approach diverges towards a more contemporary, transformer-based model, inspired by the advancements outlined by Chanin [59].

Figure 2.4: Illustration of a Frame-Semantic Parser as presented by Swayamdipta et al. [58].



At the heart of frame-semantic parsing, as conceptualized by Gildea and Juravsky [60] and formalized by the FrameNet project [57], is the identification of structured semantic frames and their arguments from natural language text. As illustrated in Figure 2.4, these frames encapsulate events, relations, or situations along with their participants, making it a critical tool in natural language understanding (NLU) tasks. The practical applications of frame semantics are broad, ranging from voice assistants and dialog systems [61] to complex text analysis [62].

The process of frame-semantic parsing constitutes three subtasks:

- **Trigger Identification**: This initial step involves pinpointing locations in a sentence that could potentially evoke a frame. It is a foundational task that sets the stage for the next components.
- **Frame Classification**: Following trigger identification, each potential trigger is analyzed to classify the specific FrameNet frame it references. This task is facilitated by leveraging lexical units (LUs) from FrameNet.

- **Argument Extraction**: The final task involves identifying the frame elements and their corresponding arguments within the text. This process adds depth to the frame by fleshing out its components and contextualizing its application within the sentence.

While frame-semantic parsers have arguably not received as much attention as other language modeling methods, three major contributions of the past few years can be highlighted. Swayamdipta et al.'s [58] approach – which is still outperforming many other implementations – presented an efficient parser with softmax-margin segmental RNNs and a syntactic scaffold, *SegRNN*. It demonstrates that syntax, while beneficial, is not a necessity for high-performance frame-semantic parsing. Kalyanpur et al. [63] explore the application of transformer-based architectures to frame-semantic parsing, employing a multi-task learning approach that improves upon previous state-of-the-art results. Most recently, Chanin [59] developed the first open-source approach which treats frame-semantic parsing as a sequence-to-sequence text generation task, utilizing a T5 transformer model. It emphasizes the importance of pre-training on related datasets and employing data augmentations for improved performance. The distinctive strength of a frame-semantic parser lies in its ability to contextualize information, rather than interpreting it in isolation. This feature is particularly invaluable in conflict analysis, where the semantics of discourse play a critical role.

Applying the Frame-Semantic Parser to Identify Cause and Effect Frames

The implementation of my frame-semantic parser involves several key steps. I begin by splitting the news article texts into sentences, resulting in a total of 1.6 million sentences. This granular approach allows me to focus on individual narrative elements within the articles and since frame-semantic parsers are reported to perform better on sentence-level [58, 59]. To reduce potentially unnecessary computational load, I further filter the set of sentences down to those containing at least 5 words and a maximum of 20 words (resulting in 826,209 sentences). Here, I make the assumptions that shorter sentences are unlikely to contain proper cause-effect relations, and that the model will have difficulty of accurately dissecting cause-effect relations of very long sentences. I acknowledge that this heuristic potentially introduces limitations to my study and that future iterations of this work should likewise evaluate particularly long sentences.

To determine the right frame-semantic parsing model for downstream analysis, I evaluate the performance of different models. I compare the *Softmax-margin SegRNN* approach by Swayamdipta et al. [58], the vanilla implementation of Chanin's [59] transformer model, and the transformer-based solution with tuned hyperparameters. I evaluate all models against the FrameNet benchmark dataset using its pre-labeled train-development-test split. I eval-

uate the models with the F1-score as performance metric, as common for frame-semantic parsing [58, 59, 63]. Additionally, I evaluate the tuned transformer-based model on a subset of the FrameNet data, which contains any of the previously determined conflict-related seed phrase – to investigate how the generically trained parsers perform on the relevant domain of study.

The best performing model will then be used to extract and filter relevant frames from the set of sentences. The extraction step captures the full text of each frame element. The filtering step then refines this data, focusing on specific frames related to causality as I am interested in these frames and not just any semantic relations. I rely on both manually identified frame names (informed by Vieu et al. [64] and Vieu [65]) and pattern-based searches in FrameNet to compile a comprehensive list of 37 relevant frames (see Appendix C for full list of frames).

To optimize the performance of the parser, I operate in batches of 16 sentences at a time. The implementation is designed to be efficient and scalable, processing batches of sentences and extracting the most relevant semantic frames. I utilize one GPU and 51 GB in system RAM for all computations. This approach enables me to parse and analyze a substantial corpus of news articles, providing a rich dataset for conflict analysis.

Lastly, I validate which cause-effect relations identified by the frame-semantic parser are relevant to the study of conflicts. News articles deal with many themes and topics that contain causal relations but are not relevant to this study. Thus, I reference the previously identified seed words semantically proximate to the term conflict and filter all identified cause-effect pairs out that do not contain a seed word in their respective effect frames.

2.3 Results and Findings

2.3.1 Identified Conflict-Specific Seed Phrases

The distribution of semantic similarity scores of the candidate seeds with the three initial key phrases are displayed in Figure 2.5. One can observe the typical S-curve shape of a cumulative distribution function (CDF), with the steepness of the curve in the middle range indicating a concentration of candidates semantic similarity around the median. As the observations are more spread out towards the top of the plot, one can visually observe that a smaller set of candidate seeds scores a higher similarity, which is to be expected as I considered a large number of uni-grams, bi-grams, and tri-grams as candidates.

I examine the top 50 candidate seeds further. Potentially driven by the fact that conflicts are usually related to specific geographic (or temporal) labels and information, a number of

Table 2.1: Alphabetical List of Semantically Most Similar, Conflict-Related Seed Words.

Candidate Seeds		
action	armed	army
attack	attacks	battle
battles	bloodshed	bloody
civil	combat	conflict
conflicts	confrontation	crisis
fight	fighting	force
forces	fought	hostilities
invasion	issue	military
occupation	peace	political
soldiers	struggle	troops
violence	war	warfare
wars		

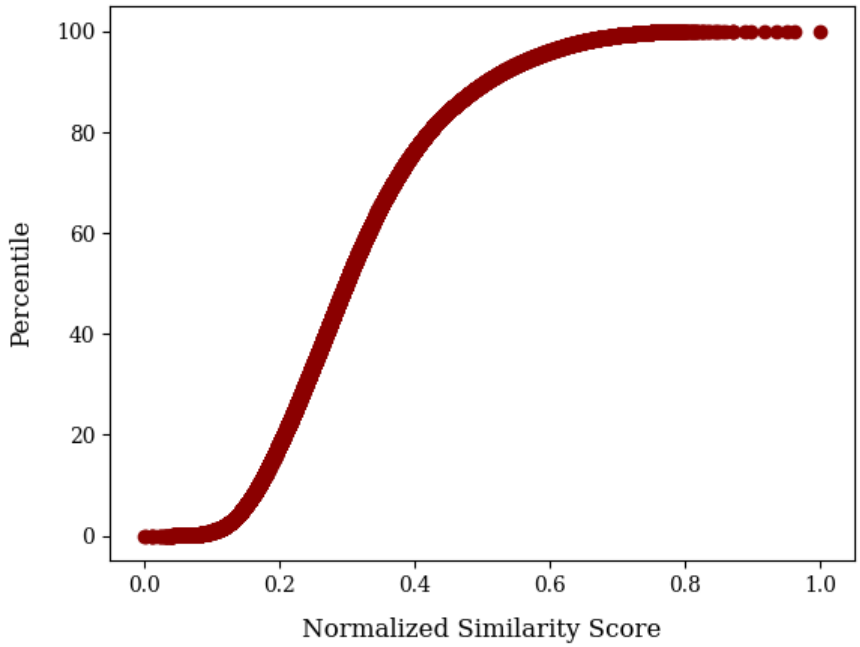
the top 50 most semantically similar phrases are geographic terms like *Afghanistan*. Since I do not want to introduce geographic biases for the downstream use of the seed phrases, I exclude these candidates and continue the analysis with 34 seeds – a comprehensive set of terms representing the different nuances in which news might refer to conflicts (shown in Table 2.1. A list of all top 50 candidate seeds can be found in Appendix C.

The results of the semantic similarity modeling also suggest that this methods is applicable to a wide array of tasks and domains. Beyond the realm of conflicts, the idea of identifying semantically proximate words and phrases can help address the nuances and ambiguities of language – where many words and expressions can be chosen to speak about the same topic. For instance, semantic similarity could potentially be a valuable support for information retrieval tasks by enhancing the relevance and coverage of search results.

2.3.2 Model Performance on Benchmarks

In line with previous studies, the *Softmax-margin SegRNN* by Swayamdipta et al. [58] performs very well across all three sub-tasks of the frame-semantic parsing process (as shown in Figure 2.6). This baseline model performs best in the frame classification task on the development set with an F1-score of **0.8974**, slightly diminishing to **0.8655** on the test set. This trend of marginal performance drop from development to test sets is also observed in the argument extraction and trigger identification tasks, indicating a consistent model

Figure 2.5: Percentile Plot of Candidate Seeds' Normalized Semantic Similarity to Conflict.



Note: The semantic similarity score was min-max normalized.

generalizability.

The untuned, vanilla implementation, as presented by Chanin [59], marks a noteworthy performance shift, particularly in the argument extraction task, where it outperforms the baseline with an F1-score of **0.6692** on the development set and a stable transition to the test set with a score of **0.6694**. Despite this advancement, the untuned implementation trails behind the baseline in the frame classification and trigger identification tasks. This divergence in performance could potentially be attributed to the inherent architectural differences, where transformer-based models leverage global dependencies, potentially offering a richer contextual understanding conducive to argument extraction but less so for the granular detection required in trigger identification.

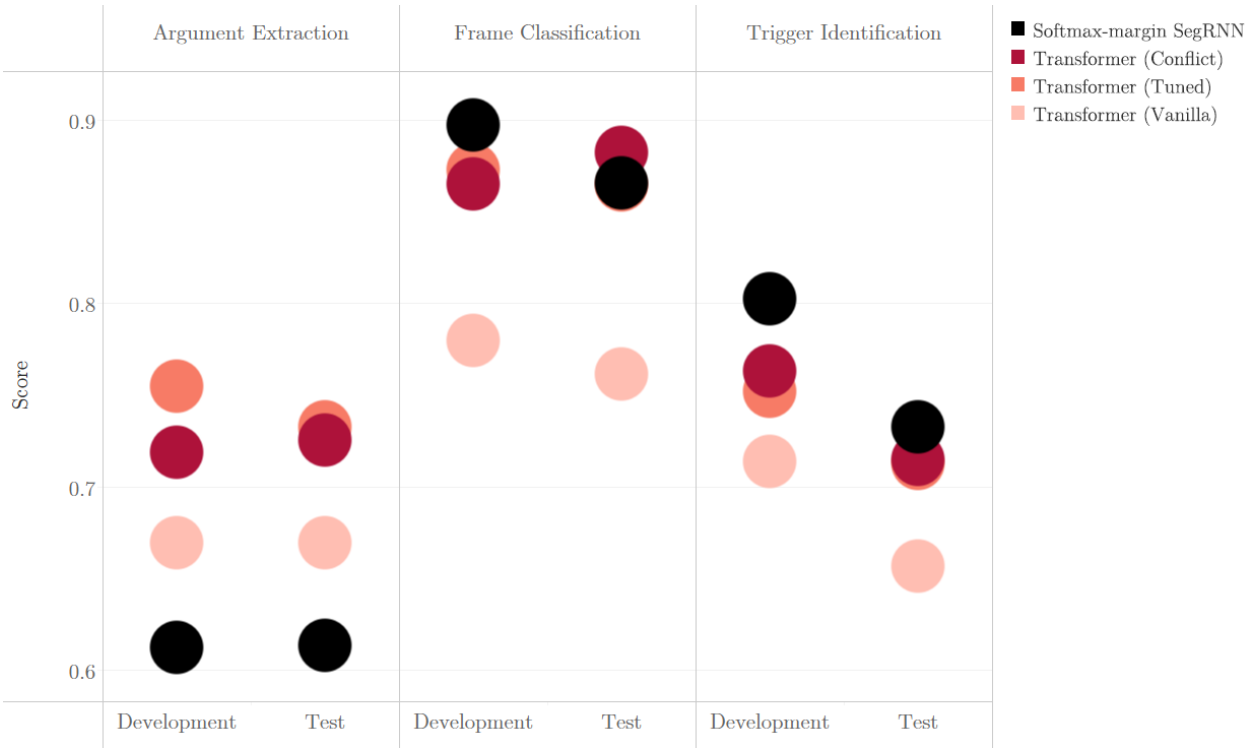
Next, the transformer-based implementation with tuned hyperparameters results in a substantial enhancement across all tasks, notably yielding an F1-score of **0.755** in the argument extraction task on the development set. This tuned model demonstrates robustness, retaining high performance on test data with scores of **0.7324**, **0.8644**, and **0.7127** for the argument extraction, frame classification, and trigger identification tasks respectively. Thereby, it is overall the best-performing model.

Lastly, the domain-specific performance of the generic transformer-based frame-semantic parser is tested in a specialized assessment on conflict-related test data. It sustains its efficacy and thereby verifies the adaptability of the model to domain-specific tasks. This finding

alleviates concerns regarding potential biases in the training data derived from FrameNet, which could have been skewed towards commercial or non-political content. The F1-score on conflict-specific data remains at **0.7253** for the argument extraction task and at **0.8824** for the frame classification task.

The empirical evidence points towards the tuned transformer-based model as the superior choice for my purposes. It outperformed the well-regarded *SegRNN* implementation, and not only maintained high performance across standard datasets but also showed no significant deviation when applied to conflict-specific content. This paves the way for its application in the subsequent phase of my analysis.

Figure 2.6: Performance of Frame-Semantic Parser Models Across the Three Sub-Tasks.



Note: Softmax-margin SegRNN as presented by Swayamdipta et al. [58]. All transformer models based on work by Chanin [59].

2.3.3 Efficacy of Frame-Semantic Parsing and Next Steps

The Value-Add of Frame-Semantic Parsing

The selected transformer-based frame-semantic parser, as determined from prior comparative analyses, is then utilized to elucidate causal connections in the conflict space – marking a critical transition from model selection to practical application.

The analysis yields compelling results. From the previously prepared corpus of 826,209 sentences, the parser identified 1,376 sentences that contained a cause-effect relationship with a conflict-related seed phrase situated within the effect frame. This discovery is significant, as it highlights the model's capacity to detect nuanced causal structures within large datasets. Furthermore, within these frames, I was able to identify 163 unique precursors, a detailed enumeration of which is presented in the Appendix C. To provide a glimpse into the nature of the identified sentences and precursors, refer to Figure 2.7, which showcases representative examples.

A critical component of this analysis is understanding the nature and thematic distribution of the identified precursors. As depicted in Figure 2.8, a notable subset of precursors pertains to various military actors – such as soldiers, army, gangs, militias, rebels, and insurgents. This finding aligns well with the understanding that reports on militarization or troop movements are potent indicators of escalating tensions. Furthermore, a related cluster of precursors revolves around the weaponization of different actors, exemplified by terms like missiles, guns, or tanks. This again points to the anticipatory nature of conflict, where preparations for potential confrontations are a tell-tale sign.

Beyond the military aspect, the findings also unveil precursors related to civil and political movements, including protests, protesters, riots, and rallies. These findings indicate the significance of civic actions as potential harbingers of conflict. Another intriguing set of precursors relates to societal developments that might incite tensions, such as propaganda, dictatorships, misinformation, or general societal discontent.

Interestingly, the analysis also brought to light a group of precursors related to humanitarian crises, including displacement, aid, famine, hunger, and viruses. The emergence of this category underscores the multifaceted nature of conflict precursors, extending beyond mere political or military dimensions to encompass broader societal and humanitarian factors. It also exemplifies the interconnectedness of human-made and nature-made crises.

The ability to identify such a diverse and relevant array of precursors from a comparatively modest sample of news articles underscores the efficacy of employing a frame-semantic parser to identify conflict precursor.

Practicalities of the Frame-Semantic Parser

As outlined earlier, this chapter was aimed at foundational research to explore and validate the potential utilization of a frame-semantic parser to identify causal relations within narratives on (violent) conflicts. My proof of concept has not only demonstrated the feasibility of such an approach but also has laid a robust groundwork for future endeavors in utilizing its outcomes for downstream tasks such as conflict prediction. Researchers in this field can now

Figure 2.7: Example Sentences with Identified Conflict Precursors.

Example 1: "Sirajul Haq said the presence of foreign troops in Afghanistan was the basic reason of instability in the region."

Example 2: "I don't think society understands enough about the role of propaganda and violent speech in provoking actual violence."

Example 3: "In fact, militarization makes police more likely to turn to violence to solve problems."

Note: The quotes identified foreign troops, propaganda and violent speech, and militarization as precursors.

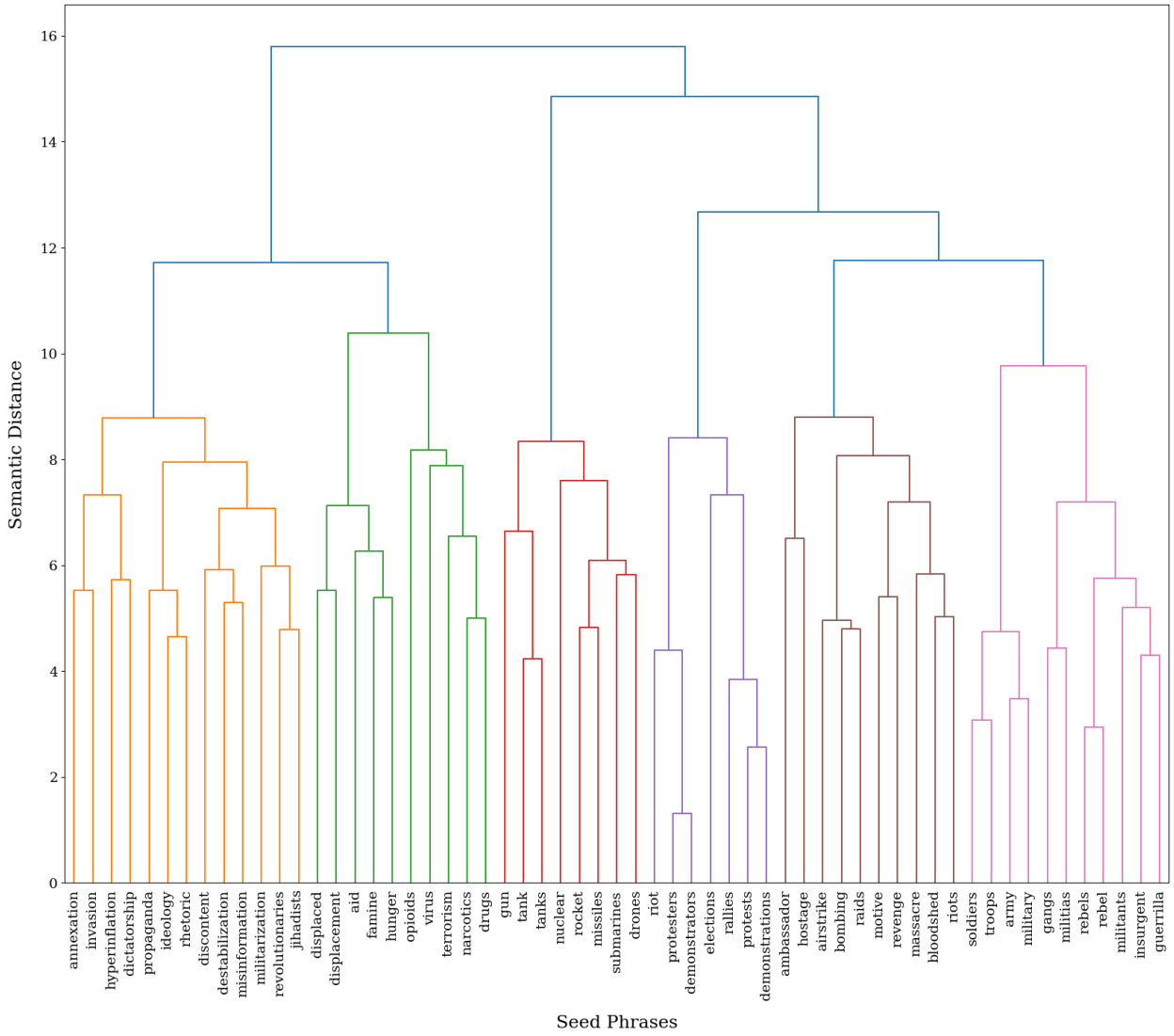
leverage this methodology to augment their predictive models with a nuanced understanding of causal precursors, thus enriching the analytical depth of their predictive features.

Practically, this advancement implies that researchers can periodically re-apply the process delineated in this study to update and refine their selection of causal precursors. Post identification of the precursors, researchers have the option to further substantiate their relevance through methods such as the Granger causality test, thereby reinforcing the empirical validity of these indicators.

Importantly, for the day-to-day applications in conflict prediction, the need to continuously rerun the entire precursor identification process is not required. Instead, researchers can focus on computing weighted counts of the identified precursors in news articles, thereby monitoring their signal. This approach enables a dynamic monitoring and allows for timely updates of features. Moreover, by integrating keyword searches for both the precursors and geographic markers – such as district or county names – researchers can achieve a high degree of granularity in their predictive features. This technique has been successfully employed in other areas, such as food insecurity predictions, as demonstrated by Balashankar et al. [42]

The inherent adaptability of this approach is one of its most salient benefits. It can be seamlessly integrated into a variety of modeling frameworks and is compatible with diverse target variables that characterize the broad spectrum of conflict prediction research outlined in Section 2.1. The versatility of this methodology not only facilitates its application across different contexts but also encourages innovative adaptations tailored to research in other thematic domains.

Figure 2.8: Hierarchical Clustering of Identified Conflict Precursors.



Note: This dendrogram visualizes the hierarchical cluster obtained with the scipy library. The Ward's method was employed for the clustering which minimizes the total within-cluster variance. It was chosen as it tends to create clusters of similar sizes that are well-balanced.

2.4 Key Takeaways on Identifying Conflict Precursors

2.4.1 Policy Implications

This chapter aids the policy work on conflict management and resolution by offering an advancement for Early Warning Systems (EWS), which are becoming increasingly utilized in this space. The general value of these systems lies in their ability to provide timely alerts about the potential onset of conflicts, thereby enabling precautionary actions by policymak-

ers. Such anticipatory actions include the formulation of targeted intervention strategies, including preemptive peace talks. Policymakers, equipped with data-driven insights, can craft more effective communication and negotiation strategies. Furthermore, this technology aids in the strategic allocation of resources. By predicting potential hotspots of conflict, resources such as humanitarian aid and peacekeeping efforts can be directed more efficiently, ensuring maximum impact where it is most needed [45].

Despite these benefits, it is crucial to acknowledge the inherent limitations and risks of quantitative models. Firstly, while this project represents an advancement towards granular and explainable approaches, predictions – or numbers in simpler terms – do not necessarily capture the complexities and diversities inherent in conflicts. As such, they should be viewed as indicators rather than definitive outcomes. Secondly, there is the inherent risk of biases in machine learning models. The data, from which they learn patterns, may carry societal biases, inadvertently reflecting and reinforcing existing prejudices and perspectives. Contested topics that our society grapples with right now could potentially be perpetuated through seemingly objective data. A particularly delicate concern arises from the fear of self-fulfilling prophecies of conflict predictions. The dissemination of causal precursors and the prediction about an imminent conflict might inadvertently escalate tensions, leading to the very conflict the prediction sought to prevent. This risk underscores the need for careful management of predictive information and its dissemination.

Addressing these challenges necessitates a strong commitment to ethical considerations and responsible use of predictive technologies. It is imperative to continuously monitor the prevalence of biases in the quantitative study of a topic that quite literally deals with life and death. Ethical use also involves considering the potential impact of predictions on affected communities and the broader societal implications.

2.4.2 Limitations and Future Research

Despite the promising advances of the study of (violent) conflicts via natural language processing methods, my study is not without its limitations, which pave the way for future research opportunities.

The efficacy of the frame-semantic parser to identifying causal precursors to conflicts is inherently contingent on the **data dependency** of the input news article dataset. The quality and geo-temporal diversity of the dataset are critical, as biases in media reporting or constraints in the breadth of covered articles may introduce skewness in my analysis, potentially impairing the ability to identify and monitor precusors over time. An extended iteration of this project, given adequate funding, might benefit from leveraging the compre-

hensive Factiva dataset presented in Section 2.2.1, which could provide a richer and more varied corpus for analysis.

Additionally, while I showcased that the domain-inspecific implementation of a transformer-based parser performed comparably well on a niche domain as the study of (violent) conflicts, in future iterations of this project I would explore approaches to **tuning the model** further. Due to the nature of frame-semantic parsing, the required data is relatively complex in structure. However, machine learning concepts such as active learning, which require limited data, have shown great promises and are a noteworthy opportunity.

The method's success in conflict discourse analysis invites the question of its **applicability to other domains**, particularly those suffering from the same constraints in data points and indicators. To ascertain the versatility of my approach, further research is necessitated to apply and evaluate the methodology across disparate thematic domains.

While the frame-semantic parser satisfies the goal of an interpretable and explainable modeling solution, the current architectures are relatively complex. These inherent complexities can pose challenges in terms of scaling and deployment. Future work could focus on simplifying these models without compromising their performance - for instance via pruning and quantization.

My successfully pioneered use of **semantic similarity modeling** for the curation of domain-specific seed phrases allowed this study to capture linguistic diversity and structural nuances. However, there exists potential for advancement. The implementation of more sophisticated semantic analysis mechanisms could further refine the selection of seed phrases. Although the alternatives – like *BERT*-based models – that I explored did not confer substantial enhancements, with the continuous influx of new models, this area remains ripe for exploration.

Another prospect lies in the **integration of additional data sources**. Augmenting the current dataset with diverse forms of information and news propagation, such as social media posts, governmental reports, or scholarly articles, could provide an ever more holistic view of conflict narratives and their causal relations.

In summary, this chapter presents a substantial contribution to the confluence of natural language processing and the study of conflicts. By surmounting the present limitations and furthering the groundwork laid by this research, future studies can enhance and extend the capabilities of semantic analysis in this dynamic and consequential domain.

Chapter 3

Social Media Communication of Political Leaders in the Face of Adversity

3.1 Theory

3.1.1 The Relevance of Studying Political Communication

In the digital era, the proliferation of data and its integration into various aspects of society is undeniable, revolutionizing how information is disseminated and consumed. The rapid evolution of technology has brought about a significant surge in internet accessibility worldwide, particularly in developing nations [19]. This widespread adoption of the internet across the globe has mirrored societies digitally, effectively creating "digital twins" of communities and populations worldwide. A major component of this development has been the explosive growth of social media usage [66]. Social media platforms – whether Facebook, WhatsApp, or Twitter – have opened up a new venue of communication both for regular citizens and people of the public eye. The consequences could have not been any starker: while social media platforms are oftentimes credited to promote the ease of organizing democratic movements and activists [67], they may have also facilitated state-coordinated misinformation campaigns in Myanmar inciting the genocide of the Rohingya minority [66, 68]. It becomes apparent that the communication of political actors on social media platforms is a relevant angle to understanding conflicts and crises.

Social media is an increasingly important communication channel for political leaders when wanting to engage with their citizens [69]. By 2020, leaders from at least 163 countries had a personal presence on Twitter, with many leveraging the platform for pivotal announcements about global events such as the COVID-19 pandemic [70]. Social media also represents a uniquely direct form of communication between leaders and their citizens. Un-

like press releases and even televised speeches, social media communication is relatively well insulated from the biases associated with media coverage and commentary. Political leaders can determine exactly what, when, and how frequently they want to post on social media platforms without worrying about the agendas of traditional media outlets, thereby forging a direct connection with citizens.

However, there is reason to believe that the utilization of social media in politics varies amongst leaders of different types of political systems. Leaders in democratic and authoritarian regimes face distinct political pressures and cater to different audiences, reflecting in the way they disseminate information, also via social media. In democratic nations, leaders predominantly utilize social media as a channel to disseminate information directly to the general populace. This approach underscores the participatory nature of these societies, where leaders are directly accountable to their citizens. Conversely, in less democratic regimes, the use of social media often serves as a tool for reinforcing regime narratives and interfacing with select elite groups. Here, the focus is less on widespread public engagement and more on consolidating support within specific power structures [19].

I undertake a comprehensive analysis of social media usage patterns among 13 African political leaders by examining 19,518 tweets sourced from their official Twitter accounts, spanning the period from January 2018 to December 2021. This investigation aims to uncover the variances in the use of social media across different governmental regimes. To achieve this, I assess the frequency of leaders' tweets, offering insights into the contrasting communication frameworks. Additionally, I delve into the sentiment conveyed in these tweets by applying an advanced natural language processing (NLP) technique specifically designed for Twitter, as developed by Barbieri et al. [71]. My findings reveal a notable disparity in the Twitter activities of leaders from more democratic nations compared to those from less democratic ones.

Contrary to my expectations and the empirical findings of Bulovsky [72], there is no evidence that the democratic leaders tweeted significantly more frequently than their authoritarian counterparts. Quite the opposite is the case. Authoritarian leaders, it seems, have not only caught up to their democratic counterparts, but they even tend to tweet more often. However, the statistical significance of this pattern depends on the prevalence of Twitter usage among the populace. In terms of the tone of their tweets, democratic leaders generally maintain a more balanced, positive sentiment. Specifically, they demonstrate a higher level of positivity in their tweets on average, a significant decline in positivity correlated with the emergence of the COVID-19 pandemic.

This research enriches the existing body of knowledge regarding political communication across regime types. The intersection of communication technologies, particularly social

media, with political discourse has not been extensively explored, as evidenced in my literature mapping (see Figure 2.1). In the realm of social media research within social sciences, the primary focus has been on its use by the general public, with studies such as those by Chunly [73] and Ruijgrok [74] highlighting this trend. When it comes to the exploration of social media usage by political elites, much of the existing research is centered on political candidates [e.g. 75], or focuses on the lower echelons of government [e.g. 76]. To my best knowledge, there is a scarcity of research that specifically addresses the use of social media by heads of government. A notable exceptions is a study by Bulovsky [72], while all other approaches predominantly features single-case analyses [see 77] or are limited to comparisons within democratic regimes, exemplified by Rivas-de-Roca and Pérez-Curiel [78].

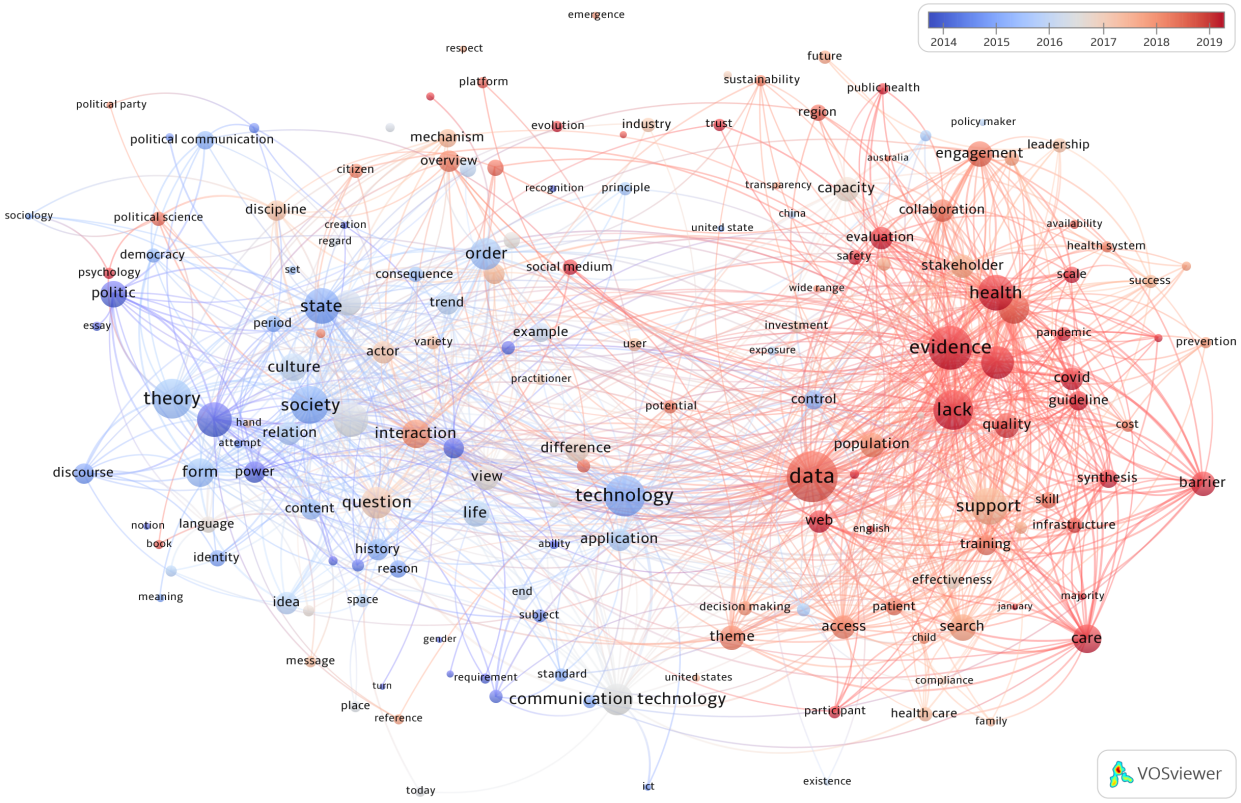
I also improve on existing research about the social media use of political elites with regards to the methodology. I apply recent advances in NLP to assess the sentiment of tweets rather than focusing exclusively on structure. Language indicates how speakers feel, how they perceive their environment, and the emotions they seek to spark in their audience. However, obtaining reliable measures of sentiment for large data sets remains challenging. The multilingual, transformer-based *XLM-roBERTa-base* model developed by Barbieri et al. [71] was fine-tuned to study the sentiment of tweets and represents an important step forward in this regard. By showcasing the relevance and significance of this multilingual model on my smaller, English-speaking case selection, I further validate its applicability and open up a new potential venue for future, multilingual research.

Lastly, my research has real-world significance. Leaders' public communication affects the thoughts and actions of their citizens, hence shaping the dynamics of the different political systems. People might either be swayed to believe in the agenda presented by leaders on social media or be encouraged to reflect on their willingness to seek change given a euphemistic presentation of harsh realities [e.g. 79]. In an increasingly connected world, the government's self-portrayal in social media may also alter foreign powers' perceptions [80]. It becomes apparent that it matters what and how political leaders communicate on social media. As a result, these platforms have emerged as a relevant political communication venue worthy of study.

3.1.2 Political Incentives, Engagement Patterns, and Sentiments

In the current era, leaders at the highest levels of government are tasked with maneuvering through the intricate dynamics of a hybrid media landscape. This environment encompasses both conventional and modern platforms, where diverse gatekeepers play a pivotal role in shaping and propagating their communication, as Chadwick et al. [81] observe. Simultane-

Figure 3.1: Bibliometric Network of Most Relevant Terms in Literature on Political Communication over Time.



Note: The graph was created with VOSviewer. The approach is described in Appendix A.

ously, these leaders are presented with an unparalleled opportunity to engage in cost-effective, continuous, and unmediated dialogue with their citizens via social media. This digital arena empowers them to independently generate and tailor content, delivering their messages directly to the public on their terms, circumventing the traditional inter-mediation of news media.

In this chapter, I contend that the use of social media by high-ranking politicians – in particular by heads of state – mirrors the strategic environment dictated by their political systems. Specifically, I focus on two critical dimensions that vary with a country's level of democracy: accountability and information control. Accountability pertains to the entities capable of displacing political leaders from power. According to Bueno de Mesquita et al. [82] leaders in more democratic settings are answerable to a wider segment of the population compared to those in less democratic regimes. This disparity significantly influences the primary audiences targeted by political communications. Typically, democratic leaders aim to connect with and engage a more extensive and diverse audience than their non-democratic

counterparts. The second aspect, information control, involves the leader's capacity to sway the information accessible to citizens within their domain. The impact of different messaging types is partly contingent on whether recipients can readily encounter alternate perspectives. The interaction between the state and traditional media, along with the extent of information manipulation, varies with the degree of democracy [83]. In essence, as nations progress towards greater democracy, the focus of accountability shifts from the elites to the general public, and the grip of the state over information tends to weaken. I propose that these variations in incentives, rooted in the level of democracy, are reflected in the patterns of social media communication by political leaders.

I assert that a fundamental aspect of social media communication is its frequency — the regularity with which leaders engage with citizens through these platforms alike other media channels. Consistent with Bulovsky's [72] findings, I hypothesize that leaders from more democratic countries are likely to post on social media more frequently than their counterparts in less democratic nations. Drawing upon another study on social media usage by democratic political leaders [84], I recognize Twitter as a crucial medium for influencing citizen perceptions. From this perspective, social media is instrumental for reaching out to those who may not engage with traditional media. Other studies point to the specific role it plays in building trust in the government [85] and governmental responsiveness [86]. There are, it seems, clear incentives for democratic leaders to engage with citizens through social media communication. The greater the accountability of leaders to their citizens, the more they endeavor to establish direct connections with their populace, enhancing the value of social media for engaging such audiences. In contrast, authoritarian leaders, accountable mainly to a limited elite, find broader communication channels less efficient for their purposes. As Bulovsky [72] articulates, the lack of electoral incentives leads to a more inactive social media presence among authoritarian leaders. This observation forms the basis of my first hypothesis.

Hypothesis 1. *Leaders of more democratic countries tweet more frequently than leaders of less democratic countries.*

I recognize that frequency is just one dimension of how social media communication varies across different types of regimes. If leaders adapt their communication strategies to align with their unique political motivations, variations in style are also likely. One key stylistic element I examine is the tone, or sentiment, used by political leaders in their social media posts. I propose that the presence of alternative information sources affects the overall positivity of these posts. In non-democratic settings, leaders often have the means to present their messages and actions in any light without significant fear of repercussions. Guriev and

Treisman [87] shed light on this phenomenon, explaining how modern autocrats increasingly resort to subtle forms of information manipulation. They suggest that these rulers maintain power not through coercion or indoctrination but by persuading citizens, albeit misleadingly, of their competence and dedication to public welfare. In their comparative analysis of speeches, they observe that such "informational autocrats" focus more on touting economic achievements and service delivery rather than using threats or violence. That said, authoritarian regimes are often associated with a single, charismatic leader. For example, based on a study of 98 states, von Soest and Grauvogel [88] find that authoritarian regimes are more likely to legitimate their rule by appealing to the leaders' personal authority and charisma. Such leaders may also be inclined towards extreme communication styles. Nai and Maier [89] show that political candidates in the 2018 US Senate midterms who displayed the "dark" personality traits of psychopathy, narcissism, and Machiavellianism are more likely to post negative and uncivil tweets. While recent years have seen the personalization of political communication in democracies [90], I expect the pattern of personalized, raging, unfiltered communication more present in authoritarian regimes. In the same vein, on average, I expect democratic leaders to exhibit a more positive tone in their communications.

Hypothesis 2a. *Leaders of more democratic countries have a more positive style of social media communication, on average, than leaders of less democratic countries.*

Power is more concentrated in authoritarian regimes than democratic ones. This limits the constraining influence of parties and media advisors, resulting in a less scripted communication style. In contrast, democratic leaders operate under the constraints imposed by a more liberated media environment, a defining characteristic of democracy [67]. In scenarios involving unfavorable events, authoritarian leaders have the option to employ censorship, thereby excluding challenging topics or narratives from their social media discourse, effectively using these platforms as part of a broader, orchestrated information campaign. In contrast, democratic leaders do not have the same capacity to control the narrative. Despite the fact that adverse events are a commonality across all political contexts, I anticipate a discernible difference in the social media communications of democratic leaders compared to their authoritarian counterparts. When confronted with objectively negative circumstances, I predict a more pronounced decrease in the positivity of posts from democratic leaders, reflecting their need to confront and acknowledge the realities of their situation.

Hypothesis 2b. *Leaders of more democratic countries respond to negative events with a greater decline in positivity than leaders of less democratic countries.*

3.2 Methodology

3.2.1 Case Selection

To evaluate the hypotheses, I utilize an original dataset comprising tweets posted by political leaders in Sub-Saharan Africa from January 2018 to December 2021. This region, characterized by a diverse range of political regimes, provides a rich context for examining variations in regime type, my study's main independent variable. The selection of my dataset, while aiming to be as comprehensive as possible, is inevitably constrained by factors such as the prevalence of Twitter usage and language barriers.

Firstly, not every political leader in Africa is active on Twitter, a trend consistent globally. To compile my dataset, I identified African political leaders who met three criteria during the period from 2018 to 2021: (1) active Twitter usage, (2) operating a verified government media account, and (3) tweeting predominantly in English. It is important to note that Twitter's requirement for verifying an email address linked to an official government domain provides assurance regarding the legitimacy of the accounts included in my study. I chose a four-year span for my dataset to balance several practical considerations. On one hand, this time frame makes the study robust to variations in electoral incentives, as it encompasses election periods in most countries. On the other hand, a more extended period could introduce complexities due to evolving Twitter usage trends, making comparisons less straightforward. Hence, a shorter duration enhances the comparability of tweets across the sample.

Secondly, the language of communication varies among political leaders. Out of the 30 leaders identified as active on Twitter during my study period, 12 exclusively used English, 9 used only French, 1 used only a different language (João Lourenço of Angola, Portuguese), and 8 communicated in multiple languages. For practicality, my focus is on leaders who primarily tweet in English, as detailed in Table 3.1.

3.2.2 Dependent Variables

For this analysis, I source Twitter data through the platform's Academic API, which yielded two primary types of information: account details and tweets themselves. The account information included the date of account creation, profile bios, follower and following counts, and total tweet counts. I extracted these details using the *tweepy* Python package. The second aspect of my dataset comprised all tweets posted from January 1st, 2018, to December 31st, 2021. Due to the limitations of *tweepy*, which restricts data extraction to a maximum of 3,200 tweets per account, I employed a social networking services (SNS) scraper, *snscrape* [91], to access a more extensive range of tweets.

Table 3.1: Overview of Political Leaders.

Name	Country	Avg. Democracy Score	Twitter Users (% Pop)	Age	Number of Tweets
Cyril Ramaphosa	South Africa	0.72	5.99	Mid	7,030
Nana Akufo-Addo	Ghana	0.71	3.06	Old	1,023
Mokgweetsi Masisi	Botswana	0.64	4.72	Mid	1,948
George Weah	Liberia	0.62	0.35	Young	81
Juliu Maada Bio	Sierra Leone	0.55	0.44	Mid	194
Muhammadu Buhari	Nigeria	0.53	1.65	Old	2,002
Adama Barrow	The Gambia	0.51	1.26	Young	169
Hakainde Hichilema	Zambia	0.38	1.00	Mid	163
Edgar Lungu	Zambia	0.35	1.05	Mid	1,282
Yoweri Museveni	Uganda	0.30	0.65	Old	4,112
Paul Biya	Cameroon	0.30	0.73	Old	850
Emmerson Mnangagwa	Zimbabwe	0.29	2.24	Old	654
Paul Kagame	Rwanda	0.24	1.10	Mid	212

Notes: Leaders active on Twitter are manually identified through the official government account labels. Electoral Democracy Score and Twitter Users (% Pop) are averaged over the observations for the users. Age is based on the average age in the data, split as follows: young - under 55, mid - 55 to 69, old - 70 and older. Only English tweets are considered. All information was last accessed January 4th, 2024.

From the initial collection of tweets, I refined the dataset to include only those tweets that were posted during the leaders' tenures in office, resulting in 32,792 tweets. Further narrowing down to tweets in English, my final dataset comprised 19,518 tweets with corresponding metadata from 13 political leaders. Following the extraction process, I meticulously removed all irrelevant information from the dataset and systematically formatted the relevant features to facilitate ease of analysis and interpretation.

To measure the frequency of social media communication, I compute the daily number of tweets posted by each leader.

To analyze the sentiment of tweets – essentially gauging how positive or negative their language is – I employed a multilingual *XLM-roBERTa-base* model, which has been pre-trained on nearly 200 million tweets, as detailed by Barbieri et al. [71]. This model generates sentiment scores for each tweet on a scale ranging from -1 to 1, which I re-scale to a range from 0 to 2 (with 2 being the most positive score, see Figure 3.2). The model's fine-tuning for assessing the emotional tone of text, as Barbieri et al. [71] explain, makes it particularly adept

at handling Twitter data. It is an advancement of *RoBERTa*, which itself is a more robust iteration of Google's *BERT* model [92]. Like the original *BERT* model, the *XLM-roBERTa-base* model I used is based on transformers, which are crucial for considering long-range dependencies in sequential data, such as the words in a sentence [93].

Figure 3.2: Examples of Tweets' Compound Sentiment Score (XLM-roBERTa-base Model, 0 to 2 Range).

(a) Positive Tweet. Sentiment Score = 1.906

(b) Negative Tweet. Sentiment Score = 0.055

For example, if a tweet from a country leader reads "it was magnificent," the model can discern the reference of "it," even if its context appears earlier or later in the sentence. The bidirectional nature of *BERT*-based models, like the one employed in my study, allows for the consideration of each word's dependencies with every other word in the sentence, both preceding and following. This attribute renders the model particularly effective in interpreting various sentence structures, as Devlin et al. [93] highlight.

This methodology improves on previous work that utilizes dictionary-based sentiment analysis approaches. The latter often fall short in handling the unstructured nature of social media content and struggle to factor in contextual nuances [94]. For instance, Figure 3.3 in my data showcases two posts that the *VADER* algorithm erroneously categorized. The model developed by Barbieri et al. [71] surpasses other common methods in various sentiment analysis benchmarks, instilling confidence in the reliability of this advanced tool for my analysis.

Given that the tweets I analyzed are from official governmental accounts and are generally well-structured, I did not deem further data pre-processing necessary. These official tweets tend to exhibit less noise compared to typical social media content. Moreover, the transformer-based model is adept at interpreting contextual elements like punctuation, effectively capturing the distinctive style of social media communications as discussed by Barbosa and Feng [95] as well as Derczynski et al. [96]. The rationale behind my choice of model and its comparative performance against the *VADER* algorithm is thoroughly detailed in Appendix B.

Sentiment analysis is certainly not novel to political science. For example, Chambers et

Figure 3.3: Tweets of Former Nigerian President Buhari Incorrectly Labeled by VADER Model.

(a) Tweet Incorrectly Labeled as Negative

(b) Tweet Incorrectly Labeled as Positive

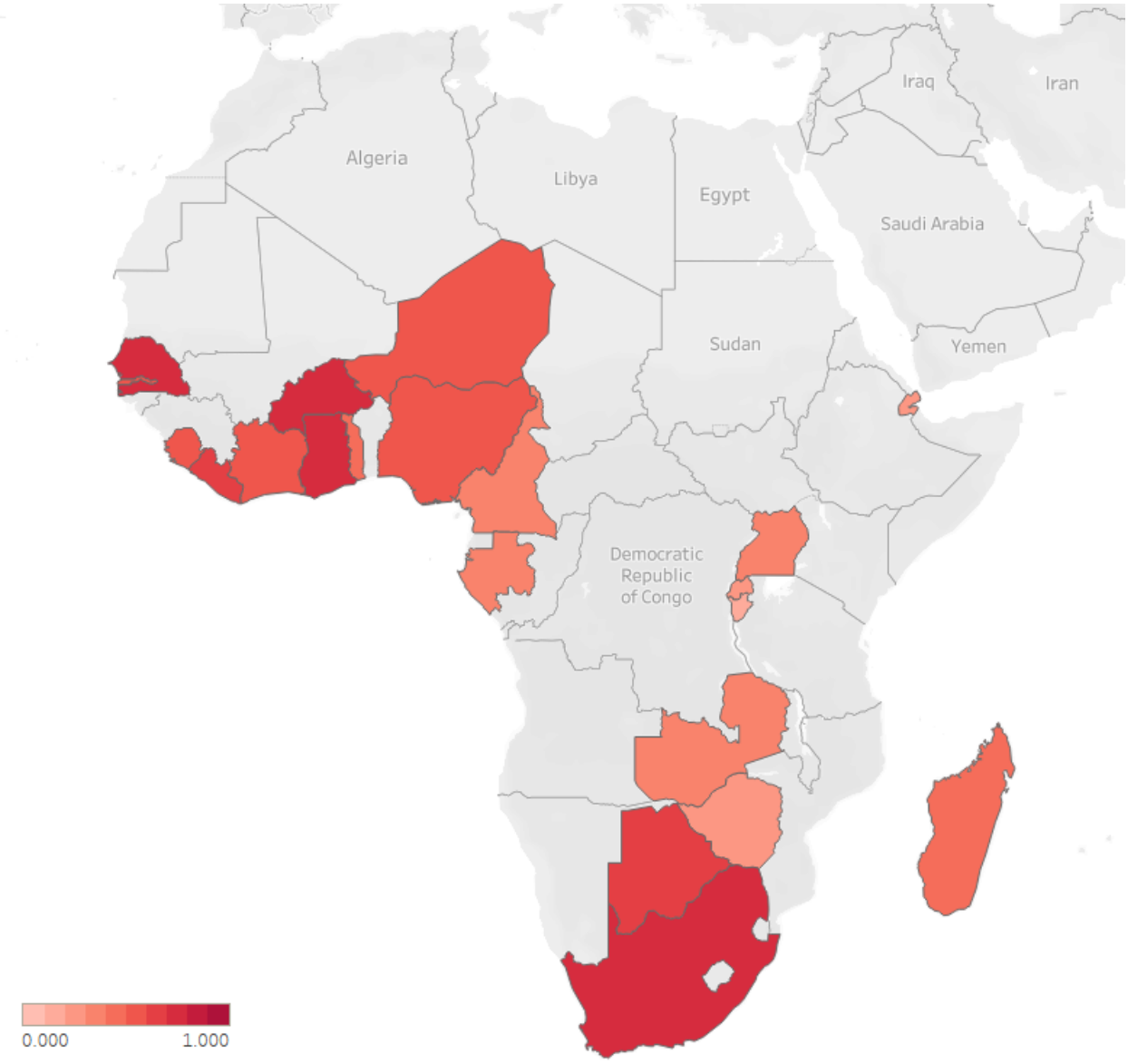
al. [97] study the alignment of citizens' opinions with country's formal international relations, and Ceron et al. [98] find that sentiment can help monitor electoral campaigns. However, my study is the first one known to me that employs this natural language processing approach in political leaders' online communication with citizens.

3.2.3 Independent Variables

My primary independent variable captures the regime type or level of democratization. To this end, I utilize the electoral democracy score from V-DEM, which is accessible for the entire study period and facilitates the examination of variations in democratic degrees, as illustrated in Figure 3.4. This approach is particularly suitable given that the mechanisms I propose are not binary. For instance, as leaders increasingly rely on citizens over elites for authority, I expect a corresponding rise in their efforts to maintain frequently engaged with the audience. Similarly, as their control over information wanes, I hypothesize a more measured approach in their social media communications in times of adversity.

To gauge the significance of Twitter within each country, I incorporate a variable reflecting the platform's usage rate in the population. Drawing on insights from Haman and Školník [99], who found a notable impact of overall Twitter usage on the adoption of the platform by members of parliament in 32 European countries, my measure accounts for the

Figure 3.4: Map of Cases Colored by Avg. Electoral Democracy Score (0 to 1 Range).



percentage of the population using Twitter. As shown in Table 3.1, Twitter usage varies widely in my sample, yet some countries exhibit usage rates comparable to larger economies in the Global South, such as India (1.7%) and Brazil (8.0%). While Twitter coverage is relatively low, the platform remains the primary social media on which many political leaders are actively present with a certified public account – making it an appropriate entry point for the study of leaders' social media engagement.

The sample includes tweets from the period before and during the COVID-19 pandemic. To consider the impact of this global crisis on political communication, I introduce a variable

indicating the confirmation of the first COVID-19 case in each country. This variable serves to account for shifts in the political climate during such distressing times. Furthermore, it allows me to directly investigate changes in sentiment within countries when the news cycle turns objectively more negative. I actively choose not to utilize more granular COVID-19 measures, such as the number of cases or deaths, due to the inconsistencies in reporting, particularly in the Global South [100].

Furthermore, I control for several characteristics of the leaders in my study. I consider the leaders' age, which could introduce biases linked to their familiarity with technology. The educational level of the leaders is another factor I take into account, as identified by Krcmaric et al. [24] and Baturo [101] as pertinent in the context of political leadership studies. Following the categorization by Ellis et al. [102], I classify educational levels on a scale from 0 to 3 (0 = Primary, 1 = Secondary, 2 = University, and 3 = Graduate).

3.2.4 Modeling Choice

This study employs two types of regression models. First, I utilize a negative binomial regression to explore the relationship between the level of democratization in a country and the frequency of tweets by political leaders (*Hypothesis 1*). This regression type is appropriate for count data with an excessive number of zeros, known as zero inflation, in addition to over-dispersion [103]. The negative binomial regression model is particularly useful when the data exhibit variance that exceeds the mean, which is common in social media activity data where many users may have periods of inactivity resulting in zero counts [103]. Thereby, the model is an extension of the Poisson regression, adding an extra parameter to account for the over-dispersion. The expected log count of tweets is modeled as a linear combination of the independent variables:

$$\log(\mathrm{E}[\text{Tweet Frequency}|X]) = \log(\mathrm{E}[Y|X]) = X\beta \tag{3.1}$$

where Y represents the count of tweets, X is the matrix of independent variables, and β is the vector of coefficients. The variance function for the negative binomial model is:

$$\mathrm{Var}(Y|X) = \mathrm{E}[Y|X] + \alpha(\mathrm{E}[Y|X])^2 \tag{3.2}$$

The term α denotes the over-dispersion parameter, which allows the variance to exceed the mean. When $\alpha = 0$, the model simplifies to a Poisson model with equal mean and variance (equidispersion). I construct three variations of the regression model. The dependent variable in all models is the daily *number of tweets*, as previously described. The independent variables include the *Electoral Democracy Score*, the *Twitter usage (% Pop)*, the prevalence

of *COVID-19*, the leaders' *age*, and their *level of education* (see Equation 3.3).

$$\begin{aligned}\text{Tweet Frequency} = \beta_0 &+ \beta_1(\text{Elec. Dem. Score}) \\ &+ \beta_2(\text{Twitter Users (\% Pop)}) \\ &+ \beta_3(\text{COVID-19 Active}) \\ &+ \beta_4(\text{Age}) \\ &+ \beta_5(\text{Education}) \\ &+ \epsilon\end{aligned} \tag{3.3}$$

Models (2) and (3) introduce interaction terms to capture additional dynamics not reflected in Model (1). These interactions allow me to examine how the relationship between democratization and tweet frequency is moderated by factors such as Twitter usage in the population and the impact of the COVID-19 pandemic. Specifically, Model (2) includes the interaction between the *Electoral Democracy Score* and the *Twitter usage (% Pop)* (see Equation 3.4), while Model (3) includes the interaction between the *Electoral Democracy Score* and the prevalence of *COVID-19* (see Equation 3.5). These additions help to explore the nuances in how political leaders' social media behavior might be influenced by these interacting factors.

$$\begin{aligned}\text{Tweet Frequency} = \beta_0 &+ \beta_1(\text{Elec. Dem. Score}) \\ &+ \beta_2(\text{Twitter Users (\% Pop)}) \\ &+ \beta_3(\text{COVID-19 Active}) \\ &+ \beta_4(\text{Age}) \\ &+ \beta_5(\text{Education}) \\ &+ \beta_6(\text{Elec. Dem. Score} \times \text{Twitter Users (\% Pop)}) \\ &+ \epsilon\end{aligned} \tag{3.4}$$

$$\begin{aligned}\text{Tweet Frequency} = \beta_0 &+ \beta_1(\text{Elec. Dem. Score}) \\ &+ \beta_2(\text{Twitter Users (\% Pop)}) \\ &+ \beta_3(\text{COVID-19 Active}) \\ &+ \beta_4(\text{Age}) \\ &+ \beta_5(\text{Education}) \\ &+ \beta_7(\text{Elec. Dem. Score} \times \text{COVID-19 Active}) \\ &+ \epsilon\end{aligned} \tag{3.5}$$

In these equations, β_0 represents the intercept, β_i (where $i = 1, 2, \ldots$) are the coefficients of the respective independent variables and their interactions, and ϵ is the error term.

Secondly, for the analysis of tweet sentiment (*Hypothesis 2a* and *2b*), I employ linear regression models to examine the relation of democratization and the positivity of tweets by political leaders. Linear regression is suitable for modeling continuous dependent variables, such as sentiment scores, which typically do not exhibit the count-based over-dispersion characteristic of the frequency data addressed in *Hypothesis 1*. The dependent variable in these models is the sentiment score of each tweet, representing the positivity of the communication. The independent variables remain consistent with those utilized in the previous models.

Model (1) examines the main effects of these variables on tweet sentiment (see Equation 3.6). In Models (2) and (3), interaction terms are introduced to explore more complex relationships. Model (2) investigates how the interaction between *Electoral Democracy Score* and *Twitter usage (% Pop)* in the population might affect tweet sentiment (see Equation 3.7), while Model (3) looks at the interaction between *Electoral Democracy Score* and the prevalence of *COVID-19* (see Equation 3.8).

The equations for these models are expressed as follows:

$$\begin{aligned}\text{Tweet Positivity} = \beta_0 &+ \beta_1(\text{Elec. Dem. Score}) \\ &+ \beta_2(\text{Twitter Users (\% Pop)}) \\ &+ \beta_3(\text{COVID-19 Active}) \\ &+ \beta_4(\text{Age}) \\ &+ \beta_5(\text{Education}) \\ &+ \epsilon\end{aligned} \tag{3.6}$$

$$\begin{aligned}\text{Tweet Positivity} = \beta_0 &+ \beta_1(\text{Elec. Dem. Score}) \\ &+ \beta_2(\text{Twitter Users (\% Pop)}) \\ &+ \beta_3(\text{COVID-19 Active}) \\ &+ \beta_4(\text{Age}) \\ &+ \beta_5(\text{Education}) \\ &+ \beta_6(\text{Elec. Dem. Score} \times \text{Twitter Users (\% Pop)}) \\ &+ \epsilon\end{aligned} \tag{3.7}$$

$$\begin{aligned}\text{Tweet Positivity} = \beta_0 &+ \beta_1(\text{Elec. Dem. Score}) \\ &+ \beta_2(\text{Twitter Users (\% Pop)}) \\ &+ \beta_3(\text{COVID-19 Active}) \\ &+ \beta_4(\text{Age}) \\ &+ \beta_5(\text{Education}) \\ &+ \beta_7(\text{Elec. Dem. Score} \times \text{COVID-19 Active}) \\ &+ \epsilon\end{aligned} \tag{3.8}$$

In the above equations, β_0 represents the intercept, β_i (where $i = 1, 2, \ldots$) are the estimated coefficients of the independent variables and their interaction terms, and ϵ denotes the error term. The linear models assume homoscedasticity and normally distributed error terms, with robust standard errors clustered at the user to account for within-user correlation in sentiment expression.

3.3 Results and Findings

3.3.1 Frequency Differences Across Regime Types

I initiate my analysis by examining *Hypothesis 1*, which posits that democratic leaders are more active on Twitter compared to their non-democratic counterparts. To measure this, I consider the number of tweets posted by a political leader on any given day. In addressing this hypothesis, I utilize a negative binomial model, which is particularly suited for my dataset due to the prevalence of days where certain users did not tweet at all, resulting in a substantial number of zeros.

Contrary to what *Hypothesis 1* suggests, my initial findings indicate a negative correlation between the electoral democracy score and the daily tweet count. Yet, this relationship does not reach statistical significance in my baseline model, as evidenced in Model (1) of Table 3.2. This unexpected trend might be influenced by the nature of authoritarian regimes, which are often characterized by more "newsworthy" events such as military conflicts and economic challenges, as Yu and Jong-A-Pin [104] suggest. Additionally, I consider the contrasting self-representation styles of authoritarian and democratic leaders. Authoritarian leaders might engage in more self-illustration or positive portrayal, whereas democratic leaders often exhibit a sense of humbleness and might face repercussions for overly frequent and positive depictions of their work. Regarding the study of conflicts, the finding also suggests that leaders of less democratic countries could likely push hateful narratives that incentivize discrimination or at worst violence, without significant fear of repercussions [66, 68].

To understand the observations better, I explore two interactions. In both cases, the negative effect of democratization on tweet frequency becomes statistically significant.

First, I interact the electoral democracy score with the measure of country-level Twitter use (see Table 3.2, Model (2)). The positive interaction term suggests that as the proportion of Twitter users increases, the negative impact of a lower democracy score on the frequency of the dependent variable becomes smaller. In other words, in cases with a higher percentage of Twitter users, the difference in frequency between more democratic and less democratic countries is not as pronounced as it is in countries with fewer Twitter users. Essentially, the presence of Twitter users is moderating the relationship between democracy score and frequency, making the outcomes more similar across countries with varying levels of democracy.

This conditionality is illustrated in Panel (a) of Figure 3.5, which depicts the average marginal effect (AME) of electoral democracy score for different levels of Twitter usage. When Twitter use is low, the predicted effect of electoral democracy on tweet frequency is negative. However, when Twitter use is roughly greater than the midpoint in the data, the difference in tweet frequency is mitigated.

Secondly, I introduce an interaction between the electoral democracy score and the prevalence of COVID-19, as delineated in Model (3). As for Twitter use, the positive coefficient of this interaction term indicates a diminished disparity in tweet frequency associated with varying levels of democracy in the context of the pandemic. This observation suggests that the dynamics or responses elicited by the pandemic are less influenced by a country's democratic score.

My interpretation of this finding considers several factors: in challenging times like a pandemic, democratic leaders might become more communicative as they are not constrained to

Table 3.2: Relationship between Democracy and Tweet Frequency.

	DV: Frequency		
	Model (1)	Model (2)	Model (3)
Electoral Dem. Score	−2.029	−5.832**	−3.118*
	(1.643)	(2.356)	(1.655)
Twitter Users (% Pop)	0.439***	−1.432**	0.402***
	(0.116)	(0.693)	(0.109)
COVID-19 Active	0.095	0.121	−1.362**
	(0.354)	(0.242)	(0.602)
Age	0.076***	0.066***	0.074**
	(0.029)	(0.023)	(0.029)
Education	0.198	0.131	0.211
	(0.261)	(0.283)	(0.268)
Elec. Dem. x Twitter Users (% Pop)		2.845***	
		(1.002)	
Elec. Dem. x COVID-19 Active			2.972**
			(1.466)
Constant	−5.584**	−2.504	−4.839**
	(2.383)	(2.387)	(2.331)
Unit of Analysis	User-Day	User-Day	User-Day
Observations	15,442	15,442	15,442

*p<0.1; **p<0.05; ***p<0.01

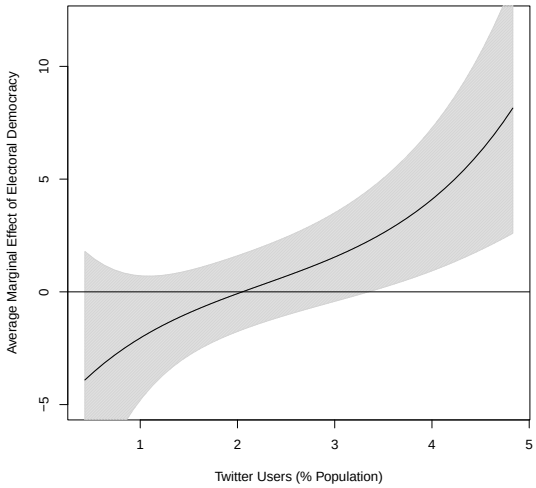
Note: Both models are negative binomial models and have robust standard errors, clustered at the user.

downplay adverse events. Conversely, authoritarian leaders may be inclined to downplay the pandemic's severity or its impact in their countries, perhaps even attempting to understate the existence of the crisis. This behavior could imply that, owing to their differing motivations, the tweeting behaviors of leaders from more and less democratic countries converge during a pandemic.

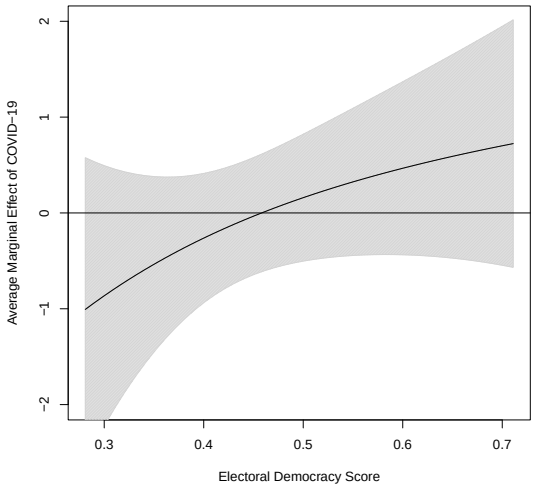
This conditional relationship is visually represented in Panel (b) of Figure 3.5, which displays the Average Marginal Effect (AME) of COVID-19 across various regime types. In

Figure 3.5: Level of Democracy and Tweet Frequency.

(a) Overall Frequency



(b) Effect of COVID-19 on Frequency



Note: Average Marginal Effects are calculated from Models (2) and (3) in Table 3.2. Vertical bars represent 95% confidence intervals.

this analysis, the impact of COVID-19 on tweet frequency is predicted to be negative in less democratic nations. In contrast, in more democratic countries, the presence of COVID-19 correlates with an increase in tweet frequency, albeit with diminishing effects.

Further, there are two interesting observations from the control variables. Across all model specifications, leader age is positively related to tweet frequency and statistically significant. At first sight, it is counter-intuitive that older leaders are more active on social media, but maybe there is a simple reasoning? A potential explanation would be that older leaders are technologically less savvy, wherefore they are more likely to gather a social media team that is specifically dedicated for an engaged online presence. Also, I do not find a statistically significant impact of education on tweet frequency, which potentially exemplifies that the barriers to use social media are negligible [72].

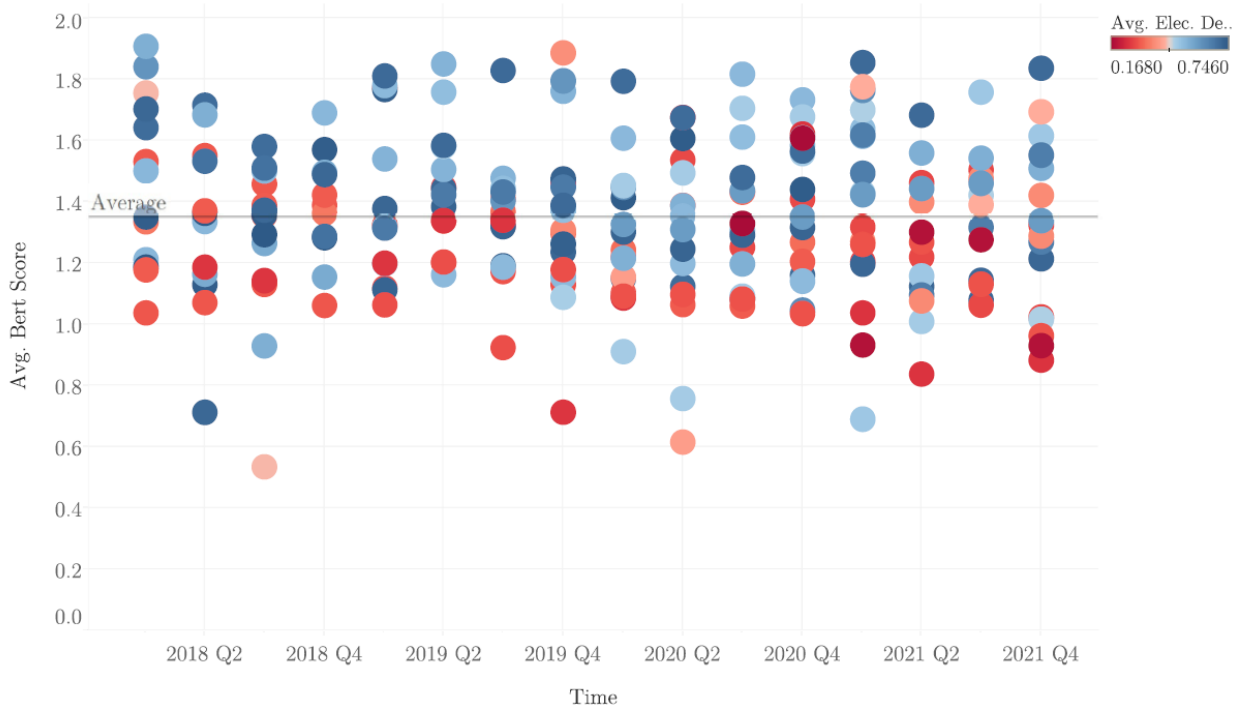
3.3.2 Sentiments of Leaders' Social Media Presence

Next, I examine the relationship between electoral democracy and tweet sentiment. The dependent variable is the tweet-level positivity score. *Hypothesis 2a* predicts that the sentiment of social media posts by democratic leaders is more positive than those by non-democratic leaders. For this hypothesis, I regress (OLS) the sentiment variable on the electoral democracy score. According to *Hypothesis 2b*, I also expect a more significant decline in positivity after the advent of COVID-19 in more democratic countries. I use an indicator variable for

whether the tweet was sent after the advent of COVID-19 to explore this expectation.

As was the case for *Hypotheses 1*, Model (1) of Table 3.3 shows there is not a significant correlation between electoral democracy and tweet positivity. However, I include the interaction with my measure of Twitter users in Model (2). The negative and significant coefficient on the interaction term again suggests that the positive effect of democratization on tweet positivity is moderated in countries with higher Twitter penetration. This is illustrated in panel (a) of Figure 3.7: conditional on a significant Twitter audience, the predicted AME of electoral democracy is negative. Potentially, this suggests that when Twitter becomes a mainstream form of political communication in democratic countries, leaders become more moderate in their style and tone. Figure 3.6 visualizes this pattern via the average sentiment over time; with less democratic leaders being predominantly located below the sample average.

Figure 3.6: Distribution of Avg. Sentiment over Time (0 to 2 Range).



Note: Excluding the statistical outlier of Rwandan President Paul Kagame.

Model (3) in Table 3.3 examines the effect of the objectively negative news that COVID-19 had reached the country in question. The presence of the COVID-19 pandemic appears to moderate the relationship between the level of democracy in a country and the positivity of tweets from its leaders, as reflected by the interaction term in Table 3.3, Model (3). Panel (b) of Figure 3.7 illustrates this relationship. This relationship exists even without restricting

attention to countries with significant Twitter penetration.

Table 3.3: Relationship between Democracy and Tweet Sentiment.

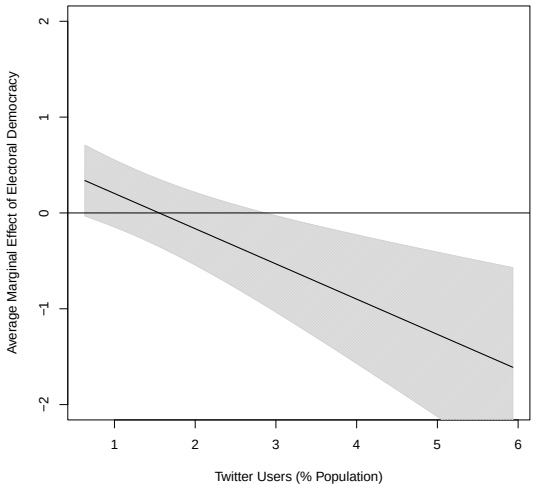
	DV: Positivity		
	Model (1)	Model (2)	Model (3)
Electoral Dem. Score	0.133	0.571**	0.210
	(0.322)	(0.222)	(0.285)
Twitter Users (% Pop)	−0.026	0.228***	−0.016
	(0.021)	(0.076)	(0.023)
COVID-19 Active	−0.022	−0.029	0.134***
	(0.030)	(0.031)	(0.046)
Age	−0.010***	−0.008***	−0.010***
	(0.003)	(0.003)	(0.003)
Education	0.032	0.036	0.027
	(0.034)	(0.033)	(0.034)
Elec. Dem. x Twitter Users (% Pop)		−0.368***	
		(0.107)	
Elec. Dem. x COVID-19 Active			−0.312***
			(0.117)
Constant	1.914***	1.475***	1.828***
	(0.263)	(0.303)	(0.250)
Unit of Analysis	Tweet	Tweet	Tweet
Observations	19,518	19,518	19,518
Adjusted R^2	0.028	0.041	0.031
Note:			*p<0.1; **p<0.05; ***p<0.01

Note: All models are linear models and have robust standard errors, clustered at the user.

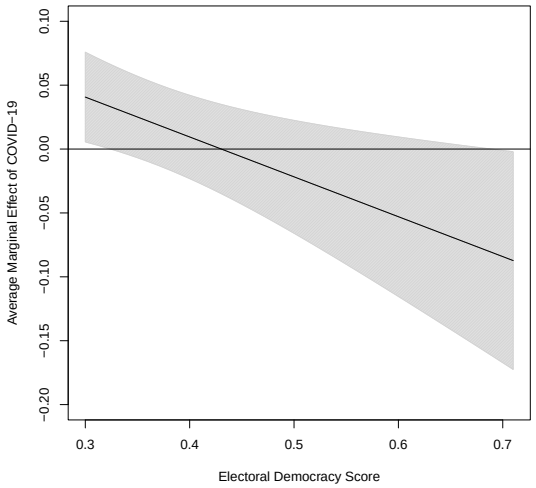
The negative interaction term between the level of democratization and the prevalence of COVID-19 suggests that the direct relationship between a country's level of democracy and the positivity of its leaders' tweets is altered when considering the impact of the pandemic. Specifically, it implies that the higher positivity typically associated with leaders from more democratic nations is reduced in the context of COVID-19. One interpretation

Figure 3.7: Level of Democracy and Tweet Sentiment.

(a) Overall Positivity



(b) Effect of COVID-19 on Positivity



Note: Average Marginal Effects are calculated from Models (2) and (3) in Table 3.3. Vertical bars represent 95% confidence intervals.

of this moderation effect is that during the pandemic, leaders from democratic countries may have felt compelled to adopt a more somber tone to reflect the gravity of the situation, despite generally being more positive in their communications. This shift could align with the expectations of transparency and the provision of accurate information in democracies, especially under the scrutiny experienced during a crisis.

Conversely, leaders from less democratic countries, where there is often less expectation for transparency and where information might be more tightly controlled, would not necessarily need to alter their tone significantly in response to COVID-19. While authoritarian leaders do not face the same scrutiny of the public as their democratic counterparts, crises like the COVID-19 pandemic test their ability to govern and portray successful governance. Negative repercussions for the country like COVID-19 deaths or a recessing economy are signs of weakness which they seek to avoid, which explains my model's observation. Consequently, the difference in tweet positivity between more and less democratic regimes could be less distinct during the pandemic than it would be otherwise. It suggests that the association between democratic governance and social media communication is context-dependent, with the pandemic serving as a significant exogenous factor that influences how leaders communicate with the public. While authoritarian leaders can potentially influence the narrative on a pandemic, they can hardly prevent its existence – as COVID-19 showed. Therefore, it is a great example to study how leaders of different regime types act in the face of adversity.

As demonstrated in the models within Table 3.3, I also observe noteworthy trends among

the control variables. Notably, the age variable consistently shows that younger leaders tend to express more positive sentiments in their tweets compared to older leaders. This could reflect generational differences in communication styles or varying degrees of engagement with social media platforms. The education level of the leaders does not appear to significantly influence the sentiment of their tweets. This finding is consistent across the models and suggests that educational attainment may not be a pivotal factor in shaping the positivity of political communication on social media.

In summary, I find support with respect to *Hypotheses 2a* and *2b*.

3.4 Key Takeaways on Leadership Communication

3.4.1 Structure and Tone Differ Across Regime Types

Political leaders from various regimes utilize social media platforms as a vital channel to reach out to their populace. Yet, the nature of this communication exhibits a systematic variation in line with the degree of democratization. Institutions within democratic settings encourage leaders to engage with a wide range of citizens, discussing policies and pertinent events in a factual, yet positive, manner. Conversely, in authoritarian regimes, where there is a reduced imperative to interact with the general citizenry and where information is more tightly controlled, communication tends to be more propagandistic, less frequent, and less amiable.

In this chapter, I analyze the official Twitter communications from the heads of 13 African nations to shed light on these divergent communication strategies. My analysis reveals systematic differences in both the structure and style of communication that are linked to the level of democracy. Specifically, the level of democracy within a country is negatively associated with the frequency of leaders' tweets but is inversely related to the positivity of the tweets. Interestingly, these correlations are moderated where Twitter usage is more widespread. The intrinsic value of social media in democratic societies stems from its capacity to connect with an audience that is of concern to the leaders [90, 105]. In contexts where social media does not effectively reach this audience, the motivation to engage in transparent, inclusive, and regular dialogue wanes, leading to a convergence in the communication incentives of democratic and non-democratic leaders.

The implications of my research extend to the broader field of comparative political communication. The evidence indicates that in non-democratic regimes, social media communication is a cog in the larger machinery of authoritarian information control and narrative shaping. In environments where alternative sources of information are scarce, the narratives

presented on social media platforms can significantly shape citizens' beliefs [25]. Grasping the nuances of these communications is essential for deconstructing the political dynamics within more opaque societies.

I have also contributed methodological advancements to the study of political communication via social media. I showcase a cutting-edge approach for sentiment analysis in tweets for political science research. The model by Barbieri et al. [71] surpasses traditional dictionary-based sentiment analysis methods, a significant advancement given the limitations of previous methods, as outlined in detail in Appendix B. As evidenced in my research, this model is an invaluable asset for future political science studies that aim to analyze tweet sentiment.

Lastly, while my study investigates the role of Twitter in contemporary political discourse, it is essential to acknowledge that the platform may experience substantial transformations, especially with changes in its leadership. Moreover, the emergence of alternative platforms, each with their unique terms of service, could also redefine the landscape. This evolution has the potential to reshape the online political conversation, influencing the strategic communication choices of political leaders across different regimes. Future research should be mindful of such developments when examining the dynamics of language, sentiment, and politics on social media platforms.

3.4.2 Limitations and Future Research

While offering a comprehensive analysis of political communication, I acknowledge limitations that pave the way for further research.

Firstly, the multilingual nature of social media communication presents both a challenge and an opportunity for deeper analysis. The current study has primarily focused on English-language tweets, and incorporating French-speaking leaders from the African continent would enrich the understanding of how language nuances influence political messaging. In an upcoming iteration of this project, I collaboratively with my co-authors incorporate French tweets and respective African political leaders in a study. Further extensions of the line of research could include a wider array of languages – including local languages – deepening the appreciation of cross-cultural political communication dynamics.

Secondly, while this study provides insightful findings within the African context and allows for a focused assessment, extending the research to include political leaders from other continents could offer a more global perspective. Comparative analysis across different geopolitical regions may uncover how cultural, social, and political variables uniquely shape social media use by political leaders.

An intriguing dimension adjacent to sentiment is the toxicity of communication, especially in relation to crises and conflicts. While I tested a common modeling solution for measuring toxicity, Google's Perspective API[1], it did not yield satisfactory results. Future studies might explore alternative tools and methods to assess this dimension. This could also extend to examining the spread of hateful speech by political figures, if applicable, providing critical insights into the responsible use of platforms for political discourse and the potential role leaders play in disseminating content that sparks violence.

Furthermore, the insightful findings on different leaders' response to exogenous events like the COVID-19 pandemic awakens the interest of studying other events that shape and influence society and the tenure of political leaders. As such, a natural extension of this project would be investigating the effect of electoral periods on leaders' communication frequency and style. I anticipate a significant shift in the social media behavior of incumbents around election times, when political communication assumes heightened importance. A comparison to leaders of the respective opposition parties would further introduce the assessment if incumbency matters.

While Twitter and other social media platforms are increasingly pertinent to political communication studies, they represent just one avenue through which leaders engage with the public. A logical extension of this research could involve a comparative analysis with other forms of official communication, such as television, radio, or press statements. For the latter, the recently launched United Nations' Diplomatic Pulse[2] platform offers a promising resource for comparing social media discourse with official press releases, potentially uncovering discrepancies or consistencies in messaging across different communication channels.

In conclusion, the findings from this chapter underscore the importance of studying political communication through social media in times of crises and conflicts. They suggest that this is a fertile area for scholarly inquiry, with significant implications for understanding contemporary political dialogue. As the landscape of digital communication continues to evolve, there is a clear mandate for ongoing research to keep pace with these developments, ensuring a nuanced and informed understanding of the interplay between political leaders, citizens, and the media in the digital age.

[1]Google Perspective API: https://perspectiveapi.com/, last accessed January 4th, 2024.
[2]United Nations Diplomatic Pulse: https://diplomaticpulse.org/en/, last accessed January 4th, 2024.

Chapter 4

Misinformation in Times of Crises

4.1 Theory

4.1.1 What is Misinformation?

In the landscape of modern information exchange, misinformation has emerged as a critical issue, drawing heightened attention from researchers, policymakers, and technology companies due to its pervasive impact [106]. Guess and Lyons [107] define misinformation as assertions that conflict with generally accepted and verifiable information. Disinformation specifically relates to a category of misinformation that is intentionally spread. These concerns, while not new, have intensified in the digital age [108]. Pennycook and Rand [109] emphasize that while misinformation has historical precedents, such as the 'Great Moon Hoax' of 1835, its contemporary manifestation is far more complex and pervasive. The proliferation of misinformation in recent years is particularly evident in the context of major political events. The 2016 United States Presidential Election and the British 'Brexit' referendum, for example, have been turning points in understanding the role and impact of 'fake news' in political discourse, as highlighted by Lazer et al. [110].

The global scope of misinformation is vast, transcending national boundaries and impacting various aspects of society [111]. The COVID-19 pandemic, termed an "infodemic" by many researchers [111–116], exemplifies the worldwide nature of misinformation. This period saw an unprecedented spread of false information ranging from unproven cures to conspiracy theories, significantly affecting public health initiatives and individual behaviors. van der Linden et al. [117] further illuminate the broader implications of misinformation, such as its capacity to undermine support for essential issues like climate change and to fuel vaccine hesitancy, illustrating its far-reaching consequences. Similarly, Ahmed et al. [7] draw a direct line between online hate, at times exacerbated by misinformation, and severe

offline consequences, including youth suicides and violent crimes. A significant factor in the rise of misinformation is the transformation of the information landscape in the digital age. Aslett et al. [118] discuss the dramatic decrease in the cost of producing and distributing information online, which has led to a massive increase in the volume of available content. This shift has fundamentally altered the role of traditional gatekeepers, placing more emphasis on search engines and algorithms for information sorting and validation. Guay et al. [119] highlight how this has led to the self-reinforcement of distrust as exemplified by the declining trust in U.S. news media. Distrust does not only complicate the public's ability to distinguish between true and false information, but it also causes skepticism which can lead to a reluctance to believe all content, exacerbating the challenge of identifying misinformation [119, 120]. The spread of misinformation is not just a technological issue but also a cognitive and emotional one. Pennycook and Rand [121] explore various motivations behind sharing misleading news on social media, from emotional responses to identity affirmation, as suggested by Brady et al. [122] and Chen et al. [123]. They suggest that misinformation often gains traction because it resonates emotionally or aligns with pre-existing beliefs, rather than due to a lack of critical evaluation. Moreover, Pennycook and Rand [121] argue that the tendency to use social media for relaxation exacerbates the issue. Pennycook and Rand [109] add that misinformation often thrives when individuals favor intuition over analytical thinking, a cognitive bias that can lead to the acceptance of false information.

Combating misinformation is a multifaceted challenge; particularly due to the impracticality of monitoring every online post for misinformation [6]. Aslett et al. [118] emphasize that while social media has been a significant focus in the fight against misinformation, other components of the digital information ecosystem have received less attention. This oversight suggests a need for a more comprehensive approach that encompasses various digital platforms and information sources. On a positive note, Pennycook and Rand [109] observe that users' susceptibility to false information is less driven by political beliefs than commonly feared. Moreover, most users report that they fact-check information they encounter online at least once per day [118]. In a similar vein, Constine [124] highlights the efforts of social media companies, civil society organizations, and government agencies in promoting campaigns that encourage users to verify news through search engines. Verifying the information which they encounter online via additional sources empowers individuals to critically evaluate information and reduce the influence of misinformation.

4.1.2 Combating the Propagation of Misinformation

The digital age has seen a surge in misinformation, prompting an extensive body of research dedicated to understanding who shares false content and the potential measures to counteract this trend. Guess et al. [125] and Osmundsen et al. [126] highlight the increasing attention on identifying the profiles of individuals who believe and disseminate misinformation, as well as exploring effective interventions. A key challenge in the field, as identified by Guay et al. [119], is the lack of focus on fundamental aspects essential for evaluating the effectiveness of interventions and understanding different groups' susceptibility to misinformation. They observe varying research designs, such as surveys where respondents rate false content's believability and likelihood of sharing [127, 128] or studies that mix false and true content [125, 126]. Guay et al. [119] stress the need for a unified approach to advance the field, proposing a framework that includes a mix of true and false content in evaluations and emphasizes discernment analysis – rather than only analyzing responses to false items.

Pennycook and Rand [121] explore the potential of accuracy nudges as an innovative intervention to improve the quality of content shared on social media. These nudges address the disconnect between belief and sharing, primarily driven by inattention rather than intentional dissemination of false information. They argue that nudges are advantageous due to their quick administration and no requirement for prior knowledge of news stories' accuracy. This approach contrasts with more traditional interventions like debunking or fact-checking, which have been extensively reviewed in other studies [129, 130].

Emphasizing digital literacy is another vital strategy in mitigating misinformation spread. Arechar et al. [111] find that simple digital literacy tips positively impacted the accuracy of news people were willing to share. Additionally, Aslett et al. [118] highlight the importance of media literacy programs, pointing out the significant yet often overlooked role of search engines in the information environment. They highlight that despite common beliefs, empirical evidence is lacking on whether searching online reduces belief in misinformation. Investigating the potential of users utilizing online search as a tool for validating information and preventing the propagation of misinformation, Aslett et al. [118] present findings that challenge common perceptions. Their research indicates that using online search to evaluate the truthfulness of false news articles can, in fact, increase the likelihood of believing them. This effect is particularly pronounced among individuals who encounter lower-quality information through their searches, pointing to the risk of entering "data voids" where misinformation is corroborated by low-quality sources. It appears that users seeking to validate potential misinformation with online searches are more likely to encounter lower-quality information than when they inquire about truthful news [118].

4.1.3 Methods of Identifying Misinformation

While the above-mentioned strategies of combating misinformation propagation are promising, a core aspect of research endeavors remains the ability to identify false claims before users are exposed to them. Fact-checking, as defined by Guo et al. [131], is the process of verifying the truthfulness of claims in written or spoken language. This practice is deeply rooted in journalism and has been an integral part of the media ecosystem for promoting accurate reporting.

Traditional fact-checking methods, which involve **manual verification** by organizations like PolitiFact, are also a common procedure in the publishing industry, where newspapers, magazines, and books undergo fact-checking before publication. This practice is not only crucial for ensuring the accuracy of reporting but also plays a significant role in maintaining public trust in media outlets.

DeVerna et al. [132] assert that fact-checking has shown effectiveness in reducing the belief in and the intention to share misinformation. This effectiveness spans across various cultural contexts, underscoring the universal value of fact-checking in combating misinformation. However, they note the limitations in scalability due to the overwhelming volume of online information and the time fact-checking takes [109], highlighting a critical challenge in the widespread implementation of fact-checking.

Martel et al. [133] bring to light the potential of involving non-expert individuals in the fact-checking process. They emphasize that active users, who are often the first to encounter new information online, can provide rapid responses to misleading content. This approach is rooted in the concept of the "wisdom of crowds," where the collective judgment of laypeople can effectively identify low-quality news sources and inaccurate posts. Research summarized by Martel et al. [133] demonstrates that crowd ratings are strongly correlated with professional fact-checker ratings, suggesting the viability of this method. However, they also caution against the risks posed by bad actors who might strategically undermine crowd-based evaluations.

In spite of its effectiveness, the scalability of manual fact-checking remains a significant issue [106]. They note that third-party fact-checkers are capable of evaluating only a small portion of the daily online content, indicating the need for more scalable solutions. The time-intensive nature of fact-checking, often taking hours or even days [131], adds to these challenges. Furthermore, Kazemi et al. [134] point out that the scalability problem is even more pronounced in non-English contexts. Martel et al. [133] also observe that the growth of professional fact-checking organizations has not kept pace with the rapid expansion of online content.

To enhance the scalability of fact-checking efforts, some researchers have shifted focus to evaluating (news) domains. Lin et al. [106] discuss the use of news domain quality ratings as a scalable alternative to evaluating individual pieces of content. By comparing different sets of expert ratings, they found a general consensus on the quality of news domains, suggesting that domain-level assessments can be a valuable tool for evaluating news quality. However, they also acknowledge that different rating systems may assess domains on varying criteria, creating potential inconsistencies in domain evaluations. Therefore, Lin et al. [106] imputed and synthesized the varying news domain datasets using Principal Component Analysis (PCA) — resulting in a rating set of 11,520 domains.

In response to the overwhelming volume of online content, there has been an urge towards developing **automated methods** for fact-checking. Various studies note that both researchers and social media platforms are focusing on developing such automated systems for detecting misinformation [131, 135–137]. Determining if a claim has already been fact-checked is part of the attempt to develop scalable solutions, in order not just to handle the influx of new information but also to efficiently manage and reference existing fact-checks to avoid duplication of effort [138].

Zeng et al. [139] articulate the requirements for a comprehensive automated fact-checking system. Such a system must excel in several key areas: identifying claims that need verification, retrieving relevant evidence, accurately assessing the claims, and providing justifications for its conclusions. These components ensure that automated fact-checking can match the thoroughness and reliability of manual processes.

The advancement of Large Language Models (LLMs) represents a significant breakthrough in automated fact-checking. Ye et al. [140] and Qin et al. [141] highlight that these models, trained on extensive and diverse datasets, have shown exceptional performance in key fact-checking tasks. For instance, recent studies have demonstrated that models like ChatGPT can effectively rate the credibility of news sources [142] and perform fact-checking tasks [143], showcasing their potential as comprehensive tools in misinformation detection. At the same time, DeVerna et al. [132] caution against the uncritical use of machine learning in fact-checking, highlighting the potential harms when AI-generated fact checks inaccurately label news items. Their findings advocate for cautious policy considerations and underscore the need for human oversight in AI applications for fact-checking. This notion boils down to the trade-off of human fact-checking's effectiveness and AI-generated checks' resource-efficiency as time remains a critical factor in combating misinformation.

Kazemi et al. [134] introduce the concept of claim matching as a means to scale up fact-checking efforts. This approach involves identifying content shared online, like text messages or posts, containing claims that are found in datasets of already fact-checked claims. The

method of claim matching is strongly related to the idea of semantic textual similarity (STS). STS, as a measure of the similarity in meaning between sentences, is an actively researched area with dynamic benchmarks like the Semantic Textual Similarity Benchmark. However, a high semantic similarity does not necessitate a high similarity of meaning. For illustrative purposes, assume the correct claim *"The Nigerian economy grew by about 3% in 2022"* [144]. A social media post stating *"The Nigerian economy grew by 20% in 2022"* might score high in semantic similarity but would arguably qualify as misinformation. Moreover, claim matching extends beyond just matching meanings, focusing on the applicability of a single fact-check to multiple claims.

In an increasingly interconnected and multilingual world, the ability to process and group similar claims in various languages is vital. Kazemi et al. [134] discuss their work on creating datasets in languages ranging from high-resource ones like English and Hindi to low-resource languages like Bengali, Malayalam, and Tamil. This multilingual approach is crucial for global fact-checking organizations aiming to extend their reach and enhance their efficiency in different linguistic contexts.

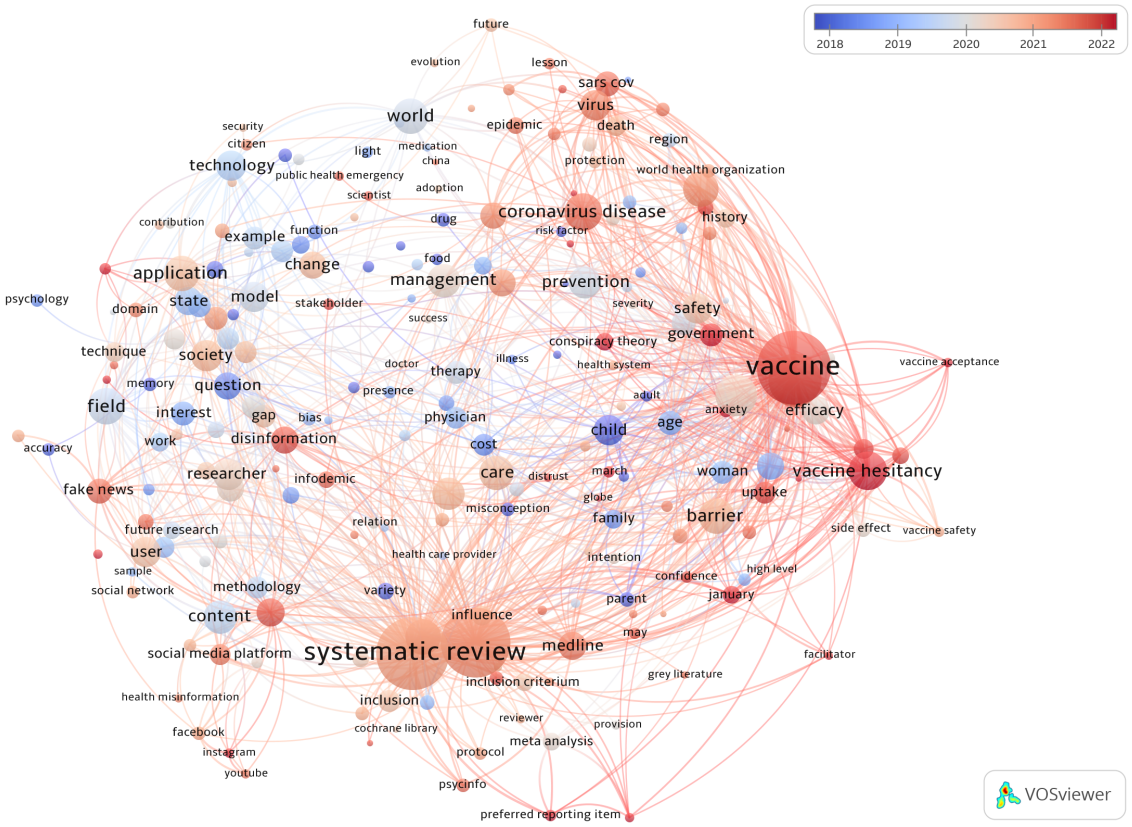
Guo et al. [131] address the critiques of fine-grained labels used by fact-checking organizations. Labels like "mostly true" often represent composite claims with varying degrees of accuracy, which can complicate the automated fact-checking process. Understanding the complexity of these meta-ratings is crucial for developing automated systems that can accurately assess composite claims.

4.1.4 Misinformation Research in the Context of Aid, Development, and Conflicts

Misinformation's impact in critical areas like development, aid, and conflict is profound and multi-dimensional. Tucker et al. [15] and Van Bavel et al. [16] highlight how misinformation goes beyond just spreading false information — it can deepen political divisions and even threaten the very foundations of democratic societies. In conflict zones or during humanitarian crises, misinformation can intensify stigma against marginalized groups and misrepresent critical situations like natural disasters, thereby complicating and impeding effective aid delivery and peace-building efforts. Unfortunately, with the notable exception of COVID-19 and related topics, very little research has looked into the intersection of misinformation and crises (see Figure 4.1).

Martel et al. [133] point out that even major social media platforms, including Facebook, Instagram, and Twitter, predominantly rely on professional manual fact-checkers in their efforts of combating misinformation. The limited number of fact-checkers, especially

Figure 4.1: Bibliometric Network of Most Relevant Terms in Literature on Misinformation over Time.



Note: The graph was created with VOSviewer. The approach is described in Appendix A.

in languages and countries with fewer resources, presents a significant challenge in efficiently moderating content on a global scale. As a great share of conflicts and natural disasters particularly affect these resource-pressed countries and regions, the pernicious risks of misinformation are exacerbated. Arechar et al. [111] note that misinformation's negative impact is starkly more pronounced outside of North America and Western Europe. In cases like Myanmar and India, misinformation on social media platforms has had dire consequences, from contributing to genocide against minority groups to inciting mob violence [66]. The study by Arechar et al. [111] across 16 countries reveals significant cross-cultural variations in susceptibility to misinformation, influenced by factors such as cognitive styles, democratic values, and individualistic versus collectivistic tendencies. These findings underscore the need for interventions that are finely attuned to the diverse cultural and social landscapes of the Global South.

The variability in how different cultures respond to misinformation is crucial to understanding its global reach. Pennycook and Rand [121] emphasize the necessity of cross-cultural

research in this domain, given the international scope of online misinformation. Arechar et al. [111] further illustrate this point by showing that individuals from countries with more open political systems and lower power distance tend to be better at distinguishing between true and false news, highlighting the interplay between cultural, political, and systemic factors in shaping responses to misinformation.

Badrinathan and Chauchard [145] critically assess the current state of misinformation research, pointing out a glaring gap in its focus on the Global South, home to the majority of the world's population. They argue that the majority of existing studies are centered in the Global North, often overlooking the unique contexts of the Global South. This research gap is especially problematic given the differences in digital platforms used, information dissemination mechanisms, and cultural norms surrounding information sharing in these regions. Milmo [146] further adds to this narrative by discussing the challenges of content moderation in non-English languages and in the Global South.

The eminent needs of advancing misinformation research outside of North America and Western Europe culminates in this chapter's central inquiry, that I will discuss with a geographic focus on Nigeria in the following sections.

Research Question of Chapter 4. *How, if so, can natural language processing techniques be employed to effectively understand and identify misinformation on social media?*

4.2 Methodology

4.2.1 Data

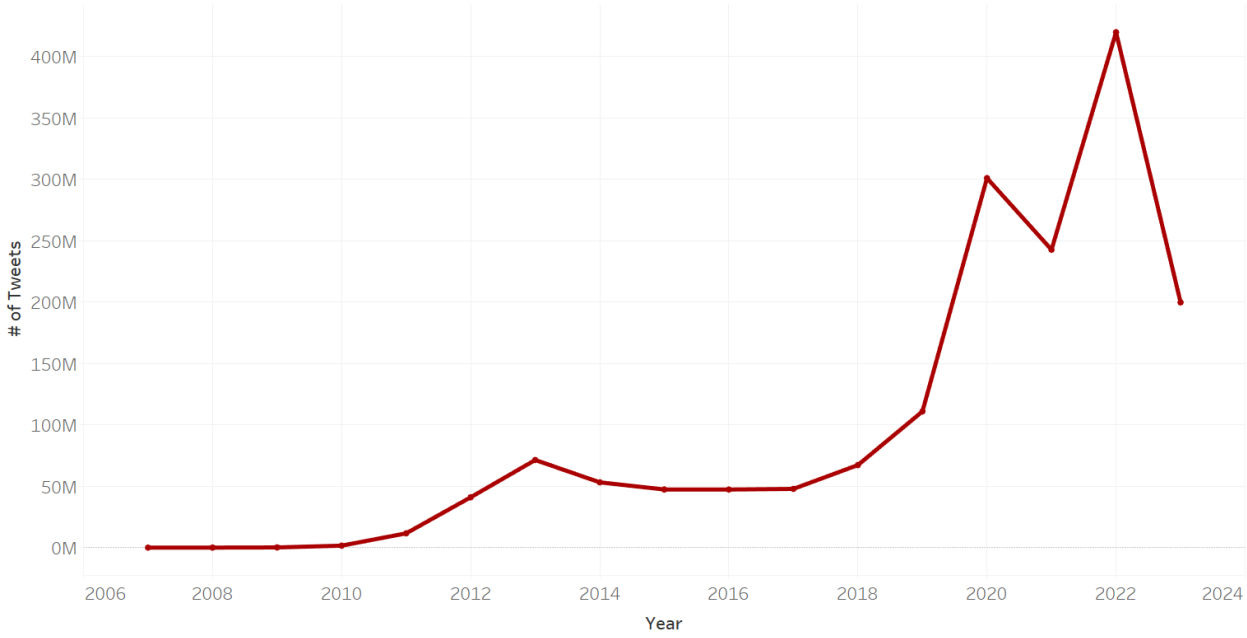
This project utilizes three primary data sources: Twitter data, fact-checked claims, and news domain quality ratings. Each source provides unique insights into the landscape of misinformation and its dissemination on social media, particularly focusing on Nigeria.

Tweets and External Domains

The rationale for choosing Twitter data hinges on the platform's significant influence in shaping policy debates, attitudes, and behaviors. Twitter stands out for its ability to amplify the voices of a few influential individuals, thereby impacting millions both directly and through its influence on other social media platforms. The platform's real-time data accessibility and diverse user demographics make it an ideal source for analyzing misinformation spread across various communities.

In my work, I turn to a comprehensive dataset collected by a research team at the World Bank with which I have been collaborating. For data collection, they harness the Twitter

Figure 4.2: Number of Tweets over Time on Nigerian Twitter.



Note: The drop in annual tweets in 2023 is explained by two factors. First, the study only considers data collected until October 2023. Second, the recent changes to the Academic Twitter API have restricted the ability to collect data.

API. This dataset encompasses the timelines of users with at least one tweet in the Twitter *Decahose* [1] and with an inferred profile location in Nigeria. Using snowball sampling, one can curate a dataset of approximately 2 million Nigerian Twitter users, a substantial representation of the total Nigerian Twitter user base as per Africa Check [147]. This dataset includes around 1.6 billion tweets on Nigerian Twitter (12 billion tweets in total) from March 2007 to July 2023, offering a voluminous and rich source for analysis. As Figure 4.2 reveals, the number of tweets in the dataset has increased over time, which aligns with my expectations for various factors. First, this increase can be attributed to the exponential growth in the volume of information available on the internet, as discussed by Aslett et al [118]. Furthermore, as social media platforms, including Twitter, continue to expand globally [72, 148], there is a corresponding rise in user engagement and content creation. This similarly applies in the developing world, where, according to Hunter and Biglaiser [19], there has been an increased penetration of social media. Second, the observed data trend might also be influenced by the challenges in accessing older tweets. As noted by Almuhimedi et al. [149], tweets tend to disappear from users' timelines and from search results upon deletion, either due to account deactivation or individual tweet deletions. Additionally, Jhaver et al. [150] highlight that content or users may be blocked or removed for various reasons over time,

[1] The Twitter Decahose refers to a dataset provided by Twitter that includes a 10% random sample of all public tweets in real-time.

which can affect the availability and visibility of older tweets in the dataset.

Thus, this initial descriptive analysis of the tweets data reveals a notable increase in prevalence and relevance of Twitter as a medium of communication in Nigeria. Approximately 75% of these tweets are in English, with the remaining showing a blend of English, Nigerian Pidgin, Yoruba, Igbo, and Hausa. This multilingual mix provides a unique window into the social media behaviors and trends in Nigeria.

Likewise, tweets are processed to obtain a collection of external domains and links shared on Twitter. From the greater Twitter dataset (12 billion tweets), I am able to extract about 2 billion (16%) links to external domains of which 924 million are unique links. A significant proportion of those external domains are shortened links, such as tinyurl.com. After expanding these shortened links and narrowing down the dataset towards the subset of tweets geographically linked to Nigerian Twitter users, I remain with a total of 7.5 million unique domains. Some of the most frequently shared domains are illustrated in Table 4.1.

Table 4.1: Most Frequently Shared Domains in Tweets on Nigerian Twitter.

Domain Name	**# of Links in Tweets**
instagram.com	92.0 million
youtube.com	87.3 million
facebook.com	67.4 million
...	...
blogspot.com	9.8 million
spotify.com	8.6 million

News Domain Ratings

As outlined in Section 4.1.3, the use of domain quality ratings recently emerged to combat the scalability challenge in information validation. As a continuation of the collection of domains shared amongst the tweets dataset, I collect a comprehensive set of news domains involved compiling an extensive list of 15,000 news domains via the open-source platform Media Cloud and public lists. This extensive list is designed to encompass a wide array of news outlets, ranging from well-established mainstream media to lesser-known niche publishers. The objective is to capture the broad spectrum of news sources that could potentially influence public opinion and discourse on Twitter.

Having compiled this list, the next step involves cross-referencing it with the Twitter dataset. This allows me to identify which news domains are being cited or referenced in tweets. By matching the domains' prevalence within the tweets (like the links in tweets mentioned above), I am able to pinpoint the news sources and other external domains actively engaged by Twitter users in Nigeria.

To expand this dataset to include quality ratings, I employ the — to my knowledge — most comprehensive list of domain quality ratings, presented by Lin et al. [106]. These ratings are based on a range of criteria, including journalistic standards and factual reporting. Additionally, I incorporate assessments from the Nigerian non-governmental fact-checking organization Centre for Democracy and Development Fact Check[2], which has specifically identified 10 news domains as frequent sources of misinformation on Nigerian social media. This local perspective is invaluable, offering a context-specific lens through which the news landscape can be scrutinized.

Ground-Truth Fact-Checked Claims

Fact-checked claims are essential in this study as they offer a verified benchmark against which Twitter content can be compared. I aggregate 250,000 fact-checked claims from 1995 to 2023, primarily sourced from the Google Fact Check aggregator, which covers approximately 170 fact-checking organizations worldwide. Quelle et al. [151] point out that while the Google Fact Check aggregator is a good starting point for fact-checked claims, it still has limited coverage in some regions of the world. Therefore, with a specific focus on Nigeria, I supplement the collection by scraping fact-checks from three African fact-checkers (DUBAWA[3], Centre for Democracy and Development Fact Check[4], and The Cable[5]). By adding 1,260 fact-checked claims relevant to the Nigerian context, this allows me to extend the number of English fact-checked claims to 83,429.

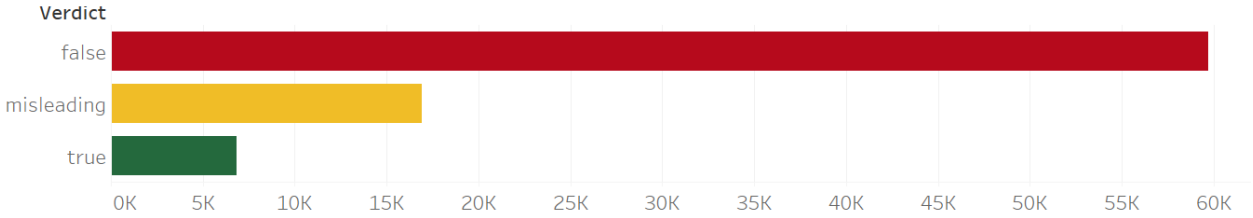
To ensure the credibility of these fact-checked claims, I verify the certification of all fact-checking organizations with the International Fact-Checking Network (IFCN)[6]. The IFCN is known for its stringent criteria in certifying fact-checking entities, thus ensuring that the claims sourced are from reputable and reliable organizations. This step is essential in mitigating potential biases that could arise from less credible sources. The importance of this validation step becomes particularly apparent through the incident of a Mexican fact-checking organization which emerged on the fact-checking landscape a few years ago and

[2]Centre for Democracy and Development: https://cddfactcheck.org/, last accessed January 4th, 2024.
[3]DUBAWA: https://dubawa.org/, last accessed January 4th, 2024.
[4]Centre for Democracy and Development: https://cddfactcheck.org/, last accessed January 4th, 2024.
[5]The Cable: https://www.thecable.ng/, last accessed January 4th, 2024.
[6]International Fact-Checking Network: https://www.poynter.org/ifcn/, last accessed January 4th, 2024.

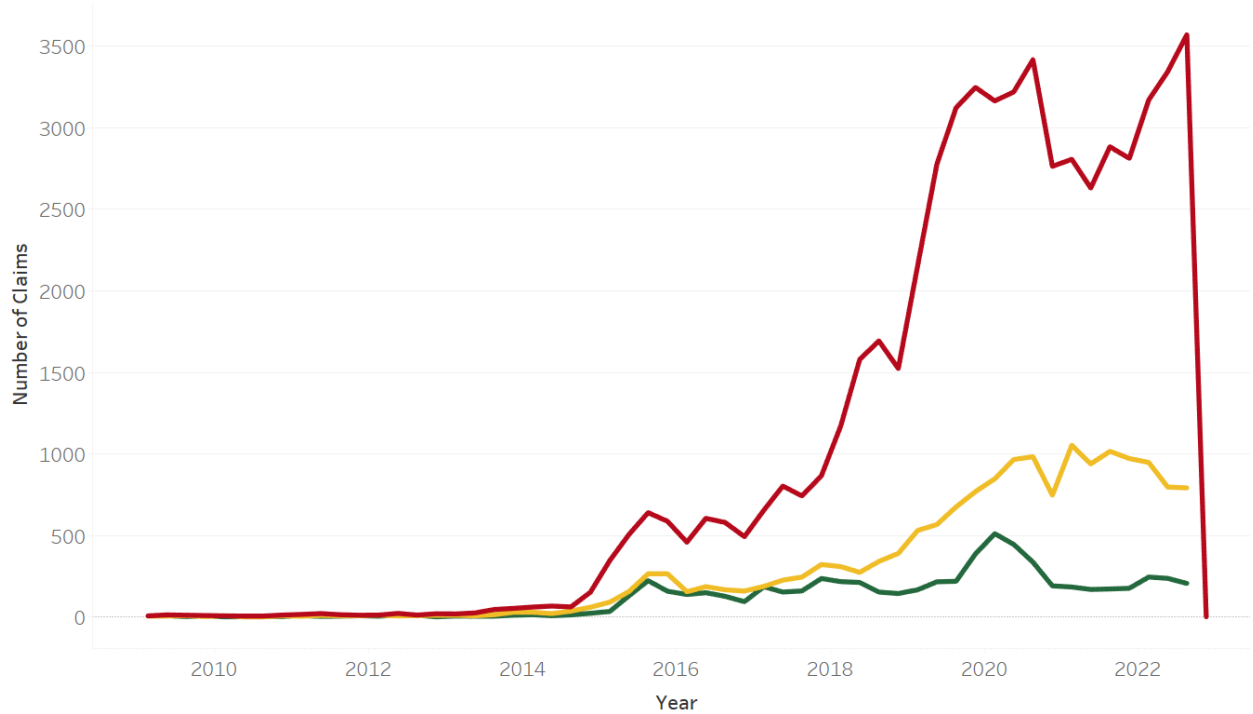
was accused of being biased in its assessments partly due to its ownership by no other than the Mexican president López Obrador [152]. The number of claims collected through the Google API is not affected by this validation – potentially due to Google performing a similar validation step since they actively collaborate with the IFCN [153]. Of the relevant Nigerian fact-checking organizations not covered by the Google API, I choose not to scrape the fact checks by Roundcheck as the organization is not verified by the IFCN. Thereby, I assemble a comprehensive yet highly valid dataset of fact-checked claims, both fake and true, for Nigeria.

Figure 4.3: Verdict Labels of Ground-Truth Fact-Checked Claims.

(a) Overall Verdict Distribution



(b) Verdict Distribution over Time



Note: False claims in red, misleading claims in yellow, and true claims in green. The drop in claims in 2023 is explained by the data being collected earlier in the year.

As illustrated in Figure 4.3, the dataset of fact-checked claims is skewed towards false claims (59,712), followed by misleading (16,891) and true claims (6,826). This distribution

is not surprising given that I expect fact-checking organizations whose data I am utilizing to be focused on identifying and highlighting false information. Oftentimes, they approach their work with the aim of debunking misleading or false claims rather than "confirming" true claims that seem unlikely to be true. For the later purposes of identifying fake claims amongst social media posts, this imbalance is not problematic since I compute the semantic similarity and mention of fake-claim-prone domains rather than prompting a trained model to identify if a claim is true or false (such as past work by DeVerna et al. [132]). I also observe an increase in the number of evaluated claim-verdict pairs over time in recent years. The dataset peaks with 3,571 ground-truth false claims recorded alone in the third quarter of 2023. While I observe an overall increase for all types of claims (that is, true, false, and misleading), the effect is particularly pertinent to observed false claims.

Figure 4.4: Verdict Label Distribution across Fact-Checking Organization over Time.

Publisher	Verdict	2019	2020	2021	2022	2023
factcheck.afp.com	false	980	1,554	1,732	1,268	1,111
	misleading	167	536	564	501	227
	true		1	1		
africacheck.org	false	280	56	64	108	69
	misleading	76	47	69	77	49
	true	93	59	61	97	77
dubawa.org	false		41	196	119	212
	misleading		16	76	32	69
	true		6	47	15	47
thecable.ng	false			36	69	54
	misleading			16	20	17
	true				4	11
cddfactcheck.org	false					92
	misleading					26
	true					39

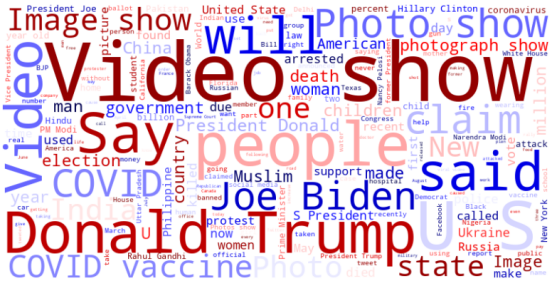
Note: The heatmap displays the yearly number of fact-checked claims per fact-checking organization.

In Figure 4.4, I zero in on a select group of fact-checking agencies with significant presence in West Africa, particularly Nigeria. These include prominent organizations such as those scraped, Agence France-Presse Fact Check[7], and Africa Check[8]. These entities not only provide fact-checks but also contribute to the growing awareness of misinformation in the region. Similar to the entire dataset with international scope, the Nigerian fact-checking agencies likewise focus on false and misleading claims. Furthermore, one can observe that more agencies have become active over the years, indicating the increasing importance of fact-checking.

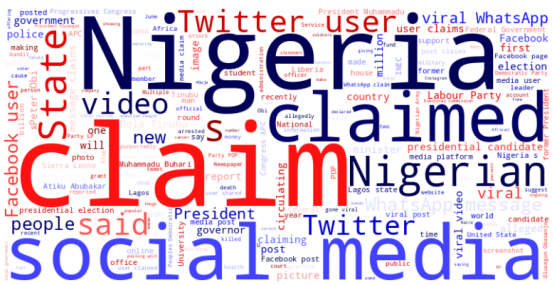
[7]Agence France-Presse: https://factcheck.afp.com/, last accessed January 4th, 2024.
[8]Africa Check: https://africacheck.org/, last accessed January 4th, 2024.

Figure 4.5: Word Clouds of Fact-Checked Claims.

(a) All Claims



(b) Claims Verified by Nigerian Fact-Checkers



Note: The word clouds were created using the wordcloud Python package. English stopwords were removed before the visualization.

A first glance at the frequency of word tokens in the claims dataset reveals some more information on the nature of fact-checked claims (see Figure 4.5). To draw a comparison between the overarching dataset on fact-checks and a geographically more narrow perspective on Nigeria, I map the word clouds of all claims and only those by the Nigerian agencies whose fact-checking repositories I scraped. I observe the US-centric focus of the international dataset through tokens like Donald Trump and Joe Biden. As for the Nigerian perspective, social media appears to be a relevant theme with specific platforms such as WhatsApp and Twitter – likely because many claims emerge and are being propagated via social media.

4.2.2 Modeling

As outlined earlier, this chapter seeks to assess if one can accomplish two goals with the use of natural language processing: (1) identify topics amongst fake claims and track the topical distribution over time, and (2) identify fake claims amongst social media posts. I proceed to detail my approach below.

Identifying Themes Amongst Fake Claims

In my exploration of the claims dataset, I move beyond the preliminary insights provided by word clouds to adopt a more systematic and rigorous approach to uncovering thematic structures within the data. To delve deeper, I implement a process that can be described as a semi-supervised approach to topic modeling.

The data consists of claims lacking "ground-truth" topic labels that would allow for straightforward categorization. This necessitates an approach that can infer structure in an unsupervised manner but also incorporates human judgment in the final designation of topics. To this end, I employ *Latent Dirichlet Allocation (LDA)*, a widely-used topic

modeling technique that identifies themes from a collection of texts by grouping together words that frequently co-occur. *LDA* differs from more sophisticated methods like *BERTopic* in several key aspects:

- **Probabilistic framework**: Unlike *BERTopic*, which generates dense vector representations, *LDA* is a generative probabilistic model that assumes each document is a mixture of a small number of topics and that each word's creation is attributable to one of the document's topics.

- **Interpretability**: *LDA* focuses on maximizing the co-occurrence of words within topics, which often results in more interpretable clusters of words, forming the basis for topic identification. In contrast, transformer-based models like *BERTopic* produce dense vector representations that, while powerful for capturing nuanced semantic relationships, do not translate directly into easily interpretable topics (as I observed when working with the claims data).

- **Scalability and simplicity**: *LDA* is generally more straightforward to implement and interpret compared to deep learning approaches, making it suitable for datasets where deep contextual embeddings might not be essential – as is the case in this task.

Moreover, the nature of the dataset in question — claims data — may not necessitate the depth of context that *BERT* provides. *LDA*'s bag-of-words assumption, while a simplification, is often sufficient for capturing the thematic essence of documents where context beyond immediate word co-occurrences is less critical. In cases where thematic breadth is more informative than depth of meaning in individual sentences, *LDA*'s approach is more aligned with the research objectives.

In the specific context of my research, where the goal is to discern broad themes and track their evolution over time, *LDA* provides a balance of depth, interpretability, and computational tractability. The semi-supervised aspect of my methodology, which combines *LDA*'s unsupervised learning with subsequent human judgment, leverages *LDA*'s strengths and compensates for its limitations, ensuring that the resulting topics are both statistically sound and thematically meaningful.

After pre-processing the text data with the common steps of lemmatization, lowercasing, removing stopwords, and removing special characters, I convert the text into a matrix of token counts with the help of a *CountVectorizer* since *LDA*, being a probabilistic model, requires numerical input. *LDA* models require the choice of the number of clusters as an input, wherefore I experiment with varying model settings (from 5 to 15 topic clusters).

The model outputs a set of topics, each represented as a collection of words statistically determined to be most informative and expressive of the respective topic, which I then investigate to understand the predominant themes in the claims. In the step of inferring these topic labels, the supervised aspect of the methodology comes into play. I pay particular attention to exclude words that, while potentially frequent in the corpus, do not contribute to the thematic significance of topics. This step ensures that the topics reflect the substantive content of the claims, rather than being skewed by frequently occurring but thematically irrelevant terms. After the topic labels are assigned, one is able to analyze the topic occurrence and distribution over time and geography discussed in Section 4.3.

Detecting Fake Claims Amongst Social Media Posts

In order to assess the ability to utilize NLP methods to detect misinformation amongst tweets, I employ a combinatorial analysis utilizing the domain quality ratings obtained by Lin et al. [106] and the ground-truth fact-checked claims. Essentially, I evaluate every tweet in a two-step process akin an ensemble method.

First, I examine the prevalence of news domains classified as non-credible within the links shared by Nigerian Twitter users in their tweets. If a domain is matched with a tweet's text sub-string, the tweet is flagged with its respective domain rating. The ground-truth dataset of 11,000 news domains has a quality rating on a scale from 0 to 1 for each domain. These ratings serve as a proxy for the credibility of the tweets which share them, with lower scores indicating a higher likelihood of misinformation. Since the domain rating is a high-level assessment of the tweets' credibility, I expect lower precision for this part of the ensemble. On a positive note, this higher-level evaluation likely yields a relatively high coverage. This component of the methodology is motivated by prior endeavours of Lazer et al. [110] and Grinberg et al. [154].

Second, I utilize the concept of semantic similarity as most recently employed by Kazemi et al. [134] and Kazemi et al. [138]. The goal is to identify tweets that are semantically similar to the collected fact-checked claims in the ground-truth dataset. In contrast to the domain ratings, this approach has a lower coverage but also has much higher precision due to its claim-tweet-level granularity. There has been an increasing number of transformer-based contextual embedding models attempting to enhance the goal of semantic similarity computation [134]. Vo and Lee [155] show that semantic similarity algorithms are a potential pathway to identify misinformation. While I focus on individual claims, they tackle the related task of identifying fact-checked reports amongst multimodal media. My starting point is the Massive Text Embedding Benchmark (MTEB)[9] leaderboard by the open-source

[9]Hugging Face: https://huggingface.co/spaces/mteb/leaderboard, last accessed January 4th, 2024.

community Hugging Face. At the time of computation in October 2023, the *large* version of the *General Text Embedding (GTE)* model presented by Lin et al. [156] leads the rankings of the Semantic Textual Similarity (STS) metric, which evaluates the performance for semantic similarity tasks. Moreover, the *GTE-large* model is applicable to various downstream tasks of text embeddings since it is trained on a large-scale corpus of relevance text pairs across various domains and scenarios [156]. It can cater English texts and lengthy texts of up to 512 tokens which is sufficient since the tweets are relatively short due to Twitter's character limits. For now, the methodology focuses only on English claims and tweets, wherefore I actively chose the faster, English-only *GTE-large* model rather than multi-lingual alternatives with comparable performance for semantic similarity tasks. In future iterations of this study with labelled claims in the relevant local languages — such as Hausa, Igbo and Yoruba in Nigeria – I would opt for a multilingual model like the *Language-agnostic BERT Sentence Embedding (LaBSE)* presented by Feng et al. [157].

I adapt the *GTE-large* model for the task's specific needs with a 'few-shot learning' strategy of a targeted set of 30 cases. This approach is predicated on the premise that a carefully curated small dataset can be incredibly potent for model fine-tuning [158], especially when dealing with nuanced tasks such as identifying misinformation. Each case in this dataset is meticulously chosen to address some of the linguistic and contextual intricacies I expect to influence the model's ability of infering the similarity of tweets and fact-checked claims accurately. [10]

Firstly, the fine-tuning process is designed to heighten the model's sensitivity to geographic and socio-demographic references pertinent to Nigeria — such as cities, states, or persons of public interest like politicians -– enabling a sharper focus on the region-specific elements that often accompany false claims.

Secondly, I aim to refine the model's responsiveness to exaggerations – a common tactic in the propagation of misinformation [159]. For instance, an authentic claim might state that the *"GDP grew by 5%"* while a misleading assertion could exaggerate this figure to *"GDP grew by 20%."* Despite the semantic proximity of these statements, their implications are markedly different. The fine-tuned model is thus being calibrated to discern these subtleties, understanding that even slight numerical distortions can significantly alter the veracity of a statement.

Thirdly, I am enhancing the model's ability to recognize and interpret negations, which are often used to flip the truth value of a statement subtly. Negations can transform a factually accurate claim into misinformation, making their detection critical in the identification of false claims.

[10] The list of claim pairs used for the model fine-tuning are listed in Appendix C.

The resultant model is then utilized to compute the embeddings of tweets and claims in order to be able to calculate their respective semantic similarities. In preparation, I determine the subset of tweets relevant for the semantic similarity computation. I filter the greater Twitter dataset of 2 billion tweets to English tweets and remove those tweets that only contain a URL since I capture those with the domain rating step. The data is further subset to tweets containing at least 3 word tokens, which removes empty tweets and still contains those that could potentially match some of the shorter claims in the fact-checked dataset like *"Alcohol kills coronavirus."* The final dataset contains 1.3 billion tweets.

Determining the semantic similarity across 83,000 claims times 1.3 billion tweets is computation is intensive. Therefore, I split the dataset into 1,500 batch files which I run via a job array of size 1,500 with 4 GB memory each on a cluster environment. I utilize the open-source JIT compiler *Numba*, which translates the subset of my Python code into faster machine code and allows me to calculate the cosine similarities between the embeddings of claims and tweets more efficiently. To maintain a scalable approach of storing the semantic similarity pairs of tweets and claims, I only store those pairs with a similarity score >0.9 (on the scale from 0 to 1 with 1 being identical). The cut-off threshold was chosen based on an initial test on a subset of the data and the status quo of the research community in this space [160, 161]. The embeddings generated by transformer-based models in the realm of semantic similarity are prone to only occupying a small part of the embedding space – due to the isotropic distribution of embeddings – wherefore similarity distributions usually have a high peak at about 0.7 and reasonable matches no lower than at 0.9 [160, 161]. As shown in Table 4.2, a semantic similarity score of 0.9 does not necessitate a match, but rather can I safely assume that tweet-claim pairs with a lower similarity are no matches. The findings with a detailed evaluation of the scores is discussed in Section 4.3.

4.3 Results and Findings

4.3.1 Identified Misinformation Topics

In the supervision step of the topic modeling that I outline in Section 4.2.2, I investigate the most relevant words of each topic cluster in the claims dataset – that is, the words statistically determined to be most informative or expressive of the topics. Therefore, the identified topics are a direct outcome of the claims data rather than a representative "ground-truth" topic landscape online. Interestingly, I observe specialized clusters that represent *Indian Politics* (key terms include Modi and BJP) and *US Politics* (key terms include Trump, Biden, and Obama). The set of identified topics aligns with my observations that the Google API

Table 4.2: Examples of Semantic Similarity Computation.

Fact-Checked Claim	Tweet	Similarity
A Facebook post claims the husband of the popular gospel singer, Osinachi Nwachukwu, has been sentenced to death.	Over the past few days, multiple social media posts, especially on Facebook, had claimed that Peter Nwachukwu, husband of late Osinachi, popular gospel singer, has been sentenced to death by hanging.	0.967
The rise in inflation rate is a global challenge due to COVID-19 and is not peculiar to Nigeria.	Inflation is worldwide and caused by the COVID-19 pandemic, not tied to Muhammadu Buhari's administration	0.933
A Twitter handle tweeted a claim that the education minister, Adamu Adamu, said the ongoing ASUU strike will be called off within the next one week.	A WhatsApp post claims ASUU has suspended its strike action indefinitely.	0.900

– the main source of verified claims – contains many fact checks from Indian-based and US-based fact-checking agencies. Aside from these two topics, the model's clusters cover a comprehensive topic space, as illustrated in Figure 4.6, wherefore I believe the below-illustrated findings are insightful in understanding the misinformation landscape.

With the topics delineated and labeled, I investigate the topic distribution from two perspectives: global examination of all claims, alongside a more localized investigation focusing on claims checked by Nigerian fact-checkers. This allows me to understand both the overarching themes at play in the set of verified claims and the specific narratives that emerge within the Nigerian context.

One immediate observation for the "global" view on the claims topics is that there is a relatively even distribution of fake claims across topics, which might be partly a result of topic models seeking to form cohesive, similarly sized clusters. However, introducing the temporal disaggregation, one can make some interesting observations. As such, in the lead-up to the last US presidential elections, the number of related fact-checks more than doubled. Likewise, one can clearly identify the occurrence of the *COVID-19* pandemic and related fact-checks, as well as an increase in *Global Politics and Conflicts* in 2022. The latter can be traced back to Russia's war on Ukraine with many claims in this cluster referring to claims on military movements, battles, and other countries' actions or statements related to the conflict.

Zooming in on the fact-checks by Nigerian fact-checking organizations, I observe a national perspective on claims that differs in its thematic distribution from the international

picture. Particularly, *Social Media* related claims (such as fake claims with their origins on platforms like YouTube and Facebook) are very prevalent and increasing in number. Likewise, *Health and Education* is a very prominent theme, which I would partly attribute to the many statements of politicians on the state of the country that I found in the dataset. Ultimately, I can thereby also identify how the public dialogue changes and how countries' "development markers" move to the center of attention in times of elections, when politicians seek to position themselves – for instance, in relation to the current government's "performance."

An overarching observation I make is on the relative prevalence of false, misleading and true claims that are investigated by the fact-checking organizations. While the deviances across topics are minor, it is interesting to observe relatively more true (and also misleading) claims on the *Health and Education* topic cluster, while the specific *COVID-19* cluster exhibits similar patterns as the other claim clusters. I hypothesize that this is potentially due to many investigated claims being statements on statistics by politicians and thought leaders rather than "dubious" claims.

Generally speaking, while the results intuitively make sense and offer an interesting view on the misinformation landscape, by continuing the analysis in the future, the value of a longitudinal perspective will emerge.

4.3.2 Identified Fake-Claim Matches

Table 4.3 presents the findings of the claim matching modeling – by showcasing the distribution of fact-checked claims categorized by their veracity. A noteworthy observation is that the model identified a significant share of claims in the Nigerian Twitter data (62,565 of the total 83,429) which suggests that the employed ground-truth claims data is appropriate for identifying fake claims.

When focusing on matches with higher confidence (that is, semantic similarity scores of $\geq$0.95 and $\geq$0.99), there is a slight increase in the proportion of both true and misleading claims relative to the total number of ground-truth fact-checks utilized in this chapter. This trend suggests that while all types of claims are identified amongst tweets, those with true or misleading content are more prevalent with higher degrees of confidence. While there is no clear explanation, a potential reason might be that the themes and geographic focus of false claims is different and less relevant for the Nigerian context than the true and misleading claims in my fact-checked dataset. It warrants further investigation to clarify the exact underlying dynamics.

Moreover, it is intriguing to observe that while the dataset of verified claims contains only

Figure 4.6: Misinformation Topics Identified with LDA (over Time).

(a) All Claims

Topic	2019	2020	2021	2022	2023
COVID-19 Pandemic	293	2,110	2,640	1,230	942
Global Politics and Conflicts	541	891	1,293	1,969	1,560
Health and Education	1,047	1,343	1,812	1,738	1,584
Indian Politics	701	1,086	1,563	1,684	1,311
International Affairs and Environment	586	1,062	1,361	1,104	1,101
Public Health and Safety	757	1,086	1,381	883	709
Religious and Cultural Issues	1,166	1,998	2,510	2,929	2,413
Social Media	708	1,254	1,716	1,559	1,453
Terrorism and Security	804	894	1,089	956	873
US Politics	1,375	3,090	2,193	1,684	1,349

(b) Claims Verified by Nigerian Fact-Checkers

Topic	2019	2020	2021	2022	2023
COVID-19 Pandemic	16	15	23	11	22
Global Politics and Conflicts	9	2	19	8	14
Health and Education	153	69	135	134	143
Indian Politics	17	3	12	16	24
International Affairs and Environment	54	23	39	30	37
Public Health and Safety	28	7	8	10	21
Religious and Cultural Issues	21	10	13	15	21
Social Media	74	70	264	246	335
Terrorism and Security	35	10	19	19	35
US Politics	41	16	33	51	110

Note: The heat maps contain the number of annual claims across the identified topics.

8.18% of true claims (see Table 4.3), the proportion of true claims amongst all matched tweets is 11.46% (see Table 4.4). This phenomenon might indicate that true information is more prevalent on the platform than often perceived, in line with findings by Guay et al. [119]. As a result, a significant volume of tweets are actually not propagating misinformation, leading to a relatively higher presence of true claims in the matched data. These observations run hand in hand with the hypothesis that the nature of the discourse on Twitter might be more factual than on other social media platforms. That Twitter's user base potentially more actively seeks to validate information could be a product of the platform serving professional engagement rather than purely leisurely pursuits.

Table 4.3: Claims Perspective of Semantic Similarity Matching.

Category	True	Misleading	False	Total
All Claims	6,826	16,891	59,712	83,429
	(8.18%)	(20.24%)	(71.58%)	(100%)
Claims with $\geq$ 0.90 Similarity Score	5,616	12,888	44,061	62,565
	(8.97%)	(20.59%)	(70.44%)	(100%)
Claims with $\geq$ 0.95 Similarity Score	1,113	1,552	4,388	7,053
	(15.78%)	(22.00%)	(62.22%)	(100%)
Claims with $\geq$ 0.99 Similarity Score	29	52	109	190
	(15.26%)	(27.37%)	(57.37%)	(100%)

Table 4.4 delves into the distribution of tweets matched with ground-truth labeled claims, revealing several key insights. Between the 1.3 billion tweets and 83,429 fact-checks, the model found 5,131,367 matches. Unsurprisingly, these matches are not one-to-one matches – that is, I both observe tweets matching with multiple claims and claims matching with multiple tweets. In the presented results, I refer to the unique number of matched tweets that I respectively assign to the claim it is semantically most similar to.

Most notably, the matches with the highest level of confidence (semantic similarity $\geq$0.99) are predominantly false claims. This stark prevalence of false claims at higher similarity scores could be influenced by the greater volume of labeled false claims in the dataset.

However, the shift in distribution from matches with a similarity score $\geq$0.9 to those $\geq$0.99 is indicative of deeper underlying patterns. One possible explanation for this could be the structural characteristics of false claims. It is conceivable that false claims possess certain linguistic or syntactic features that make them more identifiable and thus more likely to achieve higher similarity scores. For instance, false claims might be more succinct or use more definitive language, making them easier to match accurately. It is also possible that false claims are more homogeneous in their composition, making them more easily recognizable by the semantic matching algorithms.

In addition, the relatively lower percentage of true claims at the highest similarity threshold might suggest a greater diversity in the expression of true claims. While this hypothesis

requires thorough investigation, true statements may encompass a wider range of linguistic styles and content variations, which could lead to lower similarity scores when matched against a standardized dataset of fact-checks.

From these findings, I posit that the nature of misinformation and its detection is intrinsically linked to the linguistic and structural elements of the claims. Moreover, these results highlight the challenges in automatically distinguishing between true and false claims, particularly in the nuanced and varied landscape of social media discourse.

Table 4.4: Tweets Perspective of Semantic Similarity Matching.

Category	**True**	**Misleading**	**False**	**Total**
Tweets with $\geq$ 0.90 Similarity Score	355,221	641,997	2,102,208	3,099,426
	(11.46%)	(20.71%)	(67.83%)	(100%)
Tweets with $\geq$ 0.95 Similarity Score	7,332	10,647	47,758	65,737
	(11.15%)	(16.20%)	(72.65%)	(100%)
Tweets with $\geq$ 0.99 Similarity Score	218	131	8,138	8,487
	(2.57%)	(1.54%)	(95.89%)	(100%)

Next, I examine the topic distribution in tweets that match fact-checked claims, as detailed in Table 4.5. This perspective offers profound insights into the thematic focus of discussions on Twitter, particularly in relation to the accuracy of information.

One of the most striking observations is the dominance of certain topics across all levels of confidence in claim identification – as evaluated via the similarity scores. Notably, the *COVID-19 Pandemic*, *Health and Education*, and *US Politics* are prevalent, with their presence becoming increasingly pronounced at higher levels of confidence ($\geq$0.99 similarity score). For instance, discussions related to the *COVID-19 Pandemic* and *Health and Education* constitute a significant portion of tweets, peaking at 46.61% and 46.71%, respectively, at the $\geq$0.99 threshold. This indicates an intense concentration on health-related issues when it comes to "clearly" false information, reflecting the global concern and attention these topics received during the pandemic era. Similarly, *US Politics* emerges as a dominant theme, especially at the lower similarity threshold ($\geq$0.90 similarity score), suggesting engagement with political discourse on the platform.

However, the distribution of topics identified with greater confidence notably shifts.

Table 4.5: Topic Distribution Across Tweets Identified via Claim Matching.

Topic	**$\geq$ 0.90 Score**	**$\geq$ 0.95 Score**	**$\geq$ 0.99 Score**
COVID-19 Pandemic	7.96%	15.53%	46.61%
Global Politics and Conflicts	7.81%	7.19%	0.67%
Health and Education	9.98%	16.53%	46.71%
Indian Politics	9.16%	6.05%	0.38%
International Affairs and Environment	6.31%	5.17%	0.45%
Public Health and Safety	7.65%	8.33%	0.52%
Religious and Cultural Issues	10.59%	7.58%	0.14%
Social Media	14.29%	9.24%	2.64%
Terrorism and Security	7.07%	10.14%	1.33%
US Politics	19.18%	14.25%	0.55%
Total	**100.00%**	**100.00%**	**100.00%**

Note: Score refers to the semantic similarity score of the identified tweets and fact-checked claims.

While some areas like *Global Politics and Conflicts*, *Indian Politics*, and *International Affairs and Environment* decrease in their prevalence for matches with higher similarity scores, others such as *Public Health and Safety* and *Terrorism and Security* show a slight increase in presence. This shift could imply that discussions in the former set of topics are more linguistically diverse and, thus, less likely to be identified as close matches by the model, whereas the latter topics might have more defined and consistent narratives aligning with the fact-checked content.

Conversely, at the lower similarity threshold ($\geq$0.90 similarity score), there is a more balanced and diverse representation across all topics. This broader range indicates that a wider spectrum of discussions is captured when the matching criteria are less stringent, encompassing a more varied range of subjects and viewpoints.

Another interesting observation is that *Religious and Cultural Issues* are less present amongst the identified tweet-claim matches, opposed to their significant share in the ground-truth dataset. A possible explanation is that Twitter, or potentially social media platforms at large, are not typical venues for the dissemination of *Religious and Cultural Issues*. As the ground-truth dataset of claims is a composition of claims originated all forms of media, potentially, *Religious and Cultural Issues* are more prevalent in other forms of media like newspapers or television.

The prominence of topics around health and politics not only reflects public interest but also points to the potential for concentrated misinformation in these areas. The variation in topic distribution at different confidence levels underscores the complexity and diversity of information dynamics on social media platforms. This analysis provides a crucial initial step in understanding how different themes are represented and engaged with on Twitter, offering valuable insights into the patterns of misinformation propagation.

4.3.3 Prevalence of Untrustworthy Domains

As the second component of the detection endeavours of this study, I analyzed the prevalence of domain links amongst tweets, including their trustworthiness. I observe that a subset of 9.1 thousand out of the 11 thousand domains rated by Lin et al. [106] appears in the Twitter dataset. This figure, while representing a mere 0.1% of all domains shared, accounts for a significant 10% of all links shared — totaling approximately 219 million links. Closer scrutiny reveals that approximately 2,000 of these domains have ratings below 0.4, which previous studies categorize as predominantly disseminating misinformation (see Figure 4.8) [106]. These suspect domains contribute to 1.6% of all links shared, equating to 34 million links — a non-trivial volume that carries the potential for substantial impact on public discourse.

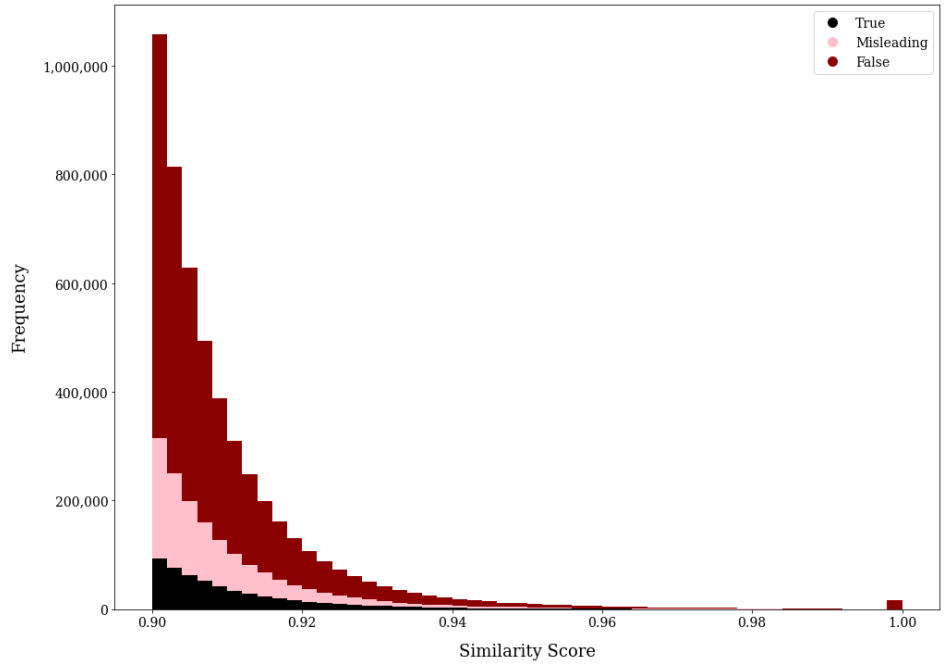
Two additional observations carry meaning.

First, more than half of all shared links are attributed to the top ten most prevalent domains (135 of 2019 million) which include *facebook.com* and *youtube.com*. With other social media platforms dominating the statistics, the interconnectivity of different platforms becomes prevalent. As users commonly move between apps to access information, any serious endeavours of combating misinformation should arguably take a cross-platform approach. Consciously evaluating this study's approach, it also introduces the challenge that one fixed domain rating is certainly not able to accurately capture the trustworthiness of millions of links shared that reference Facebook or YouTube.

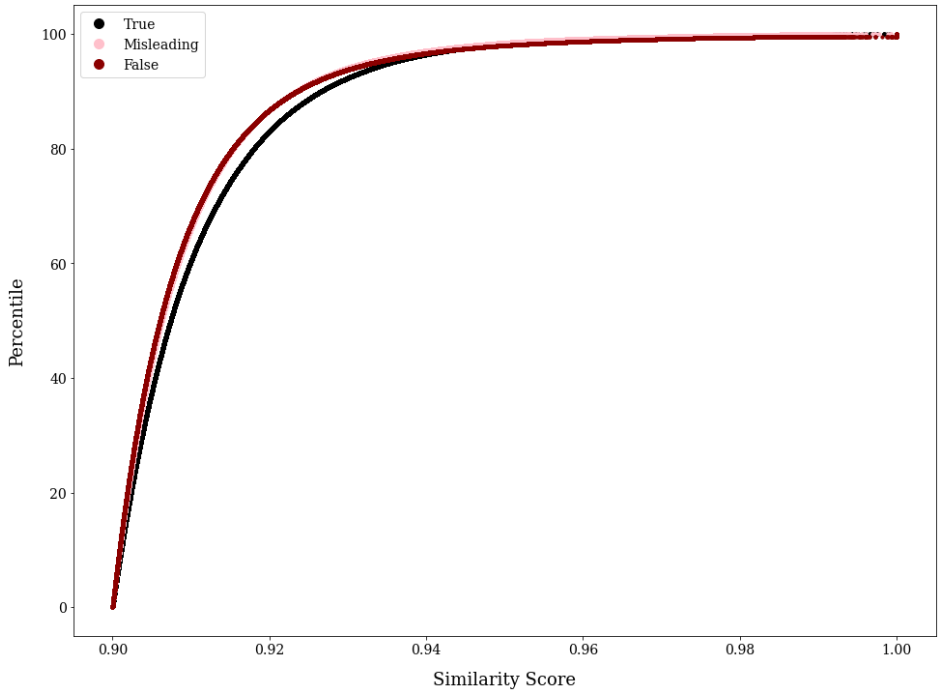
Second, various Nigerian news domains are present amongst the domain links shared in the dataset. As illustrated in Table 4.6, my modeling analysis identified a number of websites marked as not trustworthy – though limited in numbers. Facing the challenge of misinformation, on a positive note, reliable Nigerian news domains occurred significantly more often amongst the Twitter data, topped by *punchng.com* with over 2 million shared links.

Figure 4.7: Similarity Scores of Tweets Identified via Claim-Matching.

(a) Similarity Score Distribution



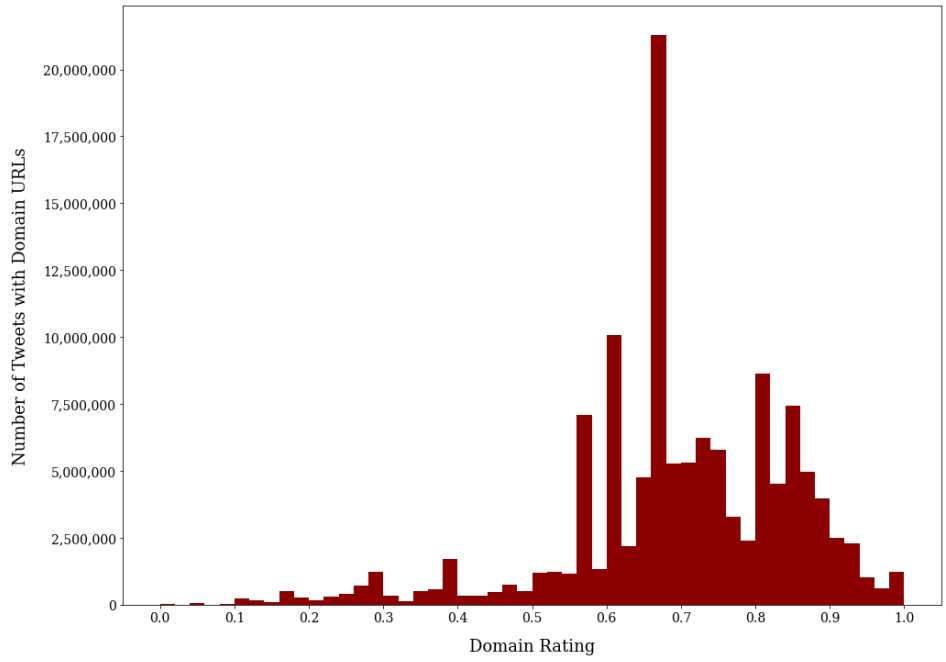
(b) Percentile Pattern



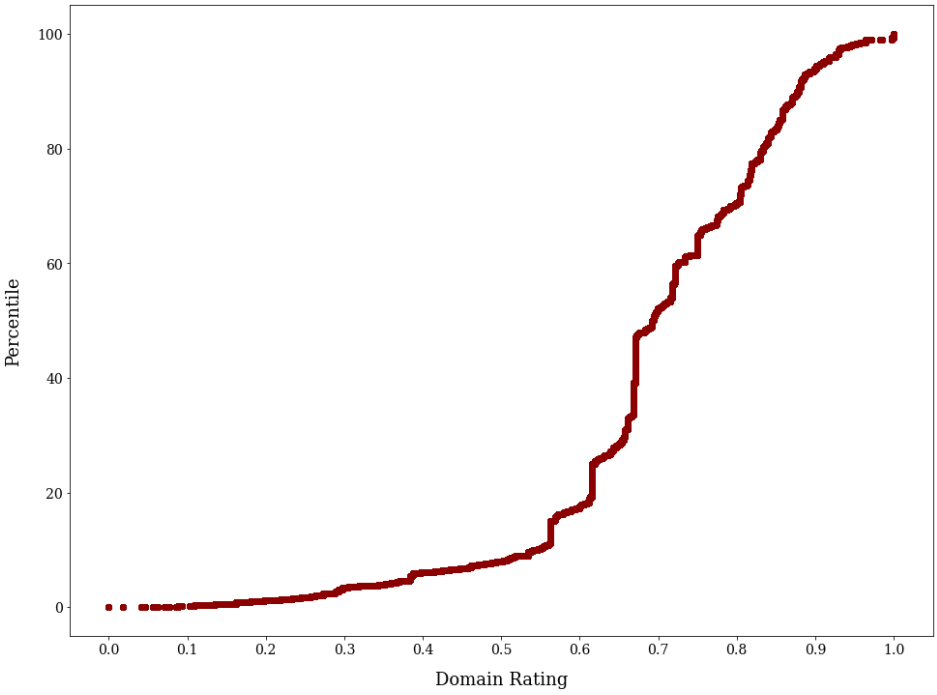
Note: The plots display the distribution patterns of all tweets that were identified to have a similarity score ≥0.9 – to any fact-checked claim.

Figure 4.8: Rating Distribution of Domains Identified in Tweets.

(a) Domain Rating Distribution



(b) Percentile Pattern



Note: The plots display the distribution patterns of all domain URLs identified in the tweets data. To prevent visual skew, the two most prevalent domains (facebook.com with 67 million and youtube.com with 27 million occurrences) are filtered here.

Table 4.6: Prevalence of Nigerian News Domains in Tweets.

(a) Untrustworthy Websites

Domain	# of Occurrences
reportera.ng	19,557
podiumsreporters.com	15,693
lagostoday.com.ng	4,625
thelagostoday.com	4,230
lagostoday.com	2,001

(b) Reliable Websites

Domain	# of Occurrences
punchng.com	2,123,137
guardian.ng	282,164
dailytrust.com	136,410
businessday.ng	92,127
blueprint.ng	29,400

4.3.4 A Composite Misinformation Probability Score

In Section 2.2.2 I delineate the contrasting nature of the two employed misinformation detection approaches: claim matching, characterized by a high precision but low coverage, and domain ratings, described by low precision but high coverage. The empirical observations from my study align with these characteristics, as evidenced by significantly more tweets mapped to domain ratings than claim matches (219 million versus 3 million matches). I observe an intersection of 158,565 tweets for which I both identified a fact-checked claim and a rated domain. Building upon these observations, I propose an ensemble approach that could potentially enhance the coverage of assessing information veracity on social media. In this approach which merges the insights from claim matching and domain ratings, I would first identify the more precise claim matching results. In a next step, for tweets without claim matches, I would employ domain ratings as a supplementary proxy of information reliability.

While this idea is straightforward and simple to implement, it is important to acknowledge the limitations inherent in domain ratings. For example, domains such as YouTube, labeled as less reliable, host a wide spectrum of content, ranging from unreliable to legitimate sources like news agencies. Consequently, relying on boilerplate domain ratings risks oversimplifying the complex landscape of online information.

To address these limitations and better capture the intricacies of information dissemination on social media, I suggest developing a probabilistic composite misinformation risk score in future research. This score would not only consider the content of individual tweets but also the underlying network structure and information propagation patterns, offering a more nuanced assessment of tweet veracity. The proposed approach for this composite score would include several steps:

1. **Initial score assignment**: Start by assigning the high-precision claim matching results as the primary source of likelihood of veracity, with similarity level S.

2. **Identify a user set** u_S: Create a set of users, u_S, who have shared tweets similar to a false claim, with a defined similarity level S.

3. **Calculate probability of domain occurrence**: For any domain d shared by users, calculate its probability of occurrence, $p(d)$, as $\frac{\text{number of tweets containing } d}{\text{total number of all tweets}}$. Given the benefits of geographically restricting the analysis, in this case one would restrict it to Nigerian Twitter.

4. **Determine probabilistic domain score**: Define the trustworthiness of a domain, $t(d)$, as $\frac{p(d|u_S)}{p(d)}$. This measures the excess probability of a domain being shared by users in u_S (that is, users who have posted tweets similar to false claims) relative to its base prevalence.

5. **Interpret probabilistic domain scores**: A score of $t(d) \leq 1$ suggests that the domain is less likely to be shared by suspicious users, indicating high trustworthiness. Conversely, high values of $t(d) > 1$ indicate lower reliability.

6. **Refine** $p(d)$ **calculation**: Optionally, refine the calculation of $p(d)$ by focusing on a sample of users who match u_S in terms of their activity patterns.

7. **Iterative refinement of domain proxies**: Continuously update and propagate the probabilistic veracity score, fine-tuning the domain proxies based on new data and patterns observed.

8. **Update user legitimacy scores**: Utilize the refined scores to reassess the legitimacy of users, considering the accounts they follow and the content they engage with.

While such a composite score was not in the scope of this chapter, it is a noteworthy extension that can also address the limits of claim matching and the domain ratings in an ever-growing pool of information on social media.

4.4 Key Takeaways on Misinformation in Crises

4.4.1 Limitations and Future Research

My approach to identifying misinformation is largely based on the value of claim matching, which while promising, presents certain limitations that must be acknowledged. Firstly,

as it requires me to employ comparatively complex models – GTE-large has 335 million parameters – it is a time-intensive process. This aspect could potentially limit the scalability and real-time applicability of the method. Secondly, there is still room for improvement in the model's implementation. For example, handling linguistic nuances such as negations is challenging and requires further refinement. While I already tune the model to address the issues, the model demands continuous fine-tuning to improve its accuracy and reliability. Thus, looking ahead, several avenues for future research emerge.

As indicated in Section 4.2.2, the expansion of this approach to other languages, especially using models like *Language-agnostic BERT Sentence Embedding (LaBSE)*, is promising. *LaBSE*'s capability to handle a number of local languages and its adaptability to the informal and hybrid linguistic nature of social media content could significantly enhance the method's performance. Another critical aspect is the inclusion of the temporal dimension in claim verification. The veracity of a claim can change over time, and this fluidity needs to be considered. For instance, a statement about a contemporary event might be false one day but become true the next. This dynamic nature of information necessitates a model that can adapt to the evolving truth of claims and track if a tweet shares this claim while it is true or false. A natural progression of this research would be to investigate the consumption patterns of tweets containing false claims or unreliable domain links. Given the detailed user data available in my Twitter dataset, it would be insightful to explore who shares misinformation, the followership network of these individuals, and the homophily between those sharing misinformation and those who do not. Considering the potential impact of misinformation on violence and conflicts, a pertinent area of study would be to examine how misinformation spreads during periods of heightened societal tensions. For example, tracking misinformation during events like the herder–farmer conflicts in Nigeria could reveal critical insights into how societal stress influences the propagation of false information.

These proposed areas of research not only address the current limitations but also open up new dimensions for understanding and combating misinformation in the complex landscape of social media.

4.4.2 Policy Implications

All in all, this chapter underscores the potential of natural language processing (NLP) in the realm of misinformation identification. By harnessing advanced NLP techniques, one can develop more effective tools for spotting and analyzing misinformation trends on social media platforms – thereby also informing policymakers.

One of the most significant takeaways is the insight provided by tracking topic distribu-

tions over time. Though technologically less complex, this aspect yields valuable information that can guide policymakers in allocating resources effectively. By understanding how different topics evolve and spread, governments and organizations can better target educational initiatives to counter misinformation and inform citizens about the prevalence of false claims.

Consistent with previous studies [109, 111, 116, 118, 119, 121], my findings reinforce the importance of misinformation as a policy issue. Policymakers must recognize the need to invest resources in combating misinformation, under the assumption that there is a genuine desire among leaders to enable the public to discern between falsehoods and truths. This is particularly crucial given that, in some cases, leaders or policymakers themselves might be directly or indirectly involved in disseminating agreeable misinformation, such as in "informational autocracies" [25, 67].

The research contributes to the ongoing debate, as framed by Tufekci [67], about whether social media acts as a curse or a blessing for democracies. Most recent societal movements have integrated digital connectivity in their tools to gain publicity and coordinate themselves [67]. However, by demonizing social media or even employing propaganda campaigns on platforms, authoritarian governments have made social media their new personal playground.

It also highlights the complexities in detecting misinformation. While it is easy to criticize big platform companies for their shortcomings in addressing misinformation, as exemplified by tragedies like the Rohingya genocide, my observations suggest that the task is far from simple. It requires complex solutions that likely still fall short of many nuances. Social media, as a digital mirror of the real world, reflects the biases and harmful views in our society and eases people's ability to spread these perspectives – at times via ideologically aligned false claims. This reality underscores a shared responsibility: It is incumbent upon society and policymakers not only to hold social media companies accountable but also to educate the public. Integrating media literacy into education curricula to teach individuals how to identify misinformation is a vital step.

Given these complexities, my research could serve as a catalyst for policymakers to take a more active role. Policymakers need to recognize the limits of technical solutions and consider broader strategies that encompass public awareness, education, and more nuanced regulatory frameworks. This multi-pronged approach offers a glimpse of hope in effectively addressing the spread of misinformation and its impact on society. The findings of this study highlight the need for a collaborative effort between technology experts, educators, policymakers, and the public to develop comprehensive solutions that safeguard information integrity in an increasingly digital world.

Chapter 5

Concluding Remarks

The use of computational techniques to analyze complex social and political phenomena has been both enlightening and challenging. This thesis explores the multifaceted potential of NLP in shaping our understanding of conflicts, providing insights into their precursors, the nature of political communication, and the challenges posed by misinformation. Each chapter, distinct in its focus, collectively reveals the intricate ways in which digital information and communication technologies interact with and affect political and social dynamics.

In the first chapter, I showcase a fame-semantic parser which reveals the potential of computational methods advancing the identification of cause-effect relations, in particular for the study and prediction of conflicts. This technological advancement enables policymakers and researchers to identify precursors with greater granularity and to respond with more targeted strategies. This predictive capacity is not just a standalone benefit but intertwines with the broader theme of how information processing and dissemination can impact conflict dynamics and our understanding thereof. This chapter also brings to light the ethical implications and limitations inherent in such predictive models, emphasizing the need for a responsible approach and cautious consideration of their insights, as cautioned by previous studies [162–164].

Building on this foundation, the second chapter shifts the focus to political communication, specifically examining how political leaders in various regimes utilize social media. Here, NLP's role extends to analyzing the sentiment and frequency of leaders' communications on platforms like Twitter, revealing systematic variations that align with the degree of democratization. Leaders in democratic regimes tend to engage more positively and transparently, contrasting with the more controlled and propagandistic communication in authoritarian settings. This finding dovetails with the first chapter's emphasis on information as a precursor to conflict, illustrating how the nature of political communication can either mitigate or exacerbate tensions [19]. The nuanced view of leaders' digital engagement with their

populace, particularly in crisis situations, adds a layer of complexity to our understanding of conflicts and political stability.

In the third chapter, I cement the potential of NLP in this line of research by presenting how it can advance the identification and monitoring of misinformation. Due to the inherent role of information veracity in steering or calming the overarching political climate [15, 16], this chapter complements the earlier discussions on predictive models and political communication. The ability to identify and analyze misinformation trends, especially on social media platforms, is crucial in the context of conflicts, where false information can exacerbate tensions or even trigger violent incidents. The interconnectivity of these forces underscores the critical role of digital information — its dissemination, reception, and manipulation -– in the landscape of nowadays crises and conflicts.

From identifying early signs of unrest, analyzing the rhetoric of political leaders, to combating the spread of misinformation, NLP serves as a pivotal tool in understanding and managing modern conflicts. Each chapter, while distinct in its focus, contributes to a collective narrative that underscores the transformative potential of computational techniques in political science. The insights gained from this thesis not only advance the methodological landscape but also uncover the complexities inherent in interpreting vast amounts of data and the ethical considerations that accompany language modeling. Looking forward, the path is ripe with opportunities for further research. Building on this work, future studies could explore additional geopolitical contexts, refine NLP methods to better capture the subtleties I study, or address the limitations encountered in this research, particularly in terms of data diversity. I acknowledge that while social media has become increasingly relevant, they are yet to reach representativeness regarding the types of users – with a bias towards urban, wealthy individuals. The reproducibility of my methods allows for future expansion to media sources that capture a more extensive share of the population, including but not limited to television and radio broadcasting.

As the world ventures further into an era where digital information increasingly shapes political and social landscapes, the findings of this thesis stress the importance of integrating computational methods with political theory to address the complexities of contemporary conflicts more effectively. Thus, the computational tools I present are characterized by their efficiency and explainability – particularly suiting the needs of the nexus of development, aid, and peace. My findings suggest that interdisciplinary research is a powerful approach, one that yields more nuanced insights and effective solutions. It is my hope that this research not only contributes to the academic discourse but also inspires continued exploration at the intersection of technology and social sciences. By breaking down traditional academic silos, one can develop more holistic and effective strategies to understand and address the mul-

tifaceted challenges of our time. Through careful design, researchers and policymakers can avert risks such as cognitive atrophy [165] and appropriate the insights derived from computational approaches. Closing on a positive note, while every day appears to be doomsday in light of recent global events, the findings suggest that there is hope at the end of the tunnel.

Appendix A

Process of Creating Bibliometric Maps

Overview: This appendix outlines the process used for constructing bibliometric networks in VOSviewer, as referenced in Chapters 2, 3, and 4. These networks were derived from data sourced from the Web of Science database, containing approximately 100 million items.

General Process

- Data Collection and Analysis:
  - Keywords and Time Frame: Focused on research papers published from January 1, 2000, to December 31, 2023.
  - Document Type: Analysis was limited to review articles.
  - Counting Method: Binary counting in titles and abstracts, excluding copyright statements.
  - Term Occurrence and Relevance: Minimum occurrence of 10 for terms, with the top 60% most relevant terms selected.
- Manual Filtering: Non-relevant terms, such as database names or common academic phrases, were manually removed.

Specific Parameters per Chapter

- Chapter 2 - Political Conflict Precursors:
  - Keyword: "political conflict" in abstracts.
  - Results: 818 papers (283 items with 21,540 links after filtering).
- Chapter 3 - Leaders and Twitter:
  - Keyword: "political and communication" in abstracts.
  - Results: 460 papers (169 items with 9,386 links after filtering).
- Chapter 4 - Misinformation:
  - Keyword: "misinformation" in abstracts.
  - Results: 607 papers (222 items with 14,574 links after filtering).

Appendix B

Modeling Choice for Tweet Sentiments

In my quest to accurately measure the sentiment of tweets, I experimented with various sentiment models, eventually finding that the XLM-roBERTa-base model was superior to the alternatives. Among the models I tested were SentiWordNetEmolex, WordNet-Affect, and SentiStrength. However, these models failed to yield insightful results, likely due to their inability to adeptly handle the unstructured nature of social media data. The most promising among the alternatives was the VADER (Valence Aware Dictionary and sEntiment Reasoner) algorithm, known for its efficacy in web-based media contexts. Despite its strengths, I noticed certain limitations in its lexical approach, as elaborated in the main text. VADER, like other lexical models, relies on a predefined dictionary of words, each tagged with sentiment scores, polarity, and subjectivity. In analyzing a tweet's sentiment, the algorithm first tokenizes the text and then matches each token to the dictionary words to compute the tweet's overall sentiment. This approach overlooks the sentiment of words not included in the dictionary and fails to consider the context that can significantly alter a word's sentiment. In contrast, the machine learning approach of the XLM-roBERTa-base model I used has the capability to integrate contextual information in its assessments. For example, it accurately identified certain tweets as positive, which VADER had misclassified as neutral, likely due to the limitations I previously mentioned.

Appendix C

Lists Relevant to Modeling

List of all Cause-Effect Frames Utilized with the Frame-Semantic Parser

causation, cause bodily experience, cause change, cause change of consistency, cause change of phase, cause change of strength, cause emotion, cause expansion, cause harm, cause impact, cause motion, cause proliferation in number, cause to amalgamate, cause to be included, cause to burn, cause to continue, cause to end, cause to experience, cause to fragment, cause to make progress, cause to move in place, cause to perceive, cause to resume, cause to start, cause to wake, contingency, evidence, explaining the facts, intentionally affect, killing, launch process, objective influence, purpose, reason, response, sign, transitive action

List of the Top 50 Most Semantically Similar Candidate Seeds

war, conflict, battle, fighting, battles, struggle, wars, fought, fight, forces, civil, conflicts, confrontation, military, end, troops, combat, army, invasion, occupation, decades, during, warfare, part, violence, between, brought, peace, against, crisis, bloody, force, since, action, political, over, the, attack, hostilities, soldiers, issue, continued, afghanistan, following, despite, armed, bloodshed, years, attacks, long

List of all Potential Conflict Precursors Identified by Frame-Semantic Parser

active hostilities, agricultural crisis, aid, air force, airstrike, ambassador, annexation, anti-government protests, armed forces, arms sales, arms shipments, army, attacks on livestock, bloodshed, boat attack, Boko Haram, bombing, border confrontations, British troops, build-up of military power, calls for additional troops, car bomb, cinema violence, civilian dispute, civilians killed, clash between religious leaders, climate crisis, collapse of peace talks, crisis in the neighboring country, demonstrations, demonstrators, deploy anti-shipping missiles, destabilization, dictatorship, discontent, displaced, displacement, dronelanding, drones, drug cartel, drugs, economic challenges, economic coercion, economic collapse, economic war, election hatred, elections, excessive force in policing, false information, famine, famine drought, federal aid, fight over succession, fire arms, foreign aid plummeted, freedom fighters, gang violence, gangs, guerrilla, gun, hateful preaching, hostage, hunger, hyperinflation, ideological outreach, ideology, ignored in the public debate, illegal drug trade, immigrant-hating, independence referendum, insurgent, interest in the US electorate, invasion, Iranian engagement,

Islamic extremism, Islamic State, Islamic State group, Islamophobic, jihadists, leader dies, limited access, low scrap prices, mass demonstrations, massacre, militants, militarization, military, military building programme, military drills, military equipment, military reinforcement, military solution, militias, misinformation, missiles, motive, narcotics, NATO, nuclear, opioids, patriotic to fight, police brutality, political chaos, political crisis, power vacuum, presence of foreign troops, presidential transitions, propaganda, propaganda violent speech, protesters, protests, public health crisis, racism hatred, raids, raised security concerns, rallies, rebel, rebels, repeated lies, revenge, revolutionaries, rhetoric, riot, riots, risks of nuclear war, rocket, sectarian lines, semi-automatic weapon, shooting, social media attacks, social stress, soldiers, soldiers killed, speech that inflicted injury, stabbing, staged incidents, straw villains, strengthen the military, submarines, suicide attack, suicide bombing, suicide bombing attack, supply of weaponry, tank, tanks, terror attacks, terrorism, terrorist attack, terrorist forces, tough rhetoric, trade conflicts, trade war, trained more recruits, troops, troops drills, US support, US troops, US-led coalition, violence being livestreamed on Facebook, violent media, virus, war on drugs, war on terrorism, water damage, water shortages

List of Claim Pairs Used for Fine-tuning the Semantic Similarity Model

1. The newly constructed roundabout is not in Aba.
   The newly constructed roundabout is not in Lagos.

2. Delta State University has made Pidgin language a compulsory course for its students.
   Delta State University has made English language a compulsory course for its students.

3. Due to abuse, the Guinness World Records (GWR) has banned Nigerians from attempting to break world records.
   Due to abuse, the Guinness World Records (GWR) has banned Germans from attempting to break world records.

4. Nigeria's GDP grew by 5% in 2020.
   Nigeria's GDP grew by almost 8% in 2020.

5. Governor Obaseki relocates the office of his deputy.
   President Buhari relocates the office of his deputy.

6. A Facebook user said Korede Bello quit his music career to join the police force.
   Someone in social media said that Korede Bello is vegetarian.

7. A Facebook page posits that tomatoes and garlic can be used to clean the prostate.
   Someone online stated that tomatoes and garlic are healthy.

8. A Facebook user claimed Mike Tyson is dead.
   A Facebook user claimed that Mike Tyson is unwell.

9. Deodorants contain aluminium, which is harmful to women.
   Some deodorants contain aluminium, which is harmful to women.

10. Marc-Vivien Foé, while representing Cameroon in the semi-final of the Confederations Cup against Colombia, collapsed at the centre of the pitch.
    Foé collapsed at the centre of the pitch.

11. Super Falcons veteran Onome Ebi is the first African to appear in six World Cup tournaments at 40.
    Against all the odds, no other African player has ever participated in six World Cup touraments before, except for Ebi now.

12. Burna Boy says that Wizkid is the (g)od of Afrobeats, the Jesus of Afrobeats, and Davido is the Joseph of Afrobeats.
    Burna Boy says that Davido is the (g)od of Afrobeats, the Jesus of Afrobeats, and Wizkid is the Joseph of Afrobeats.

13. A Facebook page post that there is a place in Jamaica called Abeokuta.
    The internet is a funny place; while you would not expect it, there exists a Abeokuta in Jamaica.

14. Kenya's national electricity grid mix is 92% green.
    Kenya's national electricity grid is very green, in fact it is almost 100% green, now at about 92% green.

15. A Twitter user claimed that avoiding shisha could help prevent breast cancer.
    It is indeed not a lie that avoiding shisha can help prevent cancer.

16. Using phones in a fuel station or near fuel at home can cause an explosion.
    While many people think it is a hoax, it is indeed not wrong that using phones in a fuel station can cause an explosion.

17. FIFA bans Lauren James for the rest of the tournament after the red card.
    FIFA bans player James for the next two games of the tournament after the red card.

18. The rise in inflation rate is a global challenge due to COVID-19 and is not peculiar to Nigeria.
    The rise in inflation rate is a challenge due to COVID-19 and peculiar to Nigeria.

19. Hushpuppi has been released from prison.
    Hushpuppi has not been released from prison.

20. Wizkid provided a free bus to gather "random people on the streets of London" for his concert at Tottenham Hotspur Stadium because he could not sell out his tickets.
    Wizkid did not provide a free bus to gather "random people on the streets of London" for his concert at Tottenham Hotspur Stadium.

21. Lagos State has implemented strict anti-corruption measures.
    Abuja has implemented strict anti-corruption measures.

22. Kano's football team emerged victorious in the West African Cup.
    Kano's football team did not perform well in the West African Cup.

23. Fela Kuti, the renowned musician, was known for his political activism in Port Harcourt.
    Fela Kuti, a celebrated musician, was estimated for his political activism.

24. The recent gubernatorial election in Ogun State was not marred by allegations of voter fraud.
    Despite controversy on social media, evidence suggests that the recent gubernatorial election in Osun State was free of voter fraud.

25. Nollywood, the Nigerian film industry, is gaining international recognition.
    Nollywood, the Nigerian film industry, is struggling to gain international recognition.

26. Kaduna's economy heavily relies on agricultural exports.
    Kaduna's economy is diversifying away from agricultural exports, although it is still very dependent on it.

27. The Rivers State government has imposed a curfew due to security concerns in Port Harcourt.
    The Rivers State government has not imposed a curfew despite security concerns in Port Harcourt.

28. Nigerian athletes have excelled in track and field events at the Olympics.
    Despite high hopes, Nigerian athletes have not performed well in track and field events at the Olympics; though they did well in swimming.

29. Only 10% of students passed the rigorous entrance exam for the University of Ibadan.
    Approximately 20% of students passed the entrance exam for the University of Ibadan, indicating a significant improvement.

References

[1] G. Urrea, S. Villa, and P. Gonçalves, "Exploratory analyses of relief and development operations using social networks," *Socio-Economic Planning Sciences*, vol. 56, pp. 27–39, 2016. DOI: 10.1016/j.seps.2016.05.001.

[2] H. Mueller and C. Rauh, "Reading between the lines: Prediction of political violence using newspaper text," *American Political Science Review*, vol. 112, no. 2, pp. 358–375, 2018. DOI: 10.1017/S0003055417000570.

[3] M. Solaimani, S. Salam, A. M. Mustafa, L. Khan, P. T. Brandt, and B. Thuraisingham, "Near real-time atrocity event coding," in *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*, IEEE, 2016, pp. 139–144. DOI: 10.1109/ISI.2016.7745457.

[4] H. Hegre, N. W. Metternich, H. M. Nygård, and J. Wucherpfennig, "Introduction: Forecasting in peace research," *Journal of Peace Research*, vol. 54, no. 2, pp. 113–124, 2017. DOI: 10.1177/0022343317691330.

[5] L.-E. Cederman and N. B. Weidmann, "Predicting armed conflict: Time to adjust our expectations?" *Science*, vol. 355, no. 6324, pp. 474–476, 2017. DOI: 10.1126/science.aal4483.

[6] A. R. KhudaBukhsh, R. Sarkar, M. S. Kamlet, and T. Mitchell, "We don't speak the same language: Interpreting polarization through machine translation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, AAAI, 2021, pp. 14 893–14 901. DOI: 10.1609/aaai.v35i17.17748.

[7] Z. Ahmed, B. Vidgen, and S. A. Hale, "Tackling racial bias in automated online hate detection: Towards fair and accurate detection of hateful users with geometric deep learning," *EPJ Data Science*, vol. 11, no. 1, p. 8, 2022. DOI: 10.1140/epjds/s13688-022-00319-9.

[8] N. Fairclough, *Language and power*. Harlow, UK: Routledge, 2001.

[9] C. Raleigh, A. Linke, H. Hegre, and J. Karlsen, "Introducing ACLED: An armed conflict location and event dataset," *Journal of Peace Research*, vol. 47, no. 5, pp. 651–660, 2010. DOI: 10.1177/0022343310378914.

[10] N. Beck, G. King, and L. Zeng, "Improving quantitative studies of international conflict: A conjecture," *American Political Science Review*, vol. 94, no. 1, pp. 21–35, 2000. DOI: 10.2307/2586378.

[11] M. D. Ward, B. D. Greenhill, and K. M. Bakke, "The perils of policy by p-value: Predicting civil conflicts," *Journal of Peace Research*, vol. 47, no. 4, pp. 363–375, 2010. DOI: 10.1177/0022343309356491.

[12] G. Schneider, N. P. Gleditsch, and S. C. Carey, "Exploring the past, anticipating the future: A symposium," *International Studies Review*, vol. 12, no. 1, pp. 1–7, 2010. DOI: 10.1111/j.1468-2486.2009.00909.x.

[13] C. Perry, "Machine learning and conflict prediction: A use case," *Stability: International Journal of Security and Development*, vol. 2, no. 3, p. 56, 2013. DOI: 10.5334/sta.cr.

[14] S. J. Cook and N. B. Weidmann, "Race to the bottom: Spatial aggregation and event data," *International Interactions*, vol. 48, no. 3, pp. 471–491, 2022. DOI: 10.1080/03050629.2022.2025365.

[15] J. A. Tucker, A. Guess, P. Barbera, C. Vaccari, A. Siegel, S. Sanovich, D. Stukal, and B. Nyhan, "Social media, political polarization, and political disinformation: A review of the scientific literature," *SSRN Electronic Journal*, pp. 1–95, 2018. DOI: 10.2139/ssrn.3144139.

[16] J. J. Van Bavel, E. A. Harris, P. Pärnamets, S. Rathje, K. C. Doell, and J. A. Tucker, "Political psychology in the digital (mis)information age: A model of news belief and sharing," *Social Issues and Policy Review*, vol. 15, no. 1, pp. 84–113, 2021. DOI: 10.1111/sipr.12077.

[17] Ş. Ş. Erçetin, A. Tekin, and Ş. N. Açıkalın, "Organized and disorganized chaos a new dynamics in peace intelligence," in *Chaos Theory in Politics*, ser. Understanding Complex Systems, Dordrecht, NL: Springer Netherlands, 2014, pp. 3–16. DOI: 10.1007/978-94-017-8691-1_1.

[18] M. Hermann, "Explaining foreign policy behavior using the personal characteristics of political leaders," *International Studies Quarterly*, vol. 24, no. 1, pp. 7–46, 1980. DOI: 10.2307/2600126.

[19] L. Y. Hunter and G. Biglaiser, "The effects of social media, elites, and political polarization on civil conflict," *Studies in Conflict & Terrorism*, pp. 1–28, 2022. DOI: 10.1080/1057610X.2022.2163461.

[20] K. Skrede Gleditsch and A. Ruggeri, "Political opportunity structures, democracy, and civil war," *Journal of Peace Research*, vol. 47, no. 3, pp. 299–310, 2010. DOI: 10.1177/0022343310362293.

[21] P. Collier, *The bottom billion: why the poorest countries are falling behind and what can be done about it*. New York, USA: Oxford University Press, 2007.

[22] T. Chadefaux, "Early warning signals for war in the news," *Journal of Peace Research*, vol. 51, no. 1, pp. 5–18, 2014. DOI: 10.1177/0022343313507302.

[23] C. Grady, R. Wolfe, D. Dawop, and L. Inks, "How contact can promote societal change amid conflict: An intergroup contact field experiment in nigeria," *Proceedings of the National Academy of Sciences*, vol. 120, no. 43, e2304882120, 2023. DOI: 10.1073/pnas.2304882120.

[24] D. Krcmaric, S. Nelson, and A. Roberts, "Studying leaders and elites: The personal biography approach," *Annual Review of Political Science*, vol. 23, no. 1, pp. 133–151, 2020. DOI: 10.1146/annurev-polisci-050718-032801.

[25] S. Guriev and D. Treisman, "Informational autocrats," *Journal of Economic Perspectives*, vol. 33, no. 4, pp. 100–127, 2019. DOI: 10.1257/jep.33.4.100.

[26] E. Newman and J. v. Selm, Eds., *Refugees and forced displacement: international security, human vulnerability, and the state*. New York, USA: United Nations University Press, 2003.

[27] D. Backer and T. Billing, "Validating famine early warning systems network projections of food security in africa, 2009–2020," *Global Food Security*, vol. 29, p. 100 510, 2021. DOI: 10.1016/j.gfs.2021.100510.

[28] A. Little and A. Meng, *Measuring democratic backsliding*, SSRN Scholarly Paper, 2023. DOI: 10.2139/ssrn.4327307.

[29] H. Buhaug and S. Gates, "The geography of civil war," *Journal of Peace Research*, vol. 39, no. 4, pp. 417–433, 2002. DOI: 10.1177/0022343302039004003.

[30] H. Buhaug and P. Lujala, "Accounting for scale: Measuring geography in quantitative studies of civil war," *Political Geography*, vol. 24, no. 4, pp. 399–418, 2005. DOI: 10.1016/j.polgeo.2005.01.006.

[31] J. D. Fearon and D. D. Laitin, "Ethnicity, insurgency, and civil war," *American Political Science Review*, vol. 97, no. 1, pp. 75–90, 2003. DOI: 10.1017/S0003055403000534.

[32] E. Gartzke, "War is in the error term," *International Organization*, vol. 53, no. 3, pp. 567–587, 1999. DOI: 10.1162/002081899550995.

[33] J. B. Londregan and K. T. Poole, "Poverty, the coup trap, and the seizure of executive power," *World Politics*, vol. 42, no. 2, pp. 151–183, 1990. DOI: 10.2307/2010462.

[34] N. B. Weidmann, "On the accuracy of media-based conflict event data," *Journal of Conflict Resolution*, vol. 59, no. 6, pp. 1129–1149, 2015. DOI: 10.1177/0022002714530431.

[35] I. Salehyan, "Best practices in the collection of conflict data," *Journal of Peace Research*, vol. 52, no. 1, pp. 105–109, 2015. DOI: 10.1177/0022343314551563.

[36] Y. M. Zhukov, C. Davenport, and N. Kostyuk, "Introducing xsub: A new portal for cross-national data on subnational violence," *Journal of Peace Research*, vol. 56, no. 4, pp. 604–614, 2019. DOI: 10.1177/0022343319836697.

[37] N. B. Weidmann and M. D. Ward, "Predicting conflict in space and time," *Journal of Conflict Resolution*, vol. 54, no. 6, pp. 883–901, 2010. DOI: 10.1177/0022002710371669.

[38] J. P. P. Font, "Chaos and political science: How floods and butterflies have proved to be relevant to move tables closer," in *Chaos Theory in Politics*, ser. Understanding Complex Systems, Dordrecht, NL: Springer Netherlands, 2014, pp. 121–141. DOI: 10.1007/978-94-017-8691-1_8.

[39] J. D. Singer, "The peace researcher and foreign policy prediction," *Peace Science Society (International)*, vol. 21, pp. 1–13, 1973. DOI: 10.4324/9780203128398-18.

[40] B. Harff, "Assessing risks of genocide and politicide," *Peace and Conflict*, pp. 57–61, 2005.

[41] T. Lynam, M. Zapata, H. Hegre, C. Bell, and C. Besaw, "Early warning and predictive analytic systems in conflict contexts: Insights from the field," *Civil Wars*, pp. 1–29, 2023. DOI: 10.1080/13698249.2023.2185377.

[42] A. Balashankar, L. Subramanian, and S. P. Fraiberger, "Predicting food crises using news streams," *Science Advances*, vol. 9, no. 9, eabm3449, 2023. DOI: 10.1126/sciadv.abm3449.

[43] M. Jahre, A. Pazirandeh, and L. Van Wassenhove, "Defining logistics preparedness: A framework and research agenda," *Journal of Humanitarian Logistics and Supply Chain Management*, vol. 6, no. 3, pp. 372–398, 2016. DOI: 10.1108/JHLSCM-04-2016-0012.

[44] G. Kovács and K. M. Spens, "Humanitarian logistics in disaster relief operations," *International Journal of Physical Distribution & Logistics Management*, vol. 37, no. 2, pp. 99–114, 2007. DOI: 10.1108/09600030710734820.

[45] UNICEF, *Unicef/wfp return on investment for emergency preparedness study*, Jan. 2015. [Online]. Available: https://www.unicef.org/reports/unicefwfp-return-investment-emergency-preparedness-study (visited on 01/01/2023).

[46] P. T. Brandt, J. R. Freeman, and P. A. Schrodt, "Real time, time series forecasting of inter-and intra-state political conflict," *Conflict Management and Peace Science*, vol. 28, no. 1, pp. 41–64, 2011.

[47] J. S. Goldstein, "A conflict-cooperation scale for weis events data," *The Journal of Conflict Resolution*, vol. 36, no. 2, pp. 369–385, 1992. DOI: 10.1177/0022002792036002007.

[48] M. D. Ward, N. W. Metternich, C. L. Dorff, M. Gallop, F. M. Hollenbach, A. Schultz, and S. Weschle, "Learning from the past and stepping into the future: Toward a new generation of conflict prediction," *International Studies Review*, vol. 15, no. 4, pp. 473–490, 2013. DOI: 10.1111/misr.12072.

[49] W. D. Nordhaus, "Geography and macroeconomics: New data and new findings," *Proceedings of the National Academy of Sciences*, vol. 103, no. 10, pp. 3510–3517, 2006. DOI: 10.1073/pnas.0509842103.

[50] K. Sims *et al.*, *Landscan global 2022*, 2023. DOI: 10.48690/1529167.

[51] S. Davies, T. Pettersson, and M. Öberg, "Organized violence 1989–2022, and the return of conflict between states," *Journal of Peace Research*, vol. 60, no. 4, pp. 691–708, 2023. DOI: 10.1177/00223433231185169.

[52] M. Reid Sarkees and P. Schafer, "The correlates of war data on war: An update to 1997," *Conflict Management and Peace Science*, vol. 18, no. 1, pp. 123–144, 2000. DOI: 10.1177/073889420001800105.

[53] L. Zeng, "Prediction and classification with neural network models," *Sociological Methods & Research*, vol. 27, no. 4, pp. 499–524, 1999. DOI: 10.1177/0049124199027004002.

[54] L. Zeng, "Neural network models for political analysis," in *Political Complexity: Nonlinear Models of Politics*, Ann Arbor, USA: University of Michigan Press, 2000, pp. 239–268.

[55] P. Vesco, H. Hegre, M. Colaresi, R. B. Jansen, A. Lo, G. Reisch, and N. B. Weidmann, "United they stand: Findings from an escalation prediction competition," *International Interactions*, vol. 48, no. 4, pp. 860–896, 2022. DOI: 10.1080/03050629.2022.2029856.

[56] R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, and Y. Choi, *Defending against neural fake news*, 2019. DOI: 10.48550/arXiv.1905.12616. arXiv: 1905.12616 [cs.CL].

[57] C. F. Baker, C. J. Fillmore, and J. B. Lowe, "The berkeley framenet project," in *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, Association for Computational Linguistics, 1998, pp. 86–90. DOI: 10.3115/980845.980860.

[58] S. Swayamdipta, S. Thomson, C. Dyer, and N. A. Smith, *Frame-semantic parsing with softmax-margin segmental rnns and a syntactic scaffold*, 2017. DOI: 10.48550/arXiv.1706.09528. arXiv: 1706.09528 [cs.CL].

[59] D. Chanin, *Open-source frame semantic parsing*, 2023. DOI: 10.48550/arXiv.2303.12788. arXiv: 2303.12788 [cs.CL].

[60] D. Gildea and D. Jurafsky, "Automatic labeling of semantic roles," *Computational Linguistics*, vol. 28, no. 3, pp. 245–288, 2002. DOI: 10.1162/089120102760275983.

[61] Y.-N. Chen, W. Y. Wang, and A. I. Rudnicky, "Unsupervised induction and filling of semantic slots for spoken dialogue systems using frame-semantic parsing," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, IEEE, 2013, pp. 120–125. DOI: 10.1109/ASRU.2013.6707716.

[62] X. Zhao, X. Walton, S. Shrestha, and A. Rios, *Bike frames: Understanding the implicit portrayal of cyclists in the news*, 2023. DOI: 10.48550/arXiv.2301.06178. arXiv: 2301.06178 [cs.CY].

[63] A. Kalyanpur, O. Biran, T. Breloff, J. Chu-Carroll, A. Diertani, O. Rambow, and M. Sammons, *Open-domain frame semantic parsing using transformers*, 2020. DOI: 10.48550/arXiv.2010.10998. arXiv: 2010.10998 [cs.CL].

[64] L. Vieu, P. Muller, M. Candito, and M. Djemaa, "A general framework for the annotation of causality based on FrameNet," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, European Language Resources Association, 2016, pp. 3807–3813.

[65] L. Vieu, "A framenet lexicon and annotated corpus as drd resource: Causality in the asfalda french framenet," in *Proceedings of the Final Action Conference TextLink 2018: Cross-Linguistic Discourse Annotation*, IRIT Computer Science Research Institute of Toulouse, 2020, pp. 172–178.

[66] J. Whitten-Woodring, M. Kleinberg, A. Thawnghmung, and T. Myat, "Poison if you don't know how to use it: Facebook, democracy, and human rights in myanmar," *The International Journal of Press/Politics*, vol. 25, no. 3, pp. 407–425, 2020. DOI: 10.1177/1940161220919666.

[67] Z. Tufekci, "Social movements and governments in the digital age: Evaluating a complex landscape," *Journal of International Affairs*, vol. 68, no. 1, pp. 1–18, 2014.

[68] P. Mozur, "A genocide incited on facebook, with posts from myanmars military," Oct. 2018. [Online]. Available: https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html (visited on 09/24/2021).

[69] J. H. Parmelee and S. L. Bichard, *Politics and the Twitter revolution: how tweets influence the relationship between political leaders and the public*. Lanham, USA: Lexington, 2011.

[70] M. Haman, "The use of twitter by state leaders and its impact on the public during the COVID-19 pandemic," *Heliyon*, vol. 6, no. 11, e05540, 2020. DOI: 10.1016/j.heliyon.2020.e05540.

[71] F. Barbieri, L. E. Anke, and J. Camacho-Collados, "XLM-T: Multilingual language models in twitter for sentiment analysis and beyond," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, European Language Resources Association, 2022, pp. 258–266.

[72] A. Bulovsky, "Authoritarian communication on social media: The relationship between democracy and leaders' digital communicative practices," *International Communication Gazette*, vol. 81, no. 1, pp. 20–45, 2019. DOI: 10.1177/1748048518767798.

[73] S. Chunly, "Social media and counterpublic spheres in an authoritarian state: Exploring online political discussions among cambodian facebook users," *Discourse, Context & Media*, vol. 34, p. 100 382, 2020. DOI: 10.1016/j.dcm.2020.100382.

[74] K. Ruijgrok, "Illusion of control: How internet use generates anti-regime sentiment in authoritarian regimes," *Contemporary Politics*, vol. 27, no. 3, pp. 247–270, 2021. DOI: 10.1080/13569775.2020.1851931.

[75] J. H. Gross and K. T. Johnson, "Twitter taunts and tirades: Negative campaigning in the age of trump," *Political Science*, vol. 49, no. 4, pp. 748–754, 2016. DOI: 10.1017/S1049096516001700.

[76] T. Graham, M. Broersma, K. Hazelhoff, and G. van 't Haar, "Between broadcasting political messages and interacting with voters," *Information, Communication & Society*, vol. 16, no. 5, pp. 692–719, 2013. DOI: 10.1080/1369118X.2013.785581.

[77] P. Norris, "Comparative political communications: Common frameworks or babelian confusion?" *Government and Opposition*, vol. 44, no. 3, pp. 321–340, 2009. DOI: 10.1111/j.1477-7053.2009.01290.x.

[78] R. Rivas-de-Roca and C. Pérez-Curiel, "Global political leaders during the covid-19 vaccination: Between propaganda and fact-checking," *Politics and the Life Sciences*, vol. 42, no. 1, pp. 104–119, 2023. DOI: 10.1017/pls.2023.4.

[79] N. B. Weidmann and E. Rød, *The internet and political protest in autocracies*. New York, USA: Oxford University Press, 2019.

[80] J. Morales, "Perceived popularity and online political dissent: Evidence from twitter in venezuela," *The International Journal of Press/Politics*, vol. 25, no. 1, pp. 5–27, 2020. DOI: 10.1177/1940161219872942.

[81] A. Chadwick, *The hybrid media system: politics and power*. New York, USA: Oxford University Press, 2013.

[82] B. Bueno de Mesquita, J. Morrow, R. Siverson, and A. Smith, "Political competition and economic growth," *Journal of Democracy*, vol. 12, no. 1, pp. 58–72, 2001. DOI: 10.1353/jod.2001.0011.

[83] R. Enikolopov, M. Petrova, and E. Zhuravskaya, "Media and political persuasion: Evidence from Russia," *The American Economic Review*, vol. 101, no. 7, pp. 3253–3285, 2011. DOI: 10.1257/aer.101.7.3253.

[84] N. Aharony, "Twitter use by three political leaders: An exploratory analysis," *Online Information Review*, vol. 36, no. 4, pp. 587–603, 2012. DOI: 10.1108/14684521211254086.

[85] M. J. Park, D. Kang, J. J. Rho, and D. H. Lee, "Policy role of social media in developing public trust: Twitter communication with government leaders," *Public Management Review*, vol. 18, no. 9, pp. 1265–1288, 2016. DOI: 10.1080/14719037.2015.1066418.

[86] S.-J. Eom, H. Hwang, and J. H. Kim, "Can social media increase government responsivenes? a case study of seoul, korea," *Government Information Quarterly*, vol. 35, pp. 109–122, 2018. DOI: 10.1016/j.giq.2017.10.002.

[87] S. Guriev and D. Treisman, *Spin dictators: the changing face of tyranny in the 21st century*. Princeton, USA: Princeton University Press, 2022.

[88] C. von Soest and J. Grauvogel, "Identity, procedures and performance: How authoritarian regimes legitimize their rule," *Contemporary Politics*, vol. 23, no. 3, pp. 287–305, 2017. DOI: 10.1080/13569775.2017.1304319.

[89] A. Nai and J. Maier, "Dark necessities? candidates' aversive personality traits and negative campaigning in the 2018 american midterms," *Electoral Studies*, vol. 68, p. 102 233, 2020. DOI: 10.1016/j.electstud.2020.102233.

[90] D. Ohr and H. Oscarsson, "Leader traits, leader image, and vote choice," in *Political Leaders and Democratic Elections*, Oxford, UK: Oxford University Press, 2011, pp. 187–219.

[91] JustAnotherArchivist, *Snscrape*, GitHub, Jan. 2024. [Online]. Available: https://github.com/JustAnotherArchivist/snscrape#snscrape (visited on 01/01/2024).

[92] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, *RoBERTa: A robustly optimized bert pretraining approach*, 2019. DOI: 10.48550/arXiv.1907.11692. arXiv: 1907.11692 [cs.CL].

[93] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *BERT: Pre-training of deep bidirectional transformers for language understanding*, 2019. DOI: 10.48550/arXiv.1810.04805. arXiv: 1810.04805 [cs.CL].

[94] C. C. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8, no. 1, pp. 216–225, 2014. DOI: 10.1609/icwsm.v8i1.14550.

[95] L. Barbosa and J. Feng, "Robust sentiment detection on twitter from biased and noisy data," in *Proceedings of the 23rd International Conference on Computational Linguistics: Posters (COLING'10)*, Association for Computational Linguistics, 2010, pp. 36–44. DOI: 10.5555/1944566.1944571.

[96] L. Derczynski, A. Ritter, S. Clark, and K. Bontcheva, "Twitter part-of-speech tagging for all: Overcoming sparse and noisy data," pp. 198–206, 2013.

[97] N. Chambers, V. Bowen, E. Genco, X. Tian, E. Young, G. Harihara, and E. Yang, "Identifying political sentiment between nation states with social media," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2015, pp. 65–75. DOI: 10.18653/v1/D15-1007.

[98] A. Ceron, L. Curini, and S. M. Iacus, "Using sentiment analysis to monitor electoral campaigns: Method matters — evidence from the united states and italy," *Social Science Computer Review*, vol. 33, no. 1, pp. 3–20, 2015. DOI: 10.1177/0894439314521983.

[99] M. Haman and M. Školník, "Politicians on social media. the online database of members of national parliaments on twitter.," *Profesional De La información*, vol. 30, no. 2, 2021. DOI: 10.3145/epi.2021.mar.17.

[100] G. T. Feyissa, L. B. Tolu, and A. Ezeh, "Covid-19 death reporting inconsistencies and working lessons for low- and middle-income countries: Opinion," *Frontiers in Medicine*, vol. 8, 2021. DOI: 10.3389/fmed.2021.595787.

[101] A. Baturo, "Cursus honorum: Personal background, careers and experience of political leaders in democracy and dictatorship - new data and analyses," *Politics and Governance*, vol. 4, no. 2, pp. 138–157, 2016. DOI: 10.17645/pag.v4i2.602.

[102] C. M. Ellis, M. C. Horowitz, and A. C. Stam, "Introducing the LEAD data set," *International Interactions*, vol. 41, no. 4, pp. 718–741, 2015. DOI: 10.1080/03050629.2015.1016157.

[103] W. N. Venables and B. D. Ripley, *Modern applied statistics with S*. New York, USA: Springer, 2002.

[104] S. Yu and R. Jong-A-Pin, "Rich or alive? political (in)stability, political leader selection and economic growth," *Journal of Comparative Economics*, vol. 48, no. 3, pp. 561–577, 2020. DOI: 10.1016/j.jce.2019.11.004.

[105] L. Helms, "Leadership succession in politics: The democracy/autocracy divide revisited," *The British Journal of Politics and International Relations*, vol. 22, no. 2, pp. 328–346, 2020. DOI: 10.1177/1369148120908528.

[106] H. Lin, J. Lasser, S. Lewandowsky, R. Cole, A. Gully, D. G. Rand, and G. Pennycook, "High level of correspondence across different news domain quality rating sets," *PNAS Nexus*, vol. 2, no. 9, pgad286, 2023. DOI: 10.1093/pnasnexus/pgad286.

[107] A. M. Guess and B. A. Lyons, "Misinformation, disinformation, and online propaganda," in *Social Media and Democracy: The State of the Field, Prospects for Reform*, ser. SSRC Anxieties of Democracy, Cambridge, UK: Cambridge University Press, 2020, pp. 10–33.

[108] S. Bradshaw and P. N. Howard, "The global disinformation order: 2019 global inventory of organised social media manipulation," 2019.

[109] G. Pennycook and D. G. Rand, "The psychology of fake news," *Trends in Cognitive Sciences*, vol. 25, no. 5, pp. 388–402, 2021. DOI: 10.1016/j.tics.2021.02.007.

[110] D. M. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, *et al.*, "The science of fake news," *Science*, vol. 359, no. 6380, pp. 1094–1096, 2018. DOI: 10.1126/science.aao2998.

[111] A. A. Arechar *et al.*, "Understanding and combatting misinformation across 16 countries on six continents," *Nature Human Behaviour*, vol. 7, no. 9, pp. 1502–1513, 2023. DOI: 10.1038/s41562-023-01641-6.

[112] E. M. Lederer, *UN chief says misinformation about COVID-19 is new enemy*, AP News, Mar. 2020. [Online]. Available: https://apnews.com/general-news-e829eddc01457c700d5541f2dc2beebc (visited on 01/01/2024).

[113] J. Roozenbeek, C. R. Schneider, S. Dryhurst, J. Kerr, A. L. Freeman, G. Recchia, A. M. Van Der Bles, and S. Van Der Linden, "Susceptibility to misinformation about covid-19 around the world," *Royal Society Open Science*, vol. 7, no. 10, p. 201 199, 2020. DOI: 10.1098/rsos.201199.

[114] C. H. Basch, Z. Meleo-Erwin, J. Fera, C. Jaime, and C. E. Basch, "A global pandemic in the time of viral memes: Covid-19 vaccine misinformation and disinformation on tiktok," *Human Vaccines & Immunotherapeutics*, vol. 17, no. 8, pp. 2373–2377, 2021. DOI: 10.1080/21645515.2021.1894896.

[115] S. Loomba, A. de Figueiredo, S. J. Piatek, K. de Graaf, and H. J. Larson, "Measuring the impact of covid-19 vaccine misinformation on vaccination intent in the uk and usa," *Nature Human Behaviour*, vol. 5, no. 3, pp. 337–348, 2021. DOI: 10.1038/s41562-021-01056-1.

[116] G. Pennycook, J. McPhetres, B. Bago, and D. G. Rand, "Beliefs about covid-19 in canada, the united kingdom, and the united states: A novel test of political polarization and motivated reasoning," *Personality and Social Psychology Bulletin*, vol. 48, no. 5, pp. 750–765, 2022. DOI: 10.1177/01461672211023652.

[117] S. van der Linden, A. Leiserowitz, S. Rosenthal, and E. Maibach, "Inoculating the public against misinformation about climate change," *Global Challenges*, vol. 1, no. 2, p. 1 600 008, 2017. DOI: 10.1002/gch2.201600008.

[118] K. Aslett, Z. Sanderson, W. Godel, N. Persily, J. Nagler, and J. A. Tucker, "Online searches to evaluate misinformation can increase its perceived veracity," *Nature*, pp. 1–9, 2023. DOI: 10.1038/s41586-023-06883-y. [Online]. Available: https://www.nature.com/articles/s41586-023-06883-y.

[119] B. Guay, A. J. Berinsky, G. Pennycook, and D. Rand, "How to think about whether misinformation interventions work," *Nature Human Behaviour*, vol. 7, no. 8, pp. 1231–1233, 2023. DOI: 10.1038/s41562-023-01667-w.

[120] M. A. Lawson and H. Kakkar, "Of pandemics, politics, and personality: The role of conscientiousness and political ideology in the sharing of fake news," *Journal of Experimental Psychology: General*, vol. 151, no. 5, pp. 1154–1177, 2022. DOI: 10.1037/xge0001120.

[121] G. Pennycook and D. G. Rand, *Nudging social media sharing towards accuracy*, 2021. DOI: 10.31234/osf.io/tp6vy. PsyArXiv: tp6vy.

[122] W. J. Brady, M. J. Crockett, and J. J. Van Bavel, "The mad model of moral contagion: The role of motivation, attention, and design in the spread of moralized content online," *Perspectives on Psychological Science*, vol. 15, no. 4, pp. 978–1010, 2020. DOI: 10.1177/1745691620917336.

[123] X. Chen, G. Pennycook, and D. Rand, "What makes news sharable on social media?" *Journal of Quantitative Description: Digital Media*, vol. 3, pp. 1–27, 2023. DOI: 10.51685/jqd.2023.007.

[124] J. Constine, *Facebook puts link to 10 tips for spotting 'false news' atop feed*, TechCrunch, Apr. 2017. [Online]. Available: https://techcrunch.com/2017/04/06/facebook-puts-link-to-10-tips-for-spotting-false-news-atop-feed/ (visited on 01/01/2024).

[125] A. M. Guess, M. Lerner, B. Lyons, J. M. Montgomery, B. Nyhan, J. Reifler, and N. Sircar, "A digital media literacy intervention increases discernment between mainstream and false news in the united states and india," *Proceedings of the National Academy of Sciences*, vol. 117, no. 27, pp. 15 536–15 545, 2020. DOI: 10.1073/pnas.1920498117.

[126] M. Osmundsen, A. Bor, P. B. Vahlstrup, A. Bechmann, and M. B. Petersen, "Partisan polarization is the primary psychological motivation behind political fake news sharing on twitter," *American Political Science Review*, vol. 115, no. 3, pp. 999–1015, 2021. DOI: 10.1017/S0003055421000290.

[127] F. Zimmermann and M. Kohring, "Mistrust, disinforming news, and vote choice: A panel survey on the origins and consequences of believing disinformation in the 2017 german parliamentary election," *Political Communication*, vol. 37, no. 2, pp. 215–237, 2020. DOI: 10.1080/10584609.2019.1686095.

[128] A. Pereira, E. Harris, and J. J. Van Bavel, "Identity concerns drive belief: The impact of partisan identity on the belief and dissemination of true and false news," *Group Processes & Intergroup Relations*, vol. 26, no. 1, pp. 24–47, 2023. DOI: 10.1177/13684302211030004.

[129] M.-p. S. Chan, C. R. Jones, K. Hall Jamieson, and D. Albarracín, "Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation," *Psychological Science*, vol. 28, no. 11, pp. 1531–1546, 2017. DOI: 10.1177/0956797617714579.

[130] C. S. Traberg, J. Roozenbeek, and S. van der Linden, "Psychological inoculation against misinformation: Current evidence and future directions," *The ANNALS of the American Academy of Political and Social Science*, vol. 700, no. 1, pp. 136–151, 2022. DOI: 10.1177/00027162221087936.

[131] Z. Guo, M. Schlichtkrull, and A. Vlachos, "A survey on automated fact-checking," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 178–206, 2022. DOI: 10.1162/tacl_a_00454.

[132] M. R. DeVerna, H. Y. Yan, K.-C. Yang, and F. Menczer, *Artificial intelligence is ineffective and potentially harmful for fact checking*, 2023. DOI: 10.48550/arXiv.2308.10800. arXiv: arXiv:2308.10800 [cs.HC].

[133] C. Martel, J. Allen, G. Pennycook, and D. G. Rand, "Crowds can effectively identify misinformation at scale," *Perspectives on Psychological Science*, 2023. DOI: 10.1177/17456916231190388.

[134] A. Kazemi, K. Garimella, D. Gaffney, and S. Hale, "Claim matching beyond english to scale global fact-checking," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, Association for Computational Linguistics, 2021, pp. 4504–4517. DOI: 10.18653/v1/2021.acl-long.347.

[135] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD Explorations Newsletter*, vol. 19, no. 1, pp. 22–36, 2017. DOI: 10.1145/3137597.3137600.

[136] N. Lee, B. Z. Li, S. Wang, W.-t. Yih, H. Ma, and M. Khabsa, "Language models as fact checkers?" In *Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER)*, Association for Computational Linguistics, 2020, pp. 36–41. DOI: 10.18653/v1/2020.fever-1.5.

[137] X. Zhou and R. Zafarani, "A survey of fake news: Fundament theories, detection methods, and opportunities," *ACM Computing Surveys*, vol. 53, no. 5, pp. 1–40, 2020. DOI: 10.1145/3395046.

[138] A. Kazemi, Z. Li, V. Pérez-Rosas, S. A. Hale, and R. Mihalcea, *Matching tweets with applicable fact-checks across languages*, 2022. DOI: 10.48550/arXiv.2202.07094. arXiv: 2202.07094 [cs.CL].

[139] X. Zeng, A. S. Abumansour, and A. Zubiaga, "Automated fact-checking: A survey," *Language and Linguistics Compass*, vol. 15, no. 10, e12438, 2021. DOI: 10.1111/lnc3.12438.

[140] J. Ye, X. Chen, N. Xu, C. Zu, Z. Shao, S. Liu, Y. Cui, Z. Zhou, C. Gong, Y. Shen, *et al.*, *A comprehensive capability analysis of gpt-3 and gpt-3.5 series models*, 2023. DOI: 10.48550/arXiv.2303.10420. arXiv: 2303.10420 [cs.CL].

[141] C. Qin, A. Zhang, Z. Zhang, J. Chen, M. Yasunaga, and D. Yang, *Is chatgpt a general-purpose natural language processing task solver?* 2023. DOI: 10.48550/arXiv.2302.06476. arXiv: 2302.06476 [cs.CL].

[142] K.-C. Yang and F. Menczer, *Large language models can rate news outlet credibility*, 2023. DOI: 10.48550/arXiv.2304.00228. arXiv: 2304.00228 [cs.CL].

[143] E. Hoes, S. Altay, and J. Bermeo, *Leveraging chatgpt for efficient fact-checking*, 2024. DOI: 10.31234/osf.io/qnjkf. PsyArXiv: qnjkf.

[144] A. O'Neill, *Nigeria: Growth rate of the real gross domestic product (gdp) from 2018 to 2028*, Statista, Nov. 2023. [Online]. Available: https://www.statista.com/statistics/382360/gross-domestic-product-gdp-growth-rate-in-nigeria/ (visited on 01/01/2024).

[145] S. Badrinathan and S. Chauchard, "Researching and countering misinformation in the global south," *Current Opinion in Psychology*, p. 101 733, 2023. DOI: 10.1016/j.copsyc.2023.101733.

[146] D. Milmo, *Frances haugen: 'i never wanted to be a whistleblower. but lives were in danger'*, The Guardian, Oct. 2021. [Online]. Available: https://www.theguardian.com/technology/2021/oct/24/frances-haugen-i-never-wanted-to-be-a-whistleblower-but-lives-were-in-danger (visited on 12/30/2023).

[147] A. Okpi, *Forty million twitter users in nigeria? how pollster's flawed figure became fact*, Africa Check, Jun. 2021. [Online]. Available: https://africacheck.org/fact-checks/reports/forty-million-twitter-users-nigeria-how-pollsters-flawed-figure-became-fact (visited on 01/04/2024).

[148] V. Peña-Araya, M. Quezada, B. Poblete, and D. Parra, "Gaining historical and international relations insights from social media: Spatio-temporal real-world news analysis using Twitter," *EPJ Data Science*, vol. 6, no. 1, pp. 1–35, 2017. DOI: 10.1140/epjds/s13688-017-0122-8.

[149] H. Almuhimedi, S. Wilson, B. Liu, N. Sadeh, and A. Acquisti, "Tweets are forever: A large-scale quantitative analysis of deleted tweets," in *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, Association for Computing Machinery, 2013, pp. 897–908. DOI: 10.1145/2441776.2441878.

[150] S. Jhaver, C. Boylston, D. Yang, and A. Bruckman, "Evaluating the effectiveness of deplatforming as a moderation strategy on twitter," in *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, Association for Computing Machinery, 2021, pp. 1–30. DOI: 10.1145/3479525.

[151] D. Quelle, C. Cheng, A. Bovet, and S. A. Hale, *Lost in translation – multilingual misinformation and its evolution*, 2023. DOI: 10.48550/arXiv.2310.18089. arXiv: 2310.18089v1 [cs.CL].

[152] C. Tardáguila, *López obrador launches its own 'verificado' and infuriates fact-checkers in mexico*, Poynter, Jul. 2019. [Online]. Available: https://www.poynter.org/ifcn/2019/lopez-obrador-launches-its-own-verificado-and-infuriates-fact-checkers-in-mexico/ (visited on 01/04/2024).

[153] O. Ma and B. Feldman, *How google and youtube are investing in fact-checking*, Google, Nov. 2022. [Online]. Available: https://blog.google/outreach-initiatives/google-news-initiative/how-google-and-youtube-are-investing-in-fact-checking/ (visited on 01/04/2024).

[154] N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, and D. Lazer, "Fake news on twitter during the 2016 u.s. presidential election," *Science*, vol. 363, no. 6425, pp. 374–378, 2019. DOI: 10.1126/science.aau2706.

[155] N. Vo and K. Lee, "Where are the facts? searching for fact-checked information to alleviate the spread of fake news," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, 2020, pp. 7717–7731. DOI: 10.18653/v1/2020.emnlp-main.621.

[156] Z. Li, X. Zhang, Y. Zhang, D. Long, P. Xie, and M. Zhang, *Towards general text embeddings with multi-stage contrastive learning*, 2023. DOI: 10.48550/arXiv.2308.03281. arXiv: 2308.03281 [cs.CL].

[157] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, *Language-agnostic bert sentence embedding*, 2022. DOI: 10.48550/arXiv.2007.01852. arXiv: 2007.01852 [cs.CL].

[158] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM Computing Surveys*, vol. 53, no. 3, pp. 1–34, 2020. DOI: 10.1145/3386252.

[159] A. Chadwick, C. Vaccari, and J. Kaiser, "The amplification of exaggerated and false news on social media: The roles of platform use, motivations, affect, and ideology," *American Behavioral Scientist*, 2021. DOI: 10.1177/00027642221118264.

[160] K. Ethayarajh, "How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Association for Computational Linguistics, 2019, pp. 55–65. DOI: 10.18653/v1/D19-1006.

[161] A. Fuster Baggetto and V. Fresno, "Is anisotropy really the cause of bert embeddings not being semantic?" In *Findings of the Association for Computational Linguistics: EMNLP 2022*, Association for Computational Linguistics, 2022, pp. 4271–4281. DOI: 10.18653/v1/2022.findings-emnlp.314.

[162] H. A. Simon, "Prediction and prescription in systems modeling," *Operations Research*, vol. 38, no. 1, pp. 7–14, 1990. DOI: 10.1287/opre.38.1.7.

[163] D. Sarewitz and R. A. Pielke Jr., "Prediction in science and policy," *Technology in Society*, vol. 21, no. 2, pp. 121–133, 1999. DOI: 10.1016/S0160-791X(99)00002-0.

[164] R. A. Pielke Jr., "The role of models in prediction for decision," in *Models in Ecosystem Science*, Princeton, USA: Princeton University Press, 2003, pp. 111–135.

[165] H. S. Sætra, "Generative AI: Here to stay, but for good?" *Technology in Society*, vol. 75, p. 102 372, 2023. DOI: 10.1016/j.techsoc.2023.102372.