Bad Apples Just for Friends: A Large Language Model Approach in Studying Police Violence Accusation Framing at US Human Rights Reports

by

Jie Lian

(Under the Direction of Khaled Rasheed and Amanda Murdie)

Abstract

Does the US government's tarnished police violence record shadow its international human rights monitoring and reporting practices? In this project, I argue that, for the domestic struggles on the police violence issues, the US government tends to take a partial standing on reporting the related violations of other governments. Through utilizing state-of-the-art pretrained Transformer-based large language models (LLM), I propose a novel text-to-network pipeline for text analysis. The proposed method allows a human-interpretable representation of the text data while effectively involving the semantic information in the output. With the help of the new method, the results show police violence accusations in the US human rights reports framed in favor of countries closer to the US. Methodologically, the proposed method shows promising potential in text analysis tasks like topic modeling. Moreover, the robust results also suggest the ability of Transformer-based LLMs to pick up the logic among words from natural language.

Index words: Large Language Model, Transformers, Text Network Analysis, Framing Bias Detection, Human Rights Report, Police Violence

Bad Apples Just for Friends: A Large Language Model Approach in Studying Police Violence Accusation Framing at US Human Rights Reports

by

Jie Lian

B.A., People's Public Security University of China, 06/2015, Beijing China
M.A., Shanghai Institutes for International Studies, 04/2019, Shanghai China

A Thesis Submitted to the Graduate Faculty of the
University of Georgia in Partial Fulfillment of the Requirements for the
Degree.

Master of Science

Athens, Georgia

2025

Bad Apples Just for Friends: A Large Language Model Approach in Studying Police Violence Accusation Framing at US Human Rights Reports

by

Jie Lian

Major Professor:	Khaled Rasheed Amanda Murdie
Committee:	Frederick Maier Gengchen Mai

Electronic Version Approved:

Ron Walcott
Dean of the Graduate School
The University of Georgia
May 2025

Dedication

To my dad, Jibin Lian; to my mom, Jing Li; to my wife, Shanshan Lian; to my son, YoYo.

Acknowledgments

I would like to express my sincerest appreciation to my advisor, Dr. Amanda Murdie. Thank you so much for trusting and supporting me to take this unknown adventure. Entering a new area was a bold decision, and the path was not always smooth. Your unwavering faith in me meant the world and helped carry me through the most challenging days. My family and I are deeply indebted to you. I look forward to continuing our academic collaboration to push the frontiers of AI and political research.

I would like to thank my advisor, Dr. Khaled Rasheed, for his guidance and support throughout my pursuit of this degree. As a particularly curious student, I deeply appreciate your patience, kindness, and the time you took to help me grow. I would like to thank Dr. Frederick Maier for introducing me to this new field and for all the guidance, patience, and kindness you have shown me. I would not have been able to enter this exciting area, discover so many remarkable developments, or receive such valuable training without your support at the very beginning.

I would like to thank Dr. Gengchen Mai for his guidance and valuable feedback on my research and also my career. I would like to thank Dr. Suchendra Bhandarkar for his recommendation and for guiding me toward the field of computer vision. I would like to thank Dr. Jason Anastasopoulos for all his help and the great opportunities for being able to work with him.

Last but not least, I would like to thank my colleague and my wife, Dr. Shanshan Lian, for supporting my wild dream. You are my love, my rock, and my truest comrade. There are exciting paths ahead, and I cannot wait to explore them by your side. And to YoYo, my son — thank you for giving me courage and serenity, even before you arrived in this world. I can't wait to welcome you onboard.

Technologies are merely tools—neither a blessing nor a curse. In the end, it is up to us to decide whether they will help build a better world or lead to the downfall of our civilizations.

Contents

List of Figures

List of Tables

Chapter 1

Bad Apples Just for Friends: A Large Language Model Approach in Studying Police Violence Accusation Framing at US Human Rights Reports

How do one country's domestic human rights issues influence its international human rights advocacy? As a long-time global human rights leader, the US has poor police abuse records. When the tragic murder of George Floyd by a police officer in Minnesota and the rising Black Lives Matter (BLM) campaign brought the world's attention to the long-standing police brutality issues in the US (Davies & Finnegan, 2020; Nebehay, 2020), a broader discussion on the US' standing as a global advocate for human rights is raised (Lynch & Gramer, 2020; Nugent & Perrigo, 2020; Wright, 2020). "The outrage about the lack of decency and the American double standard has now gone global in everyday life," according to a report in the New Yorker magazine in 2020 (Wright, 2020). Does the US's tarnished police violence record impact its reporting of human rights violations in other countries?

The US State Department's Annual Country Reports on Human Rights Practices (USSD reports) are instrumental in transnational human rights monitoring and evaluation. Despite a rich body of literature discussing potential bias in USSD reports, rare research directly investigates the potential impact of the US domestic human rights record on its international human rights monitoring and reporting activities, particularly concerning police violence. In this article, I argue that the domestic struggles encourage the US government to take a biased standing in evaluating other countries' police violence situations for both political and strategic considerations. Moreover, the political-driven bias in police violence accusations not only comes from what information is disclosed but also from how the information is organized and presented on the text level at the reports, which is often conceptualized as information "framing" in the political communication literature (Entman, 1997).

Police play an essential role in the government's atrocities like arbitrary arresting, kidnapping, and murder. Meanwhile, organization-level factors, like law enforcement duty, mismanagement, or poor training of the police officers, are often used as covers for the repressive nature of police action and the government's role as the decision-maker in police brutality. In the US context, one typical example is to frame police brutality as the result of "bad apples" (Cunningham, 2020). Mitchell summarizes the blaming of police violence on organizational-level factors as the government's "blaming management" (N. J. Mitchell, 2021, p. 25). The entanglement of the government's decision and organizational-level issues in facilitating police brutality leaves considerable discretion for the human rights monitors in composing the accusation. Distinguishing the role of the government in police atrocities from human rights reports is pivotal for researchers and human rights activists to understand the essence of the violations. Two directions in police violence accusations framing are specified in this paper: government-level framing emphasizes the role of government decision-makers in police violence, and agent-level framing focuses on the organizational-level causes for violations by individual police agents.

The current empirical pattern in studying strategic adjustments in the USSD reports tries to capture the bias through the comparison against the NGO-launched human rights reports with a specific reliance on the human-coded integrated indicators, like the Political Terror Scale (PTS) (Haschke, 2017) and the Cingranelli and Richards Human Rights Data Project (CIRI) (Cingranelli et al., 2021). While this approach has led to numerous empirical findings, the highly integrated nature of human rights indexes prevents researchers from a text-level understanding of the documents and is not suitable for detecting inconsistent framing of police violence accusations in the USSD reports. In this project, I propose a novel network representation method of text data based on the pretrained large language model (LLM). This approach integrates the strengths of conventional computational text analysis techniques, adept at handling text data on a large scale, and the advantages of text network techniques, which provide human-interpretable representations of the text data (Bail, 2016; C. W. Roberts, 2020; Segev, 2020). Furthermore, in contrast to the 'bag of words' (BoW) assumption underlying most automatic content analysis methods used in political texts[1], the proposed method leverages the advanced capabilities of pretrained Large Language Models (LLMs) in grasping the semantics within text data. The output text network effectively incorporates the semantic information from the original text. Through applying network analysis methods on the text networks generated from the USSD reports, my approach provides an effective way in detecting the framing strategies of police violence accusations in the documents. The empirical results suggest that, in the USSD report, the US government is more likely to frame the police brutality of countries it considers as friends as agent-level abuses and the police brutality of rival countries as government-level actions.

The contribution of the article is both substantive and methodological. Substantively, the uncovered framing bias concerning the police violence accusations in the USSD reports suggests the potential transnational impact of the US domestic police violence issues. For the critical role of the US in the international human rights regime (Kent, 2001), the disclosed projection mechanism could discredits the validity of the USSD reports. The research also contributes to the ongoing conversation on the bias in the human rights monitoring reports (Arnon et al., 2023; Clark & Sikkink, 2013; Haschke & Arnon, 2020; Hill Jr et al., 2013; Nieman & Ring, 2015; Poe et al., 2001). The revealed framing bias in the USSD reports shows a much subtle way for the strategic adjustment in the human rights reports, which not only involve what information is included in the reports but how the information is organized and presented.

Methodologically, the proposed LLM-based method provides scholars with a powerful computational tool for text-level analysis of the large-size corpus. Besides the application in this article, combining different network analysis tools, the network approach in text representation can be used more broadly in other conventional text analysis tasks like topic modeling and sentiment analysis. The proposed algorithmic pipeline is very flexible and could be easily embedded with various pretrained transformer-based LLMs, like Bidirectional Encoder Representations from Transformers (BERT) and Generative Pre-trained Transformers (GPT), which allows social scientists to utilize the most advanced transformer-based LLMs in their research. For the AI scholarship, the framing bias detected by the Transformers also echos existing research on the ability of LLM to capture logical connections inside of natural language (Devlin et al., 2018; Safavi & Koutra, 2021; T.-Y. Wang, 1999). Further research and experi-

[1] Simply speaking, models following the BoW assumption only focuses on the words' frequency and discards their structure or order in text analysis. Many computational methods, like topic modeling, are built on the BoW assumptions (M. E. Roberts et al., 2014; Ying et al., 2022).

ments are needed to understand the ability of connectist AIs to incorporate logic.

This paper proceeds as follows. In the following section, after briefly reviewing the literature studying political bias in human rights reports, I argue that the absence of text-level analysis in the current empirical research pattern would lead to an incomplete understanding of the issue. Following, I specify two directions for police violence accusation framing: government-level framing and agent-level framing, and argue for the importance of information framing in the USSD reports bias research. I propose that the political bias in the USSD reports is shown in how the police violence accusations are framed according to the targetted government's relationship with the US. Next, I introduce the LLM-based network approach for text data representation and how to use the text network to measure the framing inconsistency on police violence accusations across the USSD country reports. Based on the framing bias measurement, I test the political bias argument on the USSD reports, and the empirical results disclose the significant influence of one country's relationship with the US on police violence accusations framing in the reports. In the conclusion section, I summarize this article's substantive findings and methodological innovations and discuss the limitations of this research and future research directions.

1.1 Challenges of Detecting Political Bias in Human Rights Reports

Early literature has already pointed out the politicized nature of human rights reports, not only for governmental reports but also those launched by international organizations like the UN (Donnelly, 1988; Poe & Tate, 1994; Poe et al., 2001). Later research on the USSD reports discloses more sophisticated mechanisms of the political-driven bias in the documentation. Poe et al., 2001 and Clark and Sikkink, 2013 suggest that the political bias in the USSD reports is not constant and varies over time. Despite the complicated nature of the issue, most research in detecting political-driven bias at the USSD reports follows a similar empirical pattern (Clark & Sikkink, 2013; Haschke & Arnon, 2020; Nieman & Ring, 2015; Poe et al., 2001). The researchers try to capture the bias in the USSD reports by comparing them with the human rights reports launched by non-governmental organizations, like Amnesty International or Human Rights Watch, which are assumed to be less vulnerable to political considerations. Empirically, scholars rely on human-generated indicators, mainly PTS, which codes two separate indicators of the same country's annual human rights practices from the USSD reports and the World's Human Rights (WHR) report published by Amnesty International. The differences between the two indicators are taken as the proxy of the political bias in the USSD reports.

The current empirical pattern in political bias research on human rights reports raises criticisms of two aspects. First, the cross-sectional inconsistency in the human rights reports is not only caused by political considerations but also other factors like information availability (Clark & Sikkink, 2013) and transparency (Eck & Fariss, 2018). The repressive governments often tend to "hide, downplay, or dismiss information" (Clark & Sikkink, 2013, p. 545), while some governments like Sweden tend to improve the transparency of their human rights records (Eck & Fariss, 2018; Haschke & Arnon, 2020). As Eck and Fariss, 2018 suggested, using existing human rights documents and measurement projects for cross-sectional human rights practices comparison is problematic if researchers do not address the variation of information environments across countries. Early researchers have recognized that the information variation across countries, along with other factors, might lead to inconsistent evaluation references in the human rights reports, which could bias the event-based violation measurement drawing from the documents(Brysk, 1994; Poe, 2019). As a remedy, popular measurement projects like PTS and CIRI take a standard-based coding scheme that relies on field experts' perception of one country's human rights practices from the reports to evaluate the violation (Fariss, 2014). Nevertheless, instance-level information in the human rights reports still plays pivotal roles for the coders in PTS and CIRI projects in evaluating the targeted country's human rights practices [2]. This reliance on the instance-level information from the reports will cause the standard-based human rights indexes to be still vulnerable to cross-national variation of informational environment.

The second criticism comes from the comparison of the human rights indexes generated from NGO and USSD reports in political bias detection. Essentially, generating human rights indicators from reports is a data dimension reduction process in which the coders try to project unstructured text data into a one-dimensional index to reserve as much information in the text as possible. Even without considering the possible coder bias (Arnon et al., 2023; Wood & Gibney, 2010), this information projection process could cause significant loss of valuable information (Conrad et al., 2014; Cordell et al., 2022). Consequentially, researchers could easily ignore the structural differences between the NGO and USSD reports when only concerning the human rights indexes generated from the two documentations. Park et al., 2020 shows that the USSD and Amnesty International's human rights reports might emphasize different human rights topics in the documentation.[3] Moreover, the NGO and USSD

[2] For example, when coding the severity of "extrajudicial killing" of one country in the CIRI project, the experts are asked to "code based on the numbers" and score the variable of the targeted accordingly (Cingranelli & Richards, 2010, p. 8). In PTS, "coders are asked to count an instance of a more severe type of violation (e.g., an extrajudicial killing) more heavily than a less severe one (e.g., an instance of excessive use of force, an arbitrary arrest)." (Haschke, 2017, pp. 4–5).

[3] This difference could be explained by the diverse organizational goals of the monitor agencies in composing the human rights reports (Hafner-Burton & Ron, 2013; Hill Jr et al., 2013).

reports might be biased by the same set of political factors, thereby the comparison would not be helpful in political bias detection.[4]

Despite the empirical challenges, comparing government human rights practices cross-sectionally is crucial for academic research and policy-making. Being able to scrutinize the original reports closely and identify how exactly the political bias is shown at the text level, as suggested in Eck and Fariss, 2018 article, is thus essential for scholars to understand better the strategic adjustments in the USSD reports. The large-scale human rights reports challenge researchers' ability to analyze the document manually. However, advancements in computational linguistic methods have opened up valuable opportunities for text-level analysis Cordell et al., 2022, potentially leading to novel empirical findings. For example, by applying the structural topic model to the USSD reports, Bagozzi and Berliner, 2018 find that US allies receive more attention on the physical integrity rights violations than others. Considering the higher political sensitivity of the physical integrity violations Terman and Byun, 2022, this findings implies a counterintuitive fact: the USSD reports impose more severe criticism on the governments of US allies. In this project, I take advantage of the computational methods' ability to handle large-scale text data and propose an innovative approach to help scholars better inspect human rights reports on the text level.

1.2 Framing Strategies, Political Bias & Police Violence Accusations

1.2.1 Two Framing Strategies in Police Violence Reporting

In this research, I distinguish two possible directions in police violence framing. One direction emphasizes the role of the government in police violence, while the other direction tries to encourage the audience to believe agent-level factors cause observed police violence. As an important force arm of the government, the atrocities committed by the police are often from the government's direct order or encouragement. However, agent-level issues in the police department, like poor training of the officers and mismanagement, often obscure the repressive nature of the police abuse, like torturing arrested protestors. Morrow pointed out the challenge for outsiders in distinguishing the institutional-level and the agent-level human rights violations behaviors (Morrow, 2001, 2007). Moreover, the police dual identity as both government control forces and law enforcement agencies further dims the real initiators of police brutality, which could leave room for framing the government's repressive action accomplished

[4] Ron et al., 2005 show that the Amnesty International reports might be vulnerable to the influence of political factors like the US military assistance. The media profile of one country is another significant factor proven to be influential in distorting the NGO reports (Hill Jr et al., 2013; Ron et al., 2005). Coupling with the research showing the political bias in the reports of the human rights violations on different countries in the English-language media (Hafner-Burton & Ron, 2013), political factors like one country's alignment with the US, while impacting the USSD reports, might also effectively distort the NGO-launched human rights reports in a subtle way.

through the police force as agent-level issues caused by non-political factors like social inequality, poor training of the police officers, or just "bad apples."

The different role of the government in the two framing directions at the police violence accusations could also be viewed from the perspective of principal-agent theory. N. J. Mitchell, 2021 specifies how the delegation relationship between the principal and the agency could be used for blaming management. He argues that the principal seeks to transfer the blame to the agency through the delegation process (N. Mitchell, 2021). This "delegation for blaming management" theory provides a powerful theoretical lens to understand the government's intention of hiding behind the curtain in police violence by delegating repressive actions to the police. Similarly, the violation government would prefer the accusation of police violence focusing on the role of the agent (police) to avoid direct criticism for human rights violations.

Generally, the agent-level framing of police violence downgrades the accusation against the government from two aspects. First, the agent-level framing disentangles the human rights violations facts from the government's intention, which could significantly impact the information recipients' overall evaluation of the targeted government's human rights practice. As pointed out in N. J. Mitchell, 2021, the audience might be less likely to blame the government for the police atrocity under the agent-level framing. Second, the agent-level framing of police violence could cover the tension between the government and the victims. Existing research shows that the public is more tolerant of police violence when it does not involve the government's direct repression against the challengers (Jackson et al., 2018; Moore, 2010; Rejali, 2009). Shying the role of government, an agent-level framing of police violence accusations would invoke less pushback from the public against the incumbent.

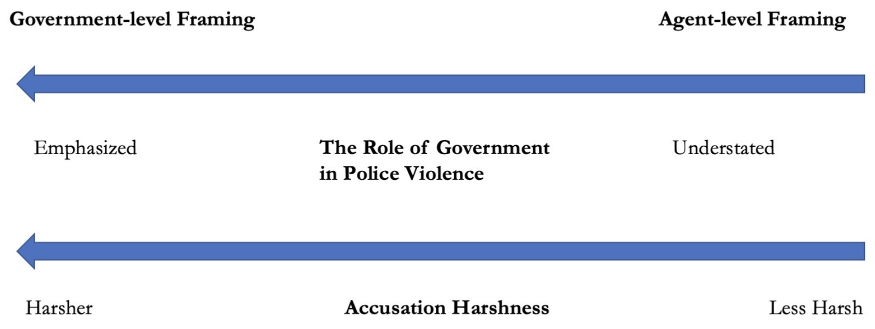


Figure 1.1: Police Violence Framing Strategies

1.2.2 Police Violence Reporting and Political Bias

Entangling with structural racism, police violence is deeply rooted in US society and has been long haunting the country (Ang, 2021; Campbell & Valera, 2020; DeVylder et al., 2020; Miller, 1998; Potter, 2013; Ritchie, 2017). I argue that the domestic struggle on the police violence issue will impact how the US government assesses other governments' similar violations in the USSD reports. Moreover, one country's political relationship with the US will play an important role in shaping the police violence accusations in the reports. Researchers already show a significant divergence in framing police brutality in the US media (Dukes & Gaither, 2017; Fridell, 2017; Porter et al., 2020). Following the political bias findings from existing literature in studying the USSD reports (Clark & Sikkink, 2013; Haschke & Arnon, 2020; Nieman & Ring, 2015; Poe et al., 2001), I expect that the police violence accusations in the USSD reports are framed in favor of those governments closer to the US. More specifically, for countries closer to the US, the US government tends to compose police violence accusations in the USSD reports nearer to agent-level framing, like in the domestic context, to avoid direct criticism against the government.

From the perspective of the US government, the partial standing in evaluating police violence situations in the USSD reports is based on two strategic considerations. On the one hand, the high political sensitivity of the police violence issue domestically encourages the US government to frame their accusation of this topic on other countries carefully. On the other hand, by framing police violence in favor of its allies, the US could also expect favor back (Terman & Byun, 2022) and ease the possible criticism from the allies against its domestic police violence issue. Based on the distinction of the police violence framing strategies specified above, the hypothesis is as follows:

- In USSD reports, the US government is more likely to report on police violence of US allies and friends using agent-level framing and report on police violence of non-allies and enemies using government-based framing.

1.2.3 Defining Framing

Political communication scholars define framing as "the process of culling a few elements of perceived reality and assembling a narrative that highlights connections among them to promote a particular interpretation." (Entman, 2007, p. 164) Two aspects of the conception are worth to be highlighted. First, framing is not equal to lie (Entman, 2007). Instead, the primary framing strategy is selecting specific aspects of a perceived reality and connecting them in a narrative to promote certain interpretations (Entman, 2010; McCombs & Ghanem, 2001; Scheufele, 2000). Second, the conceptual relation between agenda setting, what topics to talk about, and framing, how aspects of the topics are highlighted to promote specific interpretations, are still subtle (Entman, 2007; McCombs & Ghanem, 2001). In this article, I distinguish framing from the agenda-setting strategy and focus on the "information editing" nature of framing. Instead of what information on police violence issues is disclosed in the reports, I focus on how the information is organized and presented.

Framing affects both public opinion (Chong & Druckman, 2007; Iyengar, 1990; McCombs, 2002; Price et al., 2005; Rugg, 1941; Schuldt et al., 2011) and policy decisions (Baumgartner et al., 2008; Dardis et al., 2008). Researchers have shown the widespread political-driven framing bias in media reports (D'Alessio & Allen, 2000; Dietrich & Eck, 2020; Druckman & Parkin, 2005; Entman et al., 2004; Entman, 2010; Entman & Rojecki, 1993). However, the framing bias in the USSD reports has yet to be studied. As the USSD reports are highly structured and written by a large group of experts (Clark & Sikkink, 2013), the framing bias could be more subtle in the text and more challenging to detect.

1.2.4 Comparison of Police Violence Accusations Framing for Political Bias Detection

Comparing the framing inconsistency on police violence accusations also provides an effective way to detect the politics-driven bias across the USSD country reports. For the police violence issues, as long as the accusations exist across two country reports, the comparison of framing strategies will downgrade the substantive scale of the violations disclosed by the instance-level information from the reports and focus on how the violation accusations are presented in the reports and who is blamed for the violation. Empirical research shows that police violence is a widespread violation in different country regimes (Jackson et al., 2018). This fact allows an extensive pool for cross-sectional comparison on how police violence accusations are framed in the USSD reports while minimizing the influence of cross-national variations of the information environment, as presented by the differences in event-level information disclosed at each country report. Moreover, current literature uncovered the various attention on the different human rights topics across the USSD country reports (Bagozzi & Berliner, 2018; Park et al., 2020). Focusing on police violence accusation framing can narrow the empirical concern and provide a powerful lens for detecting the cross-sectional inconsistency in the USSD reports caused by clear human manipulations.

1.3 A LLM Approach for Text Representation

1.3.1 Framing Strategy Detection Methods

Although little research in social science, computer scientists have proposed many tools to capture framing bias on political text (Ajjour et al., 2019; Baumer et al., 2015; Card et al., 2015; Demszky et al., 2019; Field et al., 2018; Kwak et al., 2021; Mokhberian et al., 2020; Recasens et al., 2013; Tsur et al., 2015; Ziems & Yang, 2021). Generally speaking, two strands of methods are proposed by the CS scholars. One strand of research mainly relies on the BoW models for framing bias detection (Baumer et al., 2015; Demszky et al., 2019; Field et al., 2018; Recasens et al., 2013; Tsur et al., 2015). For example, Baumer et al., 2015 tries to identify the potential keywords to denote the existence of biased framing in political news. This method aligns with social scientists' emphasis on the role of keywords and catchphrases in framing issues (Entman & Rojecki, 1993; Gamson & Modigliani, 1989). This method aligns with social scientists' emphasis on the role of keywords and catchphrases in framing issues (Entman & Rojecki, 1993; Gamson & Modigliani, 1989). Another interesting research is from (Ziems & Yang, 2021), who examined entity-centric framing in reporting police violence in US newspapers. They find that the media coverage of police violence is highly biased along the ideological division. However, the lexicon-level analysis ignores the semantic information in original texts and might yield inaccurate measurements (Ziems & Yang, 2021). As a remedy, another strand of research tries to utilize more complicated algorithms like artificial neural networks (ANN) to involve semantic information in frame bias detection (Iyyer et al., 2014). However, the black-box nature of the ANN models heavily erodes the interpretability of these methods (De Marchi et al., 2004) and hinders the spread of the technique among social scientists.

The text network analysis has long been utilized in social science research (Bail, 2016; Carley, 2020; Kampf et al., 2015; Segev & Boudana, 2019; Segev, 2020). By converting text into a network with tokens as nodes and the co-occurrence of tokens in sentences as edges, the text network provides a humanly interpretable representation of text data, which can be applied for framing strategy detection. However, current text-to-network methods are still based on the BoW assumption and mainly use the co-occurrence of words for network construction without distinguishing the different logical connections among words (Bail, 2016; C. W. Roberts, 2020; Segev, 2020). Inspired by this technique, I propose a novel network approach for text analysis. Different from conventional text analysis methods, the proposed method utilized state-of-the-art LLMs in capturing word-level connection. The output text network provides an interpretable representation of the input text while keeping the semantic information. When combined with various network analysis tools, this approach emerges as an effective method for text-level analytical tasks, such as detecting framing biases.

1.3.2 Self-Attention Mechanism in the Transformers Model

The Transformers model is a specific type of ANN model in which multiple sub-models (called layers) are stacked together, and the output of a previous layer is the input of the next layer. Comparing to previous ANN schemes for NLP tasks [5] , the Transformers Model adopts a powerful design at each layer, called self-attention mechanism, that extends the model's capability in tracking the semantic information at sentence level during the training process and help the algorithm better "understand" the text (Vaswani et al., 2017).[6] In this project, I will utilize this self-attention mechanism in the Transformers model to capture the word-level connections in text data. For the rest of the section, I will introduce the working mechanics of the attention mechanism in the Transformers model and how it can be used to construct text networks for the original documents.

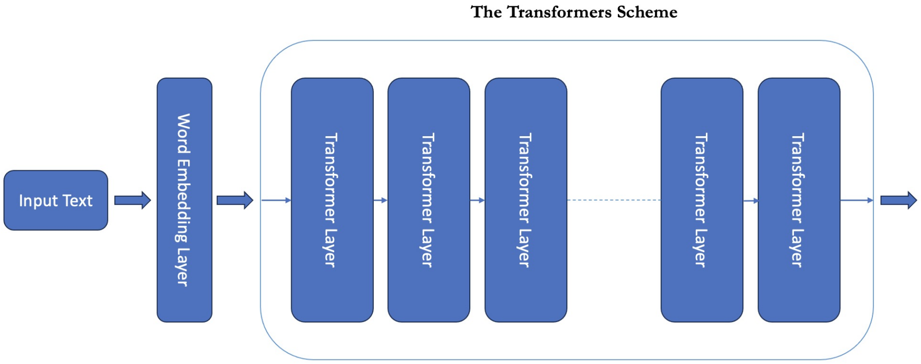


Figure 1.2: The Transformers Model Structure

Take BERT, a popular type of Transformers model, as an example. As shown in Figure 1.3, the sample sentence "The police arrested a peaceful protestor" is input into the model.[7] During training, BERT conceals one word in the in-

[5] Some examples are Recurrent Neural Network (RNN), Long Short-Term Memory Networks (LSTM), or Gated Recurrent Unit (GRU)

[6] The Transformers scheme is the backbone for most modern LLMs like Transformers (BERT) and Generative Pre-trained Transformers (GPT).

[7] The "[CLS]" and "[SEP]" strings are called placeholders, which are used to tell computer the starting and ending of one sentence.

put sentence and prompts the algorithm to use the words on both sides of the hidden word to predict it, which mechanism is called the Masked Language Model technology.[8] After rounds (called epochs) of training, BERT will be able to know how to predict the concealed word based on other words in the sentence. Like what human beings will do, BERT will pay more attention to those words closer connected with the hidden words in predicting. In our example, the word "arrested" might be weighted more than "a" in predicting the hidden word "police." To "understand" which words should be paid more attention in predicting the masked word, BERT stores the information of relevance among words through "$N \times N$" matrices (N is the number of tokens plus two placeholders). This type of "N x N" matrix produced in each layer of BERT (and other Transformer models) is called the (self-)attention matrix. Simply speaking, we can understand the attention matrix as a reference table helping Transformer models to understand the text-level word connections in the document.

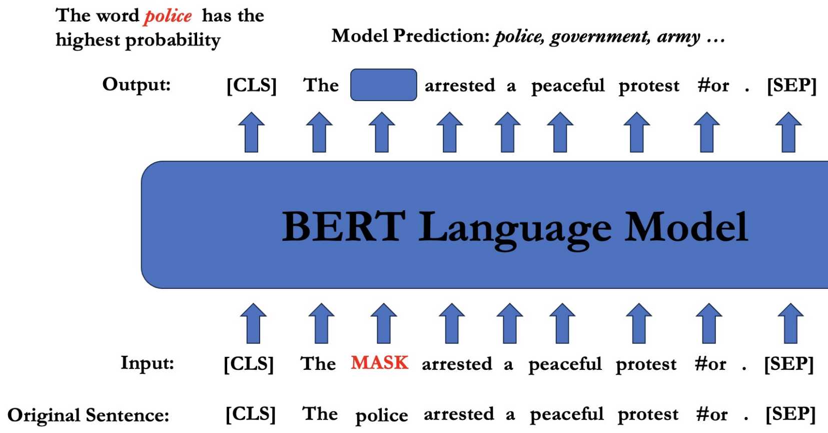

The word *police* has the highest probability
Model Prediction: *police, government, army …*
Output: [CLS] The arrested a peaceful protest #or . [SEP]
BERT Language Model
Input: [CLS] The MASK arrested a peaceful protest #or . [SEP]
Original Sentence: [CLS] The police arrested a peaceful protest #or . [SEP]

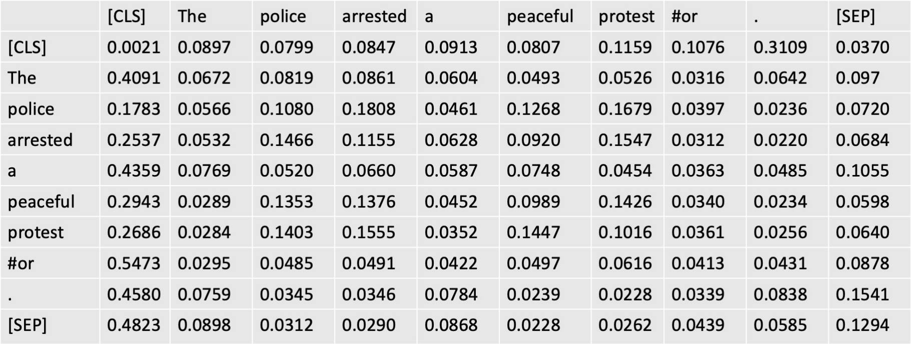
Figure 1.3: Masking Mechanism in BERT Training

Note: Through the masking mechanism, BERT learns the connection between words in one sentence and stores the word-level connection information in the attention matrix.

In the Appendix, I provide a more detailed introduction to how the transformer model calculates the attention matrix at each layer. In actual application, the Transformer models, including BERT, apply the "multi-head attention" mechanism, which simultaneously calculates multiple attention matrices in each layer to capture word-level connections at higher accuracy. The token-level connection information stored in the attention matrices allows the Transformers models to “understand” the substantive connections among words (Devlin et al., 2018). Figure 1.4 shows one self-attention matrix generated from the

[8] GPT, another popular Transformer model, has a very similar mechanism. Instead of looking at the words of both sides, GPT only allows the algorithm to look at the previous words in predicting the hidden word.

BERT model on our sample sentence by taking the average of the attention matrices generated at the final layer of the large BERT model. [9] Substantively, we can take each value in the matrix as the reference weight, suggesting the connection between the corresponding words in rows and columns.



	[CLS]	The	police	arrested	a	peaceful	protest	#or	.	[SEP]
[CLS]	0.0021	0.0897	0.0799	0.0847	0.0913	0.0807	0.1159	0.1076	0.3109	0.0370
The	0.4091	0.0672	0.0819	0.0861	0.0604	0.0493	0.0526	0.0316	0.0642	0.097
police	0.1783	0.0566	0.1080	0.1808	0.0461	0.1268	0.1679	0.0397	0.0236	0.0720
arrested	0.2537	0.0532	0.1466	0.1155	0.0628	0.0920	0.1547	0.0312	0.0220	0.0684
a	0.4359	0.0769	0.0520	0.0660	0.0587	0.0748	0.0454	0.0363	0.0485	0.1055
peaceful	0.2943	0.0289	0.1353	0.1376	0.0452	0.0989	0.1426	0.0340	0.0234	0.0598
protest	0.2686	0.0284	0.1403	0.1555	0.0352	0.1447	0.1016	0.0361	0.0256	0.0640
#or	0.5473	0.0295	0.0485	0.0491	0.0422	0.0497	0.0616	0.0413	0.0431	0.0878
.	0.4580	0.0759	0.0345	0.0346	0.0784	0.0239	0.0228	0.0339	0.0838	0.1541
[SEP]	0.4823	0.0898	0.0312	0.0290	0.0868	0.0228	0.0262	0.0439	0.0585	0.1294

Figure 1.4: Sample Self-Attention Matrix

1.3.3 Constructing Text Network from Attention Matrices

The word relations captured by attention matrices provide a convenient way to convert the original text into a weighted text network while largely keeping semantic information from the original text. Artificial intelligence researchers have recognized that pretrained transformer-based LLMs can capture the relational knowledge in text data (Safavi & Koutra, 2021). An emerging area in AI tries to extract knowledge graphs from the pretrained transformer models (AlKhamissi et al., 2022).[10] Moreover, the attention matrcies from the pretrained LLMs show their power in storing conceptual connections among the words (Vig & Belinkov, 2019; C. Wang et al., 2020; Wiegreffe & Pinter, 2019). For example, C. Wang et al., 2020 argue that the attention matrices output by the pretrained LLMs can effectively capture the logical connections among words from the given text. [11]

Accordingly, the text network constructed on the attention matrices from the pretrained LLMs could provide a comprehensive representation of the original text. In this article, I use the pretrained large BERT encoder model to extract attention matrices from the USSD reports.[12] I specify a four-step pipeline for text network construction from the BERT model. First, each USSD country report is split into sentences to feed into the BERT model. I extract on the attention matrices from the last layer of BERT to use.[13] Second, I take the average across all the extracted attention matrices at each entre and convert the output

[9] As discussed in the Appendix, the attention matrix is unsymmetric by definition (Vaswani et al., 2017).

[10] Knowledge graph stores information in graphic format in which each node is a substantive entity, and edges show logical relationship between the connected entities.

[11] Whether the attention matrix could capture the relationships among substantive entities (relationship of "things") as focused in knowledge graph research is still an ongoing conversation in AI (Jain & Wallace, 2019; Mohankumar et al., 2020; Wiegreffe & Pinter, 2019). I focus on a much simpler task, using attention matrices only for word connection recognition (relationship of "strings"). Substantive knowledge (like framing strategy) will be extracted by further querying on the text network. As C. Wang et al., 2020 implies, when only focusing on the conceptual relevance among strings without considering substantive meanings, the attention matrix from pretrained LLMs could provide enough information.

[12] The weights for the "large-bert-cased" model are available in https://huggingface.co/bert-large-cased.

matrix to symmetric by averaging it with its transpose matrix. Third, I take the symmetric matrix from step two as the adjacent matrix for sentence-level text network construction. I delete all the values in the adjacent matrix's diagonal to avoid self-loop. Each token is stemmatized and concatenated with its part-of-speech (POS) tagging through underscore as the nodes in the text network (like "kill_verb"). For simplification, in this step, I also delete all nodes of place-holders, stopwords, numbers, punctuations, and the connected edges. Finally, I combine the sentence text networks together into the document text network by merging the same nodes and adding weights of edges connecting the same pairs of nodes. Figure 1.5 shows the general process of the pipeline. [14]

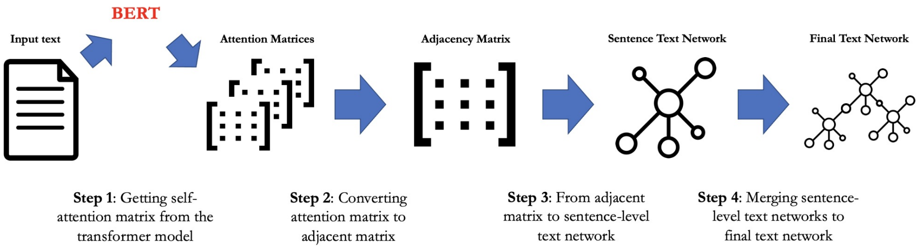

BERT
Input text
Attention Matrices
Adjacency Matrix
Sentence Text Network
Final Text Network
Step 1: Getting self-attention matrix from the transformer model
Step 2: Converting attention matrix to adjacent matrix
Step 3: From adjacent matrix to sentence-level text network
Step 4: Merging sentence-level text networks to final text network

Figure 1.5: Text-to-Network Method Pipeline

1.4 A Network Approach in Measuring the Framing Strategies in the USSD Reports

1.4.1 A Network Approach in Measuring Framing on Text-Level

On police violence accusations, the varying framing strategies in the USSD reports are characterized by the different levels of focus on the government's participation in the reported violations. In reports, when the government's role is highlighted in police violence, the semantics of the keywords "government" and "police" will be highly similar because of the two keywords' mutual-substitutive status in reported violations. Following this logic, I will utilize the semantic similarity between the the keywords "government" and "police" as the proxy measurement of the police violence accusation in the USSD reports. Moreover, as framing is achieved by the sophisticated combination of words in information presentation (Entman & Rojecki, 1993; Gamson & Modigliani, 1989), the

[13] The large BERT model has 24 layers and 16 attention heads at each layer, which leads to 384 attention matrices for each input sentence (Devlin et al., 2018). Arguably, common wisdom suggests that the deep learning model's final layer provides the highest-level feature representation for each data point. The proposed pipeline allows different layers for attention matrix extraction, especially considering the ongoing dialogue of the information intensity packed in the different layers of the deep learning models (Vig & Belinkov, 2019).

[14] The attention matrices are learned from the full text with no preprocessing which guarantees the Transformers model better captures the semantic information from the text. Text simplification is in step three when constructing the text network so that the resulting network only words bearing substantive meanings.

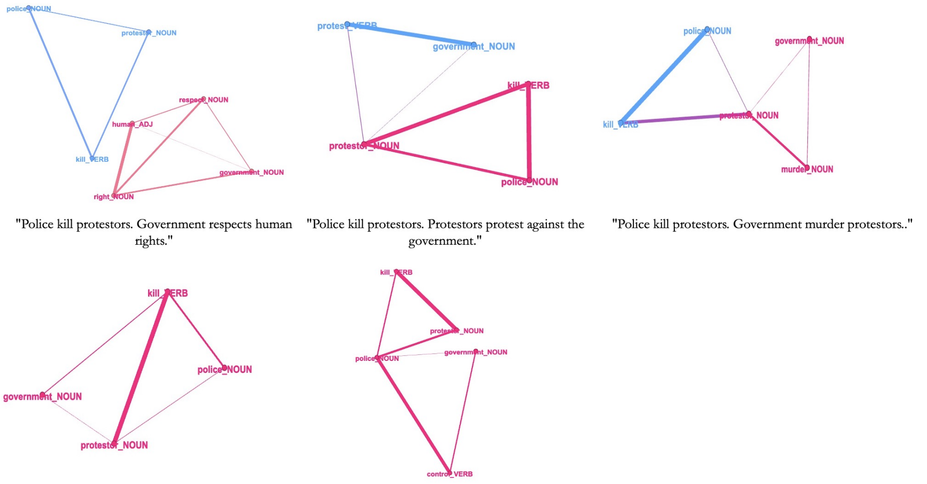
semantic similarity can be captured by analyzing the words that logically connect to these two keywords. The proposed network representation of text data provides a convenient way for this purpose.

After converting each USSD report into a text network through proposed method, I extract the subgraph around the two keywords "government" and "police" with all connected nodes and the edges (called "police-government" network). When the semantics of the two keywords are diverge and the role of government is understated in police abuse, the sets of words connecting with the "government" and "police" should be different, with minimal interconnecting edges. Conversely, when the government is highlighted in police violence accusations, the words linked to these two keywords are likely to substantially overlap, resulting in a dense network. This divergence in community structure inside of the "police-government" network can be quantified by the modularity score (Newman, 2006). Theoretically, the modularity score of a given network could take values from 0 to 1. A higher modularity score means more significant sub-modules exist in the network, which implies an agent-level framing and the role of government is downplayed in police violence accusations. A lower modularity score implies either the government is highly involved in police violence accusations or the words "government" and "police" are interchangeable in describing the relation against the same set of conceptions like specific victims or violations. Either way, the police violence accusation is framed on the government level. The extreme case could be that the police department is not taken as an independent actor, and all accusations involving police action, like arresting or torturing, default as the government's behaviors. In fact, for most USSD reports on North Korea between 1995 and 2005, the police department is not mentioned, and only the government's role in police-related brutality is highlighted.

Figure 1.6 shows examples of how the network approach could quantify police violence accusation framing strategies from text. The text networks in the graph are generated from the sample sentences using the proposed method. Applying the Louvain community detection algorithm (LDA) (Blondel et al., 2008), the colors show the detected community structure on each text network.[15] The modularity scores for the community structure of each text network are shown in Table 1.1. As shown, when the police violence accusations are separated from the government evaluation, the resulting text network has significant submodules, and the modularity score is high. As the involvement of the government in police violence becomes increasingly evident, words associated with 'government' and 'police' start to construct a denser network with

[15] When applying LDA, I set the two keywords nodes "government_NOUN" and "police_NOUN" as the starting points (seeds) for the network search.

a lower modularity score. When the modularity score is 0, the government's responsibility for police violence is clearly pointed out in sample texts.



"Police kill protestors. Government respects human rights."

"Police kill protestors. Protestors protest against the government."

"Police kill protestors. Government murder protestors.."

"Police kill protestors. Government kill protestors."

"Police kill protesters. Government controls the police."

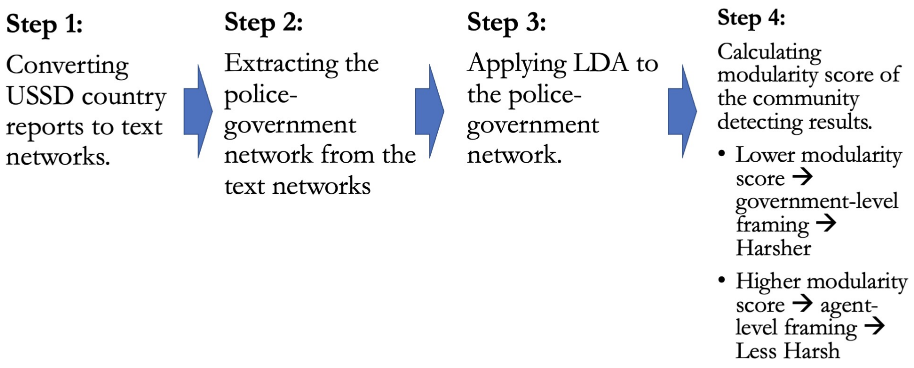
Figure 1.6: Community Detection on Sample Text Networks

Table 1.1: Modularity Scores for Text Networks

Index	Sample Sentence	Modularity Score
1	"Police kill protestors. Government respects human rights."	0.45
2	"Police kill protestors. Protestors protest against the government."	0.26
3	"Police kill protestors. Government murder protestors."	0.14
4	"Police kill protestors. Government kill protestors."	0.0
5	"Police kill protesters. Government controls the police."	0.0

1.4.2 Measuring Police Violence Accusation Framing in USSD Reports

In this research, I utilize the USSD reports corpus from the Human Rights Text dataset collected by (Fariss et al., 2015). Figure 1.7 shows the general steps of measuring police violence accusation framing strategy by the network approach stated above in the USSD reports. Each USSD country report is converted to a text network using the proposed approach, and the police-government subgraph is extracted from the output network. After applying LDA to each police-government network, the resulting modularity score for the detected community structure of each network is taken as the measurement of the framing strategy of police violence accusations in the USSD reports.



Step 1:

Converting USSD country reports to text networks.

Step 2:

Extracting the police-government network from the text networks

Step 3:

Applying LDA to the police-government network.

Step 4:

Calculating modularity score of the community detecting results.

- Lower modularity score → government-level framing → Harsher
- Higher modularity score → agent-level framing → Less Harsh

Figure 1.7: Steps in Detecting the Police Violence Accusations Framing Bias in the USSD Reports

Importantly, the proposed approach's effectiveness in measuring the framing bias in the texts relies on the utilized Transformers model's capacity to capture the semantic information from the text data. It is not guaranteed that the utilized BERT model can capture the logical connections among words with 100% accuracy. Nevertheless, the aggregation process of combining sentence text networks to a document text network helps to mitigate the potential inaccuracy for word relation detection at the sentence level. The proposed scheme also allows the embedding of more advanced LLMs for better performance.

1.5 Politics-driven Framing Bias in the USSD Reports

1.5.1 Regression Analysis

Given the measurement of police violence accusations framing, the next step is to test the political bias argument on the USSD reports. Despite the political consideration, police violence accusations in the reports could be driven by many other factors. Accordingly, I rely on statistical tools to confirm whether the variation of police violence accusations framing in the USSD reports is from political manipulations.

The unit of analysis is country-year. The response variable is the police violence accusation framing in USSD reports measured by the modularity score of the police-government networks. To measure one country's political relationship with the US, I first adopt the US alliance indicator from the Correlates of Wars (COW) database (Gibler, 2008), which provides a dichotomous measurement on whether a state constructs a formal alliance with the US in a given year. Meanwhile, I take the voting agreement variable provided by Bailey et al., 2017 as an alternative measurement, which shows the level of political agreement between one country and the US at the UN Generally Assembly voting each year, taking the value from 0 to 1 with higher values meaning more agreement. To address the potential influence of the economic relation, I include the volume of one country's annual trade with the US (Pevehouse et al., 2020) and the received US aid (USAID data).

To control the confounding impact of the reported country's actual police violence level each year, I adopt the police-related torture indicators from the Ill-Treatment and Torture (ITT) country-year dataset (Conrad et al., 2013). The inclusion of the three police-related variables from the ITT data on police violence against criminals, dissent, and marginalized individuals could also help to control the impact of different types of police violence on the accusations framing in the USSD reports. These variables are coded based on the allegations in Amnesty International's annual human rights reports, media reports, and Action Alerts, and span from 0 to 5, with higher values meaning more systematic police violation (Conrad et al., 2013). Following Jackson et al., 2018 practice, I transform the three variables into dichotomous indicators to avoid the potential influence of the preponderance of zeros in the measurements.

For other confounders, I first control one country's general human rights practices using the Fariss latent human rights protection indicators (Fariss, 2014). As the government's ability to control the police department might also impact the monitoring agency's framing of police violence, I include the state capacity variable from (Hanson & Sigman, n.d.).[16] I control one country's regime type using the polity2 variable from the POLITY V dataset (Saunders, 2010). Finally, I control the country population (log-transformed) for the potential influence of population pressure on political violence (Urdal, 2008) and the GPD (log-transformed) volume. For the limitation of data availability on the predictors and control variables, the final dataset for the regression analysis spans from 1995 to 2005. Figure 1.9 shows distribution of the modularity score in the dataset. Table 1.2 shows the country reports with the lowest modularity scores (harshest framing against the government) from the final dataset.

Fitting all variables into the ordinary least square (OLS) regression model, the results are shown in Table 1.3.[17] As shown, when one country is politically closer to the US, the modularity score of the police-government network from the corresponding USSD report is significantly higher (agent-level framing).

[16] As an alternative measurement of the central government's control of local police, I use the government mode variable from the IAEP dataset which indicates whether a country takes a unitary system or federal system (Wig et al., 2015). The main regression results remain the same and the government mode variable does not have a significant impact on the response variable.

[17] The missing values and the model diagnostic analysis are discussed in the Appendix. According to the results, no significant violation of the linear model assumptions is found.

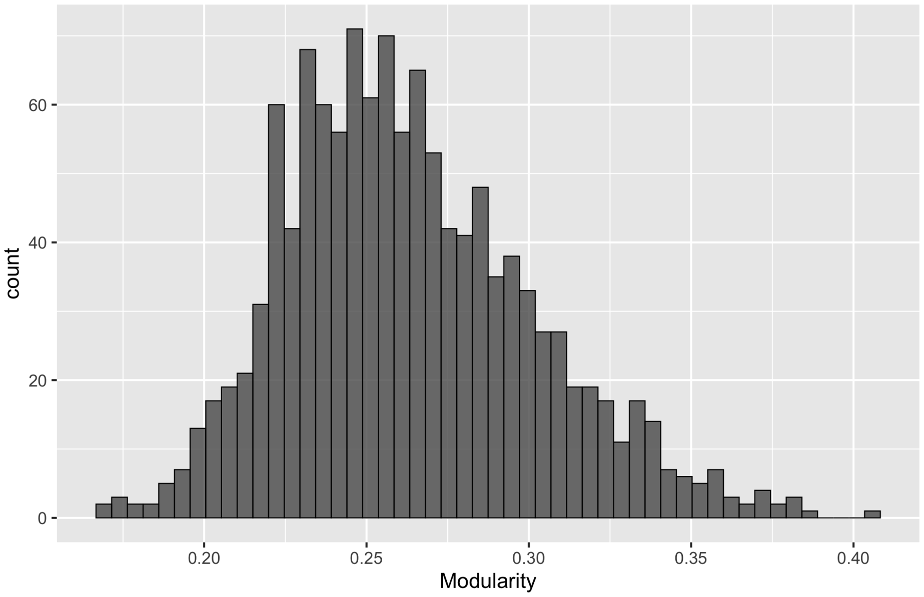


Figure 1.8: Distribution of the Moduralty Scores of the "Police-Government" Text Network from the USSD Reports: 1995 - 2005

Substantively, as the modularity score spans from 0.17 to 0.41, one country's alliance with the US is associated with the increase of modularity score by 3% in the USSD reports. From model 2, the modularity scores for countries closest to the US (UN Voting Agreement equals 1) are 34% higher than those furthest from the US (UN Voting Agreement equals 0). Both relations are statistically significant at the 99% confidence interval. These results suggest that the framing of police violence accusations is biased for countries politically closer to the US and supports the main hypothesis.

For economic relations, model 2 provides weak statistical evidences on the association of one country's enlarging trade with the US and a less harsh accusation against its government regarding the police violence topic in the USSD reports. This finding is consistent with existing suspicion on the human rights reporting bias driven by one country's economic relationship with the US (Bagozzi & Berliner, 2018; Foot, 2000; Mertus, 2008). Meanwhile, both models suggest that the US aid recipient countries would face harsher accusations against the government on police violence issues. The opposite impacts of the two economic variables could be explained by the subtle mechanisms of the US aid distribution (Demirel-Pegg & Moskowitz, 2009) and also conform with the

Table 1.2: Country Reports with Lowest Modularity Scores on the "Police-Government" Text Network: Top 20

	country_year	Modularity		country_year	Modularity
1	Russia_2001	0.17	11	Kazakhstan_1999	0.19
2	Russia_2003	0.17	12	Russia_2005	0.19
3	Russia_2004	0.17	13	Saudi Arabia_2004	0.19
4	Russia_2002	0.17	14	Saudi Arabia_2002	0.19
5	China_1999	0.18	15	Ukraine_2000	0.19
6	Russia_1999	0.18	16	Sudan_2003	0.19
7	Russia_2000	0.18	17	Belarus_2003	0.19
8	North Korea_2005	0.18	18	Israel_2005	0.19
9	Israel_2003	0.19	19	Myanmar_1999	0.19
10	China_1998	0.19	20	Israel_2001	0.19

empirical findings on the influence of US aid at the cross-national power hierarchy (T.-Y. Wang, 1999).

For other variables, the polity score is positively correlated to the response variable, which suggests the accusations of police violence against democratic governments are less harsh in the USSD reports. Interestingly, neither the human rights variable nor the police violence indicators were shown to be significantly associated with the police violence accusation framing in the USSD reports in both models. As shown in the online Appendix, when only regressing the police violence indicators against the response variable, the results suggest the police violence accusation in the USSD reports is significantly harsher when more serious police abuses are observed in the reporting countries. Similarly, one country's better human rights practice is significantly associated with less harsh police violence accusations in the USSD reports. One explanation for the nonsignificance of the police violence indicators and the human rights variable in Table 3 models is that other more significant predictors cover their covariations with the response variable. This could further suggest the existence of substantial human manipulations on the police violence accusations in the USSD reports.

1.5.2 Cases Comparison

At this point, readers may be curious about the manifestation of the framing bias on police violence accusations in the USSD reports. In this section, I aim to illustrate how the captured framing bias could appear at the text level. The case comparison could also serve as a visual validation of the proposed measurement method. I utilize a matching method following a similar logic as suggested in Ho et al., 2007 to select the cases for comparison. Focusing on the 2005 data, I compare the observations of US allies against Russia, extracting the top five countries with the smallest Minkowski distance across all other control variables.[18] Among the five cases comparable to Russia, I choose to compare the 2005 Russia Report and Brazil Report for potential substantive interests.

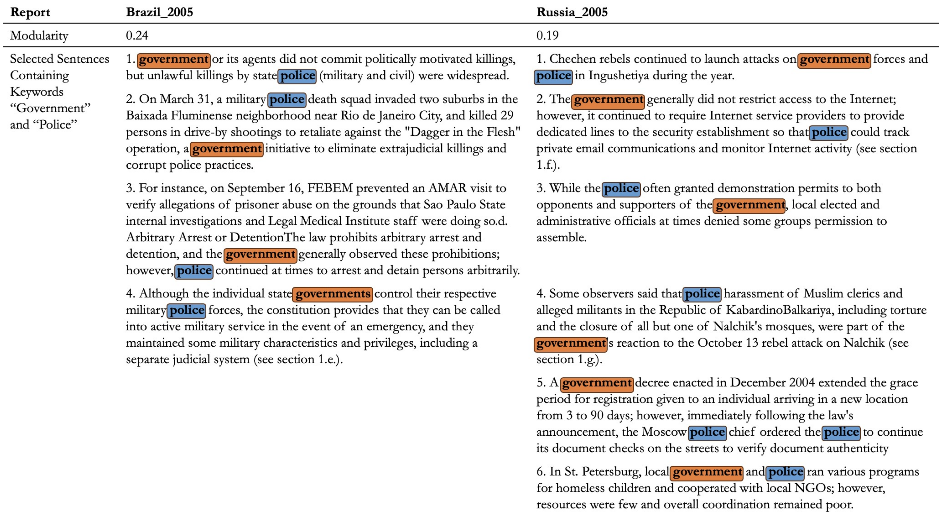
Figure 1.9 presents all the sentences from the two reports containing both the keywords "government" and "police". As shown, the role of government is understated in police violence reporting in the Brazil report and is emphasized in the Russia report. For example, in sentence one of the Brazil report, the government is pointed out as not involved in political killing even if the police killings are widespread. In sentence four from the 2005 Russia report, the role of government is highlighted in the police harassment of Muslim clerics. The different framing strategies are captured by the proposed measurement. As shown, the Russia 2005 report has a much lower modularity score compared to the Brazil report.

Importantly, the cases discussed in this section are to help readers better understand how the framing bias in the police violence accusations could be shown at the text level in the USSD reports. The proposed network approach does not rely on the keyword co-occurrence in the same sentence to measure the government's role in police violence accusations. In many reports, no sentence contains the keywords "government" and "police" simultaneously, while the proposed method could still capture the semantic similarity of the two keywords at the document level.

1.6 Robustness Check: Framing Bias in the PTS Index

As discussed above, information framing could greatly impact the audiences perceptions (Entman, 2007). If the proposed measurement is accurate, captured framing bias should influence how researchers evaluate one country's human rights practices based on the USSD reports. In the PTS human rights dataset, trained coders generate separate human rights indexes of one country from the USSD (PTS_SD) and Amnesty International's (PTS_AI) reports. If the measurement captures meaningful bias in the USSD reports, then for the country exhibiting lower modularity in the police-government network, the PTS_SD index is expected to indicate more severe human rights violations compared to the PTS_AI index.

[18] Other four cases are: Colombia, Philippines, Turkey, Ecuador. All five country reports have much higher modularity scores than the Russia 2015 report.



Report	Brazil_2005	Russia_2005
Modularity	0.24	0.19
Selected Sentences Containing Keywords “Government” and “Police”	1. government or its agents did not commit politically motivated killings, but unlawful killings by state police (military and civil) were widespread.	1. Chechen rebels continued to launch attacks on government forces and police in Ingushetiya during the year.
	2. On March 31, a military police death squad invaded two suburbs in the Baixada Fluminense neighborhood near Rio de Janeiro City, and killed 29 persons in drive-by shootings to retaliate against the "Dagger in the Flesh" operation, a government initiative to eliminate extrajudicial killings and corrupt police practices.	2. The government generally did not restrict access to the Internet; however, it continued to require Internet service providers to provide dedicated lines to the security establishment so that police could track private email communications and monitor Internet activity (see section 1.f.).
	3. For instance, on September 16, FEBEM prevented an AMAR visit to verify allegations of prisoner abuse on the grounds that Sao Paulo State internal investigations and Legal Medical Institute staff were doing so.d. Arbitrary Arrest or DetentionThe law prohibits arbitrary arrest and detention, and the government generally observed these prohibitions; however, police continued at times to arrest and detain persons arbitrarily.	3. While the police often granted demonstration permits to both opponents and supporters of the government, local elected and administrative officials at times denied some groups permission to assemble.
	4. Although the individual state governments control their respective military police forces, the constitution provides that they can be called into active military service in the event of an emergency, and they maintained some military characteristics and privileges, including a separate judicial system (see section 1.e.).	4. Some observers said that police harassment of Muslim clerics and alleged militants in the Republic of KabardinoBalkariya, including torture and the closure of all but one of Nalchik's mosques, were part of the government's reaction to the October 13 rebel attack on Nalchik (see section 1.g.).
		5. A government decree enacted in December 2004 extended the grace period for registration given to an individual arriving in a new location from 3 to 90 days; however, immediately following the law's announcement, the Moscow police chief ordered the police to continue its document checks on the streets to verify document authenticity
		6. In St. Petersburg, local government and police ran various programs for homeless children and cooperated with local NGOs; however, resources were few and overall coordination remained poor.

Figure 1.9: Cases Comparison: Brazil 2005 Report vs. Russia 2005 Report

I construct two country-year variables based on the PTS indexes from 1995 to 2005: one dichotomous "SD_HIGHER" variable signifying the cases of one country receiving a worse (higher) PTS_SD score than PTS_AI score, and one ordered "SD_AI" variable subtracting the PTS_AI score from the PTS_SD of the same country each year. As shown in Figure 1.10, for the observations with worse PTS_SD scores, the corresponding USSD reports' modularity scores are significantly lower. Figure 1.11 suggests a linear relationship in which worse PTS_SD scores is associated with lower modularity scores, even though the association does not arrive conventional significance threshold. Assuming the consistent coding standard in the PTS project (Arnon et al., 2023), these results suggest the substantive influence of the framing bias and confirm the validity of the proposed framing bias measurement.

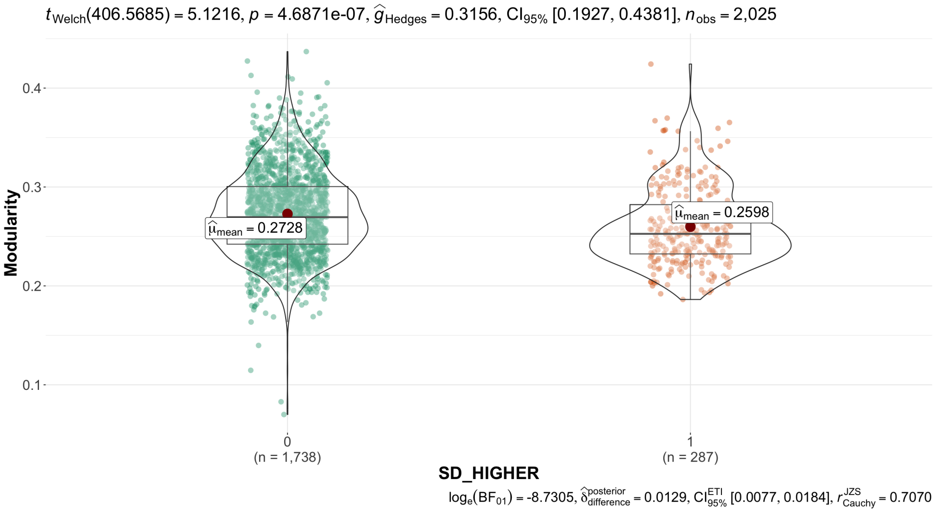


Figure 1.10: Framing Bias in the PTS Dataset: Binary Response

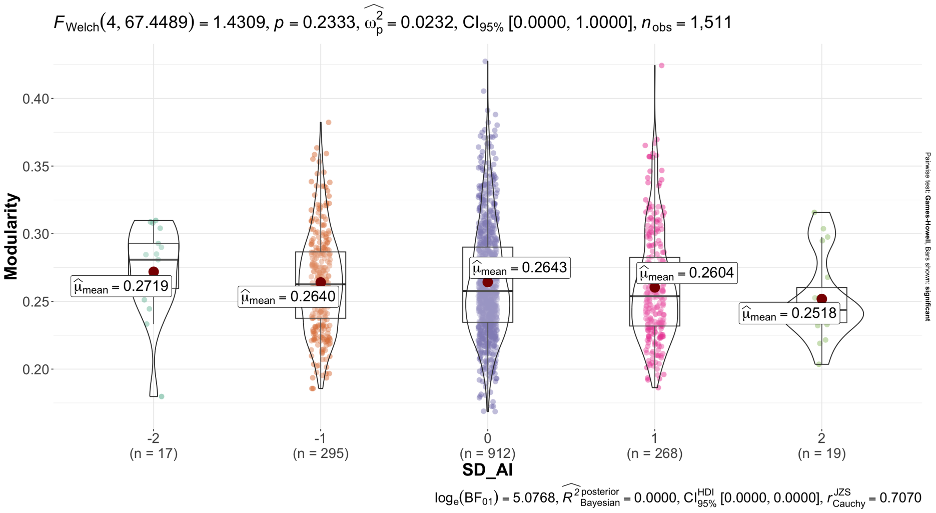


Figure 1.11: Framing Bias in the PTS Dataset: Ordered Response

Table 1.3: Empirical Results: Political Bias in USSD Reports 1995 - 2005

	Dependent variable:	
	Modularity of the Police-Government Network	
	(1)	(2)
US Ally	0.007**	
	(0.003)	
UN Voting Agreement		0.081***
		(0.008)
Received US Aid (logged)	−0.005***	−0.005***
	(0.0005)	(0.0005)
Trade with US (logged)	−0.001	0.001*
	(0.001)	(0.001)
Police Vio. against Criminals	0.003	0.00002
	(0.004)	(0.003)
Police Vio. against Dissidents	0.005	0.007**
	(0.004)	(0.003)
Police Vio. against Marginalized Indiv.	−0.0004	−0.002
	(0.003)	(0.003)
State Capacity	0.007***	0.002
	(0.002)	(0.002)
Fariss Human Rights Indicator	0.00001	0.0001
	(0.001)	(0.001)
Population (logged)	−0.008***	−0.009***
	(0.001)	(0.001)
GDP per capita (logged)	−0.012***	−0.016***
	(0.002)	(0.002)
POLITY Score	0.002***	0.002***
	(0.0002)	(0.0002)
Constant	0.569***	0.575***
	(0.024)	(0.023)
Observations	1,311	1,270
R^2	0.261	0.327
Adjusted R^2	0.255	0.322
Residual Std. Error	0.035 (df = 1299)	0.033 (df = 1258)
F Statistic	41.780*** (df = 11; 1299)	55.678*** (df = 11; 1258)
Note:	*p<0.1; **p<0.05; ***p<0.01	

Chapter 2

Conclusion & Discussion

How does one country's domestic human rights issues influence its international human rights practice? In this article, I focus on the influence of the US domestic police violence issue in its transnational human rights reporting. The nature of police force allows two directions in framing police violence: government-level framing and agent-level framing. The government-level framing represents harsher criticisms against the government for highlighting the government's direct responsibility in police abuse. Through a novel network approach for text analysis, the empirical results uncover political-driven framing bias in the police violence accusations at the USSD reports. Controlling for other variables, for countries politically closer to the US, the USSD reports tend to frame police brutality as an agent-level issue and avoid directly criticizing the government.

The theoretical contribution of this article is twofold. First, this research shows the transnational impact of US domestic police violence issues. For the importance of the USSD reports for the international human rights regime (Bagozzi & Berliner, 2018), this finding could imply a profound influence of the US police violence problem on the global level. Second, this research speaks to the ongoing literature on political-driven bias in human rights monitoring and reporting (Arnon et al., 2023; Park et al., 2020). The result suggests that political bias in human rights reports not only exists in what information is disclosed but also in how the accusation is framed. The framing bias detected in the USSD reports encourages further scholarship to inspect political texts more carefully for subtle manipulations.

This article exposes the tendency of the US government to take a slippery standing in assessing other country's human rights practices on issues with high domestic sensitivity. On one hand, this finding emphasizes the potential divergence in human rights reporting across different issue areas and violation types. If the monitor is under attack on a specific human rights issue, it might not be able to provide the most unbiased evaluations of other actors' practices in that area. For scholars, we need to also take into account the monitor's own human rights record when referring to their reports, especially concerning issues that pose challenges to the monitor itself. On the other hand, besides police violence, other issues that challenge the US, like torture, abortions, and racism, as mentioned above, might also impact how the corresponding accusations are framed in the USSD report. Further examination of the accusations on other human rights topics bothering the US in the USSD reports might help scholars better understand the biases sources of the documentation as well as the connection between the US domestic politics and its transnational human rights practices.

Methodologically, the proposed LLM-based method provides a powerful representation of text data in a network form. Different from BoW-based text analysis methods, the network representation of text data can effectively involve semantic information in analysis. Moreover, the text network provides a human-interpretable output for the automatic text processing. The text network could work with different network analysis tools for downstream tasks like topic modeling and sentiment analysis. The proposed text-to-network pipeline is compatible with different Transformer models and allows political scientists to work with the most-advanced LLMs for better text representation.

Moreover, from the application in this paper, the BERT model shows capability in capturing logic inside of natural languages. Nevertheless, based on the basic scheme of connectivist AI algorithm, it is still too early to claim that transformers-based LLMs could understand the logic in language just like human beings. The current trend in AI advancement, represented by the popularity of different generative AI models like ChatGPT, seems to become the competition on the simple neural-netowork-based model (layer) extension and computational power aggregation. Without understanding what kind of information is picked up by the billions of parameters boxed in the neural network during training, this simple competition scheme on AI development points out a rather grey picture for the AI scholarship. Assuming a "strange point" exists, in which the LLMs magically passes the Turing Test and gets the features to be defined Artificial General Intellgenece (AGI).[19] The development of connectivist AI before the "strange point" is rather boring for the simple network extension and computational power aggregation. Meanwhile, the AI algorithm after the "point" is very dangerous as we don't understand how the data is processed and the perceptions are constructed in the background. Through this project, the author hopes to contribute to the existing academic endeavors on unpacking

[19] If referring to how simple neurons could aggregate together and become the human brain, the assumption of the strange point for the final arrival of AGI through simple neural network extension and computational power aggregation is reasonable.

the blackbox of the LLMs and help scholars to understand the mechanism in the Transformers model better.

Appendix A

Bad Apples Just for Friends: A Large Language Model Approach in Studying Police Violence Accusation Framing at US Human Rights Reports

A.1 Attention Mechanism in Transformers

A.1.1 Attention Score Computation for Single Word

In this section, I will provide a more specific introduction to how the attention score of a single word in one sentence is calculated in the BERT mode. In the BERT model, the input sentence is first broken into words (called tokenization), and the word is converted to vectors of the same length for later processing, which step is known as word embedding.[20] Essentially, the word-embedding process is to use a list of values to represent a specific word.[21] The word-embedding process is only applied once on each word at the beginning of the Transformer model. After embedding, the input sentence is converted to a $n \times k$ matrix, in which each row corresponds to a token in the sentences, noted as $v_1, v_2, \ldots, v_n$, and the k as an arbitrary number for the length of each word-embedding vectors.[22]

[20] The "[CLS]" and "[SEP]" in the graph are the placeholders to identify the start and end of one sentence.

[21] In actual training, Transformers models, including BERT, will further break the words or tokens into letters. However, the training logic (masking mechanism) and the calculation of attention matrices are the same. For better illustration, I focus on the word-level mechanisms.

[22] The default value for k is 512 in (Vaswani et al., 2017).

The embedding matrix is then input into the attention layer. At every attention layer, we initiate three matrices W^Q, W^K, and W^V as the Query weight matrix, Key weight matrix, and Value weight matrix, the values of which are first given as constants and then learned from the model training. The dimensions of W^Q and W^K are $k \times d_k$, and the dimension of W^V is $k \times d_v$. [23] By multiplying v_i with the three matrices, we get three new vectors for each token i called key vector k_i, query vector q_i, and value vector v_i.

In the next step, we calculate the attention scores of each token in predicting the hidden tokens. For the example in Figure 2, the attention score for the token "The" in predicting the hidden word "police" should be $Attention_{"The"} = q_{"police"} \cdot k_{"The"}$, and for the token "arrested" in predicting police should be $Attention_{"arrested"} = q_{"police"} \cdot k_{"arrested"}$. This attention score describes how much attention should be put to the corresponding word in predicting the hidden word. For each hidden token, by calculating the attention score of every tokens from the sentence (including the hidden token itself), we can get a n-length vector in which each value is the attention score of the corresponding tokens against the targeted tokens. Through normalizing each value by the sum of the vector, the attention scores in the vector are converted to ratios, which provide a more intuitive (and computationally efficient) way to describe the weights each token got in predicting the hidden tokens. Next, for the weights of each token in the normalized attention matrix, we multiply them with the value vectors of corresponding tokens to highlight the impact of relevant tokens and drown out irrelevant tokens. Finally, we sum up the weighted value vectors, and this produces the self-attention layer at the focusing (hidden) token.

A.1.2 Attention Matrix Computation

In application, the attention layers are computed in matrix form. Following the illustration above, the dimension of the embedding matrix is $N \times k$, and each row of the embedding matrix v_i is a vector representation of corresponding words. Then, we calculate the Query, Key, and Value matrices from the pretrained weight matrices using the following function:

$$Q_{N \times d_k} = A_{N \times k} \cdot W^Q_{k \times d_k} \tag{A.1}$$

$$K_{N \times d_k} = A_{N \times k} \cdot W^K_{k \times d_k} \tag{A.2}$$

$$V_{N \times d_v} = A_{N \times k} \cdot W^V_{k \times d_v} \tag{A.3}$$

Substantively, we can understand the Query matrix (Q) as the representation of the focused word for which the context is being determined. The Key matrix

[23] Both d_k and d_v are arbitrary numbers and can take values different from N and k. The default setting is $d_k = d_v = 64$

(K) creates key vectors for each word (as each row in the K matrix), which helps the algorithm to measure the relevance between the focused word (using the corresponding query vector from the Q matrix) and other words. The Value matrix (V) is utilized for calculating the final text-aware vector representation of each word. Finally, the matrices Q, K, and V are input into the following equation to extract the attention function:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})_{N\times N}V_{N\times d_v} \tag{A.4}$$

As shown in equation (4), the production of the Query matrix (Q) and the K matrix (K) produce a $N \times N$ attention matrix, which is designed to capture the sentence-level connection among each words.[24] The $softmax$ function, which is called the activation function in the deep-learning model, is for row-wise weight normalization to convert attention score to proportion, as we discussed in the last section. After the $softmax$ function, each row of the self-attention matrix should be sum to 1.

Strictly speaking, the final attention matrix is the production of the attention weights and the value matrix, and the dimension should be $n \times d_v$, as suggested in equation (4). Each row of the attention matrix is a new representation of the corresponding words different from v_i of the embedding matrix with the involvement of the semantic information captured by the transformer model. However, most working Transformer-based algorithms, like BERT, will capture the output of $softmax(\frac{QK^T}{\sqrt{d_k}})_{N\times N}$ as the $N \times N$ self-attention matrix. [25] For the multi-head attention mechanism we discussed in the paper, there are $h = k/d_v$ number of different self-attention matrices and the $n \times d_v$ final attention matrices calculated in each layer. By concatenating the h final attention matrices together, there is a new $n \times k$ attention matrix to input to the next attention layer. In this research, I will only focus on the $N \times N$ self-attention matrices at each attention layer to extract the word-level connections. Especially as both the Q and K matrices are not symmetric, the final $n \times n$ self-attention matrices are not symmetric either.

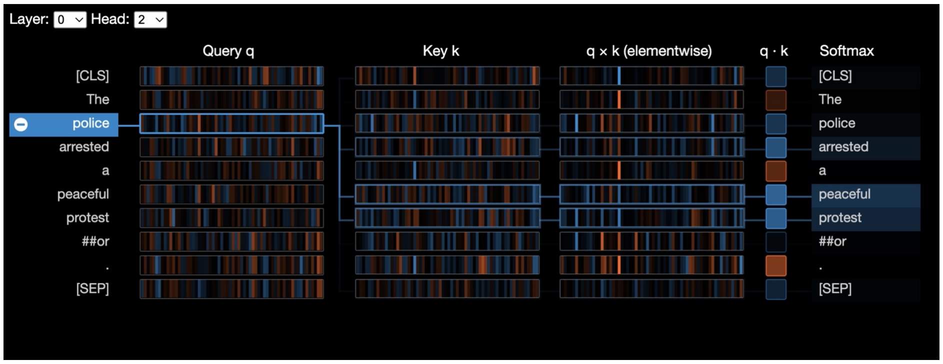
Figure A.1 [26] shows an example of how the self-attention matrix is calculated on the sample sentence from the BERT model. For each word in the sentence, there is a corresponding query vector (q) and key vector (k). To know which words are closely connected with the word "police," the model calculates the element-wise product of the query vector for "police" (q_{police}) with the key vectors of all other words ($k_{\neg police}$), as shown in the "q x k (elementwise)" column. The sums of the element-wise product for each pair of q_{police} and $k_{\neg police}$ are calculated ("$q \cdot k$" column). The words that are closely relevant to the query

[24] The $\frac{1}{\sqrt{d_k}}$ scalar is used to scale the values in the QK^T matrix. This is the original design proposed in the Vaswani et al., 2017 paper for more efficient model training.

[25] Refer to the source code of the large BERT encoder model at https://github.com/codertimo/BERT-pytorch.

[26] The BERTVIZ library published by Vig, 2019 is used to produce the graph.

word "police" will get high scores in the summation of the element-wise productions between their key vectors and q_{police}. As shown in the graph, the words "arrested," "peaceful," and "protest" are recognized to have closet relevance with "police" by BERT. Figure 3 in the paper shows the sample self-attention matrix extracted from the sample sentence by BERT. The values in the matrix show the relevance between the corresponding tokens calculated by the pretrained transformer model.



Layer: 0 Head: 2

Figure A.1: Attention Matrix Calculation Visualization

A.2 Pipeline Specification & Modularity Detection

A.2.1 Text-to-Network Pipeline

In converting reports into text networks, there are four steps in general. In step one, I extract all the attention matrices output by the large BERT encoder based on the input report. For each input document, the full text is broken into sentences. The different sentences are fed into the BERT encoder for attention matrcies extraction. Especially for simplifying the final text network, I delete all placeholders, stopwords, numbers, and dates, as well as the corresponding weights, when extracting the attention matrices. The sentence-level attention matrices for later steps only contain the words after the deletion.

In the second step, I convert the attention matrices for each sentence into a single matrix. As discussed in the article, the BERT encoder will output 16 attention matrices at each of its 24 transformer layers. I extract the 16 attention matrices from the final layer, which should provide us the best representation of the word connections according to conventional wisdom, and take an average on each cell across the matrices. The final single matrix is taken as the adjacent matrix for the input sentence for text network construction.

In step three, I construct the sentence-level text network according to the adjacent matrix obtained from the last step. In step four, I combine the sentence-level network together to get the final text network. To get the text-level network, I merge the same nodes and combine the edge weights connecting the same pairs of nodes. In the final text network, I delete all self-loops so that the edges can only connect different words. Through the four steps above, one country report is converted to a text network for further analysis.

A.2.2 Modularity Score Detection

For using modularity score as the proxy indicator for the framing of police violence accusations in the USSD reports, I first extract the subgraph in each of the report text networks centering on the keywords "police" and "government" (police-government network) Specifically, I extracted all nodes connected to the two keywords as well as the edges between any of the two selected nodes. Next, I apply the Louvain community detection (LDA) algorithm on the subgraph to detect the sub-module structure. I specify the two keyword nodes, "government" and "police," as the starting point (seeds) for the community structure search. This specification allows the community detection result to better present the frame around the keywords and avoid the potential confounding effect of other well-connected words, like the word "protest" in the second and third graphs of Figure 1.6 in the article.[27]

A.3 Empirical Analysis

A.3.1 Data Discussion

There are 2106 reports in the Human Rights Text corpus collected by (Fariss et al., 2015). I ignored the four reports from 1995 to 1998 on Tibet for the lack of country-year-level predictors. In the rest of the 2102 reports, there are 23 reports in which one of the keywords, "police" or "government," is not mentioned, as shown in Table A.1. After including all variables of interest, the situation of missing values is shown in Figure A.2. As shown, there are only 9.7% of the observations having missing values, and most missing values are concentrated on the variables from the Ill-Treatment and Torture (ITT) country-year dataset (Conrad et al., 2013). For the small proportion of incomplete observations, I argue that the missing values won't bias the statistical inference in this research.

[27] Without setting the seed words, the community detection algorithm might cluster "police" and "government" into the same subgroup.

Table A.1: Country Reports Missing One of the Keywords

	Country	Code	Year	Modularity
1681	North Korea	PRK	2004	No Report on Police
1683	North Korea	PRK	2002	No Report on Police
1217	North Korea	PRK	2001	No Report on Police
695	North Korea	PRK	2000	No Report on Police
1886	North Korea	PRK	1999	No Report on Police
11010	North Korea	PRK	1998	No Report on Police
1468	North Korea	PRK	1997	No Report on Police
1548	Eritrea	ERI	1997	No Report on Police
1858	Palau	PLW	1997	No Report on Police
389	Tuvalu	TUV	1996	No Report on Police
809	North Korea	PRK	1996	No Report on Police
1179	Somalia	SOM	1996	No Report on Police
1569	Palau	PLW	1996	No Report on Police
1840	Myanmar	MMR	1995	No Report on Police
3610	Afghanistan	AFG	1995	No Report on Police
5114	Lebanon	LBN	1995	No Report on Police
7610	Palau	PLW	1995	No Report on Police
8810	Western Sahara	ESH	1995	No Report on Police
9212	Sao Tome and Principe	STP	1995	No Report on Police
12010	North Korea	PRK	1995	No Report on Police
12510	Taiwan	TWN	1995	No Report on Government
13010	Kyrgyz Republic	KGZ	1995	No Report on Police
17212	Iran	IRN	1995	No Report on Police

A.3.2 Model Diagnostics

Figures A.3 and A.4 below show the "fitted vs. residual" plots of the two models in Table 6.2. As shown in the plots, no significant heteroskedasticity issue is detected.

Another potential issue in the regression models is the multi-collinearity. As shown in Table A.2, for the original models reported in Table 1.3 of the article, none of the predictors have a VIF score higher than 5, and thus, no significant multi-collinearity issue is identified.

A.3.3 Matching Results

In the cases comparison section, I focus on the 2005 USSD report on Russia and try to identify a suitable report on US allies to compare with it. The main

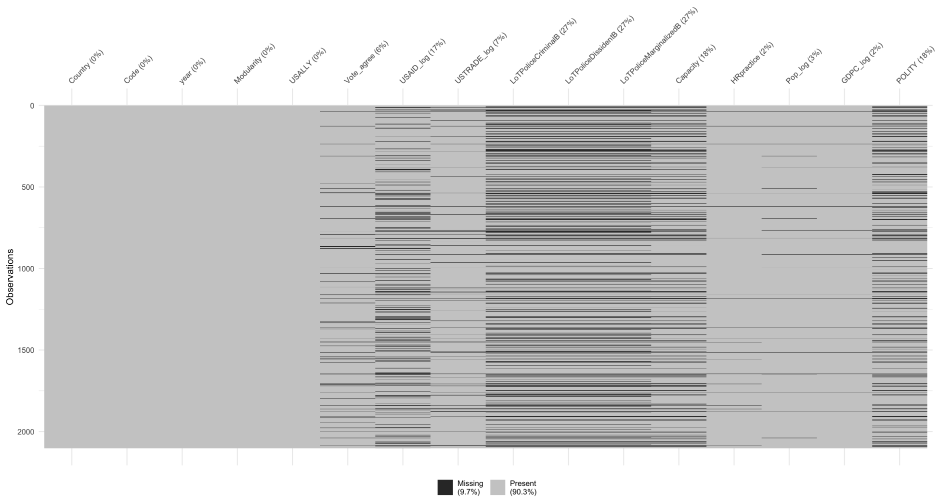


Figure A.2: Missing Values in Final Dataset

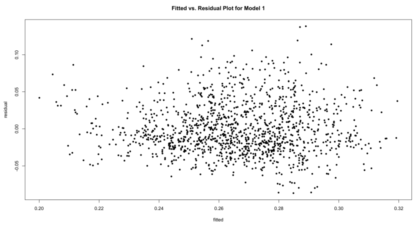


Figure A.3: Fitted vs. Residual Plot for Model 1

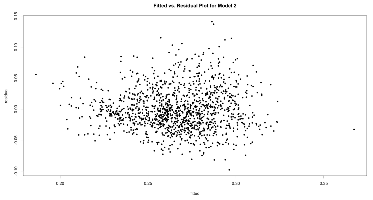


Figure A.4: Fitted vs. Residual Plot for Model 2

idea is to compare the 2005 Russia report with another country report whose target has similar features to Russia but is allied with the US so that the variations on the reports are more likely caused by political bias. For this purpose, I first extracted all the observations on US allies in 2005. Second, I focus on all the control variables (besides "US Ally" and "UN Voting Agreement") and calculate the Minkowski distance between each US ally observation and Russia in 2005. Those countries having smaller Minkowski distances share more common features with Russia. Table A.3 shows the ten countries with the smallest distance on the control variables with Russia in 2005. For potential substantive interests, I choose to compare the 2005 Brazil Report with the 2005 Russia Report in the article.

Table A.2: VIF for Regression Models

	Model 1	Model 2
US Ally	1.70	
UN Voting Agreement		1.41
Received US Aid (logged)	1.24	1.24
Trade with US (logged)	3.82	3.23
Police Vio. against Criminals	1.27	1.27
Police Vio. against Dissidents	1.18	1.19
Police Vio. against Marginalized Indiv.	1.25	1.25
State Capacity	3.67	3.87
Fariss Human Rights Indicator	2.62	2.65
Population (logged)	2.44	2.41
GDP per capita (logged)	3.77	3.80
POLITY Score	1.92	1.83

A.3.4 Police Violence & Human Rights indicators and Accusation Framing

This section presents the regression results of different police violence and human rights indicators on the modularity score. As shown in Table A.4, the presence of police violence will lead to a lower modularity score for the "police-government" network, which indicates a government-level framing of police abuses and harsher criticisms against the government. Worse human rights practice is also associated with lower modularity scores and thus harsher framing of police abuse accusations in the USSD reports. All the associations are statistically significant in intuitive directions. These significant associations ebb in the regression models shown in Table 1.3 of the article. This means that controlling the impact of confounders, the police violence accusations detected by the proposed network approach are shaped by other more influential variables, which are the political relation variables in this research.

Table A.3: Top Ten Countries with Closest Minkowski Distance to Russia in 2005

	Country	Modularity	Minkowski Distance to Russia
1	Colombia	0.23	1.99
2	Philippines	0.26	3.41
3	Türkiye	0.23	3.74
4	Ecuador	0.30	4.22
5	Brazil	0.24	4.34
6	Honduras	0.25	4.44
7	Mexico	0.26	4.53
8	Peru	0.26	4.71
9	Guatemala	0.25	4.86
10	Dominican Republic	0.25	5.28

Table A.4: Police Violence & Human Rights indicators vs. Accusation Framing

	Dependent variable:				
	Modularity				
	(1)	(2)	(3)	(4)	(5)
Police Vio. against Criminals	-0.011^{***}				-0.006^{**}
	(0.003)				(0.003)
Police Vio. against Dissidents		-0.014^{***}			-0.006
		(0.004)			(0.004)
Police Vio. against Marginalized Indiv.			-0.008^{***}		-0.003
			(0.003)		(0.003)
Fariss Human Rights Indicator				0.007^{***}	0.007^{***}
				(0.001)	(0.001)
Constant	0.269^{***}	0.268^{***}	0.268^{***}	0.266^{***}	0.268^{***}
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Observations	1,511	1,511	1,511	2,034	1,511
R^2	0.012	0.010	0.004	0.074	0.070
Adjusted R^2	0.011	0.010	0.004	0.074	0.068
Residual Std. Error	0.040 (df = 1509)	0.040 (df = 1509)	0.040 (df = 1509)	0.041 (df = 2032)	0.039 (df = 1506)
F Statistic	18.416^{***} (df = 1; 1509)	15.666^{***} (df = 1; 1509)	6.699^{***} (df = 1; 1509)	163.443^{***} (df = 1; 2032)	28.341^{***} (df = 4; 1506)
Note:					$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Bibliography

Ajjour, Y., Alshomary, M., Wachsmuth, H., & Stein, B. (2019). Modeling frames in argumentation. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2922–2932.

AlKhamissi, B., Li, M., Celikyilmaz, A., Diab, M., & Ghazvininejad, M. (2022). A review on language models as knowledge bases. *arXiv preprint arXiv:2204.06031*.

Ang, D. (2021). The effects of police violence on inner-city students. *The Quarterly Journal of Economics*, *136*(1), 115–168.

Arnon, D., Haschke, P., & Park, B. (2023). The right accounting of wrongs: Examining temporal changes to human rights monitoring and reporting. *British Journal of Political Science*, *53*(1), 163–182.

Bagozzi, B. E., & Berliner, D. (2018). The politics of scrutiny in human rights monitoring: Evidence from structural topic models of us state department human rights reports. *Political Science Research and Methods*, *6*(4), 661–677.

Bail, C. A. (2016). Combining natural language processing and network analysis to examine how advocacy organizations stimulate conversation on social media. *Proceedings of the National Academy of Sciences*, *113*(42), 11823–11828.

Bailey, M. A., Strezhnev, A., & Voeten, E. (2017). Estimating dynamic state preferences from united nations voting data. *Journal of Conflict Resolution*, *61*(2), 430–456.

Baumer, E., Elovic, E., Qin, Y., Polletta, F., & Gay, G. (2015). Testing and comparing computational approaches for identifying the language of framing in political news. *Proceedings of the 2015 conference of the North American chapter of the Association for Computational Linguistics: human language technologies*, 1472–1482.

Baumgartner, F. R., De Boef, S. L., & Boydstun, A. E. (2008). *The decline of the death penalty and the discovery of innocence*. Cambridge University Press.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, *2008*(10), P10008.

Brysk, A. (1994). The politics of measurement: The contested count of the disappeared in argentina. *Hum. Rts. Q.*, *16*, 676.

Campbell, F., & Valera, P. (2020). "the only thing new is the cameras": A study of us college students' perceptions of police violence on social media. *Journal of Black Studies*, *51*(7), 654–670.

Card, D., Boydstun, A., Gross, J. H., Resnik, P., & Smith, N. A. (2015). The media frames corpus: Annotations of frames across issues. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 438–444.

Carley, K. M. (2020). Network text analysis: The network position of concepts. In *Text analysis for the social sciences* (pp. 79–100). Routledge.

Chong, D., & Druckman, J. N. (2007). Framing theory. *Annu. Rev. Polit. Sci.*, *10*, 103–126.

Cingranelli, D. L., & Richards, D. L. (2010). The cingranelli and richards (ciri) human rights data project. *Hum. Rts. Q.*, *32*, 401.

Cingranelli, D. L., Richards, D. L., & Clay, K. C. (2021). *The CIRI Human Rights Dataset, Version 2014.04.14*. https://doi.org/10.7910/DVN/UKCPXT

Clark, A. M., & Sikkink, K. (2013). Information effects and human rights data: Is the good news about increased human rights information bad news for human rights measures? *Human Rights Quarterly*, 539–568.

Conrad, C. R., Haglund, J., & Moore, W. H. (2013). Disaggregating torture allegations: Introducing the ill-treatment and torture (itt) country-year data. *International Studies Perspectives*, *14*(2), 199–220.

Conrad, C. R., Haglund, J., & Moore, W. H. (2014). Torture allegations as events data: Introducing the ill-treatment and torture (itt) specific allegation data. *Journal of Peace Research*, *51*(3), 429–438.

Cordell, R., Clay, K. C., Fariss, C. J., Wood, R. M., & Wright, T. M. (2022). Disaggregating repression: Identifying physical integrity rights allegations in human rights reports. *International Studies Quarterly*, *66*(2), sqac016.

Cunningham, M. (2020). 'a few bad apples': Phrase describing rotten police officers used to have different meaning [Available at: https://abcnews.go.com/US/bad-apples-phrase-describing-rotten-police-officers-meaning/story?id=71201096]. *ABC News*.

D'Alessio, D., & Allen, M. (2000). Media bias in presidential elections: A meta-analysis. *Journal of communication*, *50*(4), 133–156.

Dardis, F. E., Baumgartner, F. R., Boydstun, A. E., De Boef, S., & Shen, F. (2008). Media framing of capital punishment and its impact on individuals' cognitive responses. *Mass Communication & Society*, *11*(2), 115–140.

Davies, G., & Finnegan, C. (2020). Us 'vulnerable' to accusations of hypocrisy, as china and iran criticize response to george floyd protests [Available at: https://abcnews.go.com/International/us-vulnerable-accusations-hypocrisy-china-iran-criticize-response/story?id=71021338]. *ABC News*.

De Marchi, S., Gelpi, C., & Grynaviski, J. D. (2004). Untangling neural nets. *American Political Science Review*, *98*(2), 371–378.

Demirel-Pegg, T., & Moskowitz, J. (2009). Us aid allocation: The nexus of human rights, democracy, and development. *Journal of Peace Research*, *46*(2), 181–198.

Demszky, D., Garg, N., Voigt, R., Zou, J., Gentzkow, M., Shapiro, J., & Jurafsky, D. (2019). Analyzing polarization in social media: Method and application to tweets on 21 mass shootings. *arXiv preprint arXiv:1904.01596*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

DeVylder, J., Fedina, L., & Link, B. (2020). Impact of police violence on mental health: A theoretical framework. *American journal of public health*, *110*(11), 1704–1710.

Dietrich, N., & Eck, K. (2020). Known unknowns: Media bias in the reporting of political violence. *International Interactions*, *46*(6), 1043–1060.

Donnelly, J. (1988). Human rights at the united nations 1955–85: The question of bias. *International Studies Quarterly*, *32*(3), 275–303.

Druckman, J. N., & Parkin, M. (2005). The impact of media bias: How editorial slant affects voters. *The Journal of Politics*, *67*(4), 1030–1049.

Dukes, K. N., & Gaither, S. E. (2017). Black racial stereotypes and victim blaming: Implications for media coverage and criminal proceedings in cases of police violence against racial and ethnic minorities. *Journal of Social Issues*, *73*(4), 789–807.

Eck, K., & Fariss, C. J. (2018). Ill treatment and torture in sweden: A critique of cross-case comparisons. *Hum. Rts. Q.*, *40*, 591.

Entman, R. M. (1997). Manufacturing discord: Media in the affirmative action debate. *Harvard international journal of press/politics*, *2*(4), 32–51.

Entman, R. M., et al. (2004). *Projections of power: Framing news, public opinion, and us foreign policy*. University of Chicago Press.

Entman, R. M. (2007). Framing bias: Media in the distribution of power. *Journal of communication*, *57*(1), 163–173.

Entman, R. M. (2010). Media framing biases and political power: Explaining slant in news of campaign 2008. *Journalism*, *11*(4), 389–408.

Entman, R. M., & Rojecki, A. (1993). Freezing out the public: Elite and media framing of the us anti-nuclear movement.

Fariss, C. J. (2014). Respect for human rights has improved over time: Modeling the changing standard of accountability. *American Political Science Review*, *108*(2), 297–318.

Fariss, C. J., Linder, F. J., Jones, Z. M., Crabtree, C. D., Biek, M. A., Ross, A.-S. M., Kaur, T., & Tsai, M. (2015). Human rights texts: Converting human rights primary source documents into data. *PloS one*, *10*(9), e0138935.

Field, A., Kliger, D., Wintner, S., Pan, J., Jurafsky, D., & Tsvetkov, Y. (2018). Framing and agenda-setting in russian news: A computational analysis of intricate political strategies. *arXiv preprint arXiv:1808.09386*.

Foot, R. (2000). *Rights beyond borders: The global community and the struggle over human rights in china*. OUP Oxford.

Fridell, L. A. (2017). Explaining the disparity in results across studies assessing racial disparity in police use of force: A research note. *American journal of criminal justice*, *42*, 502–513.

Gamson, W. A., & Modigliani, A. (1989). Media discourse and public opinion on nuclear power: A constructionist approach. *American journal of sociology*, *95*(1), 1–37.

Gibler, D. M. (2008). *International military alliances, 1648-2008*. CQ Press.

Hafner-Burton, E., & Ron, J. (2013). The latin bias: Regions, the anglo-american media, and human rights. *International Studies Quarterly*, *57*(3), 474–491.

Hanson, J. K., & Sigman, R. (n.d.). The state capacity dataset.

Haschke, P. (2017). The political terror scale (pts) codebook. *University Of North Carolina, Asheville*.

Haschke, P., & Arnon, D. (2020). What bias? changing standards, information effects, and human rights measurement. *Journal of Human Rights*, *19*(1), 33–45.

Hill Jr, D. W., Moore, W. H., & Mukherjee, B. (2013). Information politics versus organizational incentives: When are amnesty international's "naming and shaming" reports biased? *International Studies Quarterly*, *57*(2), 219–232.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political analysis*, *15*(3), 199–236.

Iyengar, S. (1990). Framing responsibility for political issues: The case of poverty. *Political behavior*, *12*, 19–40.

Iyyer, M., Enns, P., Boyd-Graber, J., & Resnik, P. (2014). Political ideology detection using recursive neural networks. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1113–1122.

Jackson, J. L., Hall, S. L., & Hill Jr, D. W. (2018). Democracy and police violence. *Research & Politics*, *5*(1), 2053168018759126.

Jain, S., & Wallace, B. C. (2019). Attention is not explanation. *arXiv preprint arXiv:1902.10186*.

Kampf, R., Manor, I., & Segev, E. (2015). Digital diplomacy 2.0? a cross-national comparison of public engagement in facebook and twitter. *The Hague Journal of Diplomacy*, *10*(4), 331–362.

Kent, A. (2001). States monitoring states: The united states, australia, and china's human rights, 1990-2001. *Hum. Rts. Q.*, *23*, 583.

Kwak, H., An, J., Jing, E., & Ahn, Y.-Y. (2021). Frameaxis: Characterizing microframe bias and intensity with word embedding. *PeerJ Computer Science*, *7*, e644.

Lynch, C., & Gramer, R. (2020). With scenes of police brutality, america's beacon to the world winks out [Available at: https://foreignpolicy.com/2020/06/01/protests-trump-soft-power-wanes-racial-injustice-police-violence-george-floyd-world-reaction-police-brutality/]. *FP*.

McCombs, M. (2002). The agenda-setting role of the mass media in the shaping of public opinion. *Mass Media Economics 2002 Conference, London School of Economics: http://sticerd. lse. ac. uk/dps/extra/McCombs. pdf*.

McCombs, M., & Ghanem, S. I. (2001). The convergence of agenda setting and framing. In *Framing public life* (pp. 83–98). Routledge.

Mertus, J. A. (2008). *Bait and switch: Human rights and us foreign policy*. Routledge.

Miller, M. (1998). Police brutality. *Yale L. & Pol'y Rev.*, *17*, 149.

Mitchell, N. (2021). Principals, agents, and passing the buck: How delegation is used by leaders to manage blame. *LSE European Politics and Policy (EUROPP) blog*.

Mitchell, N. J. (2021). *Why delegate?* Oxford.

Mohankumar, A. K., Nema, P., Narasimhan, S., Khapra, M. M., Srinivasan, B. V., & Ravindran, B. (2020). Towards transparent and explainable attention models. *arXiv preprint arXiv:2004.14243*.

Mokhberian, N., Abeliuk, A., Cummings, P., & Lerman, K. (2020). Moral framing and ideological bias of news. *Social Informatics: 12th International Conference, SocInfo 2020, Pisa, Italy, October 6–9, 2020, Proceedings 12*, 206–219.

Moore, W. H. (2010). Incarceration, interrogation, and counterterror: Do (liberal) democratic institutions constrain leviathan? *PS: Political Science & Politics*, *43*(3), 421–424.

Morrow, J. D. (2001). The institutional features of the prisoners of war treaties. *International Organization*, *55*(4), 971–991.

Morrow, J. D. (2007). When do states follow the laws of war? *American Political Science Review*, *101*(3), 559–572.

Nebehay, S. (2020). U.s. criticized for police brutality, racism at u.n. rights review [Available at: https://www.reuters.com/article/us-usa-un-rights/u-s-criticized-for-police-brutality-racism-at-u-n-rights-review-idUSKBN27P2W2]. *Reuters*.

Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the national academy of sciences*, *103*(23), 8577–8582.

Nieman, M. D., & Ring, J. J. (2015). The construction of human rights: Accounting for systematic bias in common human rights measures. *European Political Science*, *14*, 473–495.

Nugent, C., & Perrigo, B. (2020). 'the edge of an abyss.' how the world's newspapers are responding as the u.s. descends into chaos [Available at: https://time.com/5846698/world-reactions-george-floyd-protests/]. *Time*.

Park, B., Greene, K., & Colaresi, M. (2020). Human rights are (increasingly) plural: Learning the changing taxonomy of human rights from large-scale text reveals information effects. *American Political Science Review*, *114*(3), 888–910.

Pevehouse, J. C., Nordstrom, T., McManus, R. W., & Jamison, A. S. (2020). Tracking organizations in the world: The correlates of war igo version 3.0 datasets. *Journal of Peace Research*, *57*(3), 492–503.

Poe, S. C. (2019). The decision to repress: An integrative theoretical approach to the research on human rights and repression. *Understanding human rights violations*, 16–38.

Poe, S. C., Carey, S. C., & Vazquez, T. C. (2001). How are these pictures different? a quantitative comparison of the us state department and amnesty international human rights reports, 1976-1995. *Hum. Rts. Q.*, *23*, 650.

Poe, S. C., & Tate, C. N. (1994). Repression of human rights to personal integrity in the 1980s: A global analysis. *American political science review*, *88*(4), 853–872.

Porter, E. V., Wood, T., & Cohen, C. (2020). The public's dilemma: Race and political evaluations of police killings. In *The politics of protest* (pp. 127–154). Routledge.

Potter, G. (2013). The history of policing in the united states. *EKU School of Justice Studies*, *1*, 16.

Price, V., Nir, L., & Cappella, J. N. (2005). Framing public discussion of gay civil unions. *Public opinion quarterly*, *69*(2), 179–212.

Recasens, M., Danescu-Niculescu-Mizil, C., & Jurafsky, D. (2013). Linguistic models for analyzing and detecting biased language. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1650–1659.

Rejali, D. (2009). Torture and democracy. In *Torture and democracy*. Princeton University Press.

Ritchie, A. J. (2017). *Invisible no more: Police violence against black women and women of color*. Beacon press.

Roberts, C. W. (2020). *Text analysis for the social sciences: Methods for drawing statistical inferences from texts and transcripts*. Routledge.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., & Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American journal of political science*, *58*(4), 1064–1082.

Ron, J., Ramos, H., & Rodgers, K. (2005). Transnational information politics: Ngo human rights reporting, 1986–2000. *International Studies Quarterly*, *49*(3), 557–587.

Rugg, D. (1941). Experiments in wording questions: Ii. *Public opinion quarterly*, *5*(1), 91.

Safavi, T., & Koutra, D. (2021). Relational world knowledge representation in contextual language models: A review. *arXiv preprint arXiv:2104.05837*.

Saunders, B. (2010). Polity5, political regime characteristics and transitions, 1800-2018 dataset users' manual. *Ethics*, *121*, 148–177.

Scheufele, D. A. (2000). Agenda-setting, priming, and framing revisited: Another look at cognitive effects of political communication. *Mass communication & society*, *3*(2-3), 297–316.

Schuldt, J. P., Konrath, S. H., & Schwarz, N. (2011). "global warming" or "climate change"? whether the planet is warming depends on question wording. *Public opinion quarterly*, *75*(1), 115–124.

Segev, E., & Boudana, S. (2019). When fake news makes the news: Definitions of fake news in american newspapers. *International Conference of the European Forum on Fake News and Disinformation, May 15*.

Segev, E. (2020). Textual network analysis: Detecting prevailing themes and biases in international news and social media. *Sociology Compass*, *14*(4), e12779.

Terman, R., & Byun, J. (2022). Punishment and politicization in the international human rights regime. *American Political Science Review*, *116*(2), 385–402.

Tsur, O., Calacci, D., & Lazer, D. (2015). A frame of mind: Using statistical models for detection of framing and agenda setting campaigns. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1629–1638.

Urdal, H. (2008). Population, resources, and political violence: A subnational study of india, 1956–2002. *Journal of Conflict Resolution*, *52*(4), 590–617.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Vig, J. (2019). Bertviz: A tool for visualizing multihead self-attention in the bert model. *ICLR workshop: Debugging machine learning models*.

Vig, J., & Belinkov, Y. (2019). Analyzing the structure of attention in a transformer language model. *arXiv preprint arXiv:1906.04284*.

Wang, C., Liu, X., & Song, D. (2020). Language models are open knowledge graphs. *arXiv preprint arXiv:2010.11967*.

Wang, T.-Y. (1999). Us foreign aid and un voting: An analysis of important issues. *International Studies Quarterly*, *43*(1), 199–210.

Wiegreffe, S., & Pinter, Y. (2019). Attention is not not explanation. *arXiv preprint arXiv:1908.04626*.

Wig, T., Hegre, H., & Regan, P. M. (2015). Updated data on institutions and elections 1960–2012: Presenting the iaep dataset version 2.0: Research & politics.

Wood, R. M., & Gibney, M. (2010). The political terror scale (pts): A re-introduction and a comparison to ciri. *Hum. Rts. Q.*, *32*, 367.

Wright, R. (2020). Fury at america and its values spreads globally [Available at: https://www.newyorker.com/news/our-columnists/after-the-killing-of-george-floyd-fury-at-america-and-its-values-spreads-globally]. *Time*.

Ying, L., Montgomery, J. M., & Stewart, B. M. (2022). Topics, concepts, and measurement: A crowdsourced procedure for validating topics as measures. *Political Analysis*, *30*(4), 570–589.

Ziems, C., & Yang, D. (2021). To protect and to serve? analyzing entity-centric framing of police violence. *arXiv preprint arXiv:2109.05325*.