

Narratives and Networks of Online Hate

Joshua Uyheng

CMU-S3D-24-102

May 2024

Software and Societal Systems Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:

Dr. Kathleen M. Carley (Chair)

Dr. Patrick Park

Dr. Daniel M. Oppenheimer

Dr. James Hawdon (Virginia Tech)

*Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Societal Computing*

Copyright © 2024 **Joshua Uyheng**

This work was supported in part by the Knight Foundation and the Office of Naval Research grants N000141812106 and N000141812108. Additional support was provided by the Center for Computational Analysis of Social and Organizational Systems (CASOS) and the Center for Informed Democracy and Social Cybersecurity (IDeaS). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Knight Foundation, Office of Naval Research or the U.S. government.

Keywords: online hate, narratives, social network analysis, information operations, disinformation, computational social science

For my family

Abstract

Social media platforms have enabled the vast proliferation of online hate messages and the growth of online hate groups worldwide. Considerable efforts in the computational sciences have thus been dedicated to detecting online hate. However, common approaches are constrained by their binary classification of individual texts. Binary classification schemes collapse the unique social *narratives* of hate used to target various identities. Focusing on individual texts likewise erases the social *networks* through which hate spreads and organizes communities. Such shortcomings spill over into downstream challenges, such as tracking the manipulation of online hate or identifying its causes and effects in the offline world. Without attending to narratives and networks, these assessments are flattened into piecemeal questions of how much hate is present or which individuals are spreading it. Consequently, interventions to mitigate online hate fail to account for its deeply social character. Hence, this thesis asks: How can we characterize the narratives and networks of online hate? How can narratives and networks of online hate help us better understand its influences and impacts?

This thesis integrates social scientific theory and computational methods to address these questions. First, I use social identity theory to reframe online hate from a decontextualized linguistic phenomenon into a product of group conflict in society. Social identities set the stage for viewing online hate in terms of its wider social narratives and networks, and motivate a theory-based model to detect online hate that is faster, more generalizable, and more explainable than existing methods. Second, I introduce computational social science tools for characterizing online hate narratives and networks. I showcase the diverse semantic features of online hate narratives targeting racial, gendered, political, and religious identities. I also capture the structural features of online hate networks in terms of properties such as density, isolation, and hierarchy. Third, I develop novel measures for the manipulation of online hate narratives and networks. These indicators quantify shifts in the content of online hate narratives and the organization of online hate networks due to the activity of bots and trolls. Finally, I connect online hate in social media conversations with offline contexts of societal upheaval. I empirically investigate how variations in offline conditions trigger changes in online hate, which in turn predicts offline violence. Collectively, these contributions enhance our social scientific understanding of the nature, influences, and impacts of online hate, and underscore the importance of theoretically informed computational methods to meaningfully engage it.

Acknowledgments

First, I would like to thank my adviser, Kathleen Carley, for taking a chance on me, and for guiding me through the journey of a lifetime. With her mentorship, my eyes have been opened to a bigger world than I could have imagined; and by her example, she has shown me what it might take to change it for the better.

I would also like to thank my thesis committee: Patrick Park for the energizing conversations that have inspired new frontiers for my work, Danny Oppenheimer for the tough questions that have pushed my thinking to the next level, and Jim Hawdon for reminding me of the profound reasons we do our work as scientists in society.

I have learned and grown so much during my time at Carnegie Mellon University. In no small way, this is due to the people I am happy to call my friends and colleagues. I am especially grateful to Charity Jacobs and Samantha Phillips, for our little team that has been my rock through a pandemic and back; Christine Lepird, for being an amazing teaching partner and a role model of excellence paired with kindness; Daniele Bellutta, for surviving being my classmate and co-author several times over; Evan Williams, for bringing out the ridiculous in me; Lynnette Ng, for the karaoke and the Southeast Asian solidarity; Henry Xu, for the lunches, the walks, and the existential crises; and Janice Blane, who in many big and small ways made our lab feel like a home. I will miss everyone dearly.

Throughout this program, I have owed so much to the superhuman efforts of Sienna Watkins, whose warmth and wisdom have kept me grounded over the years. Thank you as well to Connie Herold for answering all my questions and getting me from the start to the end in one piece. Thank you to Dawn Robertson, Samantha Clark, Mike Kowalchuck, and Jeff Reminga, whose cleverness and support I have been so fortunate to lean on numerous times. Thank you to the CASOS folks that walked this path before me and kept me looking steadily ahead: Aman Tyagi, Tom Magelinski, Iain Cruickshank, Michael Yoder, and Kenny Joseph. Thank you to David Gray Widder, who was my first friend here.

Even before coming to CMU, I have been blessed with mentors who nurtured my talents and my spirit, and made it possible to dream this dream. I truly believe I would not be here without Cristina Montiel, Nico Canoy, Reena Estuar, Rofel Brion, Kamille Tabalan, Jumela Sarmiento, and Mari-Jo Ruiz. Thank you for planting in me the seeds to wonder and to be brave. I can only hope to honor your faith in me.

And of course, where would I be without my own personal Greek chorus? To Emma Guanco, Max Velazco, Matthew Limlengco, Harvey Parafina, Aeron Syliongtagay, Janelle Paris, Marco Bartolome, Jeivi Nicdao, Gil Roxas, Ica Divinagracia, Hadrian Ang, Selina Ablaza, Regine Cabato, Christian Benitez, Carissa Pobre, Nicko Caluya, Bee Leung, Martina Herras, Judee Rallos, Hanna Dy, Glyzel Sua, Emy Tan, Shiph Belonguel, Kath Khoo, Tim Rufino, Gabe Estampador, Rocco de Castro, Karl Estuart, and Angelica Sinay: thank you for witnessing me. You are my world.

Finally, I dedicate this to my beloved family. Thank you, Papa, Mama, Atin, Yampol, and Haneul. Thank you, Tita B, Tita Sue, and all the Uys, Batucans, and Hermosisimas who make up this big circle I am so lucky to call home. Thank you, Koko and Tita Isabel. Thank you, Ama and Dad. You believed in me first. This, and everything, is for you.

Contents

1	Introduction	1
1.1	Thesis Objectives	1
1.2	Literature Review	2
1.2.1	Hate as a Social Emotion	2
1.2.2	Hate and Social Identities	3
1.2.3	Computational Methods for Online Hate Detection	4
1.2.4	A Constructural View of Online Hate	5
1.2.5	Social Cybersecurity	6
1.3	Data and Tools	7
1.3.1	Datasets	7
1.3.2	Hate Speech Mega-Corpus	7
1.3.3	National and Global Covid-19 Discourse	8
1.3.4	External Databases	8
1.4	Tools	8
1.4.1	Netmapper	9
1.4.2	ORA	9
1.4.3	Hate Speech Detector	9
1.4.4	BotHunter	9
1.4.5	TrollHunter	9
1.4.6	Location Detection	10
1.5	Overview of Chapters	10
2	A Theory-Based Model of Online Hate	11
2.1	Introduction	11
2.2	Related Work	12
2.2.1	Computational Modeling for Online Hate Detection	12
2.2.2	A Psycholinguistic View of Online Hate	13
2.2.3	Principles for Social Cybersecurity Methodology	14
2.3	Data and Method	14
2.3.1	A Multilingual and Multiplatform Mega-Corpus	14
2.3.2	Tools and Models	15
2.4	Psycholinguistic Properties of Online Hate	18
2.4.1	Presence of Abuse and Social Identity Terms	18
2.4.2	Targeting of Abuse and Social Identity Terms	20

2.5	Theory-Based Detection of Online Hate	21
2.5.1	Model Assessment	22
2.5.2	Performance	23
2.5.3	Generalizability	24
2.5.4	Efficiency	25
2.5.5	Explainability	26
2.6	Discussion	29
3	Narratives of Online Hate	32
3.1	Introduction	32
3.2	Related Work	33
3.2.1	Social Stories of Hate	33
3.2.2	Characterizing Targeted Online Hate	34
3.2.3	Narratives as Meaning Systems	34
3.3	Data and Method	35
3.3.1	Dataset	35
3.3.2	Analysis	36
3.4	Results	38
3.4.1	Topic Clustering	39
3.4.2	Semantic Characterization	40
3.4.3	Validation and Computational Reasoning	43
3.5	Discussion	46
4	Networks of Online Hate	49
4.1	Introduction	49
4.2	Related Work	50
4.2.1	Social Networks and Online Communication	50
4.2.2	Online Harms and Social Structure	50
4.2.3	Multilevel Networks of Online Hate	51
4.3	Data and Methods	52
4.3.1	Dataset	52
4.3.2	Measures	53
4.3.3	Group-Level Analysis	53
4.3.4	Multilevel Analysis	55
4.4	Results	56
4.4.1	Group Targets of Online Hate	57
4.4.2	Structural Features of Online Hate Networks	58
4.4.3	Key Actors in Online Hate Networks	59
4.5	Discussion	61
5	Manipulation of Online Hate	64
5.1	Introduction	64
5.2	Related Work	65
5.2.1	Information Operation Maneuvers	65

5.2.2	Social-Cyber Actors	66
5.2.3	Manipulation of Online Hate	67
5.3	Data and Methods	68
5.3.1	Dataset and Measures	68
5.3.2	Manipulation Assessment	70
5.4	Results	72
5.4.1	General Presence of Bots and Trolls	72
5.4.2	Bot and Troll Narratives	73
5.4.3	Bot and Troll Agenda-Setting	76
5.4.4	Bot and Troll Networks	77
5.5	Discussion	79
6	Offline Contexts of Online Hate	82
6.1	Introduction	82
6.2	Related Work	83
6.2.1	Societal Theories of Hate	83
6.2.2	The Online-Offline Nexus	84
6.2.3	Hate and Violence in the Pandemic	85
6.3	Data and Methods	86
6.3.1	Datasets	86
6.3.2	Analyses	88
6.4	Results	89
6.4.1	General Strain Associations	89
6.4.2	Concurrent Associations	90
6.4.3	Lagged Associations	92
6.4.4	Integrated Intensification Model	92
6.5	Discussion	95
7	Concluding Remarks	98
7.1	Theoretical Contributions	98
7.1.1	Psycholinguistics of Online Hate	98
7.1.2	Semantic Systems of Online Hate	99
7.1.3	Interactions and Communities of Online Hate	100
7.1.4	Social-Cyber Maneuvers on Online Hate	101
7.1.5	Online-Offline Interfaces of Hate	102
7.2	Methodological Advances	103
7.2.1	Tools for Online Hate Detection	103
7.2.2	Tools for Discovery and Characterization	104
7.2.3	Tools for Social-Cyber Assessment	105
7.3	Practical Implications	105
7.3.1	For Platforms	105
7.3.2	For Policymakers and Governance	109
7.3.3	For Educators and Advocates	110
7.4	Final Notes	111

A	Supplementary Documentation	113
A.1	Model Training and Evaluation (Chapter 2)	113
A.2	Problematic Word Features (Chapter 2)	113
A.3	Error Analyses (Chapter 2)	114
A.4	Narrative Models (Chapter 3)	114
A.5	Sample Texts in Mega-Corpus (Chapter 3)	114
A.5.1	Hateful Topics	115
A.5.2	Non-Hateful Topics	116
A.6	Network Models (Chapter 4)	117
A.7	Account Similarity in Clusters (Chapter 4)	117
A.8	Manipulation Models (Chapter 5)	117
A.9	Sample Texts in Covid Dataset (Chapter 5)	118
A.9.1	Selected Topics	118
A.9.2	Unselected Topics	119
A.10	Alternate Analysis (Chapter 6)	121
B	Tool Implementation	123
B.1	Online Harms Report	123
B.2	Online Hate Detector	124
	Bibliography	127

List of Figures

- 1.1 Hate in societal context. Figure adapted from Opotow and McClelland [190]. 3
- 2.1 Visualized coefficient estimates of a multilevel logistic regression model with psycholinguistic measures predicting online hate labels in a multilingual and multiplatform dataset. 20
- 2.2 Visualized coefficient estimates of logistic regression models with psycholinguistic measures predicting online hate labels in a multilingual and multiplatform dataset. 22
- 2.3 Predictions per second made by different online hate detection models. . . 25
- 2.4 Trade-offs between performance and speed among online hate detection models. 26
- 2.5 Distribution of coefficient ranks for abusive and identity terms, versus all other attributes. 27
- 2.6 Overall average predictive scores of abusive and identity terms, versus all other attributes. 27
- 2.7 Key categories among top-ranking features in word-based machine learning models. 28
- 2.8 Performance of word-based machine learning models before and after the ablation of named entities, numbers, and web artifacts from the dataset vocabulary. The dashed line denotes the corresponding performance measure of the theory-based model utilizing Netmapper’s psycholinguistic measures. 29
- 3.1 Summary of proposed methodological pipeline for characterizing narratives of online hate. 36
- 3.2 Plate diagram of the hate-aware supervised topic model. 37
- 3.3 Prompt for large language model in validation step of hate narrative validation methodology. 38
- 3.4 Outputs of supervised topic model on the labeled subset of online hate megacorporus. **Left:** Coefficients on target variable of hate label. **Right:** Saliency of gender, race, and religion labels over topics. 39
- 3.5 Semantic network visualizations of derived online hate narratives. **A:** Degrading misogyny. **B:** White supremacy and xenophobia. **C:** Sexual domination of queer people. **D:** Anti-progressive backlash. 41

3.6	Results of validation with GPT-3. Given a text, the large language model was generally capable of distinguishing which topic it belonged to based on labels assigned to results of the supervised topic model ($N = 100$).	44
4.1	Group-level estimates of associations between hate scores and various identity targets.	57
4.2	Distribution of estimated associations for group targeting and structural features of online hate networks.	58
4.3	Individual-level predictors of online hate in group context. Top: Estimates for the United States. Bottom: Estimates for the Philippines.	60
5.1	Plate diagram of the hate-aware and manipulation-aware supervised topic model.	70
5.2	Outputs of supervised topic model on Covid-19 dataset. Top-Left: Coefficients on target variable of hate label. Top-Right: Coefficients on target variables of bot and troll scores, with linear correlation visualized as a trend-line. Bottom: Saliency of political, gendered, racial, and religious identities over topics.	73
5.3	Visualized account-level coefficients of multilevel model. Error bars represent 95% confidence intervals.	78
6.1	Summary of theoretical models examined in this chapter relating online hate to offline hate. A: General strain theory model. B: Social media as mirror and reinforcer models. C: Intensification model.	86
6.2	Visualization of estimated relationships between online hate and offline hate crime under a strain theory framework. Left: Estimated probability of hate crime occurring given level of Covid-19 infections. Right: Estimated level of online hate given whether or not a hate crime has occurred. Error bars indicate 95% confidence intervals.	89
6.3	Estimated differences in relationship between online hate and offline hate crimes given political affiliation of state governor. Error bars indicate 95% confidence intervals.	90
6.4	Estimated differences in relationship between online hate and offline hate crimes given concurrent levels of Covid-19 infection. Error bars indicate 95% confidence intervals.	91
6.5	Estimated lagged moderation models linking past and future levels of online hate and offline hate crimes. Note: $^+p < .10$, $^*p < .05$, $^{**}p < .01$, $^{***}p < .001$	93
A.1	Log-likelihood curve over varied numbers of topics and minimum term occurrence thresholds for vocabulary retention.	116
A.2	Log-likelihood curve over varied numbers of topics and minimum term occurrence thresholds for vocabulary retention.	118
B.1	Input interface for ORA Online Harms Report.	123
B.2	Multilevel directory of ORA Online Harms Report output.	124

B.3 Multilevel directory of ORA Online Harms Report output.	125
---	-----

List of Tables

- 1.1 Summary of datasets used in this thesis. 8
- 1.2 Summary of tools used in this thesis. 9

- 2.1 Summary of hate speech datasets. 16
- 2.2 Results of multilevel logistic regression over a fifteen-dataset mega-corpus of online hate. 19
- 2.3 Results of multilevel logistic regression analyses over targeted forms of online hate. 21
- 2.4 Performance of online hate detection models. 23
- 2.5 Generalization scores of online hate detection models. 24

- 3.1 Summary of datasets to characterize narratives of online hate. 35
- 3.2 Summary of extracted topics featuring positive online hate scores. 39
- 3.3 Results of chi-square tests assessing the extent to which supervised topic models produce more meaningful clusters than a vanilla topic model. 40
- 3.4 Descriptive statistics for semantic networks of online hate narratives. 42
- 3.5 Results of two-way analysis of variance in importance scores assigned by semantic network method versus standard topic modeling scores. 42
- 3.6 Semantic centrality measures of top-ranking terms. 43
- 3.7 Reasoning of GPT-3 for accurate classifications. 45
- 3.8 Reasoning of GPT-3 for misclassified examples. 46

- 4.1 Multilevel summary of measures used to characterize online hate networks. 54
- 4.2 Summary of group-level predictors of online hate. 59
- 4.3 Summary of estimated individual-level predictors of online hate. 61

- 5.1 Multilevel measures of bot and troll manipulation of online hate networks. 69
- 5.2 Summary of hate-aware and manipulation-aware supervised topics. 74
- 5.3 Semantic centrality measures of top-ranking terms. 75
- 5.4 Summary of estimates for lagged Poisson regression models. 76
- 5.5 Summary of group-level effects linking bots and trolls with the structural features of online hate networks. 77
- 5.6 Summary of account-level coefficients of multilevel model. 79

- 6.1 Summary of integrated analysis of online hate and offline hate crimes under an intensification model. 94

6.2	Summary of findings on relationships between online hate and offline hate crimes during the Covid-19 pandemic.	95
7.1	Summary of platform definitions of online hate.	106
7.2	Summary of platform interventions against online hate.	107
7.3	Examples of hate speech policies around the world.	109
A.1	Summary of hyperparameters for grid search in online hate model selection.	113
A.2	Examples of top problematic features in word-based models.	114
A.3	Examples misclassifications by online hate detection model.	115
A.4	Regression coefficients of strain model with number of hate crime victims.	121
A.5	Regression coefficients of model linking online hate to number of hate crime victims.	122
A.6	Lagged moderation models linking online hate to number of hate crime victims.	122

Chapter 1

Introduction

1.1 Thesis Objectives

Hate is a threat to societies and is on the rise¹. Our hyper-connected digital world has not only left major societal divides unresolved, but also facilitated new ways for them to manifest. From the onset of the Covid-19 pandemic, the world saw a dramatic rise in online racist narratives against Asian people alongside violence against them offline². Yet even beyond the pandemic, social media has long enabled the networked organization of hate groups³ and hate campaigns⁴ in high-stakes contexts of national and international conflict.

Numerous computational methods have been built to detect online hate at scale. But these approaches typically consider online hate in isolation from its social context through binary labels of hate versus non-hate. Such approaches are convenient in that they allow for straightforward assessments of individual texts and accounts as expressing or not expressing hate. In large-scale online conversations, such tools likewise enable broad estimations of how much hate is spreading, which can be a useful signal about the health of both online and offline social systems [52, 53].

However, when purely considered in this fashion, online hate is flattened out of its deeply social character. There is more to hateful messages than simply whether or not they contain hate. On the one hand, the expression of hate in a given social media message takes place through the construction of various systems of meaning which attach to a given target identity and attack it in a socially significant way [200, 229]. Such meaning systems draw upon extant socio-cultural resources and reflect prevailing social divisions to make hate resonate with those who express it and vilify those to whom it is directed. Reducing online hate to binary labels takes away from these more nuanced characterizations of how it might arise from or shape people's hateful beliefs and subsequent hate-inspired behaviors.

¹<https://www.justice.gov/hatecrimes/hate-crime-statistics>

²<https://www.adl.org/blog/at-the-extremes-the-2020-election-and-american-extremism-part-3>

³<https://www.nytimes.com/2019/02/20/us/hate-groups-rise.html>

⁴https://www.washingtonpost.com/world/asia_pacific/facebook-trolls-philippines-death-threats-clone-accounts-duterte-terror-bill/2020/06/08/3114988a-a966-11ea-a43b-be9f6494a87d_story.html

Online hate also spreads not solely through individual messages or individual people. Hate thrives in the context of online communities, who legitimize the expression of hate and make hate a source of belonging [58, 250]. Group settings are also critical as they dictate the way that hateful information flows. Some groups, for instance, may wish to isolate themselves from the mainstream, and such distancing tactics might allow for more extreme views to flourish. Within these groups, leaders may be influential in amplifying the spread of hate or directing how it should be targeted and expressed. Proselytizers also play unique roles in bringing others to join an online hate group. Without such consideration of the group settings of online hate, these critical dynamics are left unexplored.

Finally, especially for the latest decade of innovations in the social media environment, online hate is a prime target for organized manipulation. We have recently witnessed the rise of information operations aiming to influence cyberspace to achieve various domestic and international geopolitical objectives [24, 247, 248]. Such campaigns have utilized advances in computational technologies and the sheer ubiquity of social media platforms to reshape online conversations, spread falsehoods and incendiary messages, and sow discord in modern societies worldwide. In the wake of such developments, new demands on computational social science has given rise to multidisciplinary efforts collectively known as social cybersecurity [52, 53]. Online hate should likewise be examined through this lens to understand how its narratives and networks may be subjected to manipulation, with impacts that go beyond online harm and potentially spark offline violence.

In view of the rich social dimensions of online hate, this thesis argues that online hate needs to be understood in terms of the *narratives* by which it attacks its targets, and the *networks* within which it spreads. In this thesis, I propose methods that capture these social dimensions of online hate, assess their manipulation, and measure their offline influences and impacts. Through this work, I contribute to interdisciplinary theory around online hate, new tools for its characterization, and a range of empirical insights in a multi-lingual and multi-platform mega-corpus of online hate, as well as national and international contexts of the global Covid-19 pandemic.

1.2 Literature Review

1.2.1 Hate as a Social Emotion

To situate online hate in social scientific theory, I first turn to classical scholarship in psychology. Sternberg offers one of the foundational conceptualization of hate as an emotion featuring three components: (a) the negation of intimacy (disgust), (b) passion (anger or fear), and (c) commitment (devaluation of targets) [229]. This sets hate apart from the universal, basic emotions as defined by scholars such as Ekman [76] and Plutchik [199]. In contrast, hate is complex and learned within specific cultural, political, and economic arrangements [7, 230].

Hate is thus usefully thought of as a social emotion that arises within and between groups in society [216]. Although hate may indeed be felt on an interpersonal level, it acquires distinct features from other negative emotions and unique societal significance

when considered on an intergroup level [155]. Which social groups engage in hate, which groups are targeted, and how such hate is experienced or expressed depend on how society organizes these groups [82].

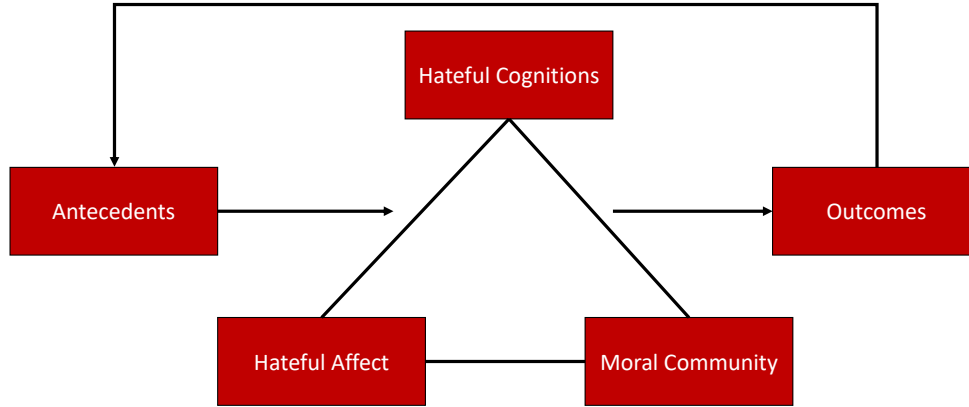


Figure 1.1: Hate in societal context. Figure adapted from Opatow and McClelland [190].

Opatow and McClelland [190] integrate considerable social scientific scholarship to produce a process model of hate that informs my approach in this thesis (see Figure 1.1). Central to the model is its recognition of hate’s cognitive, affective, and group components. Cognitive components of hate include negative or harmful beliefs about the targets of hate such as the invocation of stereotypes [83, 234]. Affective components refer to the basic constituent emotions of fear, anger, and disgust as may be experienced physiologically in response to perceived threats [101, 120]. Finally, the group component of hate here refers to the ‘moral communities’ through which hate: marks out-groups as immoral and deserving of derogation, and designates the in-group as moral and therefore justified for engaging in hate [189, 210]

Beyond hate as a phenomenon in itself, this model valuably links hate to antecedents such as societal strain and anomie, whereby group conflict is produced or heightened due to upheavals in social structure and the frustration of group needs [4, 164]. Meanwhile, outcomes encompass various behavioral manifestations of hate in the form of prejudice, discrimination, and violence against the targets of hate [60, 115, 133]. Notably, the model links outcomes back to antecedents. This implies that increased hateful behaviors can further fray intergroup relations, amplify public perceptions of conflict, and thus legitimize subsequent expressions of hate [26, 186].

1.2.2 Hate and Social Identities

Given these rich social dynamics, social identity theory emerges as a powerful lens to understand hate [2, 236]. Social identity theory explains social psychological processes using *social identities*, defined as “that part of the individuals’ self-concept which derives

from their knowledge of their membership of a social group (or groups) together with the value and emotional significance of that membership” [235, p.255].

Even without intergroup conflict, social identities drive individuals to more positively view their in-groups over out-groups [171]. When out-groups do emerge (or are perceived) as threats, group members strive to positively affirm their own groups and negatively antagonize the other [77, 225]. Affect control theory, a social identity theory of emotion, explains how socio-cultural understandings of the in-group’s and out-group’s social identities are used to generate charged meanings that reinforce the existing social hierarchy [22, 111]. Crucially, then, a social identity lens highlights both the destructive functions of hate like attacking and devaluing its targets, and its constructive functions as in strengthening and bringing groups together under a common cause [36, 210].

Remarkably, social identities already feature in regulatory definitions of hate online. For instance, Twitter defines *hateful conduct* as acts that “promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease”⁵. Similarly, Facebook/Meta defines *hate speech* in terms of “a direct attack against people—rather than concepts or institutions—on the basis of what we call protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease”⁶. By recognizing these categories as social identities, this thesis proposes a deeper understanding of the social character of online hate in terms of why it arises, what makes it harmful, and how it persists.

1.2.3 Computational Methods for Online Hate Detection

Given its severity as a social problem, significant efforts in the computational sciences have been dedicated to detecting online hate in textual form, operationalizing it as *hate speech* [86]. A primary step in this regard has been the collection of labeled corpora distinguishing between examples of hate speech and non-hate speech [201]. In their basic form, such datasets introduce dichotomous labels of hate versus non-hate [14, 69, 70, 132, 203]. Others annotate more specific hate types, often designated by their targets [59, 78, 102, 122, 262]. Many datasets also include additional labels for problematic texts that fall short of hate, such as “offensive” or “abusive” language [67, 89, 160, 191]; or labels for texts that explicitly fight back against hate, such as “counter” speech [59] or “help” speech [192]. While many datasets are in English or collected from Twitter, a growing number of corpora are also including non-English languages [59, 70, 122, 191] and other social media platforms like Reddit, Gab, and Weibo [14, 69, 132, 203].

Machine learning and natural language processing techniques have leveraged these labeled datasets for automated hate speech detection. Initial approaches used standard logistic regression models, support vector machines, and random forest classifiers to predict each text’s labels either using handcrafted features or the entire corpus vocabulary [67, 217]. More recently, researchers have increasingly favored more complex, deep learning

⁵<https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy>

⁶<https://transparency.fb.com/policies/community-standards/hate-speech/>

models such as long short-term memory networks and convolutional neural networks for greater classification performance [16, 278]. Such models have taken advantage of learned semantic representations of texts using word embeddings [166, 195] and transformer models, with BERT (Bidirectional Encoder Representations from Transformers) as the most well-known example [73].

Despite these advances, major problems have also been identified in the literature. Many hate speech models have been shown to simply feature low or inconsistent performance [86, 88]. More broadly, researchers have uncovered systematic biases across datasets and models, such as tendencies to penalize language used by minorities or amplify non-representative opinions on what constitutes hate [68, 265]. To promote accountability around these issues, some researchers have advocated a return to simpler models that can offer explainable predictions [150, 251] or new methods for post-hoc interpretation of deep learning models which otherwise act as blackboxes [260]. Others urge greater rigor and transparency in the dataset annotation process [261], or more diverse datasets of hate speech across languages and cultures [12].

Beyond these technical concerns, however, a more fundamental limitation of prevailing approaches emerges from their exclusive consideration of hate speech as a category of text. Standard computational pipelines treat texts in isolation in order to assign them binary or discrete labels (e.g., hate or non-hate) [200]. In contrast to the rich social scientific literature on hate, these methods eclipse the unique social narratives of hate used to attack various identities [159, 221]. Moreover, by isolating individual texts from their contexts of utterance, these methods erase the social networks through which hate spreads and organizes communities [146].

1.2.4 A Constructural View of Online Hate

Prevailing computational methods are thus inadequate to address the deeper social dimensions of online hate. But while social identity theory broadly lays the foundation for the conceptual links between intergroup relations and online hate [236, 237], it is constructural theory which explicitly operationalizes how such dynamics play out in terms of narratives and networks [46, 47]. Constructural theory holds that group dynamics are guided by two systems: systems of meaning and systems of interaction. Systems of meaning—also known as semantic networks, knowledge networks, concept networks—indicate who knows what and how ideas are connected together. Systems of interaction, on the other hand, define who talks to whom.

These two systems dynamically shape each other in society. How information is organized within a given social system shapes who will talk to whom [125, 163]. People tend to interact with those who share their beliefs or who might possess information that they wish to obtain. In turn, interactions between people result in changes in meaning systems. How concepts are connected may change as information is exchanged, and different people may gain access to previously unknown information [93, 94]. Over time, these two systems co-shape each other, producing complex and emergent effects of group stability and social change [47, 139].

Narratives Applied to the context of online hate, narratives capture the organization of the meaning systems which define how a given social identity is attacked. Drawing upon constructural theory, the utility of semantic network methods has been demonstrated in a variety of contexts to extract the ways that meanings are constructed through the interconnection of various concepts [48, 49]. By characterizing the ways that concepts are organized in semantic networks, the most central ideas can be identified in a given narrative.

The overall characteristics of semantic networks may likewise point to significant processes in meaning-making [51]. For instance, increases in the size of a semantic network indicate the expansion of a given narrative to include other concepts [173]. The fragmentation of a semantic network may similarly indicate the presence of peripheral variants in meaning-making whereby the same narrative might have different modes of expression in different contexts [204]. By viewing online hate through a constructural perspective, this thesis seeks to similarly characterize key systems of meaning involved in attacking a given social identity, beyond dichotomous labels of hate and non-hate.

Networks Likewise, social networks capture how systems of people interact with each other in spreading online hate. Social network analysis has been applied to a variety of online harms to showcase the importance of group structure. Well-established in the literature are various structural forms of online conflict that may manifest empirically through specific network metrics. For instance, the development of echo chambers can be captured through network measures that define the mutual isolation of two different groups [18, 273]. Trench warfare is another type of online conflict whereby cross-group interactions characterized by hostility can result in enhanced polarization [128, 246].

The applicability of a network methodology to online hate groups is particularly promising from a multilevel perspective. From a group-level, it is critical to identify the organizational signatures of online hate communities. Various structural features may be tied to the spread of online hate or increased vulnerability to it [250]. Within such online communities, network measures may also aid in identifying key actors in propagating online hate [247, 249]. Social network representations enable diverse characterizations of different types of influence that a given online actor might have, which can then be tied to their participation in producing online hate messages. This thesis therefore adopts a constructural perspective also to characterize how patterns of social interaction and group dynamics are tied to the proliferation of online hate on social media.

1.2.5 Social Cybersecurity

Finally, this thesis adopts the multidisciplinary perspective of social cybersecurity. Social cybersecurity integrates social and computational research for tackling real-world problems [53]. In contrast to more traditional cybersecurity that primarily protects technologies, social cybersecurity seeks solutions for “people as social beings”, thus requiring “socially informed, social human being led computational social science” [52, p.378]. Methodologically, social cybersecurity consequently values “operational utility” over decontextualized

performance scores like accuracy [53, p.2]. Hence, although this thesis does harness high-performance computational tools throughout its methodology [23, 116, 251, 255], the priority of this thesis lies instead in how these tools function interoperably and in line with social scientific theory to produce socially relevant insights around online hate [252].

From this perspective, online hate not only reflects natural human psychology, but is also artificially manipulated [33]. The last decade has seen an explosion of research on online information operations aiming to shape public opinion and sow discord in society [64, 268]. One prominent dimension of information operations is their use of inauthentic actors on social media, such as automated bots and disruptive trolls [80, 255]. Yet interestingly, while such agents have been well-studied in the spread of disinformation [220, 254], their links to online hate remain under-explored. This thesis examines their inorganic involvement in spreading online hate alongside organic actors [226]. Beyond actors, social cybersecurity further breaks down information operations into a collection of actions which broadly influence narratives and networks [27, 52]. Collectively known as the BEND framework [24], these social-cyber maneuvers may perform positive actions that build networks or enhance narratives, or negative actions that neutralize networks or distort narratives. This approach offers meaningful synergy with this thesis’s identification of narratives and networks of online hate.

Finally, social cybersecurity recognizes the importance of the offline world in shaping the causes and effects of online conflict [52]. Some researchers have linked the dynamics of online activities to offline politics and violence [147, 179]. Conversely, online messages and groups have likewise been linked to offline protests and other forms of political participation [106, 137]. Building on this scholarship, a social cybersecurity view of online hate recognizes that its offline antecedents and outcomes are bidirectionally interlinked [190]. Online-offline links are thus constructed in terms of complex pathways whereby offline conditions trigger online hate, online hate is manipulated by diverse online influences, and online hate in turn shape offline outcomes.

1.3 Data and Tools

1.3.1 Datasets

This thesis leverages several large-scale datasets to understand online hate in national and global settings. Online data is also linked to offline contexts through the use of external databases. All datasets are summarized in Table 1.1.

1.3.2 Hate Speech Mega-Corpus

A curated corpus is developed using fifteen existing labeled datasets of hate speech [14, 59, 67, 69, 70, 78, 89, 102, 122, 132, 160, 172, 191, 203, 262]. Featuring a total of over 300,000 examples of texts labeled as hate, non-hate, and a variety of other problematic (e.g., abusive, offensive) and counter-hate texts, this corpus serves as a comprehensive multi-lingual and multi-platform dataset upon which statistical features of online hate can

Table 1.1: Summary of datasets used in this thesis.

Data	Size	Ch. 2	Ch. 3	Ch. 4	Ch. 5	Ch. 6
Hate Speech Mega-Corpus	300K texts	✓	✓			
National Covid-19 Discourse	3M tweets			✓	✓	
Global Covid-19 Discourse	300M tweets					✓
U.S. Hate Crime Statistics						✓
U.S. Covid-19 Statistics						✓

be robustly analyzed and generalizable models of online hate can be developed. Data was drawn from a repository of hate speech benchmark datasets used in various natural language processing studies [201].

1.3.3 National and Global Covid-19 Discourse

Online conversations about the Covid-19 pandemic are studied on national and global levels to understand online hate in the context of crisis. National datasets are collected for the United States (with the hashtag #COVIDUS) and the Philippines (with the hashtag #COVIDPH), for a total of 3 million tweets over the period of March to May 2020 [247, 249]. A global Covid-19 dataset is also collected using a wider variety of general pandemic-related terms (e.g., #coronavirus, #wuhanvirus, #2019nCoV), totaling about 300 million tweets from 2020-2021 [117].

1.3.4 External Databases

This thesis links online hate to offline hate during the Covid-19 pandemic. Hate crimes in the U.S. are measured by a dataset from the Federal Bureau of Investigation (FBI), including state-level location, time information, and the identity target of the hate crime⁷. This will be matched to U.S. Covid-19 cases taken from the Centers for Disease Control (CDC) including case counts, state-level location, and time information⁸.

1.4 Tools

A series of computational tools are used throughout this thesis to identify and characterize hate as well as its associated narratives and networks (see Table 1.2). These tools are designed and deployed with the principles of social cybersecurity [52]. Hence, while these tools have been documented to show high performance, the focus of this thesis has more to do with their interoperable function in generating theoretically rich and operationally actionable insights [53, 252].

⁷<https://crime-data-explorer.fr.cloud.gov/pages/downloads>

⁸<https://covid.cdc.gov/covid-data-tracker>

Table 1.2: Summary of tools used in this thesis.

Tool	Function	Ch. 2	Ch. 3	Ch. 4	Ch. 5	Ch. 6
Netmapper	Psycholinguistic Analysis	✓	✓	✓	✓	✓
ORA	Dynamic Network Analysis			✓	✓	✓
Hate Speech Detector	Hate Speech Detection	✓	✓	✓	✓	✓
BotHunter	Bot Detection				✓	✓
TrollHunter	Troll Detection				✓	✓
Location Predictor	Location Prediction			✓	✓	✓

1.4.1 Netmapper

Netmapper is used for characterizing the psycholinguistic features of texts [54]. It operates in over 40 languages and produces lexical counts of well-studied psycholinguistic features like pronouns and emotion words [241], social identities [111], and indicators of BEND maneuvers [27]. Features extracted for Netmapper are used for understanding the properties of online hate building various machine learning models used in this thesis [251, 255].

1.4.2 ORA

ORA is an integrated dynamic network analysis software capable of handling large-scale multi-view networks [54]. It is used in this thesis to handle all general-purpose social network operations including centrality calculations, community detection, and visualizations [243, 263]. ORA is used to study how hate groups are organized and potentially manipulated [249, 250]

1.4.3 Hate Speech Detector

A hate speech detector is developed and subsequently utilized in this thesis, trained on the multi-lingual and multi-platform hate speech corpus described above [251]. The model is based on a random forest classifier that uses Netmapper features as predictors and has been evaluated to have 83% accuracy across datasets.

1.4.4 BotHunter

BotHunter is used to generate predictions on whether accounts are showing automated activity as a primary indicator of potential information operations [23]. It is a random forest classifier trained on a variety of known instances of bot-driven influence campaigns on national and international levels, and has been evaluated to have over 90% accuracy.

1.4.5 TrollHunter

TrollHunter is used to generate predictions on whether accounts are engaged in online trolling as a secondary indicator of potential information operations [255]. It is a random

forest classifier trained on a labeled dataset of trolling and has been shown to have 89% accuracy.

1.4.6 Location Detection

A location prediction algorithm is used to detect the location of accounts in online conversations [116]. The model is based on a bidirectional long short-term memory network (bi-LSTM) and has been shown to outperform major state-of-the-art baselines, achieving over 90% accuracy in the Covid-19 dataset [117].

1.5 Overview of Chapters

This thesis is primarily intended to be read linearly. Chapter 2 describes the primary online hate detection model that is used across the remainder of the chapters. It also theoretically describes the broad approach taken throughout this thesis in adopting a social cyber-security perspective to online hate.

Chapter 3 and 4 both build upon Chapter 2 by using online hate predictions to characterize how online hate is organized around narrative systems of meaning and networked systems of people. These two chapters are developed in parallel with each other, and thus, an interested reader may wish to read one before the other as desired. Meanwhile, Chapter 5 directly builds upon the work of Chapters 3 and 4, and thus should be understood as refinements of these earlier chapters.

Chapter 6 stands as a unique chapter that shifts from a purely online analysis of online hate by incorporating offline contexts and consequences. While it certainly builds upon the work of Chapter 2 and Chapter 5, it may also be read in relative independence for those more interested in how the online dynamics of hate interact with offline society.

Chapter 2

A Theory-Based Model of Online Hate

2.1 Introduction

Significant efforts in the computational sciences have been dedicated to detecting online hate in textual form, operationalizing it as *hate speech* [58, 86, 217]. To this end, machine learning and natural language processing techniques have leveraged large labeled datasets for automated hate speech detection [201]. Yet despite significant advances in this field, major problems have also been identified in prevailing approaches, ranging from theoretical to methodological issues [68, 88, 150, 200]. A fundamental limitation of prevailing approaches stems from their atheoretical consideration of hate speech as a category of text. By primarily attending to textual elements without reference to their social or psychological properties [241, 251], the models built upon such assumptions are prone to overfitting to certain datasets or optimizing for incidental phenomena rather than online hate itself. Moreover, interpretation of these models is relegated to interpretative procedures after the fact, rather than developing the model with explicit conceptual coherence from the beginning [52, 53, 252].

Throughout the thesis, a computational model is needed to identify online hate at scale. In line with the socio-theoretical approach taken in this research to understand online hate narratives and networks, methods are needed to bridge gaps between existing tools and social scientific theorizing around hate. I argue that by designing a computational model with theoretical first principles in mind, I can produce a conceptually coherent tool that also addresses the technical problems raised above. This chapter thus proposes a theory-based model of online hate, and showcases its robust methodological properties for social cybersecurity.

In sum, this chapter therefore investigates the following questions:

- How can social scientific theory be used to develop a computational model of online hate?
 - Do theoretically motivated textual measures capture differences between hate and non-hate?

- Do these measures work across datasets, platforms, and languages?
- How do theory-based model scale compare to the speed, accuracy, generalizability, and explainability of common tools?

2.2 Related Work

2.2.1 Computational Modeling for Online Hate Detection

Computational approaches for online hate detection draw primarily from the gamut of supervised machine learning models that have been used for text classification more broadly. Initial approaches used standard logistic regression models, support vector machines, and random forest classifiers to predict each text’s labels either using handcrafted features or the entire corpus vocabulary [67, 217]. Work by Davidson and colleagues [67] is typically positioned as pioneering this area of research, with their initial approach characterized by the use of a multiclass classifier to distinguish hand-labeled tweets as instances of hate speech, offensive speech, and regular speech.

While these early approaches achieved reasonable performance in the nascent stages of the field, they have increasingly fallen out of favor with the emergence of more advanced text classification techniques. In more recent years, researchers have favored more complex, deep learning models such as long short-term memory networks (LSTMs) and convolutional neural networks (CNNs) for greater classification performance [16, 138, 278]. The advantage of such techniques is their ability to learn higher-dimensional relationships between texts and labels, particularly by using models with orders of magnitude more parameters. By training with larger datasets and using more complex learning techniques, these models have typically outperformed the basic tools which introduced the natural language task of online hate detection [20].

Alongside advances in modeling approaches, another key step in online hate detection has involved the design of techniques to represent texts using various types of features. Most basic approaches have utilized bag-of-words approaches, which merely represent texts as unordered sets of tokens without any explicit correspondence with a semantic meaning [277]. With time, however, deep learning models have typically taken advantage of learned semantic representations of texts using word embeddings [166, 195]. In these high-dimensional representations, not only are the locations of specific words assigned a semantic meaning, but also their pairwise distances and mathematical operations are argued to represent meaningful semantic relationships.

Transformer models have also grown increasingly popular, with BERT (Bidirectional Encoder Representations from Transformers) as the most well-known example [73]. In these models, further contextual information is also taken into account. Word embeddings and transformer models have improved over purely word-based models through the data-intensive statistical methods employed to construct them. Such features typically use extremely large textual corpora to map words and sentences to high-dimensional representations where semantic distances between lexical units correspond meaningfully to their numerical distances in the vector space [145, 168].

But despite the vast progress made in this area, these computational techniques have not been immune to criticism. Many hate speech models have been shown to feature low or inconsistent performance across diverse datasets [86, 88]. Others have demonstrated systematic biases across datasets and models, such as tendencies amplify non-representative opinions on what constitutes hate [68, 265]. To promote accountability around these issues, some researchers have advocated a return to simpler models that can offer explainable predictions [150, 251] or new methods for post-hoc interpretation of deep learning models which otherwise act as blackboxes [260]. Others urge greater rigor and transparency in the dataset annotation process [261], or more pluralistic datasets of hate speech across languages and cultures [12].

2.2.2 A Psycholinguistic View of Online Hate

This thesis takes the view that extant methodological concerns in online hate detection are not purely technical concerns. Rather, I posit that in part, they arise from an overarching atheoretical approach to the phenomenon of online hate [200]. Rich social scientific scholarship has grown in relation to understanding the origins and mechanisms of hate in general, as well as its online expressions [190, 229, 230]. But these insights have gone relatively unnoticed in the innovation of a vast majority of computational techniques.

In particular, with respect to detecting the manifestation of online hate in linguistic form, I turn to the field of psycholinguistics. Psycholinguistics occupies an important niche at the intersection of the social, cognitive, and affective branches of psychology, and examines the systematic patterns of speech and communication which reflect underlying mental and emotional states [194, 228, 241]. Over several decades, the field has had significant success in demonstrating empirical signatures in people’s written and spoken interactions reflective of a broad range of psychological constructs, including happiness, sadness, and anger [112, 127, 198, 271]. Moreover, these same linguistic patterns have been meaningfully linked to psychological and behavioral outcomes, such as the success or failure of interpersonal relationships and organizational well-being [119, 161].

Applied to the study of online hate in particular, this chapter emphasizes two critical psycholinguistic measures: (a) abusive terms, and (b) social identity terms. Abusive terms refer to words which aim to attack, derogate, humiliate, or threaten a given target [141]. Social identity terms, on the other hand, are words which designate a person’s membership or identification with a given social group, including but not limited to groups defined by gender, politics, race, and religion [126, 236].

The use of abusive terms has been linked to a variety of negative emotions, such as anger. But in understanding hate as specifically targeted toward social groups, a psycholinguistic perspective would predict the co-presence of abusive terms with social identity terms in particular. Previous studies have proposed similar psycholinguistic features in the development of online hate detection models [131, 214], but these examples have only used them in aggregate without attending to their specific theoretical connections to online hate. Alongside a broader bank of psycholinguistic measures of emotion, cognitive processes, and intergroup communication [54, 241], I therefore highlight these two features in designing a theory-based computational model of online hate.

2.2.3 Principles for Social Cybersecurity Methodology

By grounding the design of a computational model in theoretical first principles, I also position this tool within the context of social cybersecurity deployment [252]. The field of social cybersecurity is an emerging multidisciplinary area of scholarship and practice that aims to preserve and promote a free and open online world [52, 53]. It develops new concepts and tools which are both conceptually coherent and practically responsive to the burgeoning real-world issues which evolve alongside ongoing technological and geopolitical developments.

To this end, social cybersecurity as a field values a distinct set of methodological qualities for tool development than a traditional, purely computational point of view. For instance, in the development of new models to advance the natural language processing literature, measures of accuracy are generally most highly valued. In many cases, however, to achieve marginal increases in what is considered the state of the art, researchers turn to increasingly complex, high-dimensional, black-box models [56, 100]. To a valuable extent, performance measures may be indicative of increased comprehension by the algorithm of the linguistic phenomenon being classified. However, as shown above, empirical research has demonstrated that such approaches are also vulnerable to a lack of generalizability, explainability, and practical deployability due to prohibitive computational costs [10, 238].

For social cybersecurity, these three qualities do not act as secondary criteria, but rather take center stage. With the innovation of new digital threats in tandem with volatile geopolitical tensions, tools are needed which are applicable across a broad range of contexts [88, 214]. Furthermore, the decisions made by a machine learning algorithm in these domains cannot be left to blackbox predictions, especially when subsequent questions of attribution or accountability for online harms are paramount [68, 160]. Finally, due to the high volume and speed of such issues, the methods employed to understand, assess, and respond to them likewise need to be scalable and agile [251, 252]. Obtaining such methodological properties may under various circumstances produce a trade-off with pure accuracy in the traditional sense. But in developing a theory-based model in this chapter within this broader social cybersecurity context of online hate, I posit that a desirable balance of these concerns is both achievable and necessary.

2.3 Data and Method

2.3.1 A Multilingual and Multiplatform Mega-Corpus

To test my psycholinguistic hypotheses and build a theory-based model of online hate, a curated mega-corpus is developed using fifteen existing labeled datasets of hate speech [14, 59, 67, 69, 70, 78, 89, 102, 122, 132, 160, 172, 191, 203, 262]. Featuring a total of over 300,000 examples of labeled texts, this mega-corpus serves as a comprehensive multi-lingual and multi-platform dataset upon which statistical features of online hate can be robustly analyzed and generalizable models of online hate can be developed. Data was drawn from a repository of hate speech benchmark datasets used in various natural language processing studies [201]. These datasets are summarized in Table 2.1.

In general, texts in these datasets are assigned dichotomous or categorical labels of hate versus non-hate [14, 69, 70, 132, 203]. Others identify more specific hate types, often designated by their targets [59, 78, 102, 122, 262]. Many datasets also include additional labels for problematic texts that fall short of hate, such as “offensive” or “abusive” language [67, 89, 160, 191]; or labels for texts that explicitly fight back against hate, such as “counter” speech [59] or “help” speech [192]. While many datasets are in English or collected from Twitter, a growing number of corpora are also including non-English languages [59, 70, 122, 191] and other social media platforms like Reddit, Gab, and Weibo [14, 69, 132, 203].

This dataset will be used to investigate the two research questions set forth in this chapter. First, I will test the presence of theoretically motivated psycholinguistic features in labeled instances of online hate. Then, I will use this aggregated mega-corpus as the training dataset for comparing my theory-based approach with common alternatives in the online hate detection literature.

2.3.2 Tools and Models

To develop and assess the proposed theory-based model, I use a variety of tools of different model types which rely on different categories of textual features. From a feature perspective, my methods include: (a) aggregated psycholinguistic measures, which represent my theory-based approach; (b) the words of the texts themselves, serving as atheoretical lexical tokens; (c) word vectors, which represent a standard tool in natural language processing which represents words in a semantically meaningful high-dimensional vector space; and (d) transformers, which represent the state-of-the-art in contemporary text classification.

Netmapper for Psycholinguistic Measurement Netmapper is used for theory-based characterizations of the psycholinguistic features of texts [54]. It operates in over 40 languages and produces lexical counts of well-studied psycholinguistic features like pronouns and emotion words [241], social identities [111], as well as new indicators of BEND maneuvers [27]. Features extracted for Netmapper are used for understanding the properties of online hate building various machine learning models used in this thesis [251, 255].

Because these measures are predetermined theoretically, they do not expand or contract depending on the incidental size or variety of any particular dataset [269]. This has the particular advantage of remaining stable across various types of data which may be drawn from different sources, languages, and platforms. Netmapper-based models are also less likely to overfit toward specific turns of phrase in a given conversation or event, since its measurements rely on conceptual abstractions as opposed to individual words or sentences. Conversely, however, I note that it is also less responsive to real-time innovations in online hate discourse, unless such innovations utilize conceptual categories already embedded in Netmapper’s psycholinguistic dictionaries. That said, an advantage of this approach is that such new categories may subsequently be incorporated into the Netmapper system and models built on top of it can be retrained accordingly.

Leveraging Netmapper measures, a model may be built by utilizing a simple heuristic to serve as a meaningful baseline for my experiments in this chapter. More specifically, for this heuristic approach, I classify as “hate” any text which contains at least one abusive

Table 2.1: Summary of hate speech datasets.

Dataset	Label	Count	Percentage
Chung [59]	Counter-Hate	6804	88.81
	Hate	857	11.19
De Pelle [70]	Non-Hate	1662	72.8
	Hate	621	27.2
El Sherief [78]	Hate	11325	100.00
Ethos [172]	Non-Hate	565	55.34
	Hate	456	44.66
Founta [89]	Abusive	4540	9.2
	Hateful	1828	3.7
	Normal	35154	71.2
	Spam	7849	15.9
GHC [132]	Non-Hate	24996	90.72
	Hate	2557	9.28
Hatebase [67]	Hate	1430	5.77
	Offensive	19190	77.43
	Non-Hate	4163	16.8
Hatex [160]	Hate	6854	34.02
	Non-Hate	7814	38.78
	Offensive	5480	27.2
MLMA [191]	Hate	3710	19.88
	Non-Hate	3058	16.39
	Offensive	11893	63.73
Multimod [102]	Hate	8577	5.72
	Non-Hate	141246	94.28
Qian [203]	Non-Hate	15087	64.53
	Hate	8292	35.47
RP [14]	Non-Hate	7141	50.01
	Hate	7139	49.99
Stormfront [69]	Hate	1196	11.17
	Non-Hate	9507	88.83
SWSR [122]	Non-Hate	5876	65.51
	Hate	3093	34.49
Waseem [262]	Non-Hate	7751	73.82
	Hate (Racism)	12	0.11
	Hate (Sexism)	2736	26.06

term and one identity term. In this method, I do not undertake any statistical learning based on the dataset and use only a single a priori rule. While I do not expect this technique to produce the most accurate model, I utilize it as a benchmark to demonstrate that these features are nonetheless meaningfully linked to the phenomena in question, while

also showing the value of more advanced approaches taken throughout this thesis which rely on the same features.

The methodology I highlight throughout this thesis uses machine learning models built on top of theory-based Netmapper features. I specifically test logistic regression and random forest models. The key idea in this approach is to combine a priori theory-based psycholinguistic measurements which statistically learns to optimize predictions from labeled data. This takes the advantages of conceptual coherence for enhanced explainability alongside practical considerations of increased predictive performance.

Word-Based Machine Learning One set of comparisons I draw with my proposed approach directly utilizes the words in a text. In contrast to a theory-based model that aggregates psycholinguistic measurements, word-based approaches simply take individual words in a dataset as the features [277]. From this standpoint, models identify which words are more or less predictive of labeled instances of online hate, and then infer these associations for prediction of unlabeled texts.

While the use of words is still explainable due to its direct correspondence with texts in their natural form, their use as a feature base can be unstable across different sources, languages, and platforms. Models which build their statistical associations on a given dataset are prone to overfitting toward the turns of phrase or even named entities which are present in that dataset. Additionally, the vocabulary size in a given corpus—i.e., the unique collection of words in a given dataset—can vary widely. On the one hand, they may tend toward extremely high numbers prior to reasonable filtering procedures, thus expanding the number of features and slowing prediction time. On the other hand, they are typically not robust to shifts in domain, since different words will tend to be used in different social contexts of online hate.

Embedding-Based Deep Learning To compare my proposed theory-based approach with more modern approaches to text classification, I use the pretrained GloVe representations as an exemplar case of the word embedding approach [195]. I map labeled texts to the pretrained GloVe vector space using several levels of dimensionality, ranging from 50-dimension to 300-dimension representations. I then use the PyTorch library¹ to train LSTM and CNN deep learning models with these features, utilizing standard dropout for regularization and a grid search for hyperparameter tuning.

Transformer Models Transformer models are generally considered the state-of-the-art in most natural language processing tasks including online hate detection. Here, I specifically use the Language-Agnostic BERT (LABSE) model to compare my proposed approach with what would generally be the standard method used in contemporary text analytics [79]. I perform training and testing with the HuggingFace library².

¹<https://pytorch.org/>

²<https://huggingface.co/>

Off-the-Shelf Finally, I also use an off-the-shelf model for performance comparison. While computational social science research into online hate is a growing field, not all studies necessarily build their own hate prediction models. Instead, off-the-shelf models can be taken as a standard tool to lower barriers to analysis and proceed directly with understanding other dynamics of online hate beyond merely detecting it. I specifically use the Perspective API³, a popular toxicity detection tool, as my off-the-shelf benchmark in this study [87, 103].

2.4 Psycholinguistic Properties of Online Hate

In this first study, the goal is to apply existing psycholinguistic measurements to a multilingual and multiplatform mega-corpus of labeled instances of online hate and non-hate. From these measurements, I statistically test the extent to which theoretically motivated features of abusive language and social identities are empirically associated with instances of online hate. Furthermore, given that a subset of the mega-corpus also contains labeled instances of particular forms of hate—gendered hate (sexism), racial hate (racism)—I also test the extent to which the presence of social identity terms associated with these categories are also correlated with these targeted hate types.

2.4.1 Presence of Abuse and Social Identity Terms

Using a multilevel logistic regression model, I set Netmapper features as predictors and the binary label of online hate or non-hate as the response variable, with clustered errors around the fifteen datasets for robustness. All predictors were grand mean centered and scaled by standard deviation for interpretability and between-variable coefficient comparisons. Table 2.2 summarizes coefficient estimates for this analysis. Figure 2.1 further visualizes these results to show coefficient estimates in descending order.

Results of this analysis demonstrate that, as motivated by the psycholinguistic literature, instances of online hate are most closely associated with abusive ($b = 0.054, SE = 0.001, p < .001$) and identity terms ($b = 0.045, SE = 0.001, p < .001$). This indicates that the presence of abusive terms and identity terms strongly distinguish between online hate and non-hate, more than other notable psycholinguistic features such as negative-emotion words ($b = 0.012, SE = 0.001, p < .001$). Conceptually, this highlights that while online hate is certainly a form of negative emotion, negativity alone is not sufficient to identify online hate. Identity terms, in fact, are nearly 4 times more strongly associated with online hate in comparison, underscoring the socially embedded nature of online hate as an emotion beyond merely its valence.

Several specific identity types also appeared to be associated strongly with online hate. In particular, gender identities ($b = 0.014, SE = 0.001, p < .001$), religious identities ($b = 0.016, SE = 0.001, p < .001$), and political identities ($b = 0.007, SE = 0.001, p < .001$) tended to have positive relationships with the likelihood that a given text was an example of online hate. Conversely, family-based identities ($b = -0.004, SE = 0.001, p < .001$) and

³<https://www.perspectiveapi.com/>

Table 2.2: Results of multilevel logistic regression over a fifteen-dataset mega-corpus of online hate.

Features	Estimate (All)
Abusives	0.054 (0.001)***
Identities (All)	0.045 (0.001)***
Identities (Gender)	0.014 (0.001)***
Identities (Race)	-0.014 (0.001)***
Identities (Religion)	0.016 (0.001)***
Identities (Politics)	0.007 (0.001)***
Identities (Family)	-0.004 (0.001)***
Identities (Job)	-0.032 (0.001)***
Concept Count	-0.024 (0.002)***
Reading Difficulty	0.007 (0.001)***
Named Entity	-0.004 (0.001)***
Exclusives	-0.001 (0.001)
Inclusives	-0.002 (0.001)**
Absolutists	-0.002 (0.001)***
Equivocals	-0.003 (0.001)***
Connectives	0.001 (0.001)
Positives	-0.009 (0.002)***
Negatives	0.012 (0.001)***
Pronouns (1st)	-0.005 (0.001)***
Pronouns (2nd)	0.001 (0.001)*
Pronouns (3rd)	0.008 (0.001)***
Numbers	-0.003 (0.001)***
All Caps	-0.002 (0.001)***
Question Marks	-0.002 (0.001)**

Note: * $p < .05$, ** $p < .01$, *** $p < .001$

job-based identities ($b = -0.032$, $SE = 0.001$, $p < .001$) had negative relationships with the likelihood that a text was labeled as online hate. These latter two categories may be more personal forms of identity and may thus be less likely to be linked to intergroup conflicts. Conversely, it is also interesting to note that racial identities on their own ($b = -0.014$, $SE = 0.001$, $p < .001$) were negatively associated with online hate.

This may be explained by the fact that several instances of offensive and abusive language—which stopped short of being labeled as online hate—contained racial epithets that were uttered in non-hateful contexts. This aligns with issues raised by Davidson and colleagues [67, 68] regarding the importance of contextualizing racialized terms which in certain cases are used in a targeted, hateful manner, and in other settings are meant as stylized forms of non-hateful in-group talk.

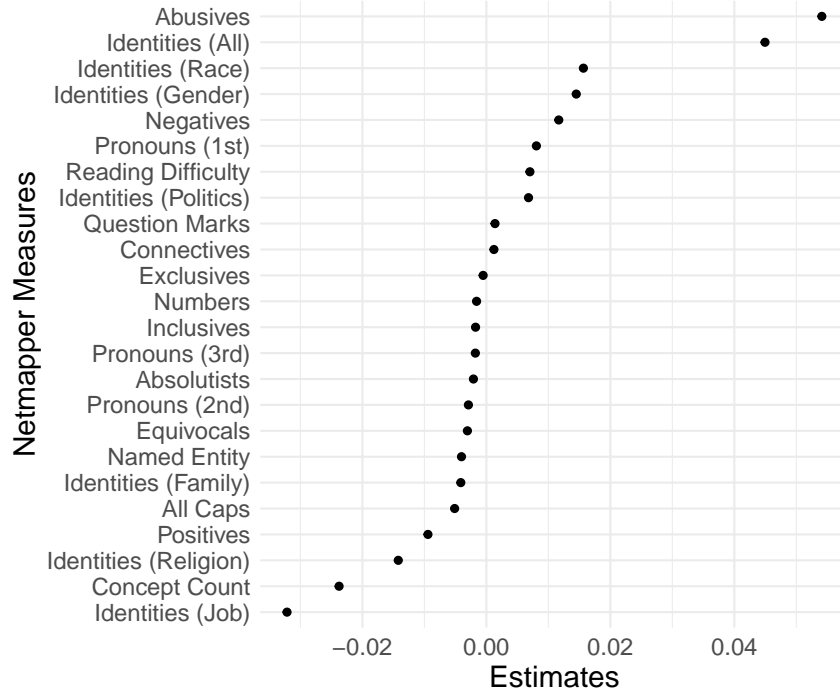


Figure 2.1: Visualized coefficient estimates of a multilevel logistic regression model with psycholinguistic measures predicting online hate labels in a multilingual and multiplatform dataset.

2.4.2 Targeting of Abuse and Social Identity Terms

Building on this analysis, I proceeded to test the presence of specific identity terms in instances of targeted forms of online hate. In the multilingual and multiplatform megacorpora are three specific sub-labels of gendered hate (sexism), racial hate (racism), and religious hate (Islamophobia). Instances of non-hate were drawn from the same datasets where positive cases were identified. Using these sub-datasets, I conducted three separate logistic regression analyses with their corresponding measures of identity terms as a focal predictor. These results are summarized in Table 2.2 and visualized in Figure 2.2.

Strikingly, as in the general case, the targeted identity terms and abusive terms consistently remain among the highest-scoring coefficients. For gender-targeted hate, the relationships are consistent with the general case, although gender identity terms are now the top-scoring coefficient ($b = 0.103, SE = 0.001, p < .001$), followed by abusive terms ($b = 0.038, SE = 0.001, p < .001$). For racial hate, racial identities have the third highest association ($b = 0.004, SE = 0.000, p < .001$) while abusive terms have the fourth highest association ($b = 0.002, SE = 0.000, p < .001$). These effects trail concept counts ($b = 0.006, SE = 0.001, p < .001$) and negative emotion terms ($b = 0.005, SE = 0.000, p < .001$), suggesting that racially targeted hate may involve the incorporation of more ideas, and its sentiments are oriented more negatively. Because the non-hate examples for this case involve other instances of racial talk, these

Table 2.3: Results of multilevel logistic regression analyses over targeted forms of online hate.

Features	Gender	Race	Religion
Abusives	0.038 (0.001) ^{***}	0.002 (0.000) ^{***}	-0.004 (0.000) ^{***}
Identities (Subtarget)	0.103 (0.001) ^{***}	0.004 (0.000) ^{***}	0.013 (0.000) ^{***}
Concept Count	0.007 (0.004)	0.006 (0.001) ^{***}	0.006 (0.001) ^{***}
Reading Difficulty	0.000 (0.001)	0.001 (0.000) ^{***}	0.003 (0.000) ^{***}
Named Entity	-0.009 (0.001) ^{***}	-0.001 (0.000) ^{**}	-0.000 (0.000)
Exclusives	0.003 (0.001) [*]	0.002 (0.000) ^{***}	0.001 (0.000) [*]
Inclusives	-0.006 (0.001) ^{***}	0.001 (0.000) ^{**}	0.001 (0.000) [*]
Absolutists	0.002 (0.001)	-0.000 (0.000)	0.000 (0.000)
Equivocals	-0.004 (0.001) ^{**}	-0.001 (0.000) ^{**}	-0.001 (0.000) [*]
Connectives	0.005 (0.002) [*]	0.001 (0.000) ^{**}	0.002 (0.001) ^{***}
Positives	-0.016 (0.003) ^{***}	-0.003 (0.001) ^{***}	-0.007 (0.001) ^{***}
Negatives	0.001 (0.002)	0.005 (0.000) ^{***}	0.006 (0.001) ^{***}
Pronouns (1st)	-0.009 (0.001) ^{***}	-0.002 (0.000) ^{***}	-0.002 (0.000) ^{***}
Pronouns (2nd)	0.003 (0.001) ^{**}	-0.000 (0.000)	-0.001 (0.000) [*]
Pronouns (3rd)	0.011 (0.002) ^{***}	0.001 (0.000) ^{***}	-0.001 (0.001)
Numbers	-0.002 (0.002)	-0.003 (0.000) ^{***}	0.000 (0.001)
All Caps	-0.009 (0.001) ^{***}	-0.001 (0.000) ^{***}	0.001 (0.000) [*]
Question Marks	-0.007 (0.001) ^{***}	-0.000 (0.000) [*]	-0.000 (0.000)

Note: ^{*} $p < .05$, ^{**} $p < .01$, ^{***} $p < .001$

distinctions are significant as they showcase what other psycholinguistic features distinguish hateful from non-hateful racial talk. Finally, religiously targeted hate had religious identities as the top-scoring coefficient ($b = 0.013, SE = 0.000, p < .001$), and a negative association obtained for abusive terms ($b = -0.004, SE = 0.000, p < .001$). Like racial hate, negative emotion terms ($b = 0.006, SE = 0.001, p < .001$) and concept counts ($b = 0.006, SE = 0.001, p < .001$) also had high associations with hate. The negative association with abusive terms may thus suggest that offensive language with religious topics might involve more such terms, but it is the particularly negative inclusion of religious identity terms alongside more invoked concepts that distinguishes instances of religious hate.

Taken together, these findings point to the nuanced importance of identity terms for online hate both in general and in its more targeted form, and highlight their intertwined role with abusive terms alongside potentially negative emotions and higher conceptual counts to rule out simplistic utterances of profanities without being oriented toward attacking a social group.

2.5 Theory-Based Detection of Online Hate

In this second study, having established the statistical prevalence of theory-based psycholinguistic measures in online hate across datasets, languages, and platforms, I now put

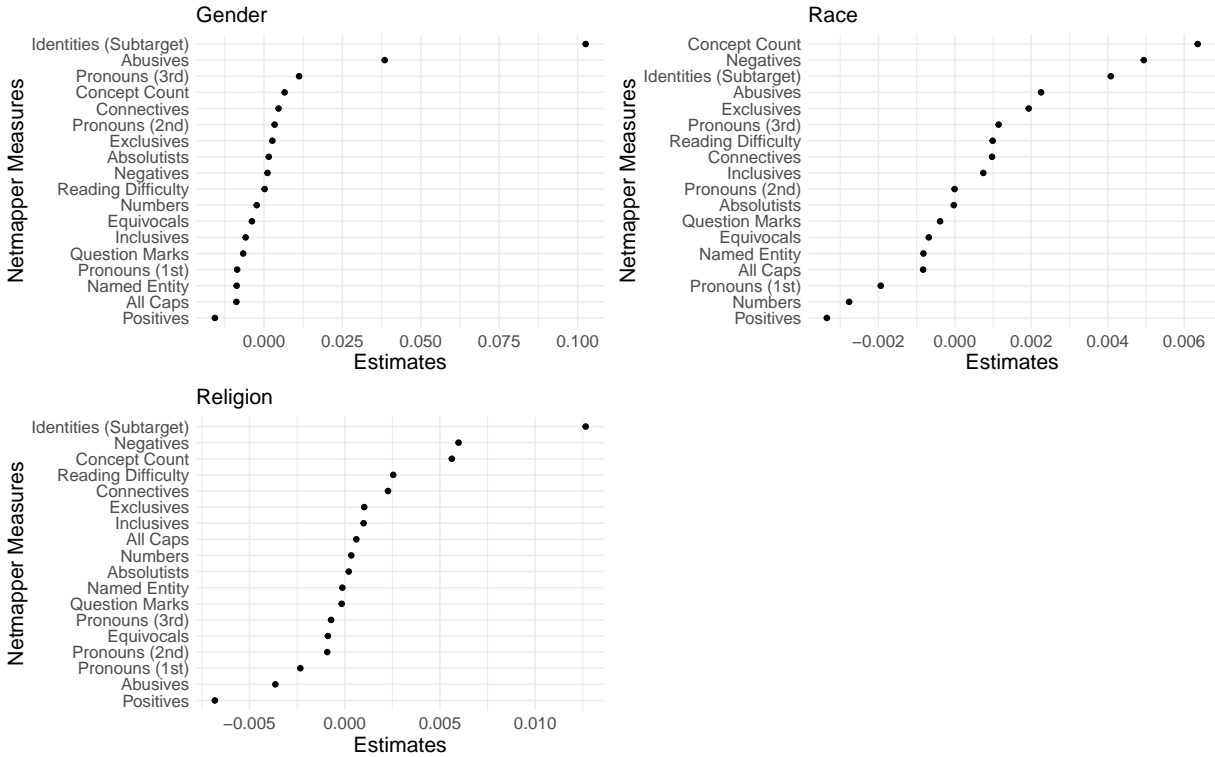


Figure 2.2: Visualized coefficient estimates of logistic regression models with psycholinguistic measures predicting online hate labels in a multilingual and multiplatform dataset.

to the test whether such relationships are practically useful. In particular, I assess the utility of online hate detection models trained using these theoretically motivated features. I perform this evaluation along a holistic collection of relevant dimensions, keeping in mind methodological considerations for practical deployment in a social cybersecurity context.

2.5.1 Model Assessment

To assess the models across feature types, standard measures of predictive performance will be assessed, focusing specifically on accuracy and the F1 score. In prioritizing a holistic analysis of model types, runtime will also be measured. For large-scale applications social cybersecurity, it is desirable that models be fast and scalable [53, 252]. Hence, I compare performance with model speed on the same machine to examine practical tradeoffs between methods.

Given documented inconsistencies in online hate detection across domains [86, 88], experiments will also be run in which the data will be stratified across datasets, languages, and platforms. Across each level, data from a single dataset, language, or platform will be held out as a test set while the remaining data will be used for training and validation. Performance on the test set will then be averaged across strata to determine generalizability across datasets, languages and platforms. Higher performance will indicate stronger generalizability.

Finally, feature importance scores for the theory-based and word-based models will also be derived to understand how these models make online hate predictions. For the proposed theory-based model, particular attention will be paid to the importance assigned to social identity terms. For the word-based model, which words are assigned high importance will also be investigated. More explainable models are expected to assign greater importance to theoretically appropriate features or terms.

2.5.2 Performance

From the perspective of predictive performance, the proposed theory-based approach performs well in comparison to standard alternatives for online hate detection. Table 2.4 summarizes the accuracy and F1 scores of the different model types when trained and tested over the entire mega-corpus.

Table 2.4: Performance of online hate detection models.

Features	Model	Accuracy	Precision	Recall	F1
Netmapper-Heuristics	Abusives and Identities	0.499	0.768	0.486	0.595
Netmapper-Model	Logistic Regression	0.739	0.909	0.738	0.815
	Random Forest	0.810	0.847	0.806	0.826
Words	Logistic Regression	0.827	0.868	0.821	0.844
	Random Forest	0.846	0.855	0.873	0.864
Word Embeddings	CNN	0.895	0.887	0.895	0.887
	LSTM	0.896	0.888	0.896	0.886
Transformers	BERT	0.825	0.681	0.825	0.746
API	Perspective	0.595	0.600	0.595	0.598

The random forest model that uses the Netmapper features performs best between the three types of models which use psycholinguistic features. Reliance on a one-rule heuristic performs relatively poorly, even if its F1 score exceeds 0.50. While the logistic regression model solely makes use of linear associations, it achieves an accuracy of 0.739 and an F1 score of 0.815, while the random forest model is able to make more complex decisions and achieves an accuracy of 0.810 and an F1 score of 0.826. For this latter model in particular, which constitutes the main model used throughout this thesis, this F1 score decomposes into a precision of 0.847 and a recall of 0.806, indicating that it may be more likely to make false negatives (Type 2 errors) than false positives (Type 1 errors).

As expected, when considered over the entire mega-corpus, word-based models, embedding-based models, and transformer models all outperform the psycholinguistic models. Notably, their improvement in accuracy ranges from about 1.7% for word-based logistic regression, to up to 8.6% when using an embedding-based LSTM. Deep learning and transformer models generally do the best, as would be expected given their more advanced stature relative to the state of the art in natural language processing.

Interestingly, however, the Perspective API performs quite poorly in identifying online hate. Its performance is relatively close to the one-rule heuristic involving Netmapper measures. This is indicative that as an off-the-shelf tool, the Perspective API should be specifically used for its intended purpose which is toxicity detection. Although it is a related task, it is not equivalent to online hate detection, and from these empirical findings, are strikingly ill-suited for this purpose.

2.5.3 Generalizability

While performance measures yield findings fairly consistent with expected results based primarily on model complexity, a generalizability analysis suggests a more nuanced assessment. In line with the social cybersecurity perspective taken here, the desired online hate detection model should be able to identify hate across a wide range of conditions, including across languages and platforms. Through data-stratified, language-stratified, and platform-stratified training and testing, I obtained the average accuracy of each model type as well as the standard deviation of their performance across these different categories. These results are summarized in Table 2.5.

Table 2.5: Generalization scores of online hate detection models.

Features	Datasets		Languages		Platforms	
	Mean	SD	Mean	SD	Mean	SD
Netmapper-Heuristics	0.621	0.272	0.702	0.151	0.695	0.158
Netmapper-ML	0.752	0.133	0.733	0.136	0.666	0.094
Words-ML	0.686	0.108	0.649	0.130	0.634	0.066
Word Embeddings-CNN	0.763	0.117	0.726	0.129	0.686	0.140
Word Embeddings-LSTM	0.767	0.122	0.729	0.135	0.694	0.139
Transformers	0.721	0.249	0.745	0.136	0.730	0.128
API-Perspective	0.709	0.196	0.729	0.140	0.723	0.157

Most importantly, it is worth noting that the generalizability of the theory-based psycholinguistic model using Netmapper measures maintains competitive performance with more advanced models. Across datasets, the best generalization is observed among the LSTM embedding-based models at 0.767, but the Netmapper model is only 1.5% behind at 0.752. Across languages, the best generalizability is seen with the language-agnostic BERT model, with the Netmapper model closely trailing by 1.2%. Platform differences were more observable with the BERT model achieving a score of 0.730, while the Netmapper model achieved about 6% less. Word-based and embedding-based models all performed less than 0.70, however, indicating that platform generalizability was best captured by the BERT model due to its unique ability to model contextualized representations. Due to the higher level of variance in message length across platforms, count-based measures such as those of words and psycholinguistic measures are more vulnerable to generalization errors. Nonetheless, given the context of the research throughout my thesis which draws upon

a single platform (Twitter), these weaknesses are less immediately relevant and may be addressed in future work.

2.5.4 Efficiency

Next, I consider the efficiency of the model. This is assessed in terms of the predictions that a model is able to produce over time, and is more holistically considered when speed is examined in a trade-off relationship with overall performance. Using the same machine across models, predictions were generated on the largemega-corpus of labeled texts and speed was assessed in predictions per second. Results are visualized in Figure 2.3.

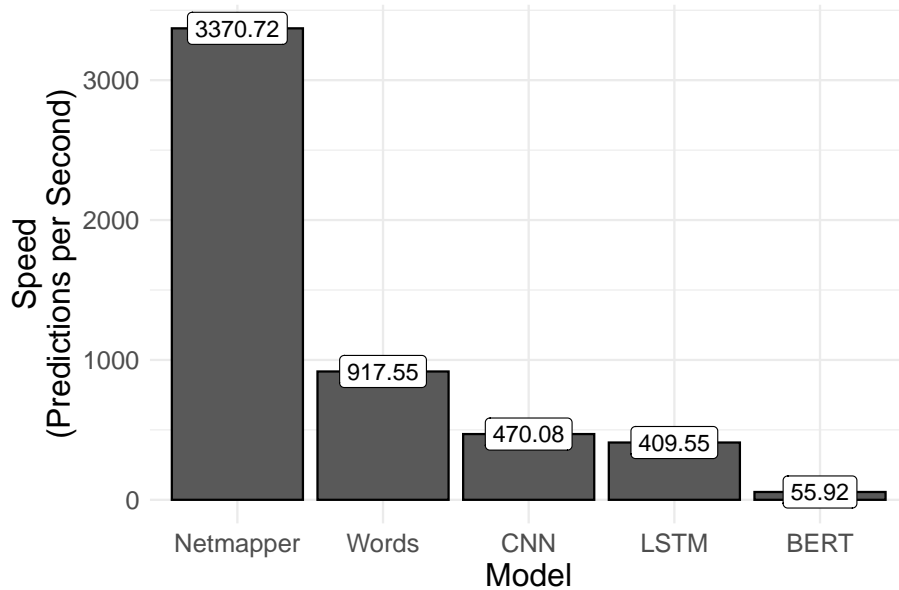


Figure 2.3: Predictions per second made by different online hate detection models.

Notably, Netmapper-based models considerably outspeed all other model types. Generating 3370.72 predictions per second, the theory-based model demonstrates its significantly more lightweight nature relative to word-based, embedding-based, and transformer models, particularly due to its tight collection of dataset-invariant parameters. By contrast, word-based models are over three times slower, generating only 917.55 predictions per second. Deep learning models which use word embeddings, on the other hand, are nearly 8 times slower, with CNNs generating 470.08 predictions per second and LSTMs generating 409.55 predictions per second. Transformer models, while representing the closest to the state of the art, were about 60 times slower at 55.92 predictions per second. The Netmapper model is therefore the fastest by far.

Considered in conjunction with its performance, a joint assessment suggests that in practical contexts of rapid and large-scale analysis, the Netmapper model demonstrates key methodological qualities for social cybersecurity deployment. In Figure 2.4, all models occupy a relatively narrow band of performance values, with Netmapper additionally

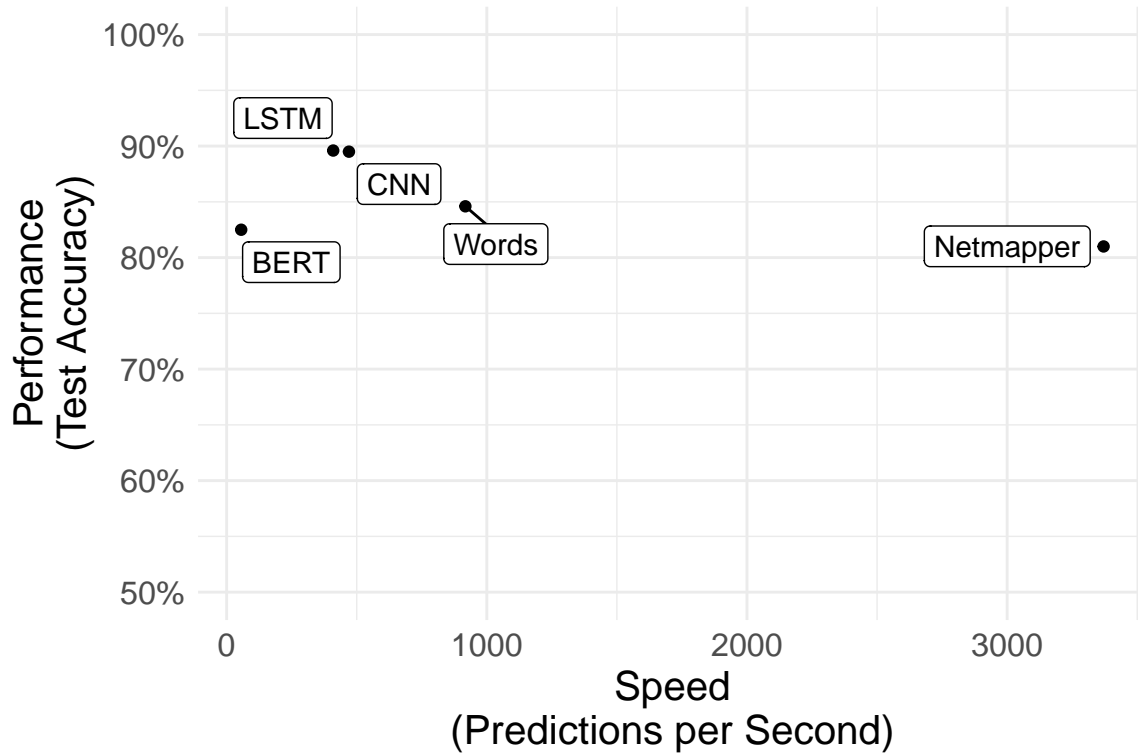


Figure 2.4: Trade-offs between performance and speed among online hate detection models.

showing competitive dataset-agnostic and language-agnostic generalizability. However, in terms of predictions per second, it is in a class of its own, even in comparison to word-based models which share a similar predictive setup with machine learning models, only with a much larger set of parameters. While achieving competitive performance, then, the theory-based model is additionally much more efficient and vastly outperforms other model types from a holistic standpoint.

2.5.5 Explainability

Finally, I consider the explainability of the theory-based model developed in this chapter. I do this in two ways. First, I examine the rank of the theoretically motivated features of abusive terms and identity terms in the highest-performing machine learning model built on Netmapper measures. Because random forest models produce importance measures of each of its features, it is naturally interpretable.

The chief competition of this model in the domain of explainability is the word-based model, since the latter remaining models are relatively blackbox given their high-dimensional and multilayered representations. The second analysis performed thus looks at the highest-ranking features for word-based models and examines whether they conform to theory-based expectations or otherwise assign high weight to construct-valid features.

Netmapper Rank Examining the Netmapper features, it appears that the distribution of feature ranks related to abusive or any class of identity terms considerably skews right, indicating that they tend to occupy the higher ranks in the feature set. Conversely, all other attributes have a much flatter distribution, more uniformly dispersed among high and low ranks. This is visualized in Figure 2.5.

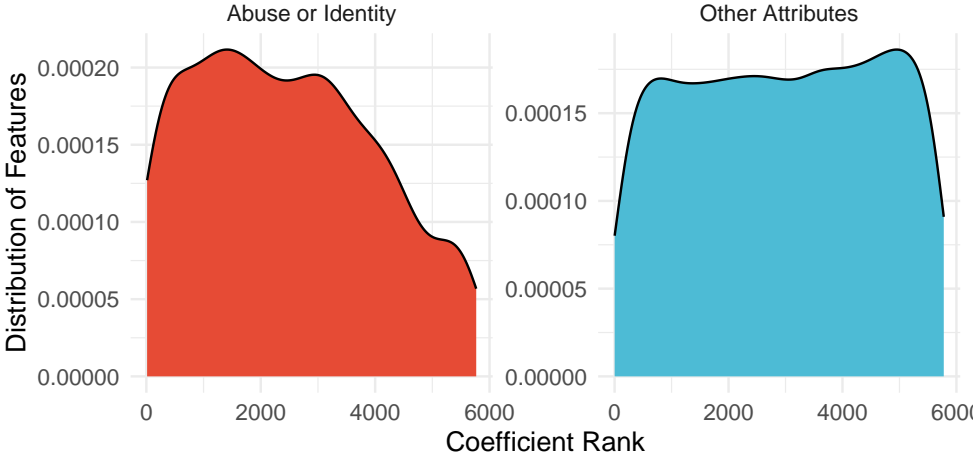


Figure 2.5: Distribution of coefficient ranks for abusive and identity terms, versus all other attributes.

On average, this difference in importance is much more pointed. In Figure 2.6, all features which involve either an abusive or identity term significantly ranks higher than all other attributes. A Welch two-sample t-test confirms that this difference is statistically significant ($t(826.09) = 3.77, p < .001$). These analyses support the explainability of the theory-based model by showing that they are among the most highly weighted measures that the model uses for predicting whether a given text is online hate or not.



Figure 2.6: Overall average predictive scores of abusive and identity terms, versus all other attributes.

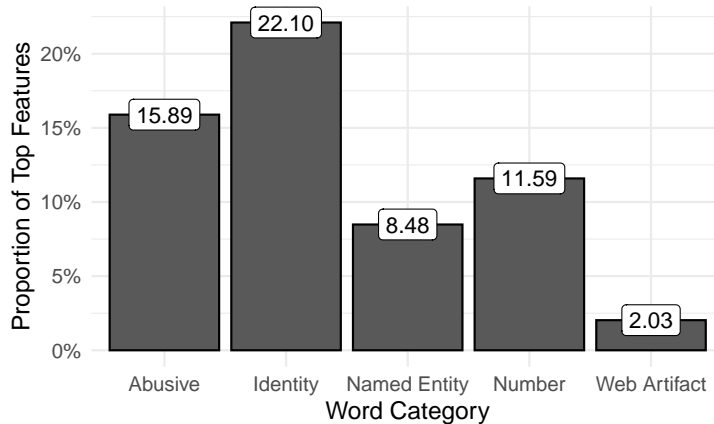


Figure 2.7: Key categories among top-ranking features in word-based machine learning models.

Problem with Word Models The results above are interesting to consider in conjunction with the corresponding analyses for words. Because the words were not pre-categorized in the word-based model, the top 1000 features by feature importance were obtained and manually annotated by two independent researchers. Open-ended coding was initially used to identify the kinds of words that occupied the top ranks. After iterative annotation, five key categories were obtained which explained over 50% of the top-ranking features, namely: abusive terms, identity terms, named entities, numbers, and web artifacts. These are summarized in Figure 2.7.

Notably, while abusive and identity terms accounted for 37.99% of the top 1000 features, another 22.10% related to noisy features. In particular, named entities—e.g., Donald Trump, Nancy Pelosi—were uniquely predictive of hate versus non-hate, which was dataset-specific rather than a general relationship with online hate as a concept. Various numbers and Web artifacts such as Twitter account handles or URLs were also related to online hate, neither of which showed any particular theoretically motivated link to the concept at hand.

Ablation experiments were thus conducted whereby these latter three categories were removed from the dataset and new models were trained. Results are summarized in Figure 2.8.

For both accuracy and F1 score, performance witnessed a noticeable drop. Once dataset-specific features and noisy features were removed from the dataset vocabulary, the predictive models became less accurate than the theory-based model, while the gap in F1 score decreased below half a percentage point. Despite considerably more parameters in the word-based model, performance was thus nearly identical to the compact and faster theory-based model after ablation. This suggests that word-based models are likely overfitting to dataset-specific aspects of the texts rather than specifically capturing key qualities of the target phenomenon of online hate.

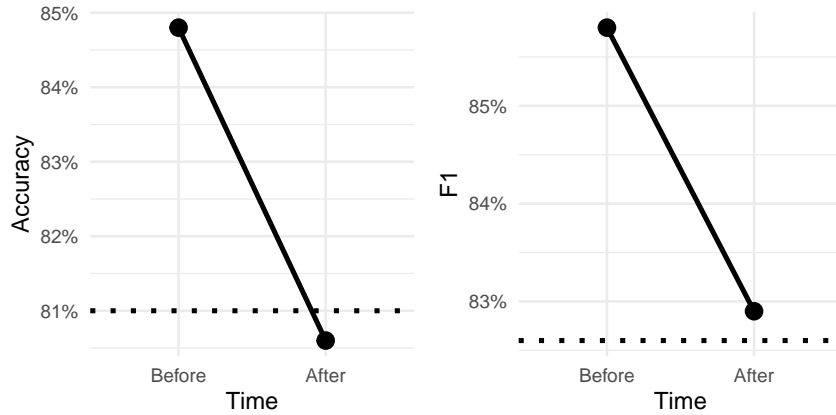


Figure 2.8: Performance of word-based machine learning models before and after the ablation of named entities, numbers, and web artifacts from the dataset vocabulary. The dashed line denotes the corresponding performance measure of the theory-based model utilizing Netmapper’s psycholinguistic measures.

2.6 Discussion

This chapter proposed a theory-based model of online hate. It proceeded in two parts. In the opening half, it statistically examined the prevalence of two theoretically motivated psycholinguistic features in a mega-corpus of labeled online hate: abusive terms and social identity terms. Multilevel regression modeling indicated that, indeed, abusive and identity terms together were the most strongly associated psycholinguistic features with online hate. These effects were moreover robust to datasets, languages, and platforms, indicating their general theoretical consistency.

Next, building on this statistical analysis, this chapter sought to develop a predictive model for online hate detection that demonstrated the practical utility of these conceptual relationships. Machine learning models were trained with the suite of psycholinguistic features and compared with common alternatives in the natural language processing literature: word-based machine learning models, embedding-based deep learning models, and transformer models. While the theory-based model showed competitive performance and generalizability with more high-dimensional models, it also featured superior efficiency as well as explainability. This showcases its desirability as a deployable and scalable tool for social cybersecurity contexts of application.

Contributions of this chapter are four-fold. First, I aggregate and standardize a large mega-corpus of labeled online hate instances across datasets, languages, and platforms. On its own, this resource provides a distinct array of diverse data points for which a battery of tests can be conducted for model development and statistical analysis such as those conducted here [201]. Scholars have pointed to major issues with online hate detection models due to the idiosyncratic particularities of individual datasets, leading to inconsistent models and untested generalizability [87, 88]. On its own, this curated dataset can be used to address such issues, and specify goals of model development along different

dimensions.

Second, using this mega-corpus, I develop and document an online hate detection model that is theoretically motivated and practically deployable for social cybersecurity analysis [53, 252]. Due to its scalability and generalizability, it can be used responsively in operational contexts where datasets of large scale and high velocity become available and insight-generation is time-sensitive. Interpretability likewise aids in this regard, especially in the context of volatile application settings where explanations are needed for predictions made. Such features may also be properly informative for when its utility in certain application settings may be limited or require additional inputs for meaningful analysis.

Third, from a conceptual standpoint, I demonstrate key relationships between general psycholinguistic categories and the distinct problematic phenomenon of online hate. This opens up future avenues of research by linking an emergent online harm with psycholinguistic constructs more well-established in the literature [194, 241]. Prior scholarship in this area shows how particular psycholinguistic features are indicative of measurable psychological states and future behavioral outcomes in the context of happiness, sadness, and anger [112, 119, 271]. Similar studies may draw upon the features analyzed here to conduct human participant research to better understand their linkages with hate as experienced beyond the online realm.

Finally, I introduce a high-level methodology for the development of theory-based tools and techniques in social cybersecurity, with online hate detection as a critical exemplar case. Development efforts for high-performing, conceptually coherent, and practically applicable models in this area may draw upon the broad methodology introduced here [252]. By incorporating theoretical considerations in an “end-to-end” fashion, other computational tools might benefit from both enhanced synergy in relation to existing social scientific work, as well as a more holistic assessment of methodological desiderata beyond an exclusive consideration of performance. As an exemplar case, this broad model development framework has also been utilized in the development of a troll detection model, similarly capturing the psycholinguistic features of trolling as an online harm, examining the validity of the model in relation to adjacent constructs, and highlighting key methodological trade-offs in relation to deep learning and transformer-based models [255].

Despite key advantages of the model proposed in this chapter, important limitations are salient. As with many approaches to text classification, problems persist when classifying texts featuring common types of distortion especially on social media. Both lexical distortions (e.g., misspellings, intentional or otherwise) as well as semantic ambiguities (e.g., polysemy, sarcasm) remain a difficult problem for natural language processing in general [92]. In addition, while the model trained here captures a large variety of online hate due to the diverse mega-corpus upon which it is trained, expressions of hate remain tied to societal dynamics which may evolve beyond the parameters of the dataset. This includes the emergence of online hate featuring new types and styles of language (e.g., neologisms) [219]. Moreover, because online hate labels also depend in part upon the socio-cultural context in which they are generated, contentious instances of online hate are also problematic for the model proposed here, as with any online hate detection model [66, 214].

Future work may opt to build upon these gaps in various ways. Additional tools specifically built to address issues of sarcasm or textual distortion may be used as an

additional data processing step at various stages of the prediction pipeline. The design of broader human-in-the-loop methodologies could also be used to address emergent forms of hate as well as guide purposeful and nuanced judgments regarding ambiguous cases. Over time, new features may also be added to the Netmapper dictionaries in line with evolving needs. Adaptive models that recognize the relative value of various feature sets across different contexts of application could also conceivably be developed and tested. While I do not tackle such issues directly in this chapter, these considerations nonetheless speak to the openness of the proposed approach to such augmentations, as well as to the broader utility of social cybersecurity principles as a methodological framework for design and deployment [52, 53, 252].

Chapter 3

Narratives of Online Hate

3.1 Introduction

Hate, like many emotions, shapes a distinct manner of making sense of the world. Affectively, hate is characterized by intense negative valence and high arousal directed toward a particular emotional target [111, 202, 223]. But cognitively, the experience of hate also operates by activating a particular system of concepts and ideas which make the targeted emotion meaningful [229, 230]. Through such systems of meaning, hate orients people’s prejudices against the objects of their hostilities, becomes a source of shared identity with an in-group, and directs particular destructive behaviors and relationships with an out-group [82, 155].

Despite this rich social scientific understanding of hate, existing research on online hate is limited by its tendency to focus on binary labels of hate and non-hate [86, 217]. To facilitate clear-cut automated detection, the complex meanings attached to hate are typically flattened out [200]. Such approaches allow for large-scale analysis of hate in a broad sense, but they elide more nuanced and contextually embedded investigation into how, in particular, online hate attacks its targets in potentially distinct ways [135, 207]. Such distinctions in meaning-making are particularly critical for understanding the underlying social conflicts which give rise to instances of online hate, as well as fashioning meaningful responses to counter or build resilience against such online hate messages [52].

This chapter utilizes the concept of *narratives* to understand how online hate engages in these cognitive processes of meaning-making around its targets. Operationally, narratives achieve such sense-making through the interlinking of various meanings in relation to the self and the other, thereby attaching hate to a specific collection of beliefs which define courses of action and contextualize broader worldviews [107, 178]. From a narrative standpoint, hate may thus be understood as a semantic system of concepts organizing how people orient toward their emotional target [48, 154]. This fundamentally shifts from a binary understanding of a given text as hate or non-hate, toward a more holistic understanding of how hateful messages perform their targeted emotional processes.

In this chapter, I set out to accomplish two key objectives. First, I introduce a pipeline of computational methods to extract hate narratives from large-scale online text. Through

a novel integration of machine learning modeling, semantic network analysis, and large language models, I develop and demonstrate a computational social science methodology for characterizing online hate narratives. Second, I apply this methodology to two a subset of the labeled mega-corpus of online hate described in Chapter 2. Taken together, this analysis showcase key distinctions in narratives of online hate in a previously labeled corpus with known subtypes of targeted hate, thereby recovering established categories as well as enriching our understanding of them. I conclude with theoretical and methodological implications of this work.

3.2 Related Work

3.2.1 Social Stories of Hate

Psychological scholarship on hate has long recognized the importance of its cognitive components [82, 84, 155]. In particular, foundational work by Sternberg characterized the affective components of hate as intertwined with specific cognitive beliefs associated with its emotional targets [229]. It is through particular beliefs that a given target of hate is perceived as different, inferior, threatening, or disgusting. In turn, it is through such judgments that hate is given motivation to thrive in individuals and in groups, and granted a distinct force in society to legitimize violence and discrimination.

Sternberg specifically referred to such cognitions as socially embedded “stories” of hate. He recognized that meanings are attached to various identities and social groups in systematic ways that constitute consistent patterns of belief associated with the emotion of hate. In his seminal theory of hate, Sternberg [229] cites numerous story archetypes which structure common hate narratives. These include: *the stranger*, which highlights the ‘otherness’ of the target of hate to justify their ostracization from the in-group; *the impure*, which sees the need to eradicate the target of hate due to their ‘contamination’; and *the thwarted*, which views targets of hate as destroying the righteous destiny of the in-group, implying that they must be stopped at all costs to maintain the in-group’s existential integrity.

More recent literature in this area has advanced this meaning-based approach to understanding hate and conflicts in general, reinforcing this basic argument about hate incorporating systems of meaning to attack its targets. In addition to identifying crystallized archetypes of hateful narratives, more recent work likewise underscores how the symbolic work of hate operates to reproduce or recreate social hierarchies. Reddi [207], for instance, considers hate as a specific form of “identity propaganda” which “exploits social orders as communication strategically designed to undermine or manipulate target populations in pursuance of a political goal through appeals about identity or identities that accord with racial and other power structures” [207, p. 5]. Similarly, KhosraviNik and Esposito [135, p. 54] understand the perpetration of online hate in terms of socio-cultural “discursive processes where boundaries of difference, uniqueness, and distinctiveness are constituted” in order to “(re)construct and (re) define social realities.”

3.2.2 Characterizing Targeted Online Hate

Given the vast proliferation of hateful content in the online world, considerable effort has been invested in detecting it computationally [16, 150, 217]. However, prevailing computational advances in detecting online hate have not seen equivalent progress in the more challenging task of characterizing *how* online hate attacks its targets [135, 200, 207]. In other words, the rich storied conception of hate in the social sciences has largely been sidelined in favor of binary approaches to merely label texts as hate or non-hate.

From a quantitative standpoint, existing approaches to move beyond these dichotomous frameworks have been fairly limited. Some attempts have sought to introduce solutions at the data annotation level, by labeling data with fine-grained typologies of hate types [122, 262]. Model development and deployment has also been reconceived in a variety of studies to specifically detect particular forms of misogyny [185, 193], racism [21, 143], and religious bigotry [134, 257]. In my own prior work, I also characterized targeted online hate by looking to the broad category of social identities mentioned in hateful texts [250]. While such approaches certainly do recognize a necessary plurality to the forms of harm online hate may inflict, they make limited progress in characterizing the meanings which facilitate hostile emotional targeting.

Interestingly, however, abundant qualitative scholarship does engage in the project of unpacking hate narratives used to target specific social identities. For instance, misogynistic movements have been shown to view masculinity as under attack to justify the harassment of women [158]. Anti-immigrant narratives position immigrants as alien threats to the status of the “authentic” people of a given nation [121]. During the COVID-19 pandemic, xenophobic narratives were identified as designating Chinese and other Asian people as carries of disease [72]. Exploiting the intersection of multiple non-dominant social identities, hateful propaganda against Kamala Harris in the 2020 U.S. elections attacked her with both gendered accusations of promiscuity and racist questions about the authenticity of her American citizenship [207]. These examples vividly illustrate the diversity of hate narratives and the unique forms of social harm they propagate. However, because such narratives have largely been identified through theoretical analyses or qualitative methods, their identification in large-scale online contexts is relatively constrained [55].

3.2.3 Narratives as Meaning Systems

Taking together the rich theoretical motivation to understand online hate narratives alongside key methodological constraints in existing approaches, I turn to narratives to operationalize systems of meaning around online hate. Narratives offer a conceptually coherent and methodologically nuanced interpretation of online hate. On the one hand, narratives align well with the storied conception of hate well-established in the social scientific literature [82, 207, 229]. On the other hand, narratives have also seen burgeoning developments in the computational sciences. Various approaches have been implemented to extract narratives from texts in an automated and scalable fashion [48, 49, 154, 205].

In this work, I specifically draw upon a constructural lens for narrative characterization [45, 46]. Within a constructural framework, narratives can be understood and operationally

modeled as semantic networks: systems of concepts and ideas represented as nodes with patterns of meaningful connection represented as the edges between them. Constructural analysis of narratives has been performed in a variety of analytical contexts such as understanding the evolution of stereotypes [125], characterizing violent rhetoric by state actors [173], and tracking plot developments in literary media [167].

Applied to the context of online hate, semantic networks allow for a quantitative characterization of how various concepts are attached to a given target or identity. In contrast to related bag-of-words models of topics, a semantic network understanding of narratives does not merely illustrate what concepts are relevant to a given meaning system [205]. It is also capable of characterizing their organizational structure. Semantic networks thus enable the identification of which concepts are more central than others, which concepts bridge various other concepts, and possibly what various clusters of concepts introduce key variations in meaning-making within a given narrative [48, 49, 154].

3.3 Data and Method

3.3.1 Dataset

For this chapter, I use a subset of the mega-corpus used in Chapter 2, with specific datasets summarized in Table 3.1. These datasets correspond to previously labeled instances of specific forms of targeted hate to demonstrate the online hate narrative methodology in analytical context.

Table 3.1: Summary of datasets to characterize narratives of online hate.

Dataset	Hate Types	Data Points
Ethos [172]	Gender, Race, Religion	382
GHC [132]	Gender, Race, Religion	1570
Multimod [102]	Gender, Religion	4703
Waseem [261]	Gender, Race	2740

From the fifteen datasets initially aggregated to form the mega-corpus, four datasets had more fine-grained labels of various hate types in addition to the broad designation of hate or non-hate. For instance, the Ethos dataset [172] and the Gab Hate Corpus [132] contained distinct labels of gender-targeted, race-targeted, and religion-targeted hate. Meanwhile, the multidimensional dataset [102] had gender and religion subtypes of hate, while Waseem [262] had labels specifically for sexism and racism. Because this first dataset has previously known labels of targeted online hate, these can be used as part of the validation step of the proposed narrative methodology. Generally, it is expected that the extracted narratives will largely align with and enrich these a priori categories. Measures of other psycholinguistic features are also derived using Netmapper [54, 126, 241].

3.3.2 Analysis

Utilizing the datasets described above, the broad analytical approach is described in Figure 3.1. I propose a novel pipeline of text analytical tools that makes use of machine learning, network analysis, and large language models. Each of the major succeeding analytical steps is described in the sections that follow.

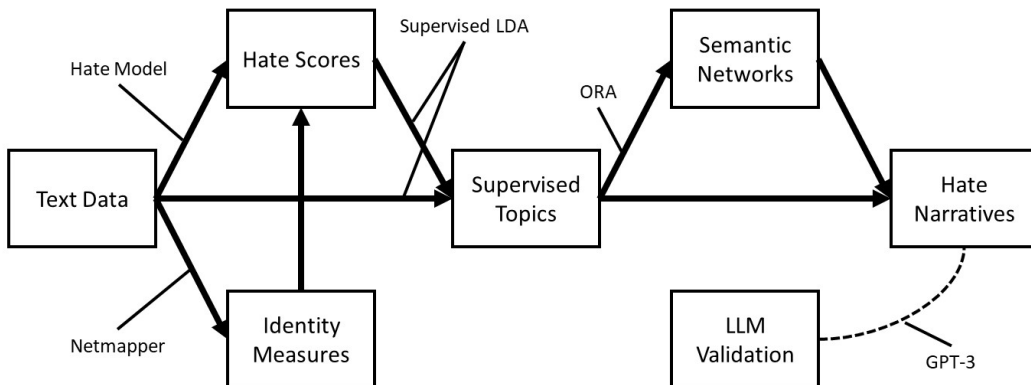


Figure 3.1: Summary of proposed methodological pipeline for characterizing narratives of online hate.

Supervised Topic Modeling Taking together the text data and its associated hate and identity measures, a supervised topic model is used to segment the collections of text into distinct subsets [162]. This serves the purpose of separating out which texts in a given dataset tend to use different sets of words, and forms the initial basis for empirically determining distinct online hate narratives.

Topic modeling approaches have been used in a variety of social scientific research applications to understand how a certain object is constructed in different ways [28, 205]. But applied to the context of online hate, a *supervised* topic model specifically offers novel methodological and conceptual value. Methodologically, it allows for the explicit modeling of hate as a target variable in the process of topic extraction, rather than treating it as an independent covariate. This aligns conceptually with understanding the segmentation of the corpus into distinct narratives as shaped by their collective expression of hate. The Tomotopy library¹ is used to execute model estimation, inference, and other auxiliary functions.

Probabilistically, the generative process for the supervised topic model is given as follows:

1. Sample the topic proportions θ_d for the documents $d \in \{1, 2, \dots, D\}$, conditioned on prior α . That is, $\theta \sim \text{Dir}(\alpha)$.
2. For a given word n out of a document d of length N , sample its topic assignments $z_{d,n}$, conditioned on θ . That is, $z \sim \text{Mult}(\theta)$.

¹<https://github.com/bab2min/tomotopy>

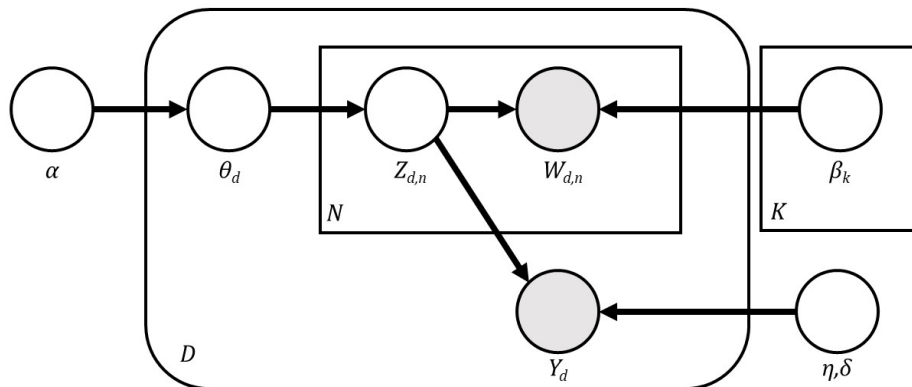


Figure 3.2: Plate diagram of the hate-aware supervised topic model.

3. Let β_k be the distribution of words for each topic $k \in \{1, 2, \dots, K\}$. Sample each word $w_{d,n}$ for the document d based on the drawn topic $z_{d,n}$. That is, $w \sim \text{Mult}(\beta_z)$.
4. Sample the response variable Y_d for document d based on the drawn topics z and parameters η, δ . In this case, Y_d corresponds to the estimated hate score for each document d . That is, $Y \sim \text{N}(\eta z, \delta^2)$.

By combining the text data with hate measures as the target variable, a supervised topic model may be estimated. Estimation is performed with collapsed Gibbs sampling which enables faster parallelization than other comparable libraries. Standard text cleaning procedures such as stopword removal, punctuation removal, and URL removal are conducted. Hyperparameters such as the number of topics and vocabulary pruning are selected using graph search and the elbow method over the log-likelihood graph for multiple runs of the algorithm [215].

Semantic Network Analysis Given the segmentation of the dataset into a set of supervised topics, I then proceed to the interpretation process. Interpretation of topic model outputs is typically performed by using a list of top-loading words for each topic. Yet while this indicates the most likely words in a given topic, these measures do not account for their interconnections. To improve upon this approach, I utilize semantic network analysis to better characterize the central concepts in a given hate narrative.

A semantic network is a network representation of a given collection of texts [48, 49]. For a given topic, I represent the set of texts which have been assigned to it using a weighted semantic network. Semantic networks for each topic are defined by a set of nodes $V = \{v_1, v_2, \dots\}$ and a set of edges $E \subseteq V \times V$ describing connections between word nodes. Edges are determined by the co-occurrence of words within a single text, and their weights are given as the number of texts which feature that co-occurrence.

Utilizing semantic network representations of the topics assigned positive hate scores, I interpret their underlying narratives by examining the most highly connected nodes within these networks. Communicative power analysis defines the strength of concepts in

Prompt Structure:

The text is: “_____”. Which of these topics is it an instance of?

- Topic label 1
- Topic label 2
- Topic label 3
- ...

Note: All abusive terms were censored with an asterisk on the first vowel.

Figure 3.3: Prompt for large language model in validation step of hate narrative validation methodology.

a semantic network in terms of their connectivity and evokability. Operationally, I take the nodes with the highest degree and betweenness centrality to identify the most important concepts in each of the narratives. A single measure is produced by taking their geometric mean. Texts are also additionally sampled to provide the original semantic context. These procedures are formalized in the Semantic Network Report of the ORA software [9].

Large Language Model Validation Finally, upon deriving semantic labels for each of the supervised topics, I perform a novel validation strategy facilitated by a large language model [156]. Specifically, using GPT-3, I provide an equal number of random samples from each topic and request that the large language model select between the pre-identified topic labels and provide an explanation for its selection². Cohen’s Kappa can then be calculated between the supervised topic model’s assignments and the assignments produced by the large language model. The prompt structure is given in Figure 3.3.

This procedure validates the analysis of hate narratives by showcasing the internal coherence of each cluster of texts determined by the supervised topic model, as well as the interpretability of the topic label derived through semantic network analysis. Moreover, due to the large language model’s ability to provide explicit reasoning behind its decision, further qualitative information is provided behind both correct classifications as well as potential misclassifications. This enriches the validation process beyond quantitative evaluation by identifying key phrases that characterize a given topic and describing gray areas between narratives that may help sharpen their interpretation [110].

3.4 Results

Utilizing the proposed integrated methodology, I analyzed a subset of the mega-corpus of labeled online hate and identified four major online hate narratives. Supervised topic modeling quantitatively clustered the text dataset into seven different topics, and identified that the first four had positive online hate scores. Semantic network analysis identified the

²<https://pypi.org/project/pyChatGPT/>

most important concepts in each hate narrative to facilitate narrative labeling, and these labels were in turn validated through the use of a large language model.

3.4.1 Topic Clustering

Figure 3.4 visualizes the results of the supervised topic model. The first four topics were identified to have positive online hate scores, and their primary targets were further classified as being (a) gender; (b) race and religion; (c) gender; and (d) gender, respectively. These additional details are summarized in Table 3.2.

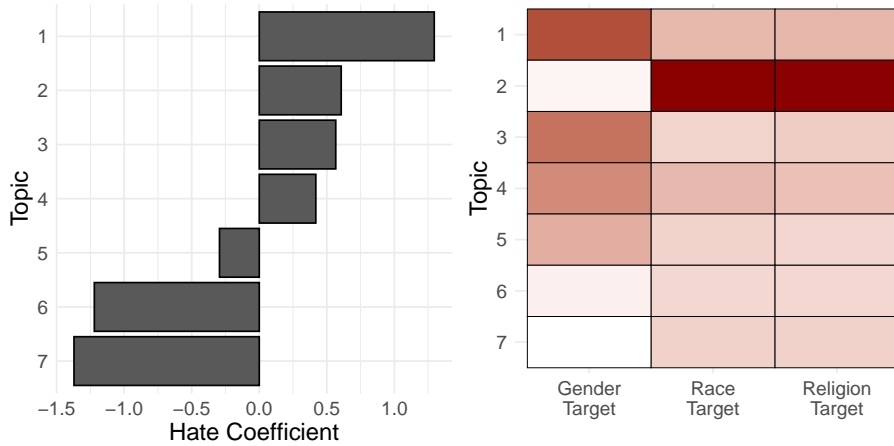


Figure 3.4: Outputs of supervised topic model on the labeled subset of online hate megacorpus. **Left:** Coefficients on target variable of hate label. **Right:** Saliency of gender, race, and religion labels over topics.

Table 3.2: Summary of extracted topics featuring positive online hate scores.

Topic	Label	Hate Norm	Identity Target(s)
1	Degrading Misogyny (DM)	1.2953	Gender
2	White Supremacy and Xenophobia (WS)	0.6066	Race and Religion
3	Sexual Domination of Queer People (SD)	0.5663	Gender
4	Anti-Progressive Backlash (AP)	0.4187	Gender

Based on the supervised topic model, topic 1 was observed to have the highest hate coefficient (1.2953). This indicates that this topic may have the most certain and explicit cases of online hate across the entire dataset. Topics 3 and 4 interestingly share primarily gender-based targets, but their hate coefficients are lower than that of topic 1. This suggests that while they both target gender identities, their use of hate-oriented language may be less explicit or clear-cut than in topic 1. Further analyses in the subsequent sections provide an interpretative basis for understanding what different identities are targeted by

these different topics, and how the narratives used to attack them may likewise showcase key qualitative distinctions.

Finally, topic 2 has the distinction of having racial and religious identities as its primary targets. This suggests that it is functionally distinct from the other three topics since it orients toward attacking an entirely different class of social identities. It is also the second highest in terms of its associated hate score. It is thus less clear-cut in its use of explicitly hateful language than the first topic, but it may be less ambiguous than topics 3 and 4.

From a model fitness perspective, it is worth noting that in comparison to a vanilla approach to topic modeling, the proposed supervised method had superior performance in: (a) separating out examples of hate versus non-hate, (b) clustering texts together which had similar targets, and (c) grouping together examples which came from different datasets. These results are reported quantitatively in Table 3.3. While distinctions by hate levels are explicitly modeled by the supervised topic model, targeting and datasets of origin are not. This provides additional support for the strength of the proposed supervised topic model for the purpose of distinguishing between online hate narratives.

Table 3.3: Results of chi-square tests assessing the extent to which supervised topic models produce more meaningful clusters than a vanilla topic model.

Clustering	Supervised LDA	Vanilla LDA
χ^2 by Hate	5128.578***	2066.485***
χ^2 by Target	8961.046***	4675.353***
χ^2 by Dataset	12771.030***	12351.130***

Note: * $p < .05$, ** $p < .01$, *** $p < .001$

Finally, of the seven topics quantitatively identified by the supervised topic model, it is worth noting that three were also associated with negative hate scores. These clusters thus group together the texts which were not hateful in the dataset. Since these are not the focus of this chapter, they are not subjected to the remaining steps of the proposed methodology. To briefly comment on their content, however, Figure 3.4 shows that topic 5 has gender targets, while both topics 6 and 7 have racial and religious targets. These corresponded to the non-hate examples in the datasets which respectively expressed: (a) the non-hateful reclamation of queer slurs presumably by LGBTQ+ speakers (e.g., “I am one sexy d*ke”); (b) texts in African-American Vernacular English (AAVE) which makes use of terms which are colloquial to the Black community but might be deemed offensive in other contexts (e.g., “what’s up, my n*gga?”); and (c) various news headlines with report on racially charged events.

3.4.2 Semantic Characterization

To produce analytical labels for each of the four narrative clusters identified, I examined their structures through semantic network analysis. Semantic networks for each of the four topics are visualized in Figure 3.5. Semantic networks in these depictions are subjected to

Leiden clustering to reduce visual load. Descriptive statistics of these semantic networks are provided in Table 3.4.

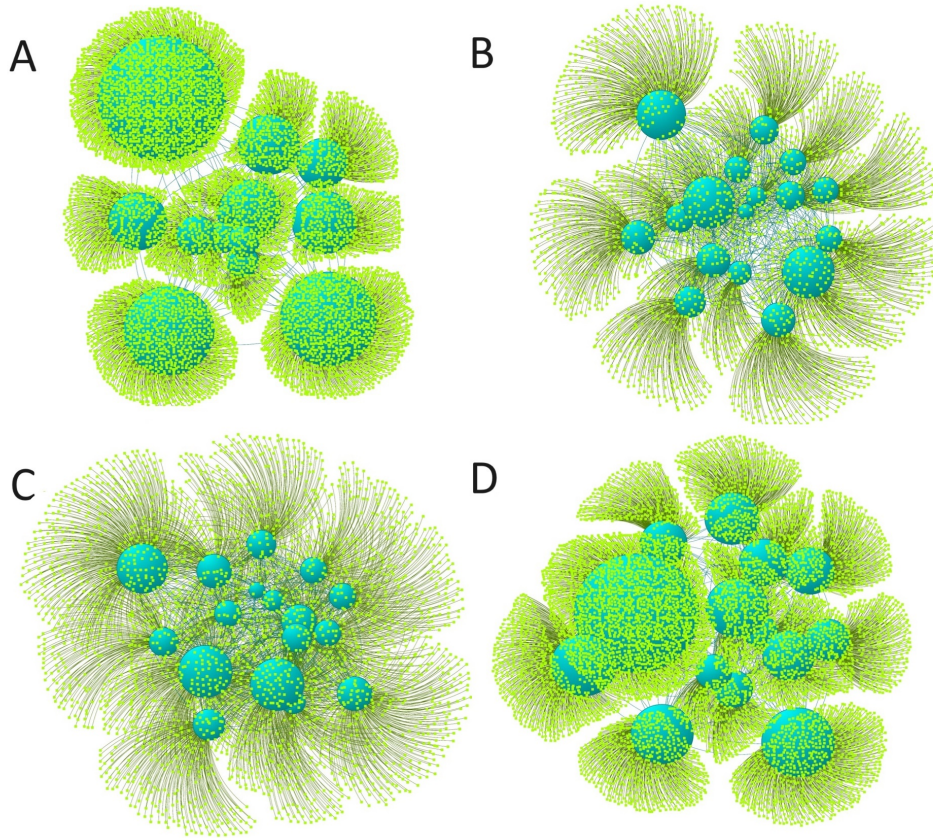


Figure 3.5: Semantic network visualizations of derived online hate narratives. **A:** Degrading misogyny. **B:** White supremacy and xenophobia. **C:** Sexual domination of queer people. **D:** Anti-progressive backlash.

Within these semantic networks, degree centrality scores were calculated to identify the most important nodes. These correspond to the core concepts within each topic, and indicate both key targets of each hate narrative as well as how these social identities are specifically attacked. Notably, the most important nodes based on semantic network degree were more likely to be identity terms than the most high-probability words based exclusively on the supervised topic model. A two-way analysis of variance on their importance scores revealed that semantic networks were more likely to assign high weights to identity terms in comparison to the probabilistic scores produced by topic modelling, as summarized in Table 3.5.

From the derived centrality measures, analytical labels could then be determined and interpreted. These calculations are summarized in Table 3.6.

Degrading Misogyny Topic 1 was interpreted to express a hate narrative of *degrading misogyny*. Two of the most central terms were specifically identified to be “c*nt”

Table 3.4: Descriptive statistics for semantic networks of online hate narratives.

Network Feature	Topic 1	Topic 2	Topic 3	Topic 4
Size	8210	5138	2920	8011
Density	0.003	0.010	0.009	0.006
Assortativity	-0.125	-0.089	-0.093	-0.132
Centralization	0.237	0.256	0.325	0.261

Table 3.5: Results of two-way analysis of variance in importance scores assigned by semantic network method versus standard topic modeling scores.

Feature	df	F-value	p-value
Identity	1	17.593	< .001
Semantic Network Method	1	0.507	> .05
Identity x Semantic Network Method	1	8.899	< .01

(0.240) and “tw*t” (0.143), both of which perform a derogatory reference to women’s genitalia as an insulting means of sexualizing them and reducing their humanity. Alongside these woman-specific abusive terms, “r*tarded” (0.097) and “animal” (0.088) were also among the highest-scoring concepts. These further reinforce the degradation of women by insulting their intelligence, and by comparing them to animals, a classic component of dehumanization rhetoric in hate narratives [229]. References to “woman” (0.105) and “girl” (0.098) were also central, highlighting the targeting of feminine social identities within this narrative.

White Supremacy and Xenophobia Topic 2 was identified as featuring a hate narrative of *white supremacy and xenophobia*. Interestingly, this topic collectively utilizes similar semantic features to target Jewish people (“jew”, 0.208), Muslims (“muslim”, 0.201; islam, 0.147), and Black people (“n*gro”, 0.132). Taken together, this showcases the shared hostility of xenophobic and white supremacy narratives against racial and religious identities deemed to be inferior and othered in the ethnoreligious hierarchy. Such targets are specifically threatened with violence (“kill”, 0.135). The texts in the dataset specifically also assert the superiority of the United States (“american”, 0.184; “usa”, 0.150) as a bastion of white identity that the othered people are positioned as not properly belonging there.

Sexual Domination of Queer People Topic 3 was aligned with a hate narrative of *sexual domination of queer people*. Like topic 1, it featured primarily gender-based targets. However, instead of targeting women, it was more specifically oriented against the LGBTQ+ population and various queer identities. Derogatory terms are used with reference to gay men (“f*ggot”, 0.307), lesbians (“d*ke”, 0.198), alongside elliptical references to trans women (“sissy”, 0.111). Furthermore, whereas topic 1 combined insults to intelligence with dehumanizing rhetoric, topic 3’s expressions of hate and violence were more

Table 3.6: Semantic centrality measures of top-ranking terms.

Topic 1 Degrading Misogyny		Topic 2 White Supremacy		Topic 3 Sexual Domination		Topic 4 Anti-Progressive Backlash	
Term	Score	Term	Score	Term	Score	Term	Score
c*nt	0.240	people	0.214	n*gro	0.334	woman	0.267
f*ck	0.192	jew	0.208	f*ggot	0.307	people	0.209
call	0.145	muslim	0.201	c*nt	0.207	man	0.197
tw*t	0.143	country	0.187	d*ke	0.198	know	0.169
n*gro	0.134	american	0.184	f*ck	0.167	agree	0.157
woman	0.105	god	0.153	a*s	0.162	sexist	0.145
sexist	0.105	usa	0.150	b*tch	0.161	right	0.127
girl	0.098	islam	0.147	d*ck	0.158	f*ck	0.106
r*tarded	0.097	kill	0.135	animal	0.135	feminist	0.102
animal	0.088	n*gro	0.132	sissy	0.111	take	0.100

particularly sexual and physical in nature. This was expressed in central terms related to genitalia and sexual acts to describe sexual violence inflicted upon gender minorities (“f*ck”, 0.167; a*s, 0.162; d*ck, 0.158). Dehumanizing comparisons to animals (0.135) were also present.

Anti-Progressive Backlash Finally, topic 4 was identified to express a hate narrative of *anti-progressive backlash*. Featuring the lowest hate score assigned by the supervised topic model, this narrative was relatively ambiguous in its expression of hate in comparison to the three other identified topics. Like topics 1 and 3, it also tended to feature gender-based identities, like “man” (0.197) and “woman” (0.267). The distinction, however, is most evident in references to “sexist” (0.145) and “feminist” (0.102), which underscore discussions of progressive values of gender equality. In making its claims against such progressive values, these texts tend to make use of relatively civil language around beliefs and argumentation (“know”, 0.169; “agree”, 0.157). Inspection of these texts in their original contexts of utterance, however, does reveal their expression of hateful sentiments which insult and threaten progressive movements and their proponents (“f*ck”, 0.106).

3.4.3 Validation and Computational Reasoning

Validation was performed with GPT-3 by providing a random sample of texts to the large language model and prompting it to characterize the text as belonging to one of the four hate narrative topics. Results of the validation analysis are visualized in Figure 3.7.

Cohen’s kappa was calculated to be 0.601, which in the literature is qualitatively characterized as “substantial agreement”. In comparison to more traditional annotation schemes in natural language processing, this value is also considerably higher than standard human annotation which in an empirical study sees a range of reliability values between 0.231 to 0.506 in the inter-rater setting. The large language model is thus capable of associating the analytical labels of the hate narratives with raw samples from the corpus, suggesting that the topics identified are coherent and interpretable.

To further strengthen the validation process, I also examined the explanations offered by GPT-3 regarding its decisions. Explanations for correctly classified texts are provided

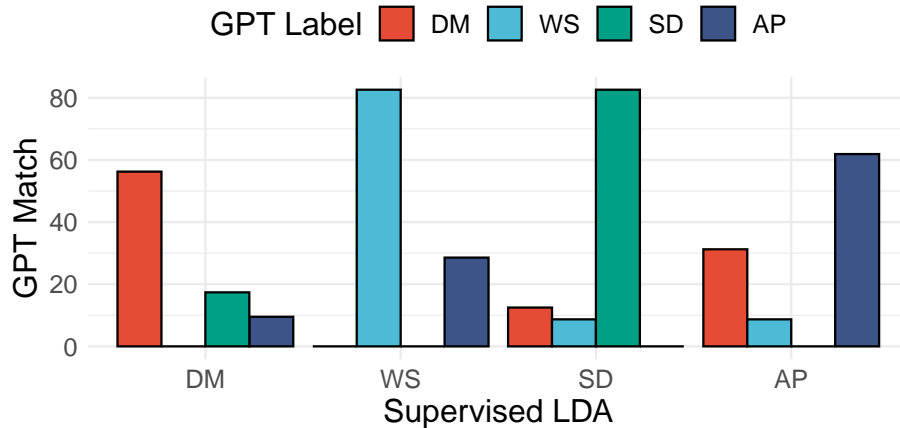


Figure 3.6: Results of validation with GPT-3. Given a text, the large language model was generally capable of distinguishing which topic it belonged to based on labels assigned to results of the supervised topic model ($N = 100$).

in Table 3.7, while explanations for misclassified examples are given in Table 3.8.

Table 3.7 shows conceptually coherent justifications for the large language model’s association of raw texts from the corpus with the analytical labels assigned to the four hate narratives. For topic 1, the narrative of degrading misogyny, GPT-3 explicitly zeroed in on the use of the terms “c*nt” and “sis” and meaningfully recognized how these terms were used in the sample text to ascribe “negative attitudes and stereotypes about women” as a reason for assigning the misogyny label. For topic 2, the narrative of white supremacy, GPT-3 lexically pointed to the sample text’s use of racial slurs as an empirical touchstone for reasoning about the label. Moreover, it provided an in-depth characterization of the text’s promotion of racist ideas through the rewriting of history to communicate the racial superiority of white people.

Meanwhile, for topic 3, the narrative of sexual domination of queer people, GPT-3 pointed to sexual orientation as the reason for the text’s expressions of hate, and also specifically mentioned violence and discrimination as the attack directed against them. This stood in contrast to the example for topic 4, the narrative of anti-progressive backlash. Here, GPT-3 identified not only that transgender people were the targets of the text, but also that a broader opposition to acceptance and gender diversity motivated the statement given. While a member of the LGBTQ+ demographic was the target of the statement, GPT-3 precisely recognized that this statement did not belong to the sexual domination narrative since the form of attack was not necessarily a physical form of violence.

Now in Table 3.8, GPT-3 showcases instances of misclassification. The reasons it provided for these judgments also shed light on boundary cases and ambiguities between hate narratives. For instance, given a text labeled under topic 1, GPT-3 decided that it might be an instance of anti-progressive backlash. This was plausible given that a hashtag of “LiberalHypocrisy” was used in the text, which might be indicative of anti-progressive values, even while engaged in degrading misogyny against “trailer trash wh*res”. For a text

Table 3.7: Reasoning of GPT-3 for accurate classifications.

Topic	Label	Reasoning
1	Degrading Misogyny	The use of the derogatory term “c*nt” to insult and demean women aligns with degrading misogyny , as it perpetuates negative attitudes and stereotypes about women. Additionally, the use of the term “sis” could be seen as an attempt to undermine and dismiss the woman being addressed.
2	White Supremacy and Xeno- phobia	The text perpetuates historical inaccuracies and promotes racist ideas, including referring to African Americans using a racial slur. It also attempts to downplay the historical role of white people in slavery while promoting anti-Indigenous and anti-Black sentiments. This aligns most closely with white supremacy as it promotes beliefs of racial superiority and attempts to rewrite history to fit a racist narrative.
3	Sexual Dom- ination of Queer People	The text provided contains highly offensive and derogatory language targeted at queer individuals. It promotes violence and discrimination based on sexual orientation. This text could be categorized as sexual domination of queer people . It perpetuates harmful stereotypes and contributes to a narrative of hatred and discrimination against LGBTQ+ individuals.
4	Anti- Progressive Backlash	The statement expresses a negative attitude towards transgender individuals, characterizing them as “sick in the head” and suggesting that their identity is a result of societal problems. It also presents a false equivalence between being transgender and other hypothetical scenarios, such as people claiming to have the right to change into animals. This contributes to an anti-progressive narrative that opposes acceptance and understanding of gender diversity.

under topic 2, GPT-3 also identified a narrative of anti-progressive backlash. While a white supremacy message was indeed embedded in the text, some ambiguity could also arise from the fact that it was an intersectionally targeted statement, specifically attacking queer people of color. GPT-3 thus believed that the statement demonstrated a broader rejection of multiculturalism and minorities, which might reasonably be considered anti-progressive.

A text under topic 3, sexual domination of queer people, was also classified by GPT-3 as engaged in degrading misogyny. Indeed, while the text itself was targeted against a lesbian woman, the text also advocated for her mistreatment due to her perceived undesirability, which aligns with topic 1. Finally, a text under topic 4, anti-progressive backlash, was directed toward feminists and their values. But because of the woman-targeted language (e.g., referring to female reproductive organs), it could also be conceivably classified as degrading misogyny.

Table 3.8: Reasoning of GPT-3 for misclassified examples.

Topic	Label	Reasoning
1	Degrading Misogyny	The text contains derogatory language (“trailer trash wh*res”) and appears to make disparaging remarks about certain individuals. It also mentions “LiberalHypocrisy,” implying a criticism of liberal attitudes or behaviors. This text could be categorized as anti-progressive backlash as it criticizes liberal views while using derogatory language towards certain groups of people.
2	White Supremacy and Xenophobia	The text expresses a longing for a time before multiculturalism and promotes racist and homophobic ideas. It refers to interracial relationships in a derogatory manner (“Blacked”) and uses offensive language to describe LGBTQ+ individuals and their choices. This text could be categorized as anti-progressive backlash due to its rejection of multiculturalism and promotion of discriminatory attitudes towards racial and sexual minorities.
3	Sexual Domination of Queer People	The text contains explicit language and promotes violence against women. It perpetuates harmful attitudes and behaviors towards women, advocating for the mistreatment of women perceived as undesirable. This text could be categorized as degrading misogyny as it promotes violence and dehumanization of women.
4	Anti- Progressive Backlash	The text provided is highly inappropriate and offensive, containing explicit and degrading language against feminists. It falls under the category of degrading misogyny . It objectifies women and reduces them to sexual organs, perpetuating harmful stereotypes and attitudes towards women. This type of language contributes to a narrative of dehumanization and disrespect towards women.

3.5 Discussion

In this chapter, I developed a methodology for characterizing online hate narratives, combining techniques in supervised topic modeling, semantic network analysis, and large language modeling [48, 49, 162]. Utilizing a previously labeled subset of the online hate mega-corpus in Chapter 2, I showed that the proposed technique meaningfully organized the texts into groups with varying levels of hate, and meaningfully distinguished texts which attacked different types of targets. In addition, I highlighted how the proposed interpretation and validation schemes could enrich analysts’ understanding of the extracted topics and their relationships through both quantitative and qualitative indicators [110].

From a high-level standpoint, this approach valuably shifts from the predominant binary framework taken in state-of-the-art models [86, 88]. While some computational models may incorporate more specific labels of targeted hate (e.g., sexism, racism), quantitative

analysis is largely limited to indicating whether a text does or does not contain hate [122, 262]. Conversely, qualitative research that explicitly turns away from a flattened view of online hate has largely relied upon in-depth engagement with small-scale samples of text, thereby inhibiting insights into larger online conversations [55, 200, 207]. Through the analytical pipeline proposed here, narrative discovery facilitates more nuanced qualitative insights of the systems of meaning employed in samples of online hate, while surfacing the empirical basis for interpretation in a principled, data-driven fashion that may apply to large corpora. As an additional methodological benefit, the entire analytical pipeline is generally language-agnostic and platform-agnostic, save for the validation step which may need to rely on updates to large language models to accommodate new languages.

Moreover, common existing methodologies like vanilla topic modeling are also capable of similar forms of data summarization, but on their own fall short of important considerations for online hate narrative discovery [28]. First, because they do not explicitly consider hate in the construction of topics, they may be more prone to group together texts that use similar words but which do not express hate. Second, while they may also be used to generate quantitative lists of topic-salient terms, these scores are generated for words in isolation and not in relation to each other. Third, inferring the collective meaning expressed by these words targeted hate is not a trivial output of the model, but still requires meaningful engagement with the data. These points constitute some of the key methodological issues addressed by the proposed analytical pipeline, which incorporates supervision into the topic modeling process, assigns word importance based on their centrality in the semantic system, and explicitly adds new post-hoc analyses to better structure the topic interpretation process.

For the specific case of the data analyzed here, the proposed methodology also usefully identified significant empirical differences which resonated with theoretical analyses of various forms of hate. It was striking to note that, despite the explicit formulation of the component datasets into binary labels of gender-based, race-based, and religion-based hate versus non-hate, the derived narrative clusters highlighted both meaningful distinctions as well as shared modes of hateful attack. In particular, for the case of gendered hate, the model foregrounded differences between different types of gendered targets (e.g., women versus sexual minorities) as well as different types of gendered attack (e.g., dehumanizing degradation versus specific sexual domination). These findings resonate with critical scholarship in queer theory and women’s studies regarding ways that regressive movements seek to put down individuals which do not conform to dominant gendered hierarchies [152, 165]. Meanwhile, by combining racial and religious targets under the same narrative, the model further indicated that these categories were being subjected to similar forms of hateful attacks. Indeed, social scientific scholarship on white supremacy highlights how similar narratives undergird hate by holding white identity as the center from which otherness is constructed and attacked [42, 98, 104, 222]. Lastly, a more general form of hate directed toward progressive values was also identified. This points to online hate’s embeddedness in major culture wars [43, 124]. While political identities were not explicitly labeled in the mega-corpus, this final topic nonetheless recovered its distinct narrative form of attack toward progressives.

All that said, this narrative methodology also features important limitations. While

the pipeline of tools does integrate key benefits of both quantitative and qualitative techniques for understanding targeted online hate, it is not capable of completely solving all their issues. Qualitative benefits allow for more meaningful understanding of quantitative patterns, whereas quantitative measures facilitate more structured interpretation [240]. However, the patterns observed speak largely to aggregated collections of text, and thus individual samples may still be subject to noise. Moreover, the need for subjective labeling remains embedded in the pipeline, even if guided by firmer computational evidence. As such, other forms of validation well-established in the qualitative literature may also be incorporated in practice to improve the narrative discovery process [239, 274]. Meanwhile, for the application described in this chapter, I also note that the subtypes of hate I examined were limited to some common forms examined in the hate speech literature, but do not exhaust all possible types of targeted hate. Thus, the empirical analysis here does not claim to be a comprehensive typology of all forms of online hate narrative, but rather a particular demonstration over a limited subset.

Chapter 4

Networks of Online Hate

4.1 Introduction

As a social emotion, hate targets and thrives in groups [190]. Social groups defined by various identity categories like gender, race, politics, and religion shape who hates whom, and how such hate is communicated. But not all hate groups are created equally. Groups are structured in specific ways that may be intentionally organized or dynamically emergent [46, 47]. Properties of community organization influence how groups are able to socially function, including their ability to coordinate, to change or remain stable, and to reinforce a sense of belonging and coherence in their members. Similarly, such group dynamics likewise apply to online hate.

Previous studies of online hate have recognized the importance of group context in its propagation and mitigation [8, 58, 123]. But prevailing studies have tended to consider groups as relatively abstract categories of membership or non-membership ascribed to individual accounts [34, 35, 206]. This ignores the explicit structural features which organize how group members interact with each other, resulting in distinct patterns of communication that might be characterized as relatively dense or diffuse, as integrated or isolated, and as hierarchical or egalitarian.

Social networks form the second pillar of this thesis’s approach to understanding online hate. Social networks have been used to describe diverse social phenomena with actors (e.g., people, organizations) as nodes connected by edges (e.g., communication, friendship) [263]. These facilitate nuanced descriptions of which individuals are influential within groups, how groups themselves form and interact, and how broader systems shape the behavior of those within it [47]. Drawing upon network science advances, this chapter proposes a distinct set of analyses based in social network methods to assess how such features are linked to the spread of online hate.

More specifically, this chapter sets out to accomplish two goals. First, I define a set of network-based measures for characterizing online hate groups. Then, in conjunction with a community detection algorithm, I identify online hate groups in the context of the Covid-19 pandemic. This enables an assessment of the structural features statistically associated with higher levels of online hate within the identified network clusters. Second, I utilize

multilevel analysis to characterize accounts within these network clusters most likely to produce online hate messages. This provides empirical evidence for understanding how online hate spreads among individuals while controlling for their local network structures. I end the chapter with theoretical and methodological implications.

4.2 Related Work

4.2.1 Social Networks and Online Communication

Online communication does not take place in a vacuum. When people propagate messages through social media, such messages may interact with the messages of others and may in turn be interacted with by others. As social interactions proliferate between online actors, larger-scale patterns are produced. Dyadic relationships evolve into groups of various sizes and shapes which likewise feature more complex properties which emerge from yet cannot be reduced to their component parts [40, 263].

Social networks offer a powerful framework to formalize, visualize, and analyze these multi-scale dynamics of online communication [50, 54]. Typically, agents on social media are represented as nodes which are connected to each other by edges which signify interactions between nodes. By leveraging such representations, graphical properties of these social networks reveal critical features of the underlying online conversation. By viewing online conversations as networked systems, key properties may be discerned across multiple levels, ranging from individual-level, to group-level, to system-level features.

On the individual level, classical measures of node centrality have been fruitfully used to understand the influence of a given account in an online conversation. Various centrality measures have been used to quantify multiple types of influence [90, 264]. For instance, basic measures of degree centrality might indicate accounts that have the highest levels of interaction with other accounts [263]. Betweenness centrality assesses the extent to which an account bridges multiple sets of accounts and acts as an information broker. Still other measures such as eigenvector centrality highlight more nuanced forms of influence, such as one's connections to other influential accounts [31].

Beyond the individual level, social networks are also capable of describing group-level properties of online conversations. As accounts interact, they form coherent clusters which separate in-group members from out-group members. Clustered accounts typically engage in more in-group interactions than out-group interactions [30, 243]. Due to such patterns, they can be empirically identified through a variety of community detection methods. The clusters themselves may also be characterized by an array of measures including their size, the extent to which they are isolated or integrated with the rest of the network, and their level of hierarchy and organization [140, 169].

4.2.2 Online Harms and Social Structure

Empirical examinations of group dynamics have proven fruitful in the computational social science literature on online phenomena, especially harmful behaviors. For instance, emerg-

ing studies show that moral emotions—especially those involving hate and animosity—diffuse at higher rates than non-moral emotions, especially among in-group networks compared to out-group networks [34, 206]. The moral emotion of outrage, while distinct from hate, is likewise amplified through social learning mechanisms over repeated social interactions with group members [35]. Intragroup and intergroup dynamics thus shape how online harms operate and evolve.

Yet while these studies acknowledge group structure, specific features of their organization remain understudied [65, 75]. In-group and out-group behaviors tend to be seen dichotomously rather than structured in specific networked ways. From this standpoint, groups are seen primarily in terms of who interacts with whom and the memberships of these interacting agents, but the broader structure which characterizes their patterns of interaction is not considered explicitly. In practice, online groups are structured across a variety of smaller, interacting clusters [243, 250]. Hate groups, in particular, tend to have leaders and followers who do not behave in identical ways [209]. Members of online hate groups also tend to feature coordination when engaged in attacking behaviors against a target [157].

Whereas the foregoing discussion of hate consistently emphasizes the role of groups, social networks offer a distinct yet under-explored view of how groups are specifically structured [211]. Across the computational social science literature on online harms, the study of echo chambers is perhaps the most well-known to engage with group structure. Echo chambers refer to pathological dynamics of intergroup (non-)communication whereby members of opposed groups tend to limit their interactions only to in-group members while remaining balkanized from the out-group. Considerable literature describes methods for identifying such harmful patterns of intergroup dynamics on social media, drawing on empirical network structures to capture the defined patterns of interaction [18, 273]. Going beyond echo chamber effects, other forms of harmful intergroup communication like trench warfare have also seen social network operationalizations in recent scholarship [128].

4.2.3 Multilevel Networks of Online Hate

Yet despite social network advances in characterizing group structures in various forms of online harm, little work has been accomplished in the area of online hate specifically. In general, empirical scholarship on online hate examines it independently of the social structures within which it propagates. Volume-based metrics characterize the amount of online hate within a given online conversation, but they do not speak to its distribution within particular subsets of online accounts or types of group activity and organization. While it can be helpful to know when online hate flares up in general [147, 179], especially when it goes viral throughout the social media platform, it is also critically informative to know how and when such fluctuations are diffuse, concentrated, or organized. It is also vital to understand how online hate might thrive within specific group structures, marking some communities as more vulnerable—or conversely, more resilient—than others.

While underexplored, recent developments in network approaches to online hate have shown promise. Empirical analyses of “hidden ecologies of hate” have shown their dynamic ability to adapt to take-down strategies by regulatory bodies, in part driven by their

resilient networked structures [123]. Other research has looked into how clustering of hate networks by ideology has contributed to the reinforcement of racist communities on social media [29], and how such online groups utilize specific cross-platform strategies to propagate their messages [197]. Critical nodes have also been identified in the networked spread of online hate messages, suggesting that the targeted removal of some accounts might be more impactful than others in disrupting the flow of hateful information online [8]. Finally, some past work has demonstrated the importance of linking online network structures to the offline politics a given hate community [41].

Building on these past approaches, this chapter proposes a multilevel framework for analyzing online hate networks [71]. Multilevel in this context refers to the joint examination of group-level and account-level dynamics. The foregoing scholarship has tended to look at these processes in relative isolation from each other. However, it is crucial not only to understand how group properties might influence the propagation of online hate, but also within these groups, whether and how individuals participate in varied amounts.

4.3 Data and Methods

4.3.1 Dataset

Online conversations around the COVID-19 pandemic were collected using Twitter’s REST application programming interface (API). Search terms were specified to obtain tweets related to the pandemic in the Philippines and the US. Both countries respectively used #COVID19PH and #COVID19US as localized hashtags for discussing the disease. Data collection for this study lasted from March 5 to May 19 of 2020 over a period of 75 days in total. In the US, this end date corresponded to a week before the #BlackLivesMatter protests.

Each dataset was stored in JSON format with user metadata, tweet metadata, and data on the interactions between users in the form of retweets, replies, quotes, and mentions. Data for each tweet also contained information about the hashtags and URLs it used. At the end of data collection, a total of 15 million tweets representing 1 million users was collected for the Philippines. For the US, a dataset of 12 million tweets representing 1.6 million users was obtained.

Both the US and the Philippines faced particular challenges in curbing COVID-19 outbreaks. At the time of analysis, the US had the largest cumulative number of confirmed cases in the world, and the Philippines the largest for Southeast Asia. Additionally, both the US [213] and the Philippines [174] were noted for their contexts of political polarization under populist leadership. These factors contribute to additional difficulties in pandemic management, but also potentially constituted conditions for exacerbated social conflicts [208].

But while these two countries shared common political features, they are also featured unique contexts for the spread of hate speech. For instance, we note that US-specific concerns with racism have vastly outlived the COVID-19 pandemic [1, 63]. Hate speech in the US may thus hinge on historical racial divisions on top of more recent spikes in

international tensions with China [25, 109]. On the other hand, the Philippines’ geographic proximity to China, and its recent history of territorial disputes with the Asian superpower, may also push particular forms of sinophobic discourse on digital platforms [175, 188].

While these past studies do not strictly suggest predictive hypotheses for the dynamics of hate speech examined here, they contextualize potential reasons for the emergence of online hate speech during the COVID-19 pandemic, especially with respect to racial groups or political figures. Other countries may feature distinct dynamics throughout the course of the pandemic as shaped by their own local societal conflicts.

4.3.2 Measures

Table 4.1 summarizes the measures involved in this chapter for analyzing online hate networks. Following the multi-level perspective adopted in this work, the features described are divided into (a) group-level features which refer to structural features of a given cluster of accounts, and (b) individual-level features which refer to properties of single social media accounts. All measurements are either scraped directly from Twitter, estimated by the online hate model [251], or computed with the Netmapper and ORA software [9, 54].

4.3.3 Group-Level Analysis

To examine the group behavior in the online conversation, social media data was represented in terms of social networks. For a given day $t \in \{1, 2, \dots, 75\}$, let $G_t = (V_t, E_t)$ be the graph representation of the online conversation. Here, V_t corresponds to the set of users in the data, represented as the set of vertices in the graph. Meanwhile, E_t represents a set of weighted, directed edges between vertices in V_t . The weight of each directed edge is given by the number of interactions originating from the source node toward the target node. To obtain edge weights, we take the sum of all forms of Twitter communication, including retweets, replies, mentions, and quotes. The ORA software was used to perform all network analysis [54].

Group Identification Community detection was performed to operationalize a localized understanding of online groups. I used a Leiden algorithm to automatically recover local clusters of users. The Leiden algorithm is an unsupervised method for community detection which iteratively refines cluster assignments with the intuitive goal of optimizing the difference between actual and expected number of edges within an assigned cluster [243]. It has been shown to be superior to the widely used Louvain algorithm by guaranteeing well-connected communities as well as faster runtime [30]. Thus, for each network snapshot, I obtained cluster assignments for all agents. Agents assigned to the same cluster were conceptualized as constituting a distinct group engaged in meaningful interaction about the pandemic. Note that for all succeeding analysis, I removed trivial clusters containing only one or two agents (i.e., isolates and pendants).

Targeting Analysis For each identified network cluster, I used the online hate detection model described in Chapter 2 to obtain an average measure of its level online hate [251]. I

Table 4.1: Multilevel summary of measures used to characterize online hate networks.

Level	Feature	Tool	Description
Group	Size	ORA	Number of agents in a given network cluster (log scale)
	Density	ORA	Number of realized in-cluster edges divided by number of possible in-cluster edges
	E/I Index	ORA	For a given cluster, let E be the number of edges between in-cluster agents and out-cluster agents and I be the number of in-cluster edges. Then $\frac{E - I}{E + I}$ is the E/I index [140].
	Bottlenecking	ORA	Control of out-group interactions by a small subset of the network cluster, measured by the Cheeger score [169]
	Group Targeting	Netmapper	Mean counts of various identity categories produced in messages by cluster members
Individual	Account Verified	Twitter	Whether the account was verified by Twitter at the time of data collection
	Account Followers	Twitter	Total number of followers of the account at the time of data collection
	Account Volume	Twitter	Total number of tweets produced by the account at time of data collection
	Hate Score	Model	Measurement of online hate by a given agent
	Individual Targeting	Netmapper	Identity terms mentioned by a given agent
	Internal Degree	ORA	Number of in-cluster edges for a given agent
	External In-Degree	ORA	Number of edges from out-cluster agents to a given agent
	External Out-Degree	ORA	Number of edges from a given agent to out-cluster agents
	Hate In-Connections	ORA	Mean hate score of accounts an individual interacts with
Hate Out-Connections	ORA	Mean hate score of accounts that interact with an individual	

also used the average Netmapper counts of identity categories mentioned by group members [126]. Using these two measurements, I constructed a slope-based index of group-level targeting of gendered, racial, political, and religious identities in online hate messages [250].

To define this index, consider time step t where a total of J_t clusters are identified. For a given cluster $j \in \{1, 2, \dots, J_t\}$, let h_j be the average hate score of its member accounts. For a given identity category d , let d_j be its average number of mentions of identities

belonging to that category among members of cluster j . The final index $T_{d,t}$ is given in Equation 4.1:

$$T_{d,t} = \frac{J_t \sum_j h_j d_j - \sum_j h_j \sum_j d_j}{J_t \sum_j h_j^2 - (\sum_j h_j)^2}. \quad (4.1)$$

A distinct measurement $T_{d,t}$ per identity category d was conducted for each time step t . This provided a semantic assessment of the contents of online hate to contextualize the structural features which form the central focus of this chapter.

Structural Analysis Finally, I considered the structural features of the identified clusters, following the hypothesis that these relate to localized levels of hate speech [136]. I was specifically interested in the following features. First, I measured cluster size, denoted by the number of unique agents assigned to the same cluster. To control for scale, I took the logarithm of cluster size in subsequent calculations. Second, I assessed density, a commonly assessed property of social networks that indicates the extent to which group members interact with each other, given all possible interactions [263]. Third, I looked at the E/I index [140]. The E/I index is a classical measure in network science which intuitively quantifies exclusive group communication [140]. Normalized between +1 and -1, higher values of the E/I index indicates high levels of communication with out-groups; lower levels suggest that the cluster communicates solely with in-group members. Low levels of the E/I index thus indicate some level of isolation from the rest of the social network. Fourth, I measured the Cheeger constant. This quantifies bottleneck behavior in the cluster, whereby a small subset of accounts control information flow [169].

4.3.4 Multilevel Analysis

To achieve the second research objective of this chapter, I performed a multilevel analysis which assessed the association between account-level measures and individual production of online hate, while controlling for cluster-level structural features.

Account Measures Three of the account-level measures are taken directly from Twitter, which quantify the account’s verified status, number of followers, and total number of tweets at the time of data collection. Using the online hate detection model in Chapter 2, an online hate score was also associated with each account [251]. Netmapper was also used to count the number of times each identity category was mentioned by each account.

Network measures were also incorporated on the individual level. Each account was associated with group-level measures describing their internal degree, external in-degree, and external out-degree. Internal degree refers to the number of connections a given account has with other accounts within the same network cluster. External in-degree refers to the number of incoming edges the account has from out-group accounts, while external out-degree refers to the number of outgoing edges the account has toward out-group accounts. These group-aware degree measures quantify the extent to which a given

account has localized influence, and controls the flow of information into or out from the group.

Combining network measures and online hate scores, I also produced additional measures which specifically had to do with an account’s exposure to other hateful accounts. By taking the average online hate score of nodes with incoming or outgoing edges, I measured the extent to which a given account interacts with or is interacted with by other hateful content.

Multilevel Model Combining account measures with group measures, I estimated the coefficients in Equation 4.2 to quantify their relationship with online hate.

$$\text{hate}_{ij} = \gamma_{00} + \sum_{q=1}^Q \gamma_{q0} x_{ijq} + \sum_{p=1}^P \gamma_{0(Q+p)} g_{jp} + U_{0j} + \sum_{q=1}^Q U_{qj} x_{ijq} + R_{ij} \quad (4.2)$$

In Equation 4.2, hate_{ij} is the target variable, the online hate score of given account i belonging to group j . The variables denoted by γ are the fixed effects estimated by this multilevel model. Here, γ_{00} is the overall fixed intercept. For the Q account-level variables, γ_{0q} is the corresponding fixed effect for account-level predictor $q \in \{1, 2, \dots, Q\}$. The value of this predictor for account i in group j is observed in the data denoted by x_{ijq} . For the P group-level variables, $\gamma_{0(Q+p)}$ is the corresponding fixed effect for group-level predictor $p \in \{1, 2, \dots, P\}$. These observations are denoted by g_{jp} . The U variables indicate the random effects, with U_{0j} specifically indicating the random intercept for cluster j and U_{qj} denoting the random slope for individual-level predictor q for cluster j . Finally R_{ij} denotes the errors.

Conceptually, this equation jointly estimates account-level and group-level associations with online hate. For a given account i belonging to a group j , its level of expressed hate is taken as a function of its account-level features such as its verified status, its individual network position, and its level of exposure to other hate messages; as well as its group-level features, such as how big or isolated its group is. This allows for a direct assessment of the extent to which one’s level of hate depends on both or either one’s own behavior, one’s interactions, and one’s communities.

For interpretability, the outcome variable of online hate was grand-mean centered. Individual account-level features were also grand-mean centered, while group-level features were centered by the average across groups. Estimations were conducted with maximum likelihood estimation and fitness was assessed with a battery of the Comparative Fit Index (CFI), the Tucker-Lewis index (TLI), the root mean square error (RMSEA), and the standardized root mean square residual (SRMR). The lavaan package was used on R Studio to perform multilevel model estimation.

4.4 Results

Structural analysis revealed key network dynamics associated with the spread and organization of hate speech surrounding the Covid-19 pandemic. Targeting analysis surfaced

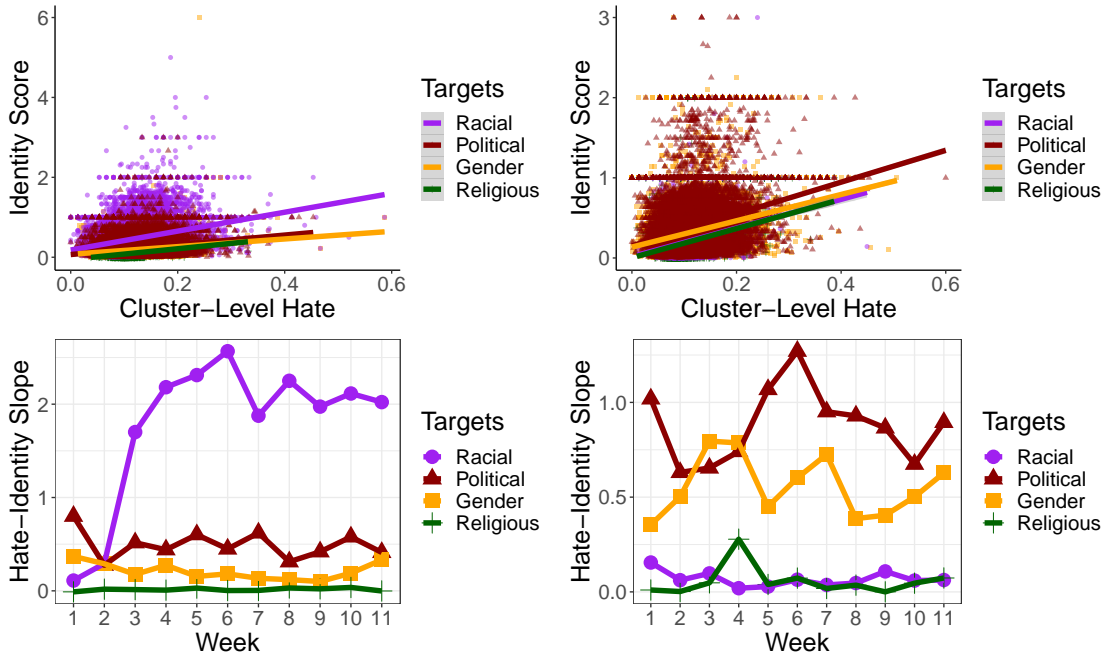


Figure 4.1: Group-level estimates of associations between hate scores and various identity targets.

the broad content of online hate in the US and the Philippines, indicating high-level distinctions in the types of online hate that spread during the health crisis. Addressing this chapter’s primary objectives, associations were estimated for group-level and account-level features in relation to online hate. Taken together, findings of the analysis suggest the importance of small, isolated, and hierarchical groups for the spread of online hate. In the context of these groups, the role of hate influencers and proselytizers is likewise underscored, as well as the power of consistent engagement with hateful content in one’s social network.

4.4.1 Group Targets of Online Hate

Targeting analysis quantifies the identities most associated with online hate in online conversations about the Covid-19 pandemic in the US and the Philippines. Figure 4.1 shows how different identity categories were dominant across societal contexts. However, in general, higher levels of hate are positively associated with more frequent invocations of all identity categories [251]. This is consistent with our analysis of online hate in the megacorporus in Chapter 2.

However, in the US, the slope is steepest for racial identities. This indicates that across all time periods, online hate is most directed toward racial identities. On the other hand, political identities have the steepest slope in the Philippines. Disaggregated over time, racial associations with online hate remain consistently dominant throughout the period examined in the US, whereas political associations generally prevail in the

Philippines. Gender associations with online hate also uniquely fluctuate to high levels in the Philippines. This suggests some notable differences in the content of online hate for the two countries during the global crisis.

4.4.2 Structural Features of Online Hate Networks

Results of group-level analysis are visualized in Figure 4.2 and summarized in Table 4.2. Estimate distributions are visualized to show uncertainty intervals associated with each group-level predictor of online hate, with peaked distributions indicating greater certainty and flatter distributions showing greater variability.

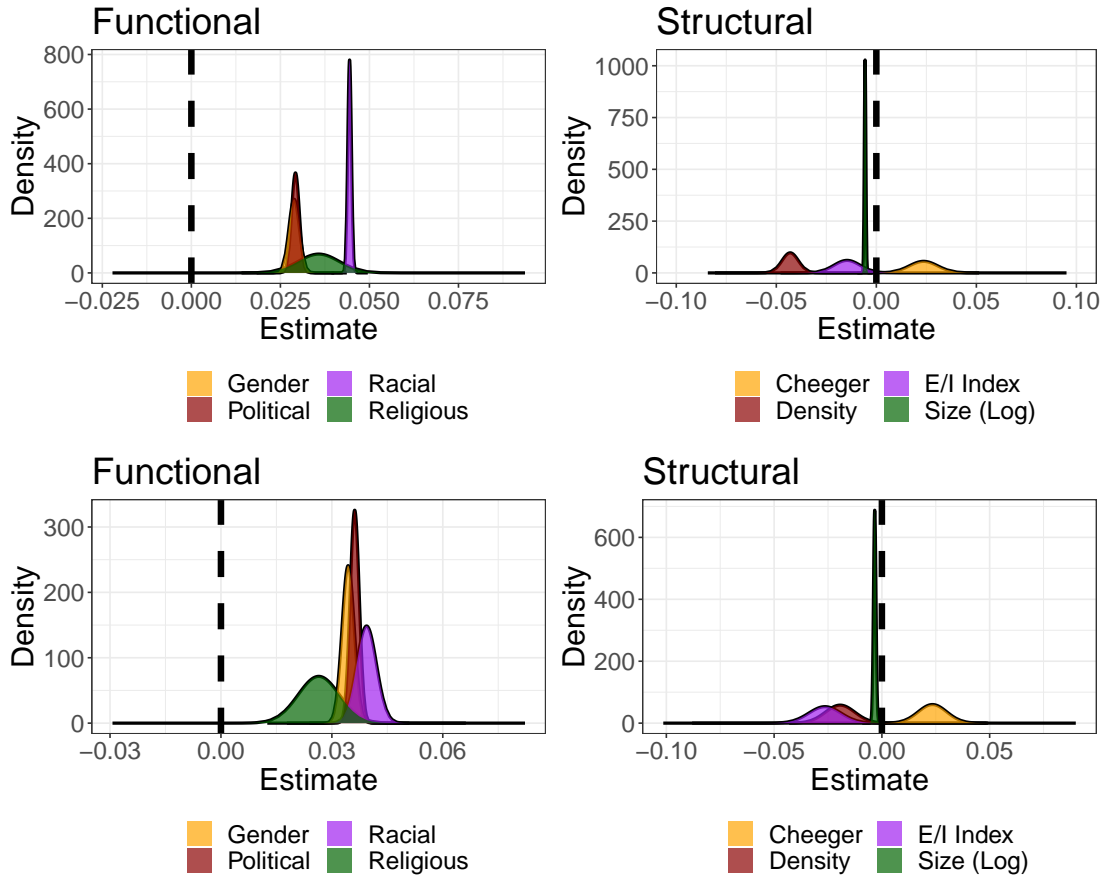


Figure 4.2: Distribution of estimated associations for group targeting and structural features of online hate networks.

Despite the broad functional differences in the online hate expressed in the US and the Philippines, analysis of group-level features suggests consistent directions in the effects of network structures. Higher levels of community-level hate are consistently predicted by clusters having smaller size (US: $-0.006, p < .001$, PH: $-0.003, p < .01$), lower density (US: $-0.043, p < .01$, PH: $-0.026, p < .01$), lower E/I index (US: $-0.015, p < .01$, PH: $-0.026, p < .01$), and higher Cheeger scores (US: $0.023, p < .01$, PH: $0.023, p < .01$). Col-

Table 4.2: Summary of group-level predictors of online hate.

Category	Predictor	Estimate (US)	Estimate (PH)
Structure	Size (Log)	-0.006 (0.000)***	-0.003 (0.001)**
	Density	-0.043 (0.004)***	-0.026 (0.007)***
	E/I index	-0.015 (0.007)*	-0.026 (0.007)***
	Cheeger score	0.023 (0.007)***	0.023 (0.007)***
Targeting	Racial	0.044 (0.001)***	0.039 (0.003)***
	Political	0.029 (0.001)***	0.036 (0.001)***
	Gender	0.029 (0.001)***	0.034 (0.002)***
	Religious	0.036 (0.006)***	0.026 (0.006)***

Note: * $p < .05$, ** $p < .01$, *** $p < .001$

lectively, these suggest that smaller communities featuring more isolated and hierarchical interactions are more likely to feature higher levels of hate over time.

All identity categories are also positively associated with higher levels of community hate. Higher community-level invocation of racial (US: 0.044, $p < .01$, PH: 0.039, $p < .01$), political (US: 0.029, $p < .01$, PH: 0.036, $p < .01$), gender (US: 0.029, $p < .01$, PH: 0.034, $p < .01$), and religious (US: 0.036, $p < .01$, PH: 0.026, $p < 0.01$) identities all had distributions concentrated on positive values. In the US, as expected, the mean value of the race distribution had the highest value, with very little variance. This reflects the extremely charged nature of racialized discourse in the US COVID-19 conversation. Interestingly, the mean of the race distribution also had the highest value in the Philippines. This indicates that when structural features and other identities are controlled for, race still matters in online hate speech in the Philippines.

4.4.3 Key Actors in Online Hate Networks

Finally, results of the multilevel modeling estimation are visualized in Figure 4.3 and summarized in Table 4.3. Here, results are less universally consistent across predictors between the US and the Philippines, but key similarities and differences yield striking insights.

First, generally it appears that account-level features are inconsistently predictive of the production of online hate. Overall, there is a slight negative relationship between how large an account is, as indicated by its verified status, number of followers, and total number of tweets. In the US and the Philippines, extremely active accounts with high volumes of tweets are both less likely to express online hate (US: -0.024 , $p < .01$, PH: -0.026 , $p < 0.01$). Interestingly, while verified status—typically granted to notable leaders, organizations, or personalities—does make online hate less likely in the Philippines (-0.021 , $p < .01$), it does not have a significant effect in the US. Total number of followers is also not significant in either context. Taken together, these patterns are likely to emerge because, especially for more active accounts, engagement in hateful online content might result in suspension, preventing accounts from getting large in the first place. Con-

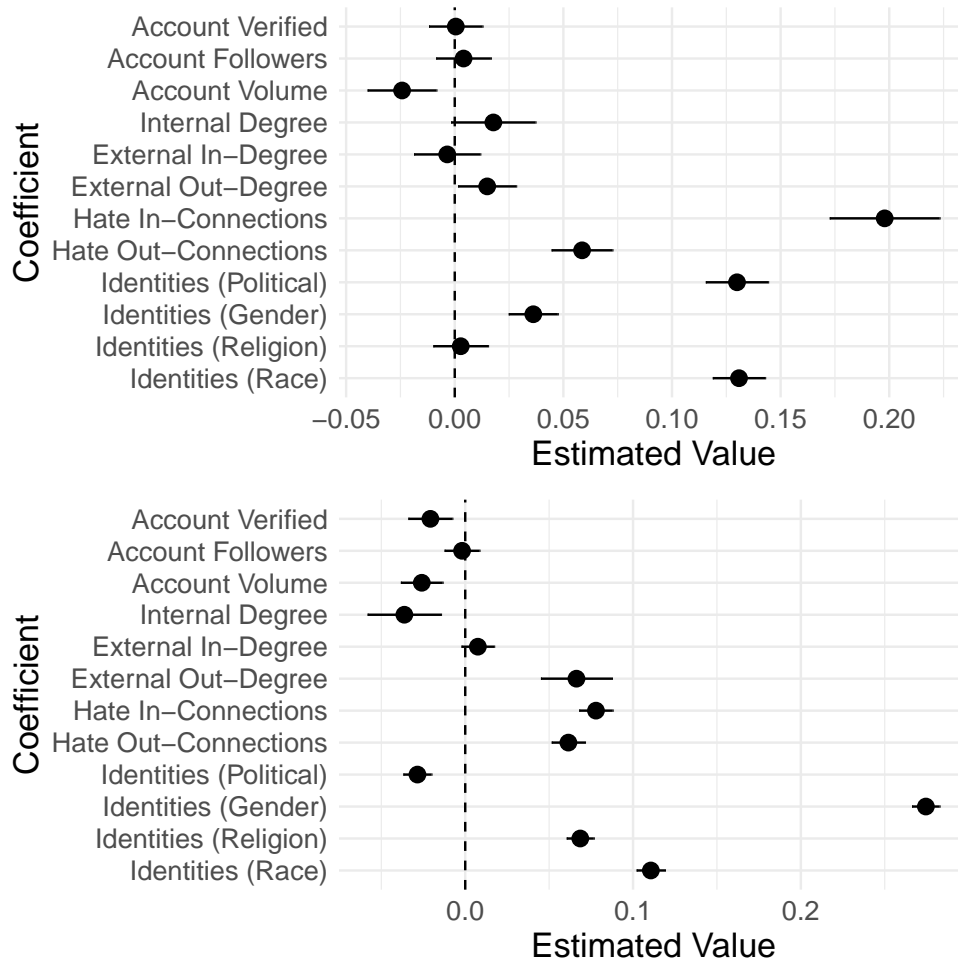


Figure 4.3: Individual-level predictors of online hate in group context. **Top:** Estimates for the United States. **Bottom:** Estimates for the Philippines.

versely, whereas in some contexts being a high-profile account—though not necessarily an active one—might mitigate the likelihood of spreading online hate, certain political settings may reward such expressions with further engagement from one’s followers [34, 35]. Thus, online hate may emerge from both low-profile and high-profile accounts, indicating that anonymity is not exclusively an enabler of online hate during a crisis like the Covid-19 pandemic.

In terms of network-based features, some features also showed mixed results while others demonstrated meaningful consistency. For the US, higher levels of online hate are predicted for accounts with high internal degree (0.018, $p < .10$), high external out-degree (0.015, $p < .01$), high hate in-connections (0.198, $p < .05$) and high hate out-connections (0.059, $p < .01$). This means that, controlling for group-level structures, the most hateful accounts are those which interact in high volumes with their in-group members, act as proselytizers by having their messages propagate to the out-group, and both interact with and are interacted with by other hateful accounts. Meanwhile, in the Philippines, higher levels of

Table 4.3: Summary of estimated individual-level predictors of online hate.

Predictor	Estimate (US)	Estimate (PH)
Account Verified	0.000 (0.006)	-0.021 (0.007)**
Account Followers	0.004 (0.006)	-0.002 (0.005)
Account Volume	-0.024 (0.008)**	-0.026 (0.006)***
Internal Degree	0.018 (0.010) ⁺	-0.036 (0.011)***
External In-Degree	-0.003 (0.008)	0.008 (0.005)
External Out-Degree	0.015 (0.007)*	0.066 (0.011)***
Hate In-Connections	0.198 (0.013)***	0.078 (0.005)***
Hate Out-Connections	0.059 (0.007)***	0.061 (0.005)***
Identities (Political)	0.130 (0.007)***	-0.028 (0.004)***
Identities (Gender)	0.036 (0.006)***	0.275 (0.004)***
Identities (Religion)	0.003 (0.006)	0.069 (0.004)***
Identities (Race)	0.131 (0.006)***	0.111 (0.004)***

Note: * $p < .05$, ** $p < .01$, *** $p < .001$

online hate are associated with lower internal degree ($-0.036, p < .01$), higher external out-degree ($0.066, p < .05$), higher hate in-connections ($0.078, p < .001$), and higher hate out-connections ($0.061, p < .001$). In this case, then, online hate is not necessarily propagated among the most active accounts within a given group. But like in the US, it is still higher among proselytizers and those who interact with and are interacted with by other hateful content. Thus, it is not just how influential an account is given its volume of interactions, but also the types of information it engages with through these interactions that is predictive of online hate.

Lastly, it is worth noting that, similar to the above analyses, individual accounts with the highest levels of online hate are also those which mention their target identities the most. Interestingly, upon controlling for group-level and account-level features, the predictive association between online hate and political identities ($0.130, p < .01$) approaches the level of racial identities ($0.131, p < .01$) in the US. Meanwhile, in the Philippines, gender ($0.275, p < .01$) outpaces race ($0.111, p < .01$) and politics ($-0.028, p < .01$).

4.5 Discussion

In this chapter, I described a framework for analyzing online hate networks. This shifted from viewing online hate in terms of isolated texts and individuals toward understanding their explicit organization in communities [58, 75]. By analyzing group-level hate in relation to the structural features of network clusters, I showed key relationships between particular patterns of community interaction and the collective production of online hate. Furthermore, through the application of a multilevel model, I demonstrated how individuals within these groups also varied in their levels of online hate depending on their position in the network. Application of these techniques to online conversations about the Covid-19

pandemic in two different societal contexts moreover indicated consistent patterns pointing to meaningful social processes meriting further investigation in future work.

Group-level analysis highlighted the importance of small size, low density, high isolation, and high bottlenecking in the organization of online hate networks. Despite differences in the content of online hate in the US and the Philippines, these shared structural features resonate with existing scholarship on group dynamics. Relatively small and isolated groups limit information flow to a select few individuals and ideas, allowing for the entrenchment of more extreme beliefs outside independent mitigating influences. This is further supported by bottlenecking behaviors, which suggest that only a handful of community members engage in communication with out-group members while most others remain sequestered within the hate group. While some of these relationships are reminiscent of commonly studied echo chamber structures [18, 273], online hate groups are also distinct from echo chambers per se by emphasizing the role of hate influencers and proselytizers who engage group members and direct the flow of information in the group [136, 209].

As for the multilevel analysis, results were less consistent between the two societal contexts, but they collectively showed the importance of interacting with high-hate accounts for predicting one’s own level of hate. For the US, in particular, localized influence within one’s group was predictive of higher levels of hate; for both national contexts, high levels of hate were also detected among those that out-group members most interacted with. Both these findings further reinforce the role of influencers and proselytizers in the networked spread of online hate, as they tend to be more engaged with in-group members and their information also actively reaches out-group members. Finally, it is worth noting that high-hate accounts tend not to be high-profile personalities, but this was not on its own one of the strongest predictors, indicating that network position and exposure had stronger relationships with levels of online hate, and that some influential social media accounts may nonetheless feature meaningful levels of online hate.

Within the context of the Covid-19 pandemic, these findings bear special implications for crisis situations. As Covid-19 infections spread throughout the world, the spread of hate and social division followed in their wake [109, 227, 244]. But the expression of such hostilities online, while pervasive, was also organized into networked communities. For the pandemic and beyond, findings from this work point to the importance of attending to how these groups are organized and the roles individuals play within these groups. Early detection of online hate groups, as well as timely identification of central hate leaders, may be critical for mitigating the spread of online hate and its attendant consequences for exacerbating an already overwhelming public health situation.

The techniques employed in this chapter are all language-agnostic and platform-agnostic, provided that text and network data are both available. While community detection can be a computationally expensive procedure [30, 243], it can be undertaken with reasonable efficiency over larger-scale conversations given theoretically meaningful network snapshots of the data over time [151]. While the structural properties of online hate groups featured key similarities between societal contexts examined here, it is not the goal of this chapter to establish these general relationships. More broadly, from an operational standpoint, the proposed analytical pipeline could be used to assess various online conversations to detect when online hate networks feature—or do not feature—the structural signatures

observed here. Specific hate groups could also be selected based on their relative position in the distributional space of the assessed structural features; using the measures identified here, specific individuals could also likewise be selected for further analysis based on their network position relative to their in-groups and out-groups. Such drill-down analyses may be of particular value in social cybersecurity contexts where known groups or individuals are of particular analytical interest.

At this juncture, it is important to situate the insights above within the limitations of this work. Because the analysis relied upon social media data, it is constrained by the quality of the text and networks it is able to extract from the platform [176]. Relatedly, while Twitter served as an important test bed in this work, the design of different social media platforms—such as Reddit and TikTok—may necessitate different types of network measures on the group and individual level. Furthermore, while the proposed analytical pipeline considered several theoretically motivated measures of group-level structure and individual-level network position, a wealth of other measures have also been developed in network science that future work may incorporate into this framework. That said, the flexibility of the analytical pipeline allows for the ready incorporation of new group-level and individual-level measures depending on the demands of unique application cases or conceptual questions. Lastly, while network effects were detected statistically over a large-scale observational dataset, they do not speak to causal effects or their underlying socio-psychological mechanisms. Future work could extend these results significantly through innovative experimental and field studies to pin such processes down [177].

Chapter 5

Manipulation of Online Hate

5.1 Introduction

Online hate on its own constitutes a significant form of online harm. When it arises organically out of extant social conflicts, it can produce intense expressions of intergroup antagonism which in turn may trigger offline behavioral manifestations of prejudice and violence [190]. Such expressions, moreover, not only cause psychological and potentially physical harm to its targeted victims, but may also further radicalize others who are exposed to it and find meaning in its message [26, 209, 224]. Online hate thus not only reflects existing social divisions, but may also further perpetuate them in a vicious cycle [189, 208].

Precisely due to its intimate connections with social order, online hate is also a major area of vulnerability for manipulation on cyberspace. Recent years have demonstrated the fertility of social media platforms for social-cyber actors to participate in and influence online conversations to achieve various strategic objectives. Employed in both domestic and international settings, information operations have been documented in the context of elections, natural disasters, delicate diplomatic junctures, and the Covid-19 pandemic [247, 249, 252]. Through the use of automated accounts known as bots or the concerted disruption instigated by troll accounts [23, 255], information operations have been linked to shifts in public opinion, the spread of misinformation, heightened polarization, and the broader decay of the social fabric. It is against this backdrop that I investigate their involvement in the dynamics of online hate.

But how specifically might online hate be manipulated? Drawing upon the earlier chapters in this thesis, this chapter proposes a narrative and network view of online hate manipulation. Conceptually, I draw upon the BEND framework of social-cyber maneuvers to link between the growing literature on information operations and the propositions presented in my foregoing work around the organization of online hate into systems of meaning and systems of people. Methodologically, I also introduce manipulation-based extensions of my proposed narrative and network analytical pipelines to examine the involvement of social-cyber actors in their production and propagation. Taken together, I aim to produce a novel computational social science framework for assessing the manipulation of online

hate narratives and networks.

In this chapter, I therefore specifically set out to accomplish the following objectives. First, for the manipulation of online hate narratives, I introduce new variables into a supervised topic model to account for the production of texts by social-cyber actors, particularly focusing on bots and trolls. I also showcase the use of lagged Poisson regression models to track the agenda-setting relationships between the online hate narratives produced by humans, bots, and trolls. Second, for the manipulation of online hate networks, I likewise introduce new variables into a multilevel model to account for individual-level and group-level exposure to the messages of bots and trolls. These measures map out how bots and trolls shape the organization of communities built around online hate. Finally, I demonstrate an application of these techniques to a case study of the online conversation around the Covid-19 pandemic in the United States. I conclude with this work’s theoretical and methodological implications.

5.2 Related Work

5.2.1 Information Operation Maneuvers

While information operations have been broadly utilized to achieve a diverse range of strategic objectives, they are accomplished on an operational level through specific and discrete actions on social media platforms. Such actions are described operationally in what is known as the BEND framework [24]. The BEND framework defines sixteen types of social-cyber maneuver along two orthogonal dimensions: positive and negative maneuvers, and narrative and network maneuvers [27, 53]. Positive narrative maneuvers are actions taken to provide additional information about or support a given message. Negative narrative maneuvers, meanwhile, are those which seek to distract from or oppose a given narrative. Among the network maneuvers, positive actions include those which build up a network, bridge networks together, or boost the perceived influence of a given actor in a network. Negative actions, on the other hand, are those which take groups apart, obliterate them entirely, or reduce an actor’s influence to irrelevancy.

From this perspective, information operations have been studied in many contexts. Perhaps most notably, in electoral settings, information operations have been documented in the spread of messages to support or oppose a political candidate, to polarize the public, and to spread falsehoods to disrupt the elections themselves [80, 247, 248]. During other forms of political upheaval, as with the #BlackLivesMatter protests, information operations took place to infiltrate activist groups [13, 231]. Such infiltration tactics served to reduce group cohesion within the movement, as well as to worsen public perceptions of racial justice activists from the outsider perspective. During moments of international tension, information operations have been deployed to weaken alliances and heighten perceptions of outside threat [252]. Information operations have also been identified in longer-term divisions, such as climate change debates, serving to push individuals to fringe information sources and increase hostilities between climate change believers and deniers [57, 246]. Throughout these diverse cases, shifts in narratives and networks have been instrumental

in manipulating the state of online conversations, with broader implications for the offline state of society.

5.2.2 Social-Cyber Actors

Information operations are specifically carried out by a set of agents on social media [52, 53]. Social-cyber actions can be performed by human accounts and even verified personalities, such as when a state leader makes an influential statement to support or oppose a cause. Their authentic followers may also engage in these actions as in organic forms of collective action. But influence campaigns are also supported by a wide array of inauthentic social-cyber actors that act as force multipliers for narrative and network influence [24, 27]. Social cybersecurity has documented a wide variety of such malign actors under a range of names which vary by function and social context, making up a menagerie of buzzers, sockpuppets, cyborgs, and astroturfers, among others. Two specific types of social-cyber actors, however, have been established as conceptually coherent and relatively context-independent in their operational properties: bots and trolls [80, 255, 275]. It is these two types of social-cyber actor which this chapter focuses on.

Bots are broadly defined as social media accounts which operate in a mostly or entirely automated fashion [80]. Leveraging the historically low barrier to entry in most social media platforms, large quantities of bot accounts have been shown to persist on various websites [64]. Moreover, while operating with relatively simple directives, the use of automated accounts is well-suited to the fast-paced propagation of pre-defined messages or engaging with social media users in a pre-designed manner. Bots have been well-studied in terms of their engagement in social-cyber maneuvers, as their automated features allow for widespread positive or negative messaging, as well as precisely coordinated group behavior [23, 181]. It is also worth noting that while bots have tended to be studied in terms of their overtly harmful effects, many automated accounts are also used for more benign purposes, such as retweeting reliable information (e.g., news updates or weather forecasts) or accomplishing niche community functions (e.g., tweeting photos of celebrities) [182, 183].

Trolls, on the other hand, are accounts specifically engaged in disruptive behavior on the internet [255, 275]. Unlike bots, which are defined by their automated programming, trolls are known by the behavior that they engage in, which includes interacting with others using offensive language (i.e., flaming) and derailing conversations with persistent non-sequitur [11]. Interestingly, trolls and bots are not mutually exclusive, so bots may also engage in trolling behavior. At the same time, because trolls may be human, they have also been associated with particular psychological traits that may predispose individuals to engage in trolling behavior [39, 170, 180]. From this standpoint, it is useful to note as well that unlike overt hate speech, trolling can be considerably more ambiguous in whether or not it violates platform regulations [255]. Trolls do not necessarily attack a protected social group or directly incite violence, but such behaviors may nevertheless promote hostility between individuals and groups, and generally lower the quality of online conversations.

Taken together, while these behaviors are what allow for the negative and large-scale influence of bots and trolls in information operations, they also facilitate their computational detection. Rich scholarship abounds around detecting social-cyber actors, specifically uti-

lizing features of their manipulative actions to detect their presence in a given online conversation [64, 80]. Bots, for instance, can be detected by using individual account-level features such as their volume of tweets relative to their age, their circadian rhythms relative to an average human's, and their use of repetitive messages and platform functions [23, 182]. Network-level features have also proved useful for bot detection, as bots are capable of engaging in unique forms of coordination beyond human norms [23, 182].

Trolls, on the other hand, have been empirically linked to the use of specific psycholinguistic features, such as simpler words, more named entities, and abusive (though not necessarily hateful) language [255, 275]. Accounting for dyadic relational activity has also been valuable for troll detection, as non-sequitur interactions can be leveraged by examining the relationship between troll messages and the messages they interact with. Leveraging these features, models such as BotHunter, BotBuster, and Botometer are some of the well-established bot detection tools in the literature [23, 182, 272]. TrollHunter has also been shown to reliably detect troll accounts in an accurate and conceptually coherent manner independent of hate speech, bot-like activity, and state-sponsored coordination [255].

5.2.3 Manipulation of Online Hate

This chapter is focused on assessing how bots and trolls shape online hate. As the foregoing scholarship on information operations suggests, the manipulation of online hate may take place at both the narrative and network level [52, 53]. More specifically, bots and trolls may be involved in these manipulative actions by contributing to the propagation of certain hateful messages over others, and by engaging in patterns of interaction to alter the structure of online hate groups.

Previous studies specifically investigating the manipulation of online hate are relatively scarce. A vast majority of the literature tends to presume the organic nature of online hate, or while acknowledging possibilities of manipulation, do not explicitly seek to estimate the extent of its effects or its particular qualities. Some studies have sought to estimate the prevalence of inorganic accounts engaged in the production of online hate messages [113, 276]. Others have also built tools trained to specifically classify inorganic accounts based on their production of hateful content [5, 17]. However, less work in this area has characterized manipulative activities beyond their sheer volume or aggregated individual actors.

While relatively under-explored, studies of the manipulation of online hate narratives and networks has largely focused on the broad identification of who or what groups are targeted by online hate messages produced by malign actors. During the Covid-19 pandemic, for instance, a study of online conversations about racism showed that bots were more likely to shift hateful messages toward the upcoming elections in the United States, entailing a shift from racial to political narratives [253]. In pre-pandemic vaccine debates, trolls were also shown to tweet more about vaccines than average social media accounts, and tended to express more hateful sentiments especially against scientists and political leaders [38]. At the time of Donald Trump's impeachment in early 2020, bots were likewise shown to contribute to the negative tone of the debate and to generally amplify right-

wing sources supportive of Trump while derogatorily attacking Democrats [95]. While these studies do begin to address the content of online hate, they do not explicitly surface deeper narrative characteristics of these messages, thus going into little detail about how hate targets were attacked. Similarly, the manipulation of online hate networks has been limited to individual-level actions of amplification, without accounting for a broader range of group-level behaviors [52, 53].

5.3 Data and Methods

5.3.1 Dataset and Measures

This chapter once again makes use of the national-level online conversation about the Covid-19 pandemic in the United States [249, 250]. This included 12 million tweets representing 1.6 million users over the 75-day period beginning from the declaration of Covid-19 as a global pandemic up to the beginning of the #BlackLivesMatter protests.

Table 5.1 summarizes the measures which are obtained on the dataset. As in Chapter 4, several theoretically motivated network properties are assessed including group size, density, the E/I index, and hierarchical bottlenecking behavior [140, 169]. Individual-level social media features such as an account’s verification status, number of followers, and total volume of tweets at the time of data collection are likewise recorded. Using the model described in Chapter 2, each account is associated with a score indicating the level of online hate it expresses [251]. The average levels of online hate expressed by the accounts that one interacts with and is interacted with by are also calculated. Levels of gendered, political, racial, and religious identities mentioned by individual accounts and on average by group are likewise measured [54, 126, 241].

Distinct to this chapter, I add new measures of bot and troll scores for each account. Bot scores are obtained using the BotHunter model, while troll scores are estimated with the TrollHunter model [23, 255]. In addition to individual-level measures, I also calculate additional indices which quantify an account’s exposure to bot and troll content, and specifically hateful content by bots and trolls. For the former quantity, the average bot and troll scores of the accounts one interacts with and is interacted with by are measured. For the latter, I take the product of each individual account’s bot scores and hate scores, as well as the product of their troll scores and hate scores. These respectively capture the interaction effect between the likelihood that one is a bot and that one expresses hate, and the interaction effect between the likelihood that one is a troll and that one expresses hate. Taking these products, I then also compute their mean values for all the accounts that one interacts with and is interacted with by. This measures that extent to which one is exposed to content that is both produced by social-cyber actors and specifically contains hate.

Table 5.1: Multilevel measures of bot and troll manipulation of online hate networks.

Level	Feature	Tool	Description
Group	Size	ORA	Number of agents in a given network cluster (log scale)
	Density	ORA	Number of realized in-cluster edges divided by number of possible in-cluster edges
	E/I Index	ORA	For a given cluster, let E be the number of edges between in-cluster agents and out-cluster agents and I be the number of in-cluster edges. Then $\frac{E - I}{E + I}$ is the E/I index [140].
	Bottlenecking	ORA	Control of out-group interactions by a small subset of the network cluster, measured by the Cheeger score
Individual	Account Verified	Twitter	Whether the account was verified by Twitter at the time of data collection
	Account Followers	Twitter	Total number of followers of the account at the time of data collection
	Account Volume	Twitter	Total number of tweets produced by the account at time of data collection
	Hate Score	Model	Measurement of online hate by a given agent
	Individual Targeting	Netmapper	Identity terms mentioned by a given agent
	Internal Degree	ORA	Number of in-cluster edges for a given agent
	External In-Degree	ORA	Number of edges from out-cluster agents to a given agent
	External Out-Degree	ORA	Number of edges from a given agent to out-cluster agents
	Hate In-Connections	ORA	Mean hate score of accounts one interacts with
	Hate Out-Connections	ORA	Mean hate score of accounts one is interacted with by
	Bot Score	BotHunter	Likelihood that a given agent is a bot
	Troll Score	TrollHunter	Likelihood that a given agent is a troll
	Bot In-Connections	ORA	Mean bot score of accounts one interacts with
	Bot Out-Connections	ORA	Mean bot score of accounts one is interacted with by
	Troll In-Connections	ORA	Mean troll score of accounts one interacts with
	Troll Out-Connections	ORA	Mean troll score of accounts one is interacted with by
	Bot-Hate In-Connections	ORA	Mean product of bot and hate scores of accounts one interacts with
	Bot-Hate Out-Connections	ORA	Mean product of bot and hate scores of accounts one is interacted with by
	Troll-Hate In-Connections	ORA	Mean product of troll and hate scores of accounts one interacts with
	Troll-Hate Out-Connections	ORA	Mean product of troll and hate scores of accounts one is interacted with by

5.3.2 Manipulation Assessment

To quantitatively capture the assessment of online hate narratives and networks, I extend the methods described in Chapter 3 and 4, respectively. By incorporating measures of bot and troll activity, I specifically capture the extent to which a given narrative is propagated by bots and trolls as opposed to human accounts, and then subsequently quantify the extent to which bot and troll narratives lead ahead of or lag behind human narratives. On the network side, I capture the extent to which bots and trolls are associated with various group-level structures and multi-level networked patterns of interaction in the production of online hate.

Narrative Manipulation

To operationalize bot and troll involvement in the spread of online hate narratives, I augment a supervised topic model with measures of bot and troll activity. In Chapter 3, online hate scores were used as an outcome variable within a supervised topic model setup. Here, with bot and troll measures available, an additional random variable may be added to jointly capture their involvement in some topics over others. This is visualized in Figure 5.1.

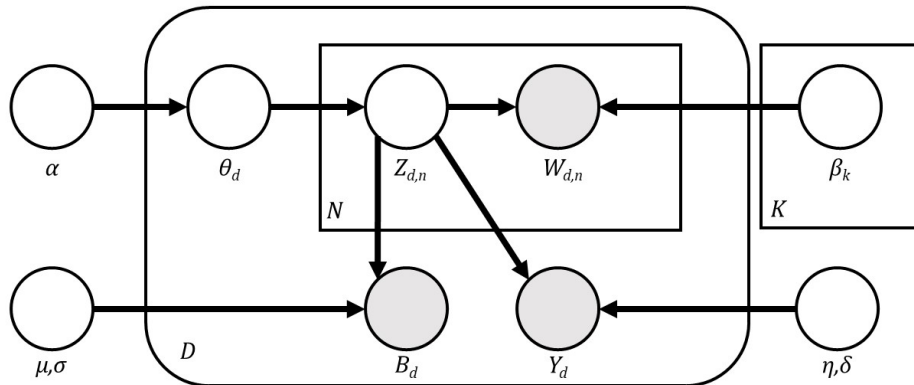


Figure 5.1: Plate diagram of the hate-aware and manipulation-aware supervised topic model.

Conceptually, Figure 5.1 shows how the extraction of topics is jointly influenced by their distribution of hate scores as well as the bot and troll scores of the agents which express them. This results in a supervised topic model that is both hate-aware as well as manipulation-aware by accounting for the salience of social-cyber actors. Probabilistically, the generative process for the supervised topic model is thus given as follows:

1. Sample the topic proportions θ_d for the documents $d \in \{1, 2, \dots, D\}$, conditioned on prior α . That is, $\theta \sim \text{Dir}(\alpha)$.
2. For a given word n out of a document d of length N , sample its topic assignments $z_{d,n}$, conditioned on θ . That is, $z \sim \text{Mult}(\theta)$.

3. Let β_k be the distribution of words for each topic $k \in \{1, 2, \dots, K\}$. Sample each word $w_{d,n}$ for the document d based on the drawn topic $z_{d,n}$. That is, $w \sim \text{Mult}(\beta_z)$.
4. Sample the response variable Y_d for document d based on the drawn topics z and parameters η, δ . In this case, Y_d corresponds to the estimated hate score for each document d . That is, $Y \sim \text{N}(\eta z, \delta^2)$.
5. Sample the response variable B_d for document d based on the drawn topics z and parameters μ, σ . In this case, B_d is a two-dimensional random vector corresponding to the estimated bot and troll scores for each document d . That is, $B \sim \text{N}(\mu z, \sigma^2)$.

Upon deriving the hate-aware and manipulation-aware supervised topics, semantic network analysis is applied to identify core concepts in the narratives of interest. As before, these are assessed by using each concept’s connectivity and evokability, concurrently measured by taking the geometric mean of each node’s degree centrality and betweenness centrality. Taken together, these measures capture the overall relevance of each concept to the online hate narrative and its ability to link together other concepts in the topic. Analytical labels are thus derived to characterize the systems of meaning associated with each topic. Parameter estimates of online hate, bot, and troll coefficients associated with each topic subsequently denote the level of hate expressed in each narrative, as well as the relative involvement of bots and trolls in each narrative’s component utterances in the online conversation.

Finally, utilizing a series of lagged Poisson regressions, I quantify the agenda-setting effects of bot and troll narratives over time. Agenda-setting measurements utilizing lagged time series have been utilized in describing whether lawmakers or citizens are more likely to shape public conversations around foreign policy, climate change, and congressional priorities [19, 148, 267]. Without establishing strict causal evidence, such measures show meaningful temporal associations between which set of actors may be shaping the broad discourse around a given social issue. Similarly, I leverage such techniques to examine the extent to which bot and troll accounts focus on some online hate narratives over others are predictive of their subsequent propagation by human accounts over time.

Network Manipulation

To assess bot and troll manipulation of online hate networks, I measure the association of individual-level and group-level exposure to bots and trolls with the structural features of online hate groups in predicting levels of online hate. In previous work, I had estimated the group-level network effects of bots on online hate groups by analyzing the interaction effect between mean bot scores and structural features of clustered communities [249]. I extend this to trolls by applying the same regression model with TrollHunter scores. I report both sets of findings below.

Next, to facilitate analysis of individual-level exposure, I also extended the multilevel model described in Chapter 4.

$$\begin{aligned} \text{hate}_{ij} = & \gamma_{00} + \sum_{q=1}^Q \gamma_{q0} x_{ijq} + \sum_{d=1}^D \delta_{d0} x_{ijd} + \sum_{p=1}^P \gamma_{0(Q+p)} g_{jp} + U_{0j} + \\ & \sum_{q=1}^Q U_{qj} x_{ijq} + \sum_{d=1}^D U_{dj} x_{ijd} + R_{ij} \end{aligned} \quad (5.1)$$

In Equation 5.1, hate_{ij} is the target variable, the online hate score of given account i belonging to group j . The variables denoted by γ are the fixed effects estimated by this multilevel model. Here, γ_{00} is the overall fixed intercept. For the Q account-level variables, γ_{0q} is the corresponding fixed effect for account-level predictor $q \in \{1, 2, \dots, Q\}$. The value of this predictor for account i in group j is observed in the data denoted by x_{ijq} . Manipulation is assessed using the D additional fixed effects on the individual level, with coefficients δ_{d0} estimated for the individual-level manipulation measures for $d \in \{1, 2, \dots, D\}$. For the P group-level variables, $\gamma_{0(Q+p)}$ is the corresponding fixed effect for group-level predictor $p \in \{1, 2, \dots, P\}$. These observations are denoted by g_{jp} . The U variables indicate the random effects, with U_{0j} specifically indicating the random intercept for cluster j , U_{qj} denoting the random slopes for individual-level predictors q for cluster j , and U_{dj} denoting the random slopes for manipulation-based predictors d for cluster j . Finally R_{ij} denotes model errors.

For interpretability, the outcome variable of online hate was grand-mean centered. Individual account-level features were also grand-mean centered, while group-level features were centered by the average across groups. Cluster-robust errors were estimated due to the networked nature of the data. The lavaan package was used on R Studio to perform multilevel model estimation¹.

5.4 Results

5.4.1 General Presence of Bots and Trolls

Throughout the dataset, a sizeable proportion of accounts were detected to be likely bots or trolls. Using a range of decision thresholds recommended for BotHunter [184], about 26.31-33.93% of the accounts for which data was collected were likely to be bots. Meanwhile, about 12.91% of the accounts were also likely to be trolls. Interestingly, despite their large prevalence in the Covid-19 conversation, bots ($r = 0.0020, p > .05$) and trolls ($r = -0.0055, p > .05$) had non-significant correlations with the spread of hate in the online Covid-19 conversation. Operationally, this indicates that many bots and trolls were not necessarily engaged in hateful activity; conversely, many humans also expressed hate in the dataset. As argued throughout this thesis, then, an individual-level assessment of online hate may not always be the most fruitful in characterizing its dynamics in an online conversation.

¹<https://lavaan.ugent.be/>

5.4.2 Bot and Troll Narratives

Results of the narrative manipulation analysis uncovered ten topics distinguished by their relative levels of expressed online hate and salience of bots, trolls, and human accounts. Figure 5.2 shows the aggregated outputs of the supervised topic model, featuring the estimated average levels of hate for each topic, the estimated levels of bot and troll activity, and the average targets of each extracted topic.

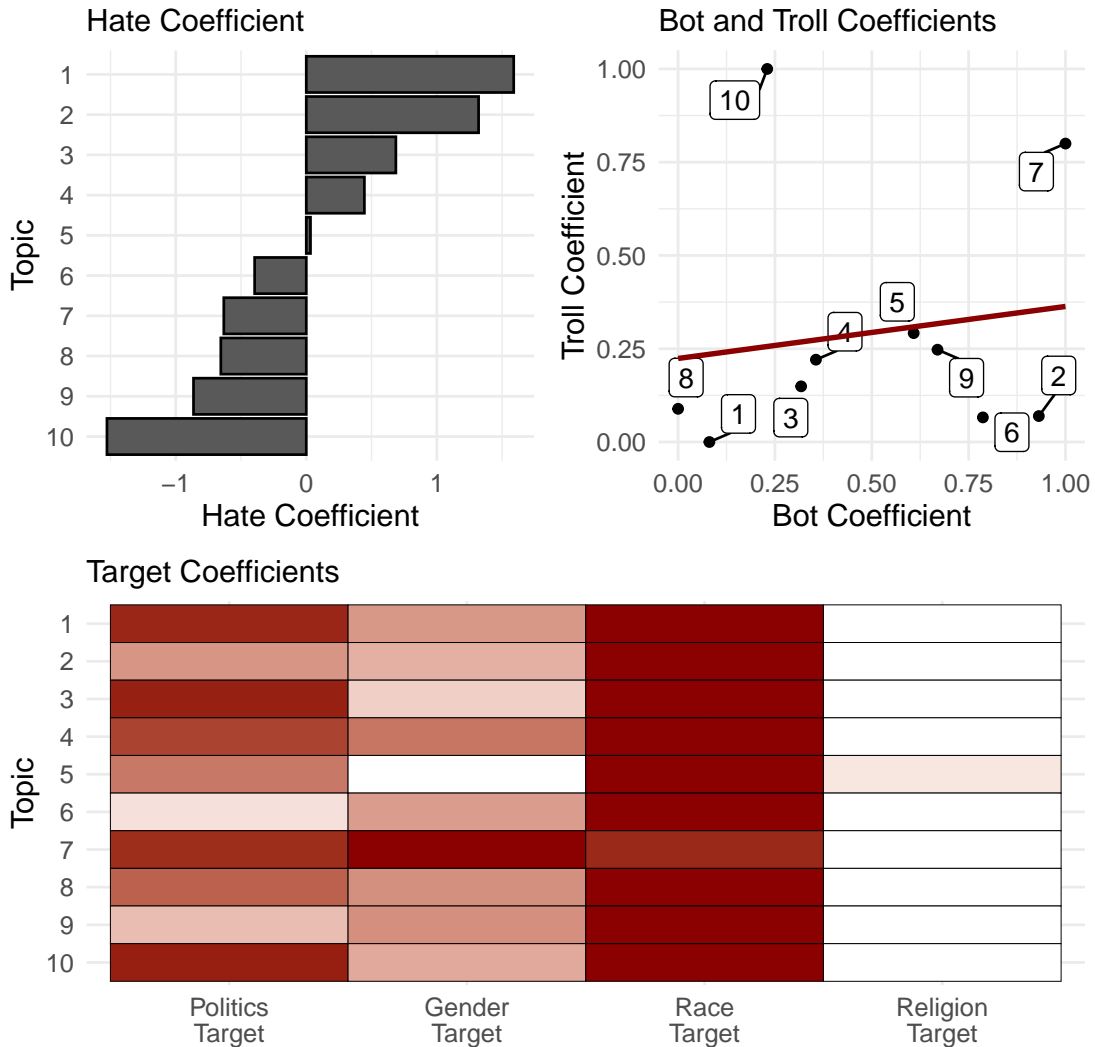


Figure 5.2: Outputs of supervised topic model on Covid-19 dataset. **Top-Left:** Coefficients on target variable of hate label. **Top-Right:** Coefficients on target variables of bot and troll scores, with linear correlation visualized as a trendline. **Bottom:** Salience of political, gendered, racial, and religious identities over topics.

Out of these ten topics, I selected four for more in-depth characterization². First were

²For completeness, I note that the remaining topics had the following content. Topic 3: Criticism of

the two topics with the highest levels of hate, each averaging over a standard deviation from the mean in the dataset. The second set were topics 7 and 10 which, while they did not explicitly feature hate, were of interest because they showed levels of bot and troll activity far beyond the norm in the dataset. Interestingly, as shown in the final subfigure of Figure 5.2, all topics featured high levels of racial targeting, with slight variations in political and gendered targeting.

Table 5.2: Summary of hate-aware and manipulation-aware supervised topics.

Topic	Label	Target	Actors	Hate Norm
1	Trump Virus	Politics and Race	Humans	1.589
2	China Virus	Race	Bots	1.320
7	Kung Flu Fighting	Politics, Gender, Race	Bots and Trolls	-0.632
10	Covid Hysteria	Politics and Race	Trolls	-1.527

Table 5.2 now summarizes the topics of interest. Topic 1, which I labeled the “Trump Virus” narrative, features simultaneous targeting of political and racial identities, and strikingly, it is primarily driven by human accounts. It is also the topic featuring the highest average levels of hate in the dataset, with the supervised topic model assigning it a normed output value of 1.589. Topic 2, which I labeled the “China Virus” narrative, features the second highest levels of online hate, with the supervised topic model assigning it a normed score of 1.320. Here, it is racial identities which take center stage, and it is driven primarily by bot accounts.

Topic 3, which I labeled “Kung Flu Fighting” was an interesting case that as it featured a low normed hate score of -0.632, but it had high levels of bot and troll activity. It also had the most diverse mix of targets, featuring political, gender, and racial identities all at once. Finally, topic 10, which I labeled “Covid Hysteria”, was the topic with the lowest normed hate score at -1.527. It primarily featured trolls that targeted political and racial identities.

In Table 5.3, I summarize the terms with the highest scores from a semantic network analysis of the topics of interest. These provide the basis for the analytical labels assigned to each supervised topic.

Trump Virus In topic 1, online hate is expressed largely toward political leaders like President Donald Trump, as well as his followers. Harsh condemnation is directed toward the president who is positioned as having not led the country effectively during the public health crisis. High-hate examples from this topic frame Trump himself as a virus and express a consequent wish for his elimination, both from office as well as from the world. As low levels of bot and troll activity were associated with these expressions of online hate, such sentiments were likely to come from organic social media accounts.

liberal politicians. Topic 4: Criticism of conservative politicians. Topic 5: Criticism of US Covid policies. Topic 6: Social activities during lockdown. Topic 8: Advice for living in quarantine. Topic 9: Explaining racial connotations of kung flu.

Table 5.3: Semantic centrality measures of top-ranking terms.

Topic 1 Trump Virus		Topic 2 China Virus	
Term	Score	Term	Score
coronavirus	0.370	chinese	0.462
trumpist	0.034	flu	0.274
donald	0.031	kung	0.069
lago	0.030	ching	0.067
usa	0.025	trump	0.040
president	0.012	coronavirus	0.011
government	0.009	wuhan	0.009
news	0.008	animal	0.008
5g	0.008	racist	0.007
gop	0.006	china	0.006

Topic 7 Kung Flu Fighting		Topic 10 Covid Hysteria	
Term	Score	Term	Score
flu	0.380	covid	0.418
fight	0.139	usa	0.054
covid	0.135	people	0.043
lightning	0.092	american	0.040
people	0.025	stop	0.032
fast	0.021	dow	0.025
trump	0.012	donald	0.023
joe	0.010	death	0.019
virus	0.009	nba	0.009
maga	0.008	coronapocalypse	0.006

China Virus In topic 2, online hate is expressed toward the nation of China as well as to people of Chinese descent. Among the less explicitly violent examples, some texts under this topic make fun of the Chinese language (“ching chong”) in reference to the virus. Others also make reference to the fact that infections originated in Wuhan to justify their use of derogatory speech. In high-hate examples under this topic, however, Chinese people are compared to animals and are threatened to be dealt with violently and eradicated in order to solve the Covid-19 crisis. Extremely high levels of bot activity are associated with this topic, suggesting that automated accounts were intent on pushing these sorts of messages to the fore of the online conversation.

Kung Flu Fighting In topic 7, low levels of online hate are detected by the model, although it is associated with high levels of bot and troll activity. In this topic, a fairly consistent motif is detected: a parodic reference to the popular 1974 song “Kung Fu Fighting” by Carl Douglas³. Here, kung fu is replaced with “kung flu” as a racially tinged reference to the pandemic. However, while topic 2 showcased explicitly derogatory refer-

³<https://genius.com/carl-douglas-kung-fu-fighting-lyrics>

ences toward China and Chinese people, messages under topic 7 appeared oriented more toward humor. Such humorously oriented messages were used in certain instances to express criticism of Trump and his followers. Bots and trolls were both strongly associated with this topic, indicating both its automated amplification as well as its use to disrupt online conversations.

Covid Hysteria Finally, topic 10 also featured low levels of online hate, but it was predominantly expressed by troll accounts. Utilizing catastrophizing terms such as “coronapocalypse”, texts under this topic magnified various social problems which arose during the Covid-19 pandemic. These included the suspension of the National Basketball Association’s official activities, economic downturns as symbolized by the Dow Jones Industrial Average, and various other miscellaneous issues such as toilet paper shortages. Trolls appeared to use such tweets also in a humorous fashion to make light of the crisis.

5.4.3 Bot and Troll Agenda-Setting

Upon characterizing the narratives of relevance, I additionally conducted analysis of the extent to which bot, troll, and human messages tended to lead or lag each other. Estimates from the series of lagged Poisson regressions are summarized in Table 5.4.

Table 5.4: Summary of estimates for lagged Poisson regression models.

Predictor	Human Messages	Bot Messages	Troll Messages
Concurrent			
Human Messages	-	0.006 (0.000)***	0.005 (0.001)***
Bot Messages	0.022 (0.001)***	-	-0.005 (0.004)
Troll Messages	0.027 (0.001)***	0.001 (0.004)	-
Lagged			
Human Messages	-0.002 (0.000)***	0.003 (0.001)***	0.003 (0.001)***
Bot Messages	0.009 (0.001)***	-0.006 (0.003)*	0.003 (0.004)
Troll Messages	0.010 (0.002)***	0.007 (0.004)	-0.011 (0.005)*

Bot and troll messages were generally associated with concurrent human messages. This indicates that any given point in time, the narratives that were salient for humans were also salient for bots and trolls. However, most striking here is that lagged bot messages and troll messages significantly predicted future human messages. Topics expressed by bot messages (0.009, $p < .001$) and troll messages (0.010, $p < .001$) at a given point in time were more likely to appear in future human messages, beyond the autoregressive effect of human messages on themselves (-0.002, $p < .001$). These effects were also stronger than lagged human messages predicting future bot messages (0.003, $p < .001$) and future troll messages (0.003, $p < .001$).

Finally, I note that across the models estimated, bot and troll messages did not predict each other, either concurrently or in lagged values. This is consistent with our understand-

ing that bots are likely to be programmed to achieve a particular information operational goal, whereas trolling accounts are likely to be driven primarily by their disruptive aims.

5.4.4 Bot and Troll Networks

In this section, I now shift to an analysis of bot and troll activities in the context of online hate networks. Here, I estimate the effects of bots and trolls on the structural features of online hate groups, as well as their relationship with individual accounts' expression of online hate in the context of these groups.

Table 5.5: Summary of group-level effects linking bots and trolls with the structural features of online hate networks.

Predictor	Bots	Trolls
Social Cyber-Actor	-0.092 (0.063)	0.464 (0.035) ^{***}
x Size	0.000 (0.000)	0.001 (0.000) [*]
x Density	0.065 (0.019) ^{***}	0.115 (0.051) [*]
x E/I Index	-0.063 (0.063)	-0.364 (0.177) [*]
x Bottlenecking	-0.017 (0.092)	0.315 (0.194)

Table 5.5 provides the estimated effects of bots and trolls on the relationship between various structural features studied in Chapter 4 and group levels of online hate [250]. For bots, it is interesting to note that their mere presence in a network cluster does not predict higher levels of online hate (-0.092, $p > .05$). Instead, they are primarily associated with intensifying the relationship between density and group-level hate (0.065, $p < .001$). Interpreted concretely, this suggests that bots are involved in densifying online hate groups by increasing interactions between members. Whereas previous results suggested that online hate groups were generally less dense than the average online community, bots may be involved in reversing that association, thereby producing online hate groups with highly engaged members.

Meanwhile, for trolls, more statistically significant effects are observed. The mere presence of trolls in a network cluster is predictive of higher levels of group-level hate (0.464, $p < .001$). Additionally, trolls have statistically significant moderating effects on the relationship between group-level hate and a group's size (0.001, $p < .001$), density (0.115, $p < .05$), and E/I index (-0.364, $p < .05$). Respectively, these estimates suggest that trolls add to the size of online hate groups, even if only very modestly. Like bots, they also contribute to the densification of online hate groups, as well as their increased isolation from the mainstream.

Next, utilizing a multilevel model, I estimated the network associations of bots and trolls on individual social media accounts' expression of online hate. Table 5.1 summarizes the estimated associations and these are visualized in Figure 5.3.

Results from this analysis showed that key effects from Chapter 4 persist even when considering the presence of bots and trolls in online hate networks. Most importantly, the

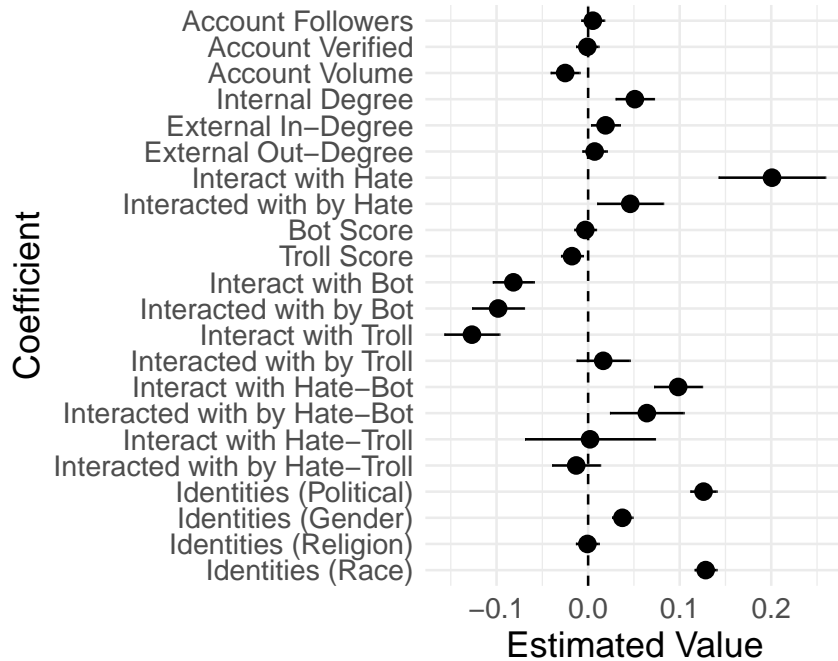


Figure 5.3: Visualized account-level coefficients of multilevel model. Error bars represent 95% confidence intervals.

most hateful accounts are those which interact with other hateful accounts (0.202, $p < .001$), and to a lesser extent, are interacted with by other hateful accounts (0.046, $p < .05$). Bots and trolls themselves, however, are not necessarily the most hateful accounts; in fact, bot scores have virtually no association with hate scores (-0.003 , $p > .05$), while trolls are less likely to express hate per se (-0.019 , $p < .01$).

Those who interact with bots in general also seem to display lower levels of online hate (-0.081 , $p < .001$), as do those that bots interact with (-0.097 , $p < .001$). Those who interact with trolls are also less likely to express hate (-0.128 , $p < .001$). This reflects how human accounts may interact with bots and trolls in a general sense to achieve a variety of functions, such as correcting misinformation or calling out offensive behavior [253]. Less hate would also be reasonably associated with such interactions especially when bots and trolls themselves are not expressing hate.

However, when considering interactions specifically with bots expressing hate, positive effects are observed. In particular, those who interact with hate-expressing bots have higher levels of online hate themselves (0.098, $p < .001$), as do those that hate-expressing bots interact with (0.064, $p < .01$). No statistically significant effects were detected for interactions with hate-expressing trolls.

Table 5.6: Summary of account-level coefficients of multilevel model.

Predictor	Estimate
Account	
Account Followers	0.005 (0.006)
Account Verified	-0.001 (0.006)
Account Volume	-0.025 (0.008)**
Internal Degree	0.051 (0.011)***
External In-Degree	0.019 (0.008)*
External Out-Degree	0.007 (0.007)
Hate Exposure	
Hate In-Connections	0.202 (0.030)***
Hate Out-Connections	0.046 (0.018)*
Social-Cyber Actors	
Bot Score	-0.003 (0.006)
Troll Score	-0.019 (0.006)**
Bot In-Connections	-0.081 (0.011)***
Bot Out-Connections	-0.097 (0.014)***
Troll In-Connections	-0.128 (0.015)***
Troll Out-Connections	0.016 (0.015)
Bot-Hate In-Connections	0.098 (0.013)***
Bot-Hate Out-Connections	0.064 (0.021)**
Troll-Hate In-Connections	0.002 (0.036)
Troll-Hate Out-Connections	-0.013 (0.013)
Targeting	
Political Identities	0.126 (0.007)***
Gendered Identities	0.037 (0.006)***
Religious Identities	-0.001 (0.006)
Racial Identities	0.128 (0.006)***

Note: * $p < .05$, ** $p < .01$, *** $p < .001$

5.5 Discussion

The aim of this chapter was to assess the manipulation of online hate narratives and networks through particular analyses of bot and troll activity. By extending the work presented in Chapters 3 and 4 of this thesis, I designed methods to specifically assess the prevalence of bots and trolls across various online hate narratives as well as their interplay with the structural features of online hate groups. Applying these methods to online conversations about the Covid-19 pandemic, I empirically characterized bot and troll narratives during the crisis as well as their group-based activities on online social networks.

Findings from the narrative analysis revealed an important distinction between organic and inorganic topics of conversation during the Covid-19 pandemic. Strikingly, while a po-

litically colored hate narrative arose from organic actors, bots appeared to focus strongly on racially charged attacks against China and people of Asian descent. While the former pointed to authentic vitriolic reactions from the public toward perceived failures of domestic leadership, the latter indicates attempts to manipulate the online conversation to heighten racial tensions within the populace as well as possible international relations with China. On top of this, it was also interesting to note that non-hateful narratives were part of bot and troll efforts to shape public discourse. For trolls, in particular, heightening perceptions of Covid hysteria did not need to rely on hostile speech in order to sow discord. Similarly, bots and trolls collectively utilizing the “kung flu fighting” tagline was itself not strictly hateful. However, through humor and popular culture references, they may have nonetheless opened the doors toward normalizing offensive stereotypes against Asian people, their association with the virus, and a broader insensitive orientation toward racial intergroup relations.

On the network side, it was critical to observe how bots and trolls either reinforced or reversed the relationships previously observed in Chapter 4 [249, 250]. Without considering bot and troll manipulation, I had shown that online hate groups tended to be smaller, less dense, more isolated, and display higher bottlenecking behavior. Bots largely appeared to densify online hate groups, whereas trolls tended to grow, densify, and further sequester them from the mainstream. Moreover, when considered from a multilevel lens, the observed relationships were more mixed. Although trolls seemed to have a greater effect on the structure of online hate groups, exposure to hate-expressing bots appeared to be more impactful at increasing other accounts’ expressed levels of online hate.

One possible explanation of these findings is that since trolls themselves generally do not express hate, they do not necessarily trigger further hate [255]. Instead, they may be effective at promoting group engagement, specifically by increasing memberships and closing off exposure to other groups. In conjunction with the narrative findings, while bots act more directly as force multipliers for amplifying hate, trolls may also be critical at spreading non-hateful content that is conducive for heightening perceptions of threat and normalizing intergroup insensitivity. Their actions may thus still be significant in building online hate groups even without expressing large amounts of hate themselves.

From a practical standpoint, these findings suggest the importance of attending to the broad range of manipulative tactics that social-cyber actors may engage in to promote the spread of online hate [24, 27]. Platforms or policies concerned exclusively with looking out for direct expressions of hate may risk a myopic view of how information operations may contribute to fraying intergroup relations and the legitimization of hostility especially in a crisis setting. Takedown efforts of influence campaigns remains significant, but they need to go beyond a binary view of organic versus inorganic, as well as of harmful versus non-harmful. When accounting for the specific narrative and network tactics employed by bots, trolls, and other influence operators, it is vital to understand how they are oriented toward manipulation even if they do not appear to explicitly express harmful messages themselves. Such broader assessments enable more holistic understanding of both the strategic objectives of such operations, as well as of key axes of vulnerability which may be salient among the public at large, both online and offline.

While this research looked at both bots and trolls as social-cyber actors associated with

online manipulation, a wide variety of other forms of manipulation may also be present in an online conversation. For instance, known influencers and public figures may be engaged in actions to strategically shape public opinion or popularize harmful sentiments, even without the cloak of anonymity [114, 187]. Other social-cyber actors are also relevant in different contexts, dependent on the digital and geopolitical environments of interest [81, 188]. While the proposed methodology in this chapter is limited by focusing on bots and trolls specifically, it is also flexible to readily accommodate their incorporation into the analysis. Because the analysis relied on social media data, the quality of the data collection pipeline is an important constraint in the kinds of inferences that may be made outside the specific case under study [176]. Similarly, because various computational tools were used for data enrichment, prediction errors may also affect the relationships observed. Finally, while this analysis showcased a variety of bot and troll actions upon narratives and networks in large-scale observational data, it does not speak to causal mechanisms. How humans are specifically affected by these actions is outside the scope of this work, but points to important future directions that may be opened up by the methods and findings presented here.

Chapter 6

Offline Contexts of Online Hate

6.1 Introduction

For the final chapter of this thesis, my goal is to situate online hate in its offline societal context. Throughout the previous chapters, I have characterized online hate in terms of its organization into narrative systems of meaning and networked systems of people. The broader objective of this research has been to view online hate as more than a linguistic act, and to enhance our understanding of it as a deeply social phenomenon. It is with this overarching perspective in mind that I aim to conclude this thesis by examining online hate in relation to its societal conditions and consequences.

The Covid-19 pandemic has served as a powerful case study for the foregoing chapters because, more than a purely medical or health issue, it has also been a crisis of societal cohesion. Expressions of online hate have been one manifestation of the pandemic's social dimensions of strain. Social media platforms became a lightning rod for public messages of prejudice and vitriol against various racial and political targets. But such hostilities were not brewing only online. In the United States, the Federal Bureau of Investigation reported a major uptick in hate crimes during the first year of the Covid-19 pandemic, with a 13% increase since 2019 in total hate crimes, and an over 30% increase in hate crimes motivated by race or ethnicity¹. Scholarly literature documents how such offline incidents of violence have been linked specifically to hateful ideas which arose with the pandemic. Personal accounts shared by victims of similar incidents have likewise described how perpetrators expressed their hate toward them, featuring many sentiments that have likewise proliferated online².

Social theories of emotion have long anchored the expression of hate in the context of intergroup relations, socio-cultural values, and the structural features of a given society [190, 229]. However, application of these theories to the online context has been relatively scarce. Meanwhile, debates about the interface between the online and offline worlds have persisted since the rise of social media [99, 106, 137]. Various scholars have proposed different frameworks organizing how online discourse might relate to offline conditions and

¹<https://www.justice.gov/crs/highlights/2020-hate-crimes-statistics>

²<https://time.com/5858649/racism-coronavirus/>

consequences, the direction of such a relationship, and even whether such a relationship exists at all. With a global crisis like the Covid-19 pandemic prompting spikes in both online hate and offline violence, how might we describe and explain their relationship?

In this chapter, my goal is to put these theoretical bodies of work in explicit conversation with each other to tackle the empirical problem of how online hate related to offline violence during the Covid-19 pandemic. I harness a large-scale Twitter dataset of online conversations about the Covid-19 pandemic to map out the spread of online hate over space and time. Through a series of interoperable computational social science tools, I obtain measures of online hate as well as bot and troll activity within the corpus. I then examine these online measures in conjunction with offline datasets of hate crimes in the United States as well as recorded cases of Covid-19 infection. Statistical analysis of various theoretically motivated models of online-offline dynamics reveals key insights around the relationship between online hate and offline violence. Finally, I conclude with broader implications of this work.

6.2 Related Work

6.2.1 Societal Theories of Hate

Online hate is expressed in a societal context. Hate itself has been understood in classical social scientific scholarship as arising out of societal conditions of division and scarcity. In conceptualizations of hate under strain theory, social groups that one is a part of shape one's needs and one's capacity to obtain resources to fulfill those needs [4, 164]. In a given society, groups may compete for those resources. Certain groups may not be as successful as others, causing the frustration of needs for their members. Members of out-groups may also be perceived as threats as a result of such competition.

Moments of crisis, like the Covid-19 pandemic, can heighten the scarcity of resources, intensify intergroup competition, and enhance the mutual perception of threat [3, 91]. Such developments produce strain upon the social structure. Hate becomes an especially powerful psychological and social force in these settings. Affectively, hate consists of intense negative feelings toward a target out-group and a sense of battle-readiness in in-group members [142]. Cognitively, hate also harbors stereotypical and prejudicial beliefs about the target which simplify and justify their derogation, or in extreme cases, their annihilation [234]. Alongside these destruction-oriented functions, hate additionally strengthens intragroup relations by providing a sense of belonging and purpose for the in-group, reinforcing in-group values and norms, and establishing shared motivations for collective action [62].

Moreover, taken together, these multidimensional components of hate lead to behavioral outcomes, such as verbal expressions of hate, hate crimes and physical violence, and potentially large-scale actions to eliminate the target group [189, 258]. From this standpoint, hate thus not only possesses societal antecedents, but also produces societal consequences. For some theorists, hate in societal context often bridges the two to create a vicious cycle. Under what are known as intensification theories of hate, when groups engage in

hate-driven behavior, they create additional conditions for the spread of hate [190]. Such conditions in turn further motivate and legitimize new hateful attitudes, cognitions, and behaviors, and so on.

6.2.2 The Online-Offline Nexus

To understand online hate in societal context, it is also critical to conceptualize the relationship between the online and offline worlds. Some of the earliest research into these questions surrounding social media hailed its massive potentials for impacting society. With the Arab Spring and the Occupy Movement in the early 2010s taking place in conjunction with social media's surge in popularity, researchers pointed to the latter's role in organizing individuals, spreading fast-paced messaging campaigns, and enabling social change on a massive scale [129, 153]. Conversely, other earlier theorizing around the online-offline nexus proposed the reverse: that the online and offline realms should be considered separately from each other. Such perspectives have been characterized as forwarding a 'dualistic' view of the online and offline worlds [106]. Similarly to this standpoint arose critique along the lines of what has been called the 'slacktivism hypothesis', which describes social and political actions taken online as either having no bearing on offline society, or potentially even hindering offline political impact by diverting collective efforts into inconsequential online activity.

Since these earlier propositions, scholarship around the online-offline nexus has evolved and sought to render explicit more nuanced explanatory mechanisms for their relationship. One broad area of work focuses on how the online world serves as a mirror of offline social realities. From this perspective, offline events prompt corresponding activity in the online space. People use social media as a means to express beliefs and emotions they hold online, and thus, the online activities they engage in might serve as valid traces of their offline psychological states and social relationships. Evidence along these lines has been modestly supportive, with some research affirming how beliefs in such contentious issues as gun control and climate change evolve with offline developments in these arenas [147, 246]. It is from this perspective also that some research has sought to forecast societal outcomes like election results based on aggregated online sentiment, though results from such efforts have also been mixed due to the non-representative demographics of social media [37, 118].

Conversely, others have theorized that online dynamics might influence offline outcomes. The precise mechanism governing such a relationship varies between frameworks. Some have pointed to the role of social media in making visible the actions of others. Social media users may identify with these visible others, and the behaviors modeled by the latter become accessible and reinforced to observers. Such affordances are thus seen to prompt the adoption of these behaviors in the offline space [137]. Other theories of social influence more directly point to the role of social interaction in shaping one's beliefs and emotions. In using social media, one tends to interact with others along group-based lines. Online information exchange with in-group members shifts one's cognitive and affective states in the offline world and thus prompts offline behavioral changes. From a social cybersecurity perspective [52, 53], it is within these paradigms that information operations find their strategic value. By capturing hearts and minds online, information operations

intend to influence offline beliefs and emotions, shape individual and collective actions, and overall impact societal outcomes.

6.2.3 Hate and Violence in the Pandemic

Taking the foregoing bodies of scholarship together, this chapter intends to explore their implications for hate and violence in the Covid-19 pandemic. Figure 6.1 visualizes the different types of relationships I examine. First, drawing upon general strain theories of hate, one might ask: Does the contextual strain from increased Covid-19 infections predict heightened online expressions of hate and offline perpetration of hate crimes? Evidence in this area is growing and prompts further investigation. Empirical research certainly analyzes how the emergence of hate crimes during the Covid-19 pandemic is linked to anxieties related to the crisis [105, 242]. At the same time, the rise of online hate expressions was likewise attributed to the crisis as a whole [227, 244]. However, in both sets of studies, the pandemic is largely treated as a monolithic event without examining its temporal variation. The pandemic exerted varying levels of strain upon society at different points in time, and in this chapter, I aim to examine whether such variations were linked to corresponding levels of online hate and offline hate crime.

Second, from mirror theories and social reinforcement approaches to social media, one might also ask: Do increased expressions of hate online predict the offline perpetration of hate crimes? And conversely, does the offline perpetration of hate crimes predict expressions of hate online? Scarce evidence exists drawing these connections. In relation to the Covid-19 pandemic, one study suggests that important flashpoints—such as various protests—acted as trigger events for higher levels of online hate on social media [149]. However, hate crimes were not studied explicitly. Outside the context of the pandemic, some studies have shown such relationships in line with mirror and social reinforcement theories of social media. One study showed that when comparing offline measures of prejudice in various regions in the United States, they predicted higher levels of online hostility in those regions [212]. Another study of protest groups in the United States and the United Kingdom showed that increased hostility between ideologically opposed collectives online predicted greater violence between them in offline events [96]. In a study of Islamophobia, it was found that online Islamophobic hate speech tended to follow, but did not necessarily precede, Islamophobic violence in the United Kingdom [266].

Finally, from intensification theories of hate, one might ask: Do increased expressions of online hate and increased hate crimes offline mutually predict further incidences of each other? This perspective is perhaps by far the most complex of all those described, as it entails the explicit examination of cross-lagged relationships. Theoretical articulations abound from more classical literature on hate more broadly [190], to more recent arguments which specifically account for online hate in relation to offline violence [15, 232, 259]. Empirically, I have examined similar bidirectional relationships in the context of the face mask debate during the Covid-19 pandemic [256]. Although I did not investigate hate explicitly, I showed that online expressions of stance toward face masks could be predicted based on various offline developments, including varying levels of Covid-19 infection as well as the declaration of stay-at-home mandates. In turn, I showed that these collective

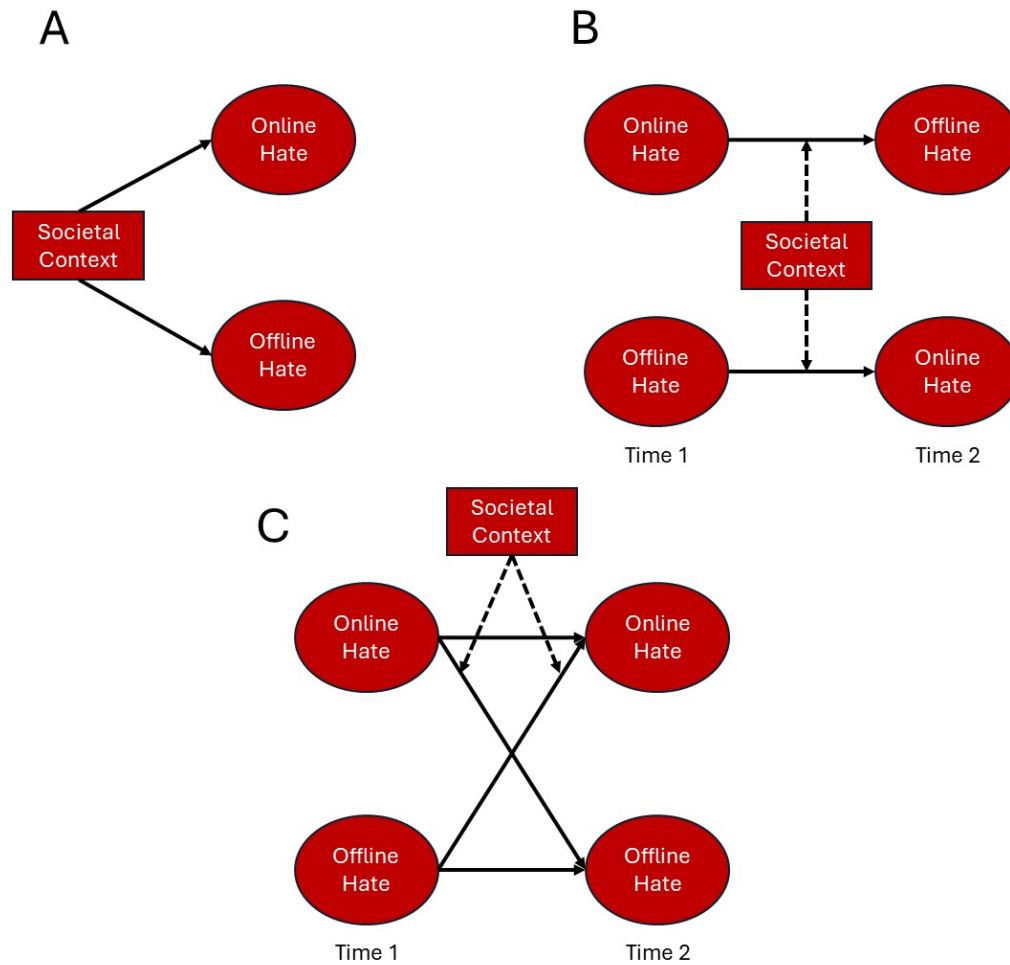


Figure 6.1: Summary of theoretical models examined in this chapter relating online hate to offline hate. **A:** General strain theory model. **B:** Social media as mirror and reinforcer models. **C:** Intensification model.

online expressions also predicted later levels of Covid-19 infection, partially mediated by actual recorded levels of mask-wearing. Previous evidence thus exists to suggest the merit of testing an intensification model of online hate in relation to offline hate crimes.

6.3 Data and Methods

6.3.1 Datasets

Three key datasets were utilized in this chapter to examine various links between online hate and offline hate crimes during the Covid-19 pandemic.

Covid Tweets A large-scale corpus of online conversations related to the Covid-19 pandemic was used for this chapter [117]. This dataset encompassed a broader set of data

collection terms and was originally intended to capture the global conversation around the crisis, including all references to “Covid19”, “ncov”, and “wuhan virus”. In total, the dataset consists of over 300 million tweets. Moreover, the dataset in its entirety spans over two years of data. For this chapter, I focused on tweets from the 2020 period, prior to the rollout of Covid-19 vaccines. This reduces the dataset size to 22,143,552 tweets [256].

For each tweet in the dataset, several measures were obtained using an interoperable pipeline of computational social science tools [252]. These measures were used as key variables in the models described below.

- **Time.** Each tweet was associated with a time stamp derived directly from the Twitter corpus indicating the date it was produced.
- **Location.** Each tweet was associated with a geolocation variable that was either derived directly from the Twitter dataset or predicted using a hierarchical location prediction algorithm with a validated accuracy of 92.1% [116]. These location variables were additionally used to filter to the dataset to just those made by users located in the United States, and subsequently, assign them to their specific states of origin.
- **Online hate.** Each tweet was associated with a measure of online hate using the model described in Chapter 2 [251].
- **Bot score.** Each tweet was associated with a score indicating the likelihood that it was from an automated account, using the BotHunter model [23].
- **Troll score.** Each tweet was associated with a score indicating the likelihood that it was from a troll account, using the TrollHunter model [255].

FBI Hate Crimes Data To quantify the perpetration of hate crimes throughout the Covid-19 pandemic, I used the dataset publicly released by the Federal Bureau of Investigation³. None of the data points contain identifiable information about perpetrators or victims. While the number of victims and perpetrators is disclosed for each data point, I focus only on the binary outcome of whether or not a hate crime has transpired at all, primarily due to the extremely skewed distribution of the count variable. Each hate incident is tagged by its time and location, which allows for matching with the incidence of online hate.

CDC Covid-19 Case Data Finally, to assess the levels of Covid-19 infection, I use records maintained by the Centers for Disease Prevention and Control⁴. Each data point provides aggregate statistics for the number of cases recorded for each day at the state level, which allows for matching with incidents of online hate and offline hate crime. I take the logarithm of actual Covid-19 case values to control for extreme differences in scale.

Other Pandemic-Relevant Variables Some of the analysis I performed also accounted for the schedule of stay-at-home orders taken from the United States government’s open

³<https://crime-data-explorer.fr.cloud.gov/pages/downloads>

⁴<https://covid.cdc.gov/covid-data-tracker>

data catalog, indicating the start and end dates for various orders ⁵. I additionally noted the official party affiliation of each state’s governor prior to the November 2020 elections⁶.

6.3.2 Analyses

To analyze the variously conceptualized links between online hate and offline hate crimes, I estimated several types of regression models. I briefly describe them below in order of increasing methodological complexity and in conjunction with the corresponding theoretical frameworks they correspond to.

Regression Models To examine concurrent relationships between online hate and offline hate crimes, I employed linear and logistic regression models. Linear regression models were used to estimate the relationship between Covid-19 cases with levels of online hate on a given day in a given state. Meanwhile, logistic regression models were used to estimate the likelihood of a hate crime occurring given concurrent levels of Covid-19 infection on a given day in a given state. For both models, linear effects of time were also estimated to detrend the target variable. I also accounted for the party affiliation of each state’s governor. All continuous variables were grand mean centered and scaled by standard deviation to allow for greater interpretability. Effects are reported and visualized with 95% confidence intervals. These models were used specifically to test general strain frameworks and concurrent versions of the social media as mirror and reinforcer frameworks.

Lagged Moderation Models Next, to test lagged relationships between online hate and offline hate crimes, I utilized lagged moderation models. As before, the specific form of the model varied with the target variable. Linear regression with lagged predictors and moderators was used when predicting online hate on a given day in a given state, whereas logistic regression with lagged predictors and moderators was used when predicting offline hate crimes on a given day in a given state. Here, I included both lagged levels of Covid-19 infection as well as the party affiliation of the state’s governor as moderators impacting the predictive effect of lagged online hate on future offline hate crimes, and of lagged offline hate crimes on future online hate. Linear effects of time were accounted for to detrend target variables. All continuous variables were grand mean centered and scaled by standard deviation to allow for greater interpretability.

Integrated Models Finally, to test the intensification model relating online hate to offline hate crimes, I employed an integrated model which jointly estimated the predictive effect of lagged online hate on future offline hate crimes and of lagged offline hate crimes on future online hate. To enrich the integrative model, I included here as additional predictors: (a) whether or not a stay-at-home order was in place, (b) lagged levels of Covid-19 cases, and (c) the state governor’s party affiliation.

⁵<https://catalog.data.gov/dataset/u-s-state-and-territorial-public-mask-mandates-from-april-10-2020-through-july-20-2021-by--7e5b8>

⁶<https://www.nga.org/governors/>

Additionally, to specifically test the potential effects of bot and troll activity on the relationships of interest, lagged values were also incorporated as predictors of both online hate and offline hate crimes. Versions of these models with and without manipulation analysis were computed to capture their marginal effects. Linear effects of time were again accounted for to detrend target variables. Robust standard errors were computed to account for the clustered nature of the data while jointly estimating associations. All continuous variables were grand mean centered and scaled by standard deviation to allow for greater interpretability. R^2 values were obtained for each target variable and reported for fitness.

6.4 Results

Overall, results of the analyses showed an asymmetrical relationship between online hate and offline hate crimes. Generally, both phenomena were correlated with each other over time, but lagged analyses showcased that online hate was predictive of subsequent offline hate crimes, while little evidence was observed for the reverse. Both online hate and offline hate crimes were also both related to offline trends in Covid-19 infections, but not consistently associated with the political affiliation of state governors. In the succeeding sections, I unpack each of the models tested according to the various theoretical frameworks relating the two phenomena in the context of the Covid-19 pandemic.

6.4.1 General Strain Associations

The first set of models I tested followed the general strain theory of hate. Here, I modeled Covid-19 cases as measures of strain exerted upon states during the pandemic, with higher levels of infection producing higher levels of societal strain. Using Covid-19 infections as a predictor, I therefore modeled their relationship with both online hate and offline hate crimes independently.

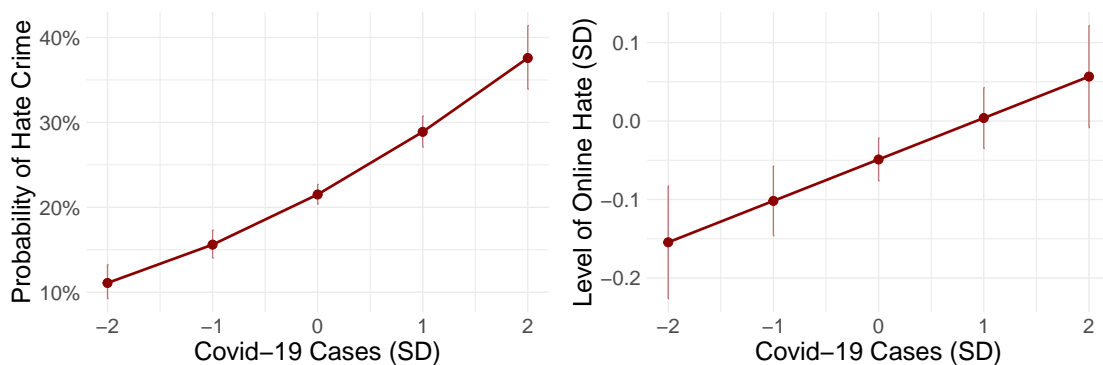


Figure 6.2: Visualization of estimated relationships between online hate and offline hate crime under a strain theory framework. **Left:** Estimated probability of hate crime occurring given level of Covid-19 infections. **Right:** Estimated level of online hate given whether or not a hate crime has occurred. Error bars indicate 95% confidence intervals.

Figure 6.2 shows that for both online hate and offline hate crimes, Covid-19 cases were positively predictive. When higher levels of Covid-19 infection were observed in a given state, the probability of a hate crime occurring likewise increased (0.409, $p < .001$). More specifically, on dates when a given state had one standard deviation fewer infections, there was typically a 15.60% likelihood of a hate crime occurring. However, when a given state had one standard deviation more infections, there was an estimated 28.89% likelihood of a hate crime occurring.

Covid-19 cases were also predictive of levels of online hate (0.109, $p < .05$), although the effect is modest. In general, when a given state had one standard deviation fewer infections, the average hate score was likewise an estimated 0.102 standard deviation lower than the mean. Conversely, when a given state had one standard deviation more infections, the average hate score was an estimated 0.004 standard deviation higher than the mean. Taken together, these results suggest that, following the argument of strain theory, higher Covid-19 infections were indeed associated with both online and offline manifestations of hate.

6.4.2 Concurrent Associations

But how did online hate and offline hate crimes relate to each other? In this section, I shifted to concurrent analyses of online hate and offline hate crimes, controlling for Covid-19 cases and the party affiliation of the governor. Below, I present estimates across these covariates.

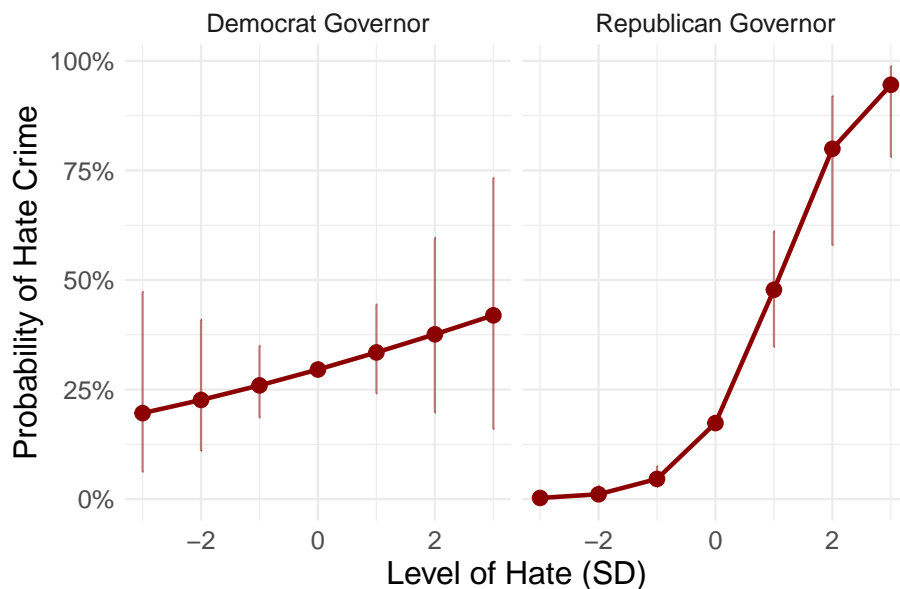


Figure 6.3: Estimated differences in relationship between online hate and offline hate crimes given political affiliation of state governor. Error bars indicate 95% confidence intervals.

In Figure 6.3, I visualize differences in the relationship between levels of online hate and the probability of an offline hate crime occurring when controlling for the party affiliation

of a state’s governor. When considering states with Democratic governors, a positive relationship was observed between online hate and offline hate crimes, but it was not statistically significant. One standard deviation higher than the mean level of online hate predicted 25.94% probability of a hate crime occurring, whereas one standard deviation of hate lower than the mean was associated with 33.47% probability of a hate crime occurring.

On the other hand, for states with Republican governors, the relationship was considerably sharper. One standard deviation higher than the mean level of online hate predicted 4.60% probability of a hate crime occurring, whereas one standard deviation of hate lower than the mean was associated with 47.77% probability of a hate crime occurring. Increases in levels of online hate were thus associated with different predictive effects on the likelihood of a hate crime occurring depending on the political leanings of state leadership.

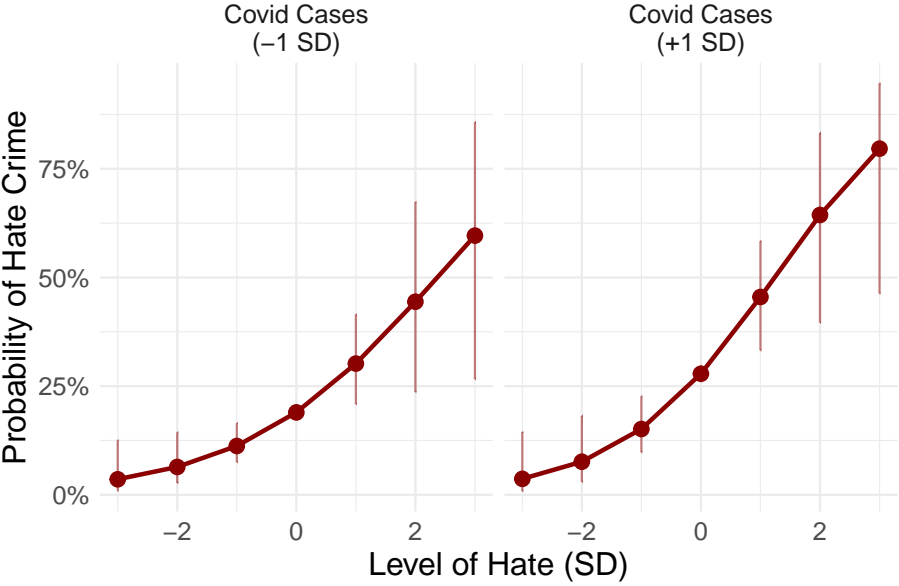


Figure 6.4: Estimated differences in relationship between online hate and offline hate crimes given concurrent levels of Covid-19 infection. Error bars indicate 95% confidence intervals.

In Figure 6.4, I visualize differences in the relationship between levels of online hate and the probability of an offline hate crime occurring when controlling for concurrent levels of Covid-19 infection. For cases with both one standard deviation fewer or more infections, the relationship was not changed substantially. On dates with relatively fewer Covid-19 infections, one standard deviation higher than the mean level of online hate predicted 30.19% probability of a hate crime occurring, whereas one standard deviation of hate lower than the mean was associated with 11.24% probability of a hate crime occurring. On dates with relatively more Covid-19 infections, one standard deviation higher than the mean level of online hate predicted 45.51% probability of a hate crime occurring, whereas one standard deviation of hate lower than the mean was associated with 15.14% probability of a hate crime occurring. Taken in conjunction with the findings above, this result suggests that while Covid-19 infections predict higher base levels of both online hate and offline

hate crime, they do not consistently strengthen or diminish the concurrent relationship between the two.

6.4.3 Lagged Associations

I now shift to lagged moderation analyses. Because of the temporal nature of the data, lagged associations explicitly aim to disentangle sequential associations between multiple time series. While outside the scope of this work, this forms an important basis for future analysis of causal relations. Moreover, moderation effects assess the extent to which given contextual covariates—in this case, the state governor’s party affiliation and lagged levels of Covid-19 infection—both predict the outcome variable as well as strengthen or diminish its relationship with other predictors of interest.

Figure 6.5 visualizes the estimated relationships in the lagged moderation models. For both models, when analyzed alongside lagged values, the concurrent relationship between online hate and offline hate crimes is no longer statistically significant. Instead, lagged predictors become statistically significant. First, let us consider the model with hate crimes as the target variable. Here, lagged values of online hate predict future incidences of offline hate crime (1.954, $p < .05$), beyond the autoregressive effect of lagged offline hate crimes on itself (1.654, $p < .05$). This lends support to the online-to-offline pathways suggested by theories of social media as a social reinforcer. In line with strain theory, lagged Covid-19 cases also predict future incidence of offline hate crime (0.425, $p < .001$). However, they do not moderate other effects. The state having a Republican governor also does not, on its own, predict greater incidence of hate crimes. But it does moderate the autoregressive relationship of lagged offline hate crimes on itself (0.555, $p < .01$). This suggests that in Republican-run states, clustered periods of hate crime occurrences—and also non-occurrences—tend to be more likely than in Democrat-run states.

Now let us consider the model with online hate as the target variable. Here, most effects are no longer statistically significant. The party affiliation of the state governor does not directly predict higher levels of online hate; nor does it moderate any of the other relationships. In addition, it is interesting to note that lagged offline hate crimes do not predict future levels of online hate (0.012, $p > .05$), particularly when considering the autoregressive effect of lagged levels of online hate on its future values (0.434, $p < .001$). This suggests that offline perpetration of violence does not necessarily lead to increased expressions of online hate in the Covid-19 pandemic. However, still in line with strain theory, lagged Covid-19 cases do still predict higher levels of online hate (0.017, $p < .001$). They do not moderate other effects.

6.4.4 Integrated Intensification Model

Finally, I consider an integrated model relating online hate with offline hate crimes. Here, I bring together the preparatory analyses in the sections above to consider a combined network of associations as suggested by an intensification model of hate during the pandemic. I specifically conduct a joint estimation of the extent to which lagged values of hate crime predict future levels of online hate, and to which lagged values of online hate predict future

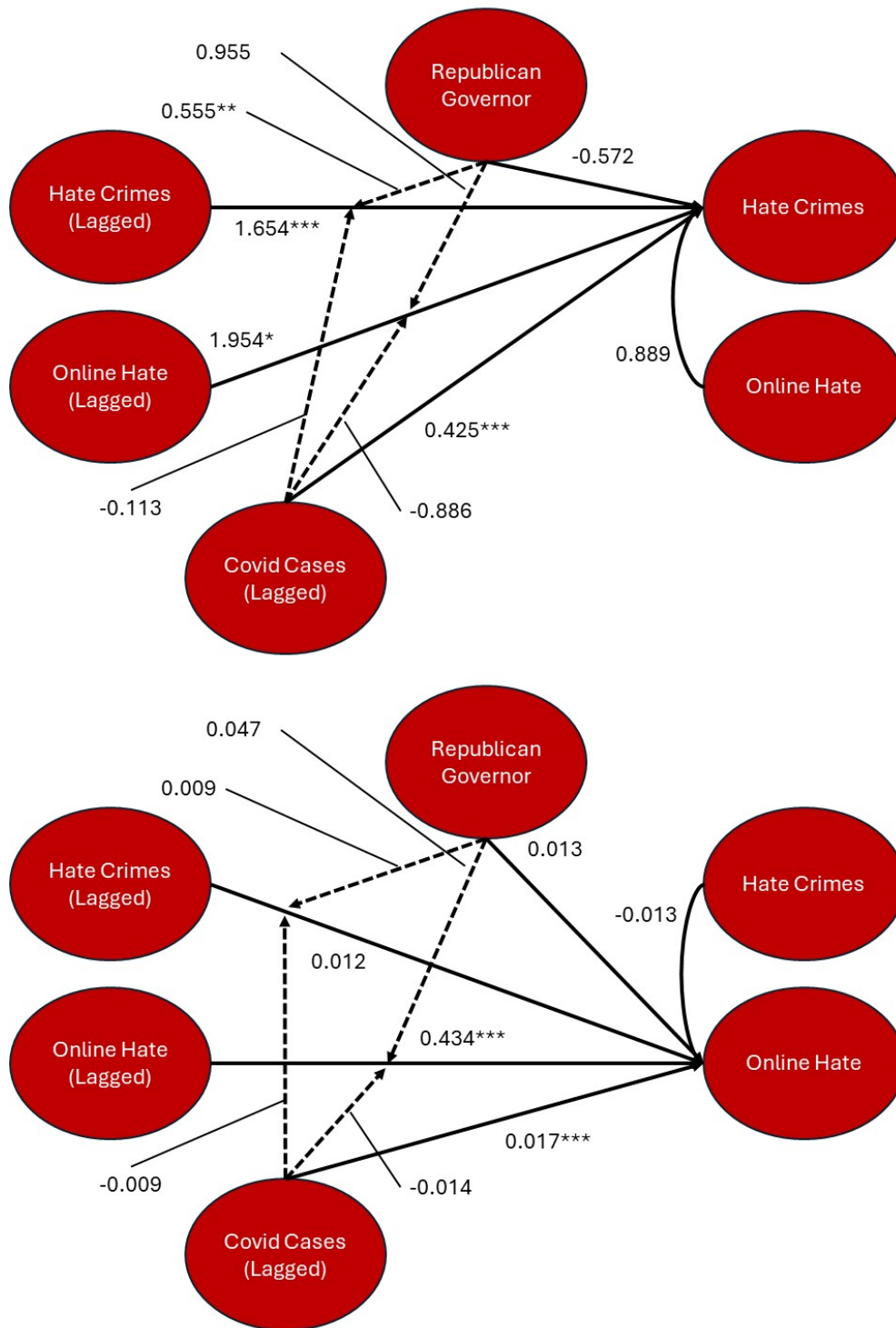


Figure 6.5: Estimated lagged moderation models linking past and future levels of online hate and offline hate crimes. Note: $^+p < .10$, $*p < .05$, $**p < .01$, $***p < .001$.

levels of offline hate crime. Two versions of this analysis are also conducted: one that does not consider the activity of bots and trolls, and the other explicitly incorporating them into the model.

Table 6.1: Summary of integrated analysis of online hate and offline hate crimes under an intensification model.

Model	Outcome	Predictor	Estimate	
No Manipulation	Hate Crime ($R^2 = 0.144$)	Hate Crime (Lagged)	0.307 (0.044)***	
		Online Hate (Lagged)	0.032 (0.008)***	
		Covid-19 Infections (Lagged)	0.053 (0.023)*	
		Republican Governor	-0.021 (0.078)	
		Stay-at-Home Policy	0.119 (0.056)*	
		Linear Time Trend	-0.160 (0.073)*	
		Online Hate ($R^2 = 0.348$)	Hate Crime (Lagged)	0.015 (0.050)
	Online Hate (Lagged)	0.615 (0.028)***		
	Covid-19 Infections (Lagged)	0.041 (0.024) ⁺		
	Republican Governor	0.004 (0.055)		
	Stay-at-Home Policy	0.067 (0.051)		
	Linear Time Trend	-0.060 (0.132)		
	With Manipulation	Hate Crime ($R^2 = 0.151$)	Hate Crime (Lagged)	0.307 (0.044)***
			Online Hate (Lagged)	0.034 (0.008)***
Bot Prevalence (Lagged)			0.018 (0.013)	
Troll Prevalence (Lagged)			-0.008 (0.010)	
Covid-19 Infections (Lagged)			0.052 (0.023)*	
Republican Governor			-0.021 (0.078)	
Stay-at-Home Policy			0.119 (0.056)*	
Linear Time Trend		-0.159 (0.073)*		
Online Hate ($R^2 = 0.666$)		Hate Crime (Lagged)	0.019 (0.050)	
Online Hate (Lagged)		1.067 (0.028)***		
Bot Prevalence (Lagged)		0.089 (0.022)***		
Troll Prevalence (Lagged)		0.072 (0.023)***		
Covid-19 Infections (Lagged)		0.041 (0.023) ⁺		
Republican Governor		0.009 (0.055)		
Stay-at-Home Policy	0.067 (0.051)			
Linear Time Trend	-0.101 (0.134)			

Note: ⁺ $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$.

Table 6.1 summarizes the results of the integrated model of online hate and offline hate crimes. Let us first consider the no manipulation model. Strikingly, when considered jointly, several relationships observed for online hate and offline hate crimes in the earlier

models remain robust. Most notably, the lagged association between online hate and offline hate crime remains asymmetrical. Whereas lagged online hate predicts higher likelihood of offline hate crime (0.032, $p < .001$) beyond its autoregressive effect (0.307, $p < .001$), the same is not true for lagged hate crimes predicting online hate (0.015, $p > .05$). Strain associations also remain statistically significant, though lagged Covid-19 infections are slightly more consistent in predicting hate crime (0.053, $p < .05$) in comparison to online hate (0.041, $p < .10$).

When considered in conjunction with bot and troll prevalence, key effects remained statistically significant. Lagged values of online hate still predict future incidence of offline hate crime (0.034, $p < .001$), over the autoregressive effect (0.307, $p < .001$). Lagged values of offline hate crime still did not predict future levels of online hate (0.019, $p > .05$). Levels of Covid-19 infection also remained predictive of future levels of offline hate crime (0.052, $p < .05$) and online hate (0.041, $p < .10$). Most noteworthy in this new model, however, is that lagged values of bot (0.089, $p < .001$) and troll prevalence (0.072, $p < .001$) were both predictive of future levels of online hate. They did not, however, directly predict offline hate crime.

6.5 Discussion

In this research, I sought to situate online hate in its offline context, specifically focusing on the relationship between online hate and offline hate crimes during the Covid-19 pandemic. To achieve this objective, I matched computational measures of online hate, bots, and trolls with offline measures of Covid-19 cases and hate crimes in the United States. Through a series of statistical analyses, I tested various temporal relationships between online and offline phenomena following strain theory, theories of the online-offline nexus for social media, and the intensification theory of hate.

Taking together the findings of the various analyses, Table 6.2 summarizes the evidence they provide for the different theoretical frameworks considered in this chapter.

Table 6.2: Summary of findings on relationships between online hate and offline hate crimes during the Covid-19 pandemic.

Theory	Prediction	Concurrent	Lagged
Strain Theory	Covid cases predict online hate.	Supported	Supported
	Covid cases predict hate crimes.	Supported	Supported
Social Media as Mirror	Hate crimes predict online hate.	Supported	Not supported
Social Media as Reinforcer	Online hate predicts hate crimes.	Supported	Supported
Intensification	Mutual cycle of online & offline hate.	-	One-sided
	Role of bots and trolls	-	One-sided

Overall, relatively consistent support was observed for the predictions of strain theory. Although effects were modest, higher levels of Covid-19 infection were consistently associated with higher levels of online hate and offline hate crime when analyzed concurrently as

well as with lagged values. This suggests that the Covid-19 pandemic exerted meaningful strain on society [4, 164], predicting both increased online expressions of hate as well as offline behavioral manifestations in the form of hate crimes.

Meanwhile, when considering theories of social media's relationship with offline society [99, 106, 137], evidence was more mixed. Online hate and offline hate crimes tended to co-occur; however, whereas lagged online hate predicted future offline hate crimes, lagged offline hate crimes did not predict future online hate. This suggests that online hate could be linked to the kind of strained intergroup relations that would lead to offline hate crime, but perhaps offline hate crimes do not precipitate future hateful expressions. Intensification theory's view of online hate and offline hate crime mutually exacerbating each other thus also has limited support [190].

One overarching explanation for these findings could be one of access to public information. Viewing social media as a credible setting for the reinforcement of social norms—especially those that might lead to hateful behavior—higher levels of online hate may be perceived by the public at large if they have access to prevailing digital discourse. However, when a hate crime is committed, public awareness is perhaps not as widespread, thus limiting its capacity to influence further hate incidents online or offline. Alternatively, another possible explanation accords greater weight to strain theory explanations and sidesteps the reinforcing effect of social media to some extent. Here, when norms and intergroup relations in a given location are strained, higher levels of both online hate and offline hate become more likely—but there are perhaps lower barriers to expressing vitriol on social media than committing violence in the offline world. Such mechanisms might reasonably explain the asymmetrical lagged relationship between online hate and offline hate crime.

Finally, in considering the role of bots and trolls, it was worth noting that while they did not directly impact offline hate crime, they did generally predict higher levels of online hate. This suggests that, in line with results from Chapter 5, their activities on social media may have resulted in the further spread of online hate. From this standpoint, intensification theory might suggest that bots and trolls could subsequently have an indirect effect on offline hate crime [190]. However, their direct effect on online hate is quite modest, and online hate's effect is in turn relatively small on future hate crime. Taken together then, there is limited evidence in this study that the activities of bots and trolls had a significant effect on the actual perpetration of hate crimes during the pandemic. That being said, this does not mean that their overall effect was insignificant. Information operations utilizing bots and trolls may not have necessarily been organized to increase offline violence, but merely to sow discord in public discourse. On that latter front, the estimated effects still remained meaningful.

From a broader methodological standpoint, this chapter presented an integrated pipeline combining social media and public government data to track interrelationships between online hate and offline violence, accounting for various temporal relationships as well as the possible influence of information operations. This toolkit and analytical framework is vital for critical questions in social cybersecurity, extending well beyond the pandemic setting [52, 53, 252]. For settings vulnerable to various forms of violence, as well as where other pertinent types of offline outcomes, these methods help to establish the extent to which large-scale online conversations could be used as an early signal to anticipate their oc-

currence, map out indices of vulnerability or resilience, and potentially design proactive mitigating responses.

From this standpoint, specific implications may be considered for the Covid-19 pandemic and similar crises. Given the overall predictive effects of strain from the Covid-19 pandemic [4, 164], efforts to mitigate both online hate and offline hate crime may need to account for how crises affect the public’s perception of threat and extant intergroup relations. Public messaging about the impacts of a crisis like the pandemic should be careful not to associate them with any particular identity or group. Instead, leaders and other actors should emphasize messages of solidarity in society to enhance societal cohesion in the face of momentous challenges to it [108, 218]. In the online space, platforms may also have a role to act in a timely manner to address the spread of online hate. Early detection and removal of such content—and the public’s subsequent exposure to it—may play an important role in slowing down the perception of social norms legitimizing violence against targeted groups. Routing manipulative actions to affect its spread should also be incorporated into such efforts. In the offline space, public reporting on targeted violence should also thus be performed sensitively so as to prevent the online-offline loop from closing into a vicious cycle [61, 74].

In closing, key limitations are also important to note. First, while this study relied on a large-scale observational dataset, causal pathways could not be explicitly tested. The psychological and social mechanisms underlying shifts in levels of online hate or offline hate crime were also not measured. While manipulation of these variables are likely outside ethical norms, offline studies directly engaged with human subjects could help clarify some of these mechanisms precisely [177]. Second, as several key measurements used in this study were obtained from computational tools, scalable insights could be inferred but these predictions are not perfectly accurate. Novel forms of hate—as well as bot and troll activity—could also emerge beyond the predictive domain of the tools utilized, especially in a novel crisis situation. Relatedly, the use of social media data also constrains the quality of the inferences in line with the quality of available data from the platform [176]. Data quality of offline records—even when taken from official sources—is also a constraining factor in this regard.

Chapter 7

Concluding Remarks

This thesis examined the problem of online hate from an interdisciplinary perspective emphasizing the integration of computational methods with social scientific theory. Anchoring my approach in *narratives* and *networks*, I investigated how online hate goes beyond a purely linguistic phenomenon, and instead encompasses broader systems of meaning and systems of people [190, 229]. It is through systems of meaning that online hate invokes shared socio-cultural resources to attack its target, and it is through systems of people that online hate becomes influential, legitimized, and forms communities. Drawing on a social cybersecurity lens [52, 53], I further showed how narratives and networks serve as key arenas for the manipulation of online hate. Lastly, in probing these social dimensions of hate, I analyzed online hate in relation to its societal contexts of offline upheaval and violence.

In this final section, I synthesize key contributions of this body of work from theoretical, methodological, and practical standpoints. I argue that, taken together: (a) the insights of this research meaningfully extend the literature on online hate in the computational and social sciences, (b) the tools and techniques developed in this work make important methodological advances in this area, and (c) the implications of this work speak to concrete applications in platform design, policymaking, and education around online hate. By engaging the rich social dimensions of online hate, I offer novel evidence and perspectives for better identifying it, understanding it, and mitigating it. I close with various recommendations for future work along these multiple planes.

7.1 Theoretical Contributions

7.1.1 Psycholinguistics of Online Hate

One principal problem surrounding the computational social science of online hate lies in its identification. As previous scholarship has argued, definitions of online hate have been plagued with issues of arbitrariness and decontextualization [68, 88, 200]. In practical settings, like social media platforms which seek to mitigate its spread, online hate has been defined in terms of attacks against laundry lists of protected characteristics and identity targets. Meanwhile, in research contexts, analysts have largely relied upon manual

annotations which have themselves been prone to inconsistency and are invariably steeped in subjectivity [86]. These approaches sidestep foundational conceptual issues and instead rely upon a “wisdom of crowds” philosophy that assumes annotators will “know it when they see it” [44, 233]. Such problems have been called out as critical to downstream methodological problems of inconsistent classifiers, overfitting, and a dependence on highly complex yet uninterpretable models [56, 214].

By exploring online hate in terms of well-established psycholinguistic categories, this research offers a distinct, conceptually embedded understanding of its empirical features. Hate as an emotion is described in precise theoretical terms in social scientific scholarship, but this work is not well-reflected in the computational literature. In utilizing a series of psycholinguistic dictionaries based in prior research [126, 194, 241], my work pins down important measurable dimensions of what makes certain texts hateful and distinct from non-hateful utterances. By demonstrating the robustness of these associations across datasets, languages, and platforms, my findings moreover suggest that these properties lie at the core of online hate as opposed to incidental features of a particular research project’s data collection or annotation procedures.

The approach taken in this thesis thus has the benefit of linking online hate to a wealth of other psycholinguistic and social phenomena. More specifically, for instance, by uncovering the systematic use of social identity terms in online hate, I point to a specific social mechanism at work in their attacks of “protected characteristics”. This further suggests explicit linkages to the wealth of social scientific research on social identities [236, 237]. This especially opens up promising dialogue with the rich scholarship that shows how social identities are leveraged in social influence and social change [62, 77, 108, 225]. Other such linkages are also worth exploring, such as online hate’s connection to pronoun usage, linguistic complexity, and other emotion-laden language, such as fear and terror.

7.1.2 Semantic Systems of Online Hate

Going beyond questions of identification, the computational social science literature has also thus far produced relatively shallow analysis beyond binary distinctions of hate and non-hate. Computational models of online hate have extended beyond this dichotomy primarily by specifying various subtypes of harmful language—such as offensive language, toxic language, and cyberbullying that falls short of hate—or by quantifying targeted forms of hate—such as misogyny, anti-semitism, or anti-Black racism [58, 87, 262]. While such methodological approaches have certainly had productive benefits in advancing the state-of-the-art in online hate detection, they can only go so far as to perform aggregate analyses of overall amounts of hate without nuancing potentially distinct modes of degradation, humiliation, and threat. This stands in contrast to foundational psychological work on hate as an emotion intimately tied to socio-culturally shared stories of alienation, othering, and dehumanization [229, 230].

The work in this thesis moved beyond these prevailing approaches by understanding online hate in terms of narratives. Through narratives, online hate is expressed in terms of a collection of social actors, to whom various characteristics and relationships are ascribed, and from which emerges a program for action. This shifted from dichotomies of hate and

non-hate to a more holistic understanding of how hate attacks its targets. Empirically, this approach specifically yielded insights into how the same collection of targets—e.g., gender-based identities—can be subjected to different types of hateful narratives, such as those which demean a woman’s intelligence versus those which assert physical and sexual violence against them. Meanwhile, apparently distinct social groups—e.g., racial and religious identities—may be the subject of a shared constellation of hateful constructions, particularly an overarching narrative of white supremacy that posits the inferiority of those deemed outside its dominant center.

Drawing on the wealth of research on cognitions and beliefs as networked phenomena [46, 234], this thesis also developed and utilized a fruitful operationalization of online hate narratives as semantic systems. Semantic network analyses have a rich tradition in understanding collective perceptions and meaning construction [48, 49]. While this thesis primarily focused on central nodes within these semantic networks, other network properties could also be explored in relation to their corresponding semantic features. Clusters within semantic networks, for instance, might indicate internal variation or core-periphery hierarchies of meaning within a given narrative [30, 243]. Properties of the network as a whole could be examined in relation to the overall complexity of a given narrative [51, 263]. Features of semantic networks could also be analyzed in relation to the contexts or effects of their constituent utterances [125]. For instance, might they predict more effective or influential hate narratives? Such questions are promising avenues of research opened up by the work presented here.

7.1.3 Interactions and Communities of Online Hate

Another characteristic of prevailing approaches to online hate has been its implicit focus on individual texts and actors in isolation. When empirically considering large-scale online conversations, past studies have typically used aggregate statistics without accounting for key forms of social variation, particularly in the sense of group structure and intergroup relationships. Some recent studies have incorporated an explicit recognition of groups in governing emotional dynamics online, but they conceptualize groups as broad categories of membership, whereby a given individual is merely a member or non-member [34, 35, 206]. This does not account for how online groups act as structured organizations, each with their own distinct size, shape, and hierarchy relative to the broader social system. This also does not acknowledge how individuals may be specifically positioned and connected to others within an online community as leaders, bridges, or peripheral participants.

The social network perspective adopted by this thesis highlights multilevel insights: that not all online hate groups are built equally, and that not all individuals within online hate groups act with equal significance. By attending to several measurable features of network clusters, I showed how higher levels of online hate could be observed in relation to key theoretically motivated patterns of group structure [250]. Furthermore, through multilevel analysis, I surfaced key actors within these groups, especially highlighting the role of leaders and proselytizers who expose other accounts to hateful content, which in turn, predicts their production of online hate [8, 253]. Because these patterns were obtained through observational study, more precise social and psychological mechanisms cannot

be causally inferred [176, 177]. But they point to large-scale regularities that indicate promising avenues of more controlled analysis.

While these findings were consistent across the contexts studied, they may not necessarily hold across all online conversations or social media platforms. Previously documented phenomena such as echo chamber and trench warfare effects illustrate how interactions between opposed groups can at times be characterized by mutual isolation, versus moments of recurrent hostility [18, 128, 273]. From a similar point of view, the network signatures identified here may point to one among a range of possible configurations of hate groups. Future work may look into what other broad structural archetypes might exist in other settings, the conditions associated with certain formations over others, and potentially how these might shape other dynamics of online hate.

7.1.4 Social-Cyber Maneuvers on Online Hate

Online hate is often discussed in conjunction with other online harms such as misinformation and disinformation [33]. But only more recently have explicit links been drawn between hate and inorganic manipulation [81]. Historically, information operations have been conceived as achieving various objectives by targeting key cognitions and beliefs [24, 64]. In contrast, their capacity to shape powerful emotions like hate remains relatively understudied. Scholarship on information operations has evolved considerably in recent years, coalescing into rich analyses of major social-cyber maneuvers targeting narratives and networks [24, 27, 52]. Conversely, however, for online hate, the study of its manipulation has largely been limited to one-dimensional assessments of inorganic increases and decreases.

In this thesis, I brought narrative and network frameworks of information operations in dialogue with a narrative and network approach to online hate. From the standpoint of narratives, I showed that certain types of messages were more likely to be produced by bots and trolls than humans [64, 80, 255, 275]. Moreover, I showed a critical distinction between: (a) organic hate narratives, (b) inorganic hate narratives, and (c) inorganic non-hate narratives which may nonetheless exert manipulative influence upon the online conversation. On the network side, I also showed how bots and trolls interacted with group structures and individual accounts within their communities. Through this multilevel analysis, I highlighted how different types of social-cyber actions lead directly to increased levels of online hate, and which might contribute more indirectly to online hate by reshaping community patterns of interaction.

Taken together, these findings emphasize that the manipulation of online hate may be understood through specific narrative and network actions. This meaningfully extends existing work on online hate manipulation and situates it in the language of the state-of-the-art in social cybersecurity scholarship [24, 27]. However, while a narrative and network framework has been shown to provide analytical utility for understanding online hate manipulation, this thesis does not yet achieve the level of generality provided in social-cyber maneuver analysis more broadly. Future work may build upon this thesis by studying online hate manipulation across a variety of contexts in order to systematically map out commonly used narrative and network actions. Consistent patterns may possibly link back

to the existing seminal understanding of social-cyber maneuvers, or potentially suggest new additions to these frameworks. Furthermore, while the present work adapts general social cybersecurity insights for an analysis of online hate, this relationship would be reversed in operational contexts of social media influence monitoring. How should analyses of online hate manipulation be incorporated into an assessment of online manipulation in its broader sense? Integrating such questions into the social cybersecurity toolkit will be key for future work in this area.

7.1.5 Online-Offline Interfaces of Hate

Finally, while various theories exist for understanding hate in its societal context, their implications for online hate have remained underdeveloped. Several frameworks have also been offered linking the online world to the offline world more broadly, but with few empirical applications. At the onset of the Covid-19 pandemic, both a rise in online hate and a spike in offline hate crimes were observed [80, 109, 208, 244]. However, their interplay has largely been left to theoretical propositions rather than empirical inquiry. The penultimate chapter of this thesis thus aimed to bridge both conceptual and empirical gaps in the literature regarding the online-offline interfaces of hate.

In this work, I specifically sought to arbitrate across different conceptual configurations linking online hate, offline hate crimes, and their broader context of societal strain from the Covid-19 pandemic. Affirming a strain theory perspective [4, 164], both online hate and offline hate crimes appeared to be positively predicted by increased Covid-19 cases in a given state. But mixed evidence was apparent for theories of intensification of hate and the online-offline nexus of social media [106, 130, 189]. More consistently, it appeared that increased online hate predicted future incidences of hate crimes, while increased hate crimes only predicted online hate under particular conditions. This asymmetry suggests the distinct importance of considering the dynamics of societal information flow and its potential effects on (perceived) social norms and intergroup relations. Whereas online public discourse is available to many at a fast pace and with low barriers to access, offline events require more concerted efforts at reportage and information dissemination. Collectively, these findings suggest critical opportunities for public communication efforts to shape the level of closure in the online-offline loop. while also managing perceptions of strain during times of crisis [61, 108, 218].

Future work in this area will benefit meaningfully from three key extensions. One path would entail an expansion of the model to specifically account for the surmised mediating effect of media coverage around hate crimes [74, 154]. Assessments of the volume of coverage as well as of framing effects would more explicitly test the inferred impact of the online-offline information loop for shaping the joint trajectories of online hate and offline hate crime. A second path would be to test these models on other contexts of upheaval and violence. While the mechanisms observed are consistent with and meaningfully enrich existing frameworks, their generality beyond the particulars of the Covid-19 pandemic is not directly addressed here. Third, the specific social and psychological mechanisms underlying these findings could also be examined through alternative methodologies. As in much of this thesis, I relied on observational analysis of large-scale data through compu-

tational social science techniques. While this approach surfaces large-scale patterns, more controlled experiments may be key for more in-depth understandings of causal linkages at play, as well as meaningful junctures in these mechanisms for intervention [176, 177].

7.2 Methodological Advances

7.2.1 Tools for Online Hate Detection

Alongside the overarching theoretical insights of this thesis, several methodological contributions are also worth noting. The primary tool used across my research is an online hate detection tool that combines psycholinguistic dictionaries with machine learning for text classification [251]. This model approaches the performance of state-of-the-art techniques in natural language processing, while also meeting crucial yet often-overlooked criteria. Due to its theory-based feature design, this model captures meaningful properties of texts that relate to its expression or non-expression of online hate in ways that cut across datasets, platforms, and languages. This contrasts with more complex, purely data-driven models that might achieve higher in-domain accuracy, but fail to generalize as well due to problems of overfitting. Furthermore, the model’s controlled parameter size enables it to run much more scalably over large datasets, with significant gains in efficiency relative to computationally expensive deep learning and transformer models. Finally, its outputs are explainable as the machine learning techniques utilized straightforwardly indicate the features used to estimate its predictions, and do not rely on blackbox or high-dimensional representations.

Based on these findings, it should be straightforward to rely on this broader methodology to model other types of online harm, such as misinformation or online extremism. Such online harms could similarly be understood at scale and with theoretical depth. Because it uses psycholinguistic features determined a priori, the model relies on a conceptually coherent—rather than data-dependent—set of measurements that characterize hate in a broad sense [126, 241].

But as societal conflicts evolve, modes of expressing hate may likewise transform. Similarly, while the model is trained on a large number of datasets that make up the curated mega-corpus [201], these too are situated in their particular socio-cultural moments and inscribe ultimately subjective assumptions about hate. Lastly, a more technical concern that plagues all text classification models are problems of text and meaning distortion, such as intentional misspellings and sarcasm [219]. To address these concerns, future work may extend the present model to accommodate socio-cultural innovations in a modular fashion. For new events that produce hate-ridden flashpoints, the model could be fitted with analyst-defined ad hoc dictionaries of new terms incorporating emergent knowledge about extant conflicts. These could be made persistent for future iterations of the model, or treated as temporary additions to the feature set. Meanwhile, on the data side, techniques such as active learning or human feedback could be used to fine-tune various versions of the model to adapt to event-specific discourse. The model could also be combined interoperably with a broader pipeline that includes new modules for multilayered cleaning of

the dataset to deal with relevant distortion problems.

7.2.2 Tools for Discovery and Characterization

Following the identification of online hate in a social media dataset, tools for discovering and characterizing its narrative and network properties are key techniques developed in this work. These enable the fundamental shift from the prevailing approach in the computational social science literature, to focus more broadly on the systems of meaning and systems of people which attend the expression and spread of hate on social media.

More specifically, one of the principal techniques developed in this thesis allows for the discovery and characterization of key online hate narratives [48, 49, 162]. The methods used here are distinctly situated as human-in-the-loop pipelines which combine the strengths of quantitative text analysis while facilitating more in-depth qualitative interpretation. In particular, the proposed narrative analysis methodology improves upon commonly used quantitative frameworks by integrating hate-aware topic modeling and semantic network analysis for identifying central concepts. In addition, qualitative engagement with the discovered text clusters is bolstered with the aid of large language models which produce human-interpretable reasoning for distinguishing between topics and highlighting their semantic overlaps [274].

The second methodological pipeline for networks also enables multilevel discovery and characterization of online hate groups and key hate actors within communities [30, 71, 243]. While the analysis presented in this thesis was specifically applied to an empirical analysis of Covid-19 data, it is also a methodology that is readily replicable for any dataset featuring both text and network data. The estimated relationships are further not assumed to hold constant across all contexts, and instead, may be estimated in new contexts to facilitate social-cyber analysis of both similar and new organizational forms of online hate.

At this time, much of the presented methodologies have already been integrated into the ORA software [9, 54]. Key features of the narrative methodology have now informed extensions of the Semantic Network Report. Through the Online Harms report, analysts are also able to obtain network-based insights into the same kinds of group properties examined here. Drill-down analyses a unique feature of the ORA reports, as they allow researchers to focus on particular texts, actors, or groups to better understand their activity. Most notably, the framing of the report in terms of online harms rather than online hate expands the methodology to apply to a range of problems on social media. In the case of the narrative methodology, other target variables may be used instead of online hate, allowing for not just hate-aware topics, but possibly also fear-aware or misinformation-aware topics. Multilevel network analyses may likewise be conducted, similar to those here, on issues such as rumor-spreading, spamming, and cyberbullying, among others [52, 58, 81].

These techniques offer flexible and theoretically rich analyses of text and network data, and future work may wish to build upon these advantages. For the narrative methodology, the analytical pipeline could be extended toward further integration of its modules. For instance, the large language model could be directly fed sample texts and results from the semantic network analysis to produce topic labels. On the network side, a wide range of network properties could be used to expand the proposed methodology. While I fo-

cused on some of the most theoretically motivated properties, other measures of centrality and group structure could add greater depth to the analysis and lead to novel analytical insights beyond those probed here. Lastly, it may be worth considering an integrated narrative-network methodology, whereby community detection and text clustering are inferred jointly. In line with a truly structural theoretical framework, distributions of word co-occurrence and group interactions would be estimated by a single statistical model that accounts for their interrelationships.

7.2.3 Tools for Social-Cyber Assessment

Alongside the functionalities provided for narrative and network analyses of online hate, techniques for manipulation assessment have also been developed in this thesis. At their basic level, they introduce straightforward extensions of the previously developed narrative and network techniques by incorporating their statistical relationships with bot and troll activity [23, 80, 275]. This specifically quantifies the types of narratives being propagated by social-cyber actors, as well as their network positions and associations with various community structures.

The network modules of this analysis have also been integrated into the ORA Online Harms report as optional parameters. Just as in the case without explicit manipulation assessment, the broad framing of this report allows for network manipulation assessment in the context of harms beyond online hate. Bot and troll manipulation of other emotions and types of information flow (e.g., disinformation, conspiracy theories) could thus also be subjected to the analyses presented here. Moreover, another flexible component of the report as it currently stands is that social-cyber actors are not limited to just bots or trolls [33, 81]. The interoperability of the report allows for any other actor characteristic to serve as the focus of potential manipulation, thus for instance allowing for manipulation analysis from the perspective of state-sponsored accounts, sockpuppets and astroturfers, or even human accounts with known influencer roles [252].

7.3 Practical Implications

7.3.1 For Platforms

This thesis suggests several practical implications for social media platforms. Social media platforms are the spaces within which online hate propagates, and while they have not successfully eradicated it, they have made considerable progress in their surveillance and mitigation efforts over the years. Insights from my research do not promise a one-size-fits-all solution to online hate across all social media platforms, but they may inform the design and deployment of innovations to improve existing measures.

Table 7.1 provides a summary of the definitions of online hate used by popular inter-

Table 7.1: Summary of platform definitions of online hate.

Platform	Definition
Twitter/X	You may not directly attack other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease.
Facebook/Meta	We define hate speech as direct attacks against people — rather than concepts or institutions— on the basis of what we call protected characteristics (PCs): race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity, and serious disease.
YouTube	Hate speech is not allowed on YouTube. We don't allow content that promotes violence or hatred against individuals or groups based on any of the following attributes, which indicate a protected group status under YouTube's policy: Age, Caste, Disability, Ethnicity, Gender Identity and Expression, Nationality, Race, Immigration Status, Religion, Sex/Gender, Sexual Orientation, Victims of a major violent event and their kin, Veteran Status.
TikTok	Hate speech and hateful behavior includes attacking, threatening, dehumanizing or degrading an individual or group based on their protected attributes. Protected attributes mean personal characteristics that you are either born with, are immutable, or it would cause severe psychological harm if you were forced to change them or were attacked because of them. These include race, ethnicity, national origin, religion, tribe, caste, sexual orientation, sex, gender, gender identity, serious disease, disability, and immigration status.
Reddit	Remember the human. Reddit is a place for creating community and belonging, not for attacking marginalized or vulnerable groups of people. Everyone has a right to use Reddit free of harassment, bullying, and threats of violence. Communities and people that incite violence or that promote hate based on identity or vulnerability will be banned. Marginalized or vulnerable groups include, but are not limited to, groups based on their actual and perceived race, color, religion, national origin, ethnicity, immigration status, gender, gender identity, sexual orientation, pregnancy, or disability. These include victims of a major violent event and their families.
Telegram	By signing up for Telegram, you accept our Privacy Policy and agree not to promote violence on publicly viewable Telegram channels, bots, etc.

net applications and social media platforms including Twitter¹, Facebook², YouTube³, TikTok⁴, Reddit⁵, and Telegram⁶. These definitions are the basis around which these platforms take action, and this language typically appears in terms of service to govern user behavior. Notably, as pointed out throughout this thesis, these definitions enumerate a list of various “protected characteristics” against which attacks, threats, and derogatory language are explicitly prohibited. Some platforms have longer lists of such characteristics, and others distinctly include victims of historical events as an additional special category for protection. Across platforms, the type of action which constitutes hate against these

¹<https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy>

²<https://transparency.fb.com/policies/community-standards/hate-speech/>

³<https://support.google.com/youtube/answer/2801939?hl=en>

⁴https://www.tiktok.com/safety/en/countering-hate?sc_version=2024

⁵<https://support.reddithelp.com/hc/en-us/articles/360045715951-Promoting-Hate-Based-on-Identity-or-Vulnerability>

⁶<https://telegram.org/tos>

Table 7.2: Summary of platform interventions against online hate.

Intervention	Twitter	Facebook	YouTube	TikTok	Reddit	Telegram
Content Removal	✓	✓	✓	✓	✓	✓
Account Suspension	✓	✓	✓	✓	✓	✓
Community Ban					✓	
Warning System	✓		✓	✓		
Rehabilitation Program			✓			
Discovery Downranking	✓			✓		
Interaction Limits	✓			✓		✓
User Reports	✓	✓	✓	✓	✓	✓
Human Moderators	✓	✓	✓	✓	✓	✓
AI Filters	✓	✓		✓		✓

groups may also vary, including more general language of “attack” and more specific mentions of “promot[ing] violence”, “dehumanizing or degrading”, and “harassment, bullying, and threats of violence”. Some platforms, like YouTube and Reddit, are more expansive in their protections, such as including “kin” or “families” of specially identified demographics. Other platforms, like Telegram, are notably more narrow in their stipulations, as in specifying only the promotion of violence on “publicly viewable Telegram channels”.

From a definitional standpoint, one possible contribution of this thesis stems from its use of a theory-based online hate detection model which emphasizes social identities [251]. The protected characteristics which tend to feature in platform definitions of online hate may be more fundamentally understood as social identity categories, thus aligning with the broad approach taken in this work. However, since different platforms emphasize different collections of social identities, the categories utilized in this thesis may need to be expanded or refined accordingly. Similarly, different types of attacks may also necessitate specialized dictionaries for abuse. Such enhancements are readily achievable due to the modular and interoperable properties of the proposed detection model. New dictionaries may be integrated in an adaptive fashion, and platforms may readily apply similar models to achieve their specific objectives in a scalable, general, and explainable fashion. These approaches could also be applied to the study of online harms more broadly, both at scale and with conceptual depth.

Meanwhile, upon identifying online hate, platforms may also use insights from this thesis to build upon existing intervention strategies. Table 7.2 summarizes measures commonly taken across social media platforms. Here, it is evident that the platforms enumerated commonly rely upon content removal and account suspension to mitigate the spread of online hate. Reddit, due to its organization into forum-based communities, additionally has a community ban option. Across platforms, various middle-ground interventions are also available based on the severity or repetition of violations. Twitter and TikTok have warning systems that prevent unwitting violators from posting certain forms of speech in the first place. Meanwhile, YouTube implements its warnings with a “three strikes” system to discourage first-time violators against repeat offenses. These warnings are coupled with mandatory rehabilitation programs that teach users how to abide by community guidelines on the website. Other intermediate mitigation efforts include engagement throttling, such

as downranking or limiting interactions with content found to be suspicious or in violation of relevant guidelines. Finally, to identify problematic content, all the reviewed platforms explicitly rely upon the reports by other user accounts, and a mix of human and algorithmic moderation.

From this perspective, insights from this work may specifically inform three key facets of existing intervention efforts. First, many of the interventions enumerated require some level of manual review by the user community or by specialized moderators. For these processes, social media platforms emphasize that manual reviews account for the “context” of relevant utterances, such as for instance, in the colloquial use of racial terms by ingroup members compared to their derogatory use by outgroup members. However, not much explicit detail is provided regarding how reviewers are meant to account for the context of utterances. Narrative techniques such as those presented here may be helpful for holistically and empirically mapping out connections between concepts in utterances under review [48, 49]. For instance, in under what narrative context is a certain racial epithet used—to dehumanize an outgroup or as an expression of ingroup solidarity? As shown here, interpretation of these broader semantic contexts—by both researchers and moderators in applied settings—are aided by the use of large language models which can serve as an informed and relatively autonomous co-annotator. These insights would also be critical for understanding the semantic context of other online harms, such as disinformation and conspiracy theories. Taken together, these could assist in mapping out the broader social stories within which harmful speech and behavior are embedded, as well as diagnose their underlying and emergent social conflicts.

Second, network techniques may be richly informative for moderating online hate in terms of its community-driven spread [250]. The interventions spelled out by social media platforms tend to look at content and accounts on a case-by-case basis, which may limit their scalability especially as online hate spreads in groups. Given the unique data platforms have access to, network techniques developed in this thesis could for instance be used to map out user-generated community reports to identify interconnected problematic accounts. A network perspective would likewise enable the identification of leaders and bridges such as those studied in this thesis, thus allowing for more impactful targeting. Across different platforms, network-based hate mitigation techniques can thus be more surgically precise: using early detection techniques to determine whether a hate group is growing around an influencer, or perhaps focusing on the social networks among haters rather than other individuals, and so reduce the likelihood of non-violating individuals becoming flagged. Similarly to the aforementioned techniques, this detection pipeline should be expanded to other forms of online harms as well.

Finally, when the social media platform detects manipulative accounts, including but not limited to bots and trolls, these techniques could also be shed greater light upon their operational lines of effort [52, 253]. Shared narratives could facilitate deeper understandings of coordinated propaganda campaigns, thus allowing for platforms to better understand what types of influence are being exerted by malign actors. Just as in this thesis, by observing the network structure of online hate groups, manipulative accounts that might be instigating hate without themselves violating the terms of service could also be subject to more nuanced review. Conversely, bots and trolls could also be better removed on

the basis of their action—such as bridging hate groups or growing an influencer’s hate communities—rather than relying solely on whether or not they are a bot or troll. By relying on a narrative and network system perspective, platforms can thus shift attention from isolated content and individual accounts to better moderate collective issues of online hate and harm.

7.3.2 For Policymakers and Governance

Outside the purview of social media platforms, governments also play an important role in addressing the problem of hate in both its online and offline forms. For the specific case of hateful speech, a range of policies have been implemented around the world. Global variation in this area of legislation is considerable, and for the specific case of online hate speech, progress has been fairly limited. Table 7.3 summarizes some features of key policies in different countries.

Table 7.3: Examples of hate speech policies around the world.

Policy Type	Example Countries
No strict regulation of speech	United States
Options for victim redress	Australia, India
Prohibition with fines/prison	Brazil, Canada, Chile, Germany, Russia
Specific forms of hate	Belgium, Finland, New Zealand, Nigeria, Singapore
Social media regulation	Germany, Singapore

In the United States, it is particularly significant that policies to regulate any speech are extremely limited. Under the First Amendment, extensive protections of speech are maintained, with jurisprudence indicating a high bar of direct incitement to violence to render a given utterance as criminally liable [32]. In other parts of the world, criminally liable hate speech is subject to different definitions, standards, and a range of penalties. For Australia and India, options are available for victims to seek redress from violators, particularly in some form of monetary compensation [270]. In countries like Brazil, Canada, Chile, Germany, and Russia, hate speech is more generally prohibited, with stricter penalties including fines and prison sentences for violators [6]. Meanwhile, in places like Belgium, Finland, New Zealand, Nigeria, and Singapore, extant regulations do not so much cover hate speech per se as they do a specific form of targeted speech [97]. For instance, some countries only cover racially charged hate as in Nigeria’s case of “ethnic hatred”, whereas Belgium’s laws are specifically concerned with Holocaust denial. Such variations tend to reflect the historical relevance of particular forms of hate in these contexts.

Laws specifically targeting hate speech on social media are considerably rarer and in their nascent stages. Two examples of particular interest are those of Germany and Singapore. In Germany, relevant laws levy substantial fines on the social media platforms themselves for failing to remove violating content within a narrow time frame [245]. Meanwhile, in Singapore, a specific ministry is given authority to decide on the mandated removal of content that violates its online standards of conduct [85]. Both examples have had mixed

results and have come into ambivalent reception, with critics especially highlighting difficulties of enforcement and blanket suppression of online material based on challenging standards.

Against this policy backdrop, as in the platform case, the narrative and network methods produced in this thesis may aid to some extent with regard to enforcement. Narrative techniques allow for a semantic contextualization of utterances of interest, allowing for a more nuanced understanding of content deemed to be problematic under these legal frameworks. Meanwhile, network techniques situate these utterances in the group context, which may specifically help with strict removal mandates as well as key distinctions between leaders and followers when dealing with online hate groups. Ultimately, however, definitions of what precisely constitutes a harmful utterance need to be carefully deliberated upon across societal contexts. The tools described are designed to operate across socio-cultural milieus, but their findings are best interpreted in conjunction with broader understandings of these factors.

Finally, when considering the impacts of hateful speech, especially in places like the United States where the standard for liability is an incitement to harm, this thesis also suggests important avenues for governance. Techniques such as those used here should be incorporated as part of public health monitoring to identify areas or time periods of particular societal risk. For future crises, joint observations of online hate and offline violence may indicate whether, when, and where one predicts the other, thus prompting appropriate action. To improve the ability to determine if a given place or time is becoming more conducive to hate, hate crimes should thus be better tracked and the data made available on severity, location, time, and possibly what media platforms the culprits used. It is vital that data collection infrastructures be strengthened to make key assessments possible. To do this will require both direct improvements to reporting procedures, data storage systems, and access mechanisms for researchers and relevant agencies; as well as decreasing barriers that result in underreporting of hate incidents by the community [196]. Moreover, when news media cover such incidents, sensationalism should likewise be avoided and framing should ensure that acts of violence are not glorified to prevent the intensification of hate [61, 74]. As crises progress within a society, government leaders, the media, and influential members of communities all have a role to play in ensuring that the experience and perception of strain has minimal impact upon intergroup relations [108, 218]. Messages that encourage solidarity and empathy, as opposed to scapegoating any particular group, will be valuable in mitigating both online and offline expressions of hate. Beyond hate, the same implications also apply for other forms of online harm linked to offline outcomes, expanding the purview of what social harm dynamics need to be monitored in relation to their offline contexts.

7.3.3 For Educators and Advocates

A third arena of practical insights from this thesis lies in the realms of education and advocacy. Previous scholarship has noted that the prevailing discourse around online hate may veer too sharply toward a regulatory understanding rather than a social one [200]. In other words, critiques have been raised regarding the construction of online hate primarily

as a problem for platform and policy regulation, thus leading to deadlocks in subjective conflicts over what constitutes hate and what should be done about it.

Efforts to communicate and teach about issues surrounding online hate may benefit from a more social perspective by highlighting its embeddedness in group dynamics. As underscored throughout this thesis, online hate arises in heated contexts of societal upheaval and conflict. Under such moments of strain, online hate becomes socially sanctioned within ingroup discourse, especially emboldened by legitimizing narratives and influential leaders [4, 164, 190]. Paradoxically, then, hateful speech and behavior may be experienced by its potential and actual propagators not as taboo but rather as an expression and a fulfillment of a given group’s values and interests. Attention can thus be fruitfully given in education efforts to these more apparently “constructive” rather than “destructive” societal functions of online hate, ideally to prompt critical in-group reflection and psychological preparedness against it, akin to inoculation efforts against misinformation [144].

Furthermore, this thesis focused on a particular fixed collection of social identity targets, and thus did not cover all possible protected categories as defined by various social media platforms or policies. But a persistent feature throughout this thesis has been its emphasis on diverse targets even within a single online conversation. From the standpoint of the targets of online hate, it can be appealing to zero in primarily on the types of attacks that affect one’s own social identities. However, while certainly differently targeted online hate may operate in distinct and nuanced ways, such tendencies can narrow down the possibilities for broader reckonings against online hate. In other words, an affiliative politics of empathy and allyship toward targets of online hate outside one’s own communities may be crucial to more effectively mitigate it in a systemic fashion [105, 209]. Taken together, these insights may allow for deeper collective reflection when conditions emerge for participation in online hate.

Beyond teaching and learning about the content and experience of online hate, these insights thus may also inform literacy, privacy, and collective action-based efforts to build grassroots resilience against online hate. Programs could be developed to aid in bottom-up identification of online hate actions on social media platforms, particularly those which take advantage of intragroup and intergroup dynamics to make individuals and groups believe in the legitimacy of participating in online hate. Sharing experiences and lessons between different communities that have been targeted by and survived hateful attacks would also be vital to enhance cohesion in the face of such forces, both organic and otherwise, which seek to divide and perpetuate conflict in human societies.

7.4 Final Notes

At the highest level, this thesis aimed to examine online hate as a social phenomenon, accounting for its linkages to shared meaning, social interactions and communities, and offline upheavals and violence. Through this work, I derived theoretical insights around these social dimensions of online hate, uncovering empirical evidence around their dynamics during a global crisis and producing novel methodologies relevant to a broad range of applications in computational social science. From a practical standpoint, implications

of this thesis also stand to inform efforts in platform design and moderation, policies for regulation and surveillance, and strategies for education and advocacy.

These multifaceted contributions notwithstanding, this thesis also featured important limitations prompting key avenues for future research. As with much previous scholarship on online hate as well as computational social science more broadly, the insights derived here are subject to the limits of data collection and annotation of labeled datasets, which are informed by imperfect and socially situated value systems [86, 201]. While these issues are addressed to some extent through multi-dataset and explainable modeling, they also demand responsive flexibility as societies evolve, and new forms of conflict and hate emerge. Furthermore, as with most research on social media, platforms do not allow unlimited access to data, and thus insights derived here are also subject to API limitations [176]. Research on other platforms, as well as explicitly multi-platform research, would also help considerably to strengthen these findings. Finally, as pointed out throughout this thesis, the computational social science approach has largely relied upon data-driven techniques based on large-scale observational studies. Human subjects research and field experiments are critical to follow up on the findings suggested here in order to provide key complementary evidence into the precise causal mechanisms at play [177].

As long as human societies are beset with conflict, they are also at risk of hate and violence. Resolving problems of online hate thus holistically goes hand in hand with broader offline efforts to improve intergroup relations and mitigate violence in society. Technology, both online and otherwise, can be incorporated into these deeply social processes as amplifiers and accelerants, but they can also hold a mirror to society that prompts reflection and response to its tensions and divisions. Moreover, the same communicative openness that allows for the large-scale proliferation of hate is also a communicative openness that has historically allowed for greater understanding and solidarity around the world. Social cybersecurity and computational social science at large aim to shed light on these complex socio-technical interactions and point the way forward toward more cohesive, secure, and resilient societies. It is hoped that this thesis contributes to this greater cause.

Appendix A

Supplementary Documentation

A.1 Model Training and Evaluation (Chapter 2)

When utilizing the entire dataset, 70% of the dataset was randomly selected for the training set, 20% was used for validation, and 10% was used for the test set. Otherwise, when performing dataset-based, language-based, or platform-based model development, one of each category was held out as the test set, while the remainder of the dataset was split between 70% for training and 30% for validation. Test accuracies and weighted F1 scores were recorded to assess prediction performance.

All model types were subjected to hyperparameter tuning at the validation stage via grid search. Best models were those which had the highest validation F1 score. The tested hyperparameters are summarized in Table A.1 below.

Table A.1: Summary of hyperparameters for grid search in online hate model selection.

Model	Hyperparameters
Logistic Regression	Lasso $C = \{0.01, 0.03, 0.1, 0.3, 1, 3, 10\}$
Random Forest	N Trees = $\{10, 30, 100, 300, 1000\}$
CNN	Embedding Sizes = $\{50, 100, 200, 300\}$ Filter Sizes = $\{32, 64, 128, 256, 512\}$ Dropout Rates = $\{0.25, 0.5, 0.75\}$
LSTM	Embedding Sizes = $\{50, 100, 200, 300\}$ Hidden Sizes = $\{32, 64, 128, 256, 512\}$ Dropout Rates = $\{0.25, 0.5, 0.75\}$

A.2 Problematic Word Features (Chapter 2)

When evaluating the explainability of word-based models, several problematic features were identified among their top-ranked features. While abusive terms and identities did also occupy some of the higher ranks in the feature space, many were also specifically

identified to include named entities, numbers, and web artifacts. Upon their removal from the feature space, word-based models notably dropped in performance relative to Netmapper-based models, indicating significant overfitting to these incidental features to training datasets. These are summarized in Table A.2.

Table A.2: Examples of top problematic features in word-based models.

Category	Examples
Named Entities	android, liverpool, barack, kardashian, wrestlemania
Numbers	8th, 2015, 24, 1969, 8220
Web Artifacts	https, www, rt, @realdonaldtrump,youtu.be

A.3 Error Analyses (Chapter 2)

To provide a general assessment of the errors made by the online hate model, Table A.3 provides several examples of key misclassifications from the online hate mega-corpus. While these do not cover all possible errors, they provide a heuristic sense of problematic areas for the model and may point to fruitful new features or auxiliary models to explore for model extensions. **Caution for readers:** Note that while abusive terms are censored, the examples for hateful topics contain extreme language.

A.4 Narrative Models (Chapter 3)

For the topic models trained, both supervised and unsupervised forms had several hyperparameters. Both were initiated with inverse document frequency weights to account to up-weight the use of rare terms and down-weight terms which appeared in most of the corpus. General hyperparameters were kept at standard values of $\alpha = 0.1$ and $\eta = 0.01$ based on the Tomotopy library. Grid search was used to determine a desirable number of topics and each unique terms’s minimum number of occurrences for retention in the vocabulary. Degenerate topics containing fewer than 100 texts were then removed to produce final sets of topics. Figure A.1 shows the log-likelihood curve for the mega-corpus, indicating a meaningful elbow for 10 topics and a 10-occurrence threshold.

A.5 Sample Texts in Mega-Corpus (Chapter 3)

In Chapter 3, several key narratives of interest were selected based on their estimated hate coefficients. Below, sample texts are provided for both topics that were subjected to further analyses and those which were not. These provide some qualitative insight into the content of each topic, both hateful and non-hateful. **Caution for readers:** Note that while abusive terms are censored, the examples for hateful topics contain extreme language.

Table A.3: Examples misclassifications by online hate detection model.

Error	Category	Sample Texts
False Positives	Non-Hateful Abuse	Don't even try and play this off on me asshole. Somebody post's about coming back from a funeral only to find that her lifelong pet had also died over the time that she was gone and you decide to try and spin it off and try to troll me?
	Discussions of Hate	Absolutely and the police have arrested him loads of times, there let him out tho islamophobia and all that crap.
	Pornographic Content	I'm so horny looking for someone who can help me to cum horny sexy milf sexy cumtribute domina gay bisexual.
	Racial Colloquialisms	All them big a*s n*ggers on the field and the kickers did all the work.
False Negatives	Implicit Hateful Sentiments	I am for this to fill our current refugee quotas but if they start voting Democrat I am gonna vote to send them back.
	Hateful Neologisms	Hi I'm fauxlocaust and I'm running for King of Utopia.
	Lacking Multimodal Context	@[user] N*gga. [url].
	Rare Variants of Epithets	City life degrades with each n*gglet born out of wedlock.

A.5.1 Hateful Topics

1: Degrading Misogyny

- Just me or hot blondes on mkr not at all hot Just average
- Can all my followers report this dumb c*nt for me
- I m not sexist But I usually walk away when women sports announcers are on Sports-center

2: White Supremacy and Xenophobia

- some people do behave like animals and its funny how certain groups or races so to speak have a really distinct way of behaving that is indeed more primitive or advanced than others although its not something we use a lot to talk about people being racist is calling black people monkeys
- those days are gone sadly we are stuffed we have now Burkas hijabs and Barbarians all over the place we need arevolution and start thining them out IMMEDIATELY
- white privledge is a myth hate speech is a myth and islam fosters terrorism happy sunday everyone

3: Sexual Domination of Queer People

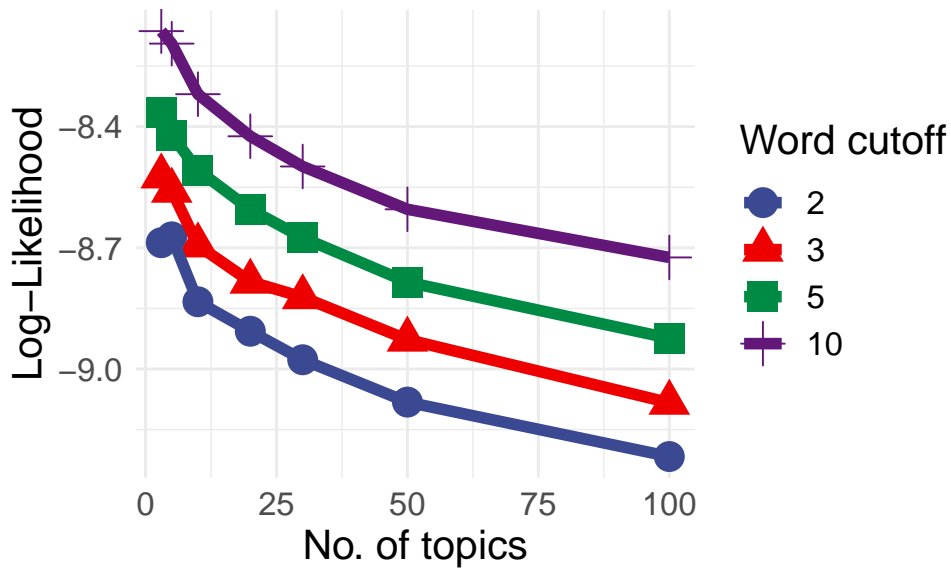


Figure A.1: Log-likelihood curve over varied numbers of topics and minimum term occurrence thresholds for vocabulary retention.

- Look at that loser f*ggot go Choking on d*ck like a good s*ssy boy
- F*ggot suck a d*ck you mad cause I got offered to Harvard you thot
- I m ready are you f*ggot barebacked chastity

4: Anti-Progressive Backlash

- I notice that the LGBT cancer is starting to slowly affect SA countries like Japan
- transgenders need help they are sick in the head Civilization has messed up the world Soon enough people will claim to have the right to change to animals and it will be seen as human rights or right to choose
- Who isn t now labeled racist now if white unless your a virtue signaling c*ck wear that shit like a badge of honor and pick up swastika and start boot stopping some n*gger f*ggot commie face in

A.5.2 Non-Hateful Topics

5: Reclamation of Queer Slurs

- LGBTQtwitter people call me d*ke like it hurts my feelings
- BOYS I m a submissive s*ssy f*ggot who loves real men
- Not to be a dramatic d*ke but I will die for her

6: African-American Vernacular English

- N*gga pass me my Tims
- So I m at the game and I just wanna smack this n*gga head so bad to hear what it sound like
- This hood n*gga got me twerking for the d*ck

7: Racially Charged News Coverage

- Triggered Redneck In Texas Yells At Couple Protesting A Confederate Monument
- Van Dyke Defense Opens with Testimony of Laquan McDonald’s Past
- Will Dick van Dyke Revive Bad Cockney Accent In Mary Poppins Returns

A.6 Network Models (Chapter 4)

In previously published work [250], group-level estimates were conducted with integrated nested Laplace approximation in order to obtain coefficient distributions and quantify uncertainty. In this thesis, more standard regression models are utilized. Group-level regressions are performed with group-level hate as the target variable and group-level features as the predictors. Multilevel regressions, on the other hand, incorporate both account-level and group-level features with account-level hate scores as the target variable. These are conducted with cluster-robust standard errors to account for systematic variation between accounts in different clusters.

A.7 Account Similarity in Clusters (Chapter 4)

Exploratory analysis was conducted to examine the similarity of accounts belonging to the same network cluster. While community detection is a standard procedure in network science, the groups it finds are based on empirical interaction patterns. As argued in Chapter 4, this is a strength of the approach in this thesis, as it accounts explicitly for group structure as opposed to viewing communities in terms of abstract membership. However, using a random sample of account pairs ($N = 4000$) which shared and did not share network clusters, it was observed that members of the same network cluster also have similar speech patterns as indicated by psycholinguistic cues. Specifically utilizing hate scores and account identity terms, account pairs of the same cluster had an average cosine similarity of 0.0278 while account pairs from different clusters had an average cosine similarity of 0.0195. A Welch two-sample t-test showed that this difference was significant ($t(3870.8) = 2.540, p < .05$).

A.8 Manipulation Models (Chapter 5)

Supervised Topic Model The supervised topic model, which featured both bot and troll scores alongside online hate scores, was initiated with inverse document frequency

weights to account to up-weight the use of rare terms and down-weight terms which appeared in most of the corpus. General hyperparameters were kept at standard values of $\alpha = 0.1$ and $\eta = 0.01$ based on the Tomotopy library. Grid search was used to determine a desirable number of topics and each unique terms’s minimum number of occurrences for retention in the vocabulary. Figure A.2 shows the log-likelihood curve for the mega-corpus, indicating a meaningful elbow for 10 topics and a 10-occurrence threshold.

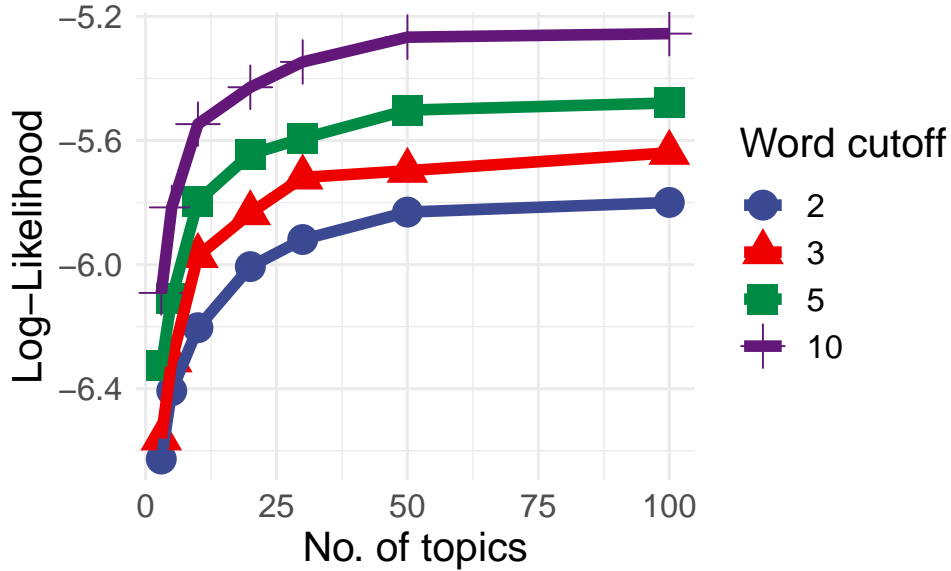


Figure A.2: Log-likelihood curve over varied numbers of topics and minimum term occurrence thresholds for vocabulary retention.

Lagged Poisson Regression For the lagged Poisson regression models, counts were used directly. Lags were set to one day and both concurrent and lagged values were included as predictors. All topics were treated as dummy variables. A linear time component was also included as a covariate to remove temporal trends.

Multilevel Network Model As in the case without bots and trolls, a multilevel model was used to incorporate both account-level and group-level features with account-level hate scores as the target variable. These are conducted with cluster-robust standard errors to account for systematic variation between accounts in different clusters. The primary difference is that bot and troll features of accounts as well as the level of bot and troll interactions are included as account-level features.

A.9 Sample Texts in Covid Dataset (Chapter 5)

A.9.1 Selected Topics

1: Trump Virus

- You should resign and let an adult handle this before more people die from your incompetence TrumptheWorstPresidentEVER TrumpVirus
- Trump at Mar a Lago last weekend has tested positive for Coronavirus maybe there is a God after all COVID19US Covid19
- trump withheld testing starting in mid January in an attempt to keep COVID19US numbers low in order to benefit his re election chances The death of thousands or millions will be on his watch and his hands

2: China Virus

- Its the chinese wuhan kung flu I just wanted to get all the racist xenophobic terms in one sentence
- It s Wu Flu Kung Flu Wuhan virus Bat Soup Bug Sweet n Sour Sicken or China virus Period
- Your normal life is being property of the state Also I already had the Kung flu Winnie the Flu if you will Anyways yes I can protect myself because I m not a slave Too bad Hong Kong wasn t armed Cheers commie propagandist

7: Kung Flu Fighting

- Everybody was Kung Flu fighting Covid was fast as lightning It s a little bit frightening To think my Mah might be dying
- You mean the wuhan virus You mean the kung flu You mean winnie the flu
- Everybody was Kung Flu Fighting Them cats was fast as lightning In fact it was a little bit frightening But they fought with expert timing With apologies to Carl Douglas

10: Covid Hysteria

- We have all been drafted to be a soldier and health care professionals stay strong people coronavirus
- You will know its beyond capabilities in the US when the usual social media activity of the MAJOR politicians and TV media begins to dwindle down I am not seeing much confidence in the current pressers given by the experts
- Here s the timeline of trump coronavirus Statements No test no true count of infections coronapocalypse

A.9.2 Unselected Topics

3: Criticism of Liberal Politicians

- That is cliché You must realize the democrats actually are happy for Kung flu now they have an actual chance to take the White House in 2020
- Impeachment was more important to the democrats than the economy the border the opioid crisis or the Kung flu

- Flu has killed more people so far this year than the kung flu I honestly believe this is blown way out of proportion We have a lot more test which means more people will test positive but the death rate should drop way lower than projected democratHoax

4: Criticism of Conservative Politicians

- Look at these jokers Two with sh*t eating grins during a crisis clueless in the back and blowhard in the center Their only concerns are the stockmarket and getting reelected COVID19US stockmarket gopcovidfailure
- This is unacceptable during GLOBAL pandemic housegop senategop senatemajldr Pass the bill negotiated by Pelosi
- Hope the sniveling coward Republicans in the Senate do their d*mn jobs SenateGOP

5: Criticism of US Covid Policies

- Rumor has it the senatemajldr Mitch McConnell and GOP in US Senate is blocking passage of the COVID 19 Relief package Perhaps it s their way of giving the President lower poverty levels
- Hey MedTwitter let s help realDonaldTrump TeamPelosi senatemajldr US govt out brainstorm some things on response to Covid19US CoronavirusUSA that aren t thinking about eg predict the future thread
- Senate WILL NOT TAKE UP Coronavirus Bill Until AFTER Senate Recess How Many People Will Have Died By Then SpeakerPelosi

6: Social Activites During Lockdown

- Mum told me a beautiful story about one of her friends doing groceries with her youngest son in times of COVID19 COVID19US Her other kids were due to come home for the weekend and they for some reason eat a lot of eggs
- Had chinese the other night no Kung Flu here AOC
- Social distancing has everyone up in arms Meanwhile Introverts are like We ve been training our whole lives for this moment coronavirus SocialDistancing

8: Advice for Living in Quarantine

- As long as Kung Flu keeps kids in school we good The second that grade schools are sent home that is going too far
- I ve been tweeting a lot about COVID19 and wanted to compile all of the good reasonable advice I ve seen in one place So here s a short ish thread of coronavirus Dos and Don ts coronavirus
- What are you listening to Add your Covid 19 music to my shared playlist

9: Explaining Racial Connotations

- Thanks for the feedback dude with 91 followers I will take it into consideration and yep I m done taking it into consideration I took it into consideration realized that I am not an idiot or an Antifa and you calling a DEADLY disease Kung Flu makes you the a*shole
- chinese virus kung flu a white person in office said this y all n*ggas been racist and the president just proved it
- COVID 19 may have originated outside US but highlighting it as a foreign virus rather than just owning up that u didn t respond fast enough potus plz thanks for compromising our safety

A.10 Alternate Analysis (Chapter 6)

One alternate specification explored was to use the number of victims as the measure of offline violence, as opposed to the binary outcome of whether or not a crime had occurred at all. While this would allow for a more particular analysis of arguably bigger as opposed to smaller incidents, larger incidents were orders of magnitude fewer than one-victim incidents. In particular, while 19.84% of the dataset had one-victim incidents, incidents with two or more victims only comprised 3.74% of the data points.

Nonetheless, it was worth noting that some results were consistent with those reported in Chapter 6, whereas others featured noteworthy distinctions. Overall, however, the effects were still more modest, and so they are relegated to this Appendix.

Strain Effects Because the number of victims is a count variable, a Poisson regression model was used to link it to Covid-19 infections (logged). A linear time trend and the state governor’s political affiliation were included as covariates. Conceptually consistent results were observed.

Table A.4: Regression coefficients of strain model with number of hate crime victims.

Coefficient	Estimate
Intercept	-1.524 (0.101)***
Covid-19 Infections (Log)	0.302 (0.037)***
Republican Governor	-0.549 (0.051)***
Linear Time Trend	-0.687 (0.955)***

Relationship with Online Hate Similarly, when linked with online hate, a Poisson regression model was used. Concurrent analysis with just levels of online hate also showed conceptually consistent results with those reported in Chapter 6. A linear time trend was included as a covariate.

Table A.5: Regression coefficients of model linking online hate to number of hate crime victims.

Coefficient	Estimate
Intercept	-1.079 (0.047)***
Online Hate	0.583 (0.128)***
Linear Time Trend	-0.381 (0.089)***

Lagged Moderation Finally, for an integrated analysis with the alternate variable, lagged moderation analysis was performed. Interestingly, the online-to-offline results are largely non-significant in this case when considering time lags. This shows that focusing on the number of victims is a weaker signal in general for the link between online hate and offline hate crimes, especially given their relatively smaller level of variation. While strain effects remain consistently significant, the main exception of note here is that the lagged number of victims may be more predictive of future online hate in Republican-run states. While this effect is not particularly strong, it is noteworthy since it departs from the asymmetrical findings reported in Chapter 6.

Table A.6: Lagged moderation models linking online hate to number of hate crime victims.

Coefficient	Estimate (Online to Offline)	Estimate (Offline to Online)
Intercept	-1.506 (0.129)***	-0.817 (0.054)***
Online Hate (Lagged)	0.114 (0.107)	0.149 (0.044)***
No. of Victims (Lagged)	0.192 (0.095)*	0.044 (0.053)
Covid-19 Infections (Lagged)	0.305 (0.046)***	0.174 (0.019)***
Republican Governor	-0.517 (0.063)***	0.055 (0.029) ⁺
Linear Time Trend	-0.725 (0.123)***	0.627 (0.057)***
Hate x Covid (Lagged)	-0.024 (0.035)	0.066 (0.015)***
Hate x Rep. Gov. (Lagged)	0.146 (0.057)*	0.094 (0.028)***
No. of Victims x Covid (Lagged)	0.031 (0.032)	-0.012 (0.019)
No. of Victims x Rep. Gov. (Lagged)	-0.009 (0.038)	0.052 (0.032) ⁺

Appendix B

Tool Implementation

B.1 Online Harms Report

The Online Harms report can be run on ORA (v 3.0.9.171) using any dataset that contains both text and network information. Given this dataset, Netmapper cues need to be added as attributes to both actor and text nodesets. Harm attributes should similarly be added to both actor and text nodesets, with the additional option of including social-cyber actor predictions to actor nodesets to assess manipulation.

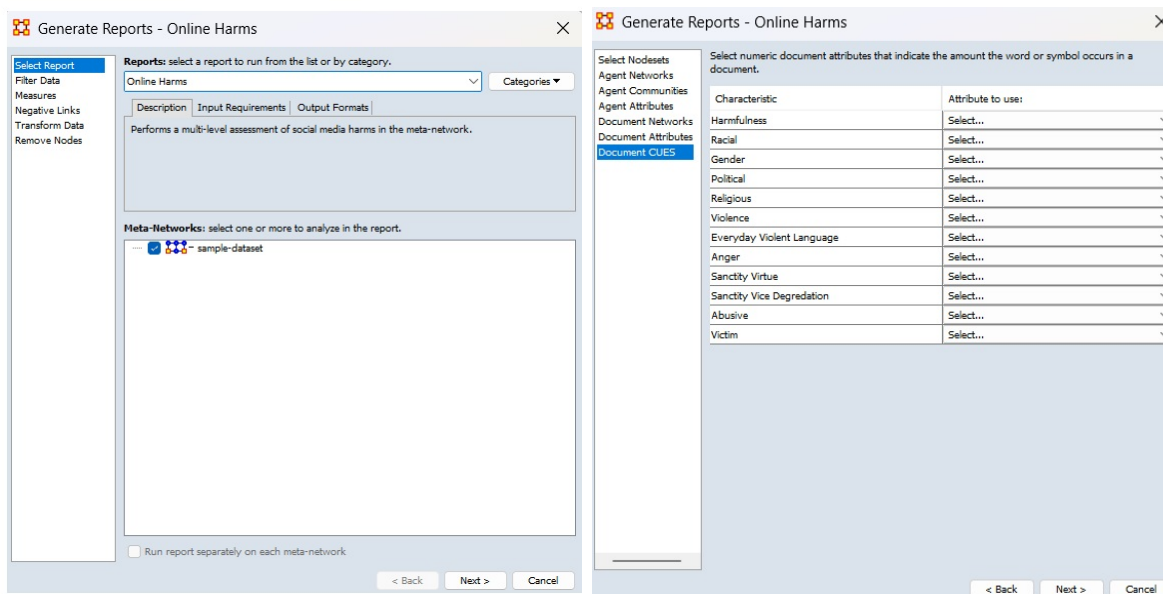


Figure B.1: Input interface for ORA Online Harms Report.

Once attributes have been added, ORA automatically pre-populates most fields in the report interface. However, the harm field needs to be explicitly chosen in the Agent Attributes and Document CUES tabs. The report is then run and produces a multilevel output file, allowing both general analyses of the dataset as a whole, as well as group-level and agent-level drill-downs.

Network-Level Analysis

Click the link below for a network-level analysis.

[Network-Level Analysis](#)

Community-Level Analysis

Click the link below for an agent community-level analysis.

[Community-Level Analysis](#)

Agent-Level Analysis

Click the link below for an agent-level analysis.

[Agent-Level Analysis](#)

Figure B.2: Multilevel directory of ORA Online Harms Report output.

B.2 Online Hate Detector

The online hate detector is set up as an executable Python script that calls on a pickled machine learning model to produce online hate predictions in probability or binary form. It is currently set up as part of the interoperable computational pipeline of the Center for the Analysis of Social and Organizational Systems (CASOS).

A directory is set up for iterative retraining of the machine learning model. This is particularly critical as updates to Netmapper dictionaries are continually made. Multiple models can be saved to the directory and selected as needed depending on the features chosen for the input data. Training data is also provided in pre-processed form and may be fed through successive Netmapper versions as they are developed. A single line of code is then used to retrain the model:

```
python3 train-hatedetector.py [cues TSV filename] [Netmapper version]
```

For a given input file, the Netmapper cues need to be produced and stored in the input folder. Outputs will be saved in the output folder. A single line of code is then used to feed the cues through the selected machine learning model:

```
python3 runrollhunter.py [model version] [cues file]
```

Document Harmfulness Scores

Average harmfulness score:	0.385	
Number of documents with harmfulness score:	48,928	45.83% (of total documents)
Number of High-Harm documents (documents with score above the mean):	22,856	21.41% (of total documents)
Number of High-Harm documents with Gender content	2,887	2.70% (of total documents)
Number of High-Harm documents with Political content	6,794	6.36% (of total documents)
Number of High-Harm documents with Racial content	7,418	6.95% (of total documents)
Number of High-Harm documents with Religious content	7,439	6.97% (of total documents)

Agent Group Harmfulness Correlations

This section reports the correlations of group agent harmfulness and group features. Group agent correlation is the average harmfulness score of agents in the group.

Correlation Description	Value
Group E/I score:	0.650
Group Density:	-0.397
Group Cheeger score:	0.568
Group BOT score:	0.443
Group agent mentions targeting:	-0.466
Group agent propagation targeting:	0.613

Community-Level Analysis

This section gives statistics on each Agent Community calculated. The values in the table below for Harmfulness, BOT score, and targeting are computed as the average value of the group's agents.

Agent communities are defined by the Leiden grouping algorithm run on the union of the agent interaction graphs.

Copy Print Search:

Group	Size	Density	Cheeger Value	E/I Index	Avg. Harmfulness	Avg. BOT Score	Avg. Gender	Avg. Political	Avg. Racial	Avg. Religious
1	5138	0.005	25.873	-0.043	-274.341	0.615	0.157	0.242	0.017	6.018e-04
2	1378	0.012	61.245	0.576	-368.998	0.470	0.109	0.145	0.015	8.813e-04
3	1165	0.022	32.979	0.127	-416.727	0.295	0.097	0.107	0.008	5.936e-04
4	845	0.011	74.359	0.769	-385.406	0.420	0.103	0.146	0.012	4.748e-04
5	582	0.033	30.256	0.217	-425.687	0.285	0.062	0.080	0.007	0.001
6	412	0.071	31.221	0.033	-402.502	0.391	0.099	0.107	0.010	0.003
7	386	0.067	12.091	-0.362	-729.834	0.365	0.089	0.055	0.016	0
8	186	0.102	30.871	0.239	-424.296	0.382	0.066	0.090	0.005	1.734e-04
9	182	0.052	53.967	0.701	-312.874	0.429	0.081	0.141	0.018	7.849e-04
10	99	0.166	44.909	0.463	-443.990	0.404	0.086	0.166	0.025	1.656e-04

Showing 1 to 10 of 14 entries

Previous 1 2 Next

Agent-Level Analysis

This section gives statistics for the agents that have the highest harmfulness score.

Copy Print Search:

Rank	Agent ID	Agent Label	Harmfulness	BOT Score	Mentions Targeting	Propagation Targeting	Gender	Political	Racial	Religious
1	1226517348116336640		1	0	4	2	0	0	0	0
2	93637779		1	0	1	0	0	0	0	0
3	1249361337244844032		1	0	78	3	0	0	0	0
4	1290599170974183424		1	0	42	4	0	0	0	0
5	1057444065682305024		1	1	294	71	0.158	0.140	0	0
6	229440231		1	0	104	35	0.100	0.100	0	0
7	1293621103013900288		1	0	14	6	0	0	0	0
8	703437218258382848		1	0	2	1	0	0	0	0
9	620571475		1	0	20	0	0	0	0	0
10	553440611		1	0	4	1	0	0	0	0

Showing 1 to 10 of 84 entries

Previous 1 2 3 4 5 ... 9 Next

Figure B.3: Multilevel directory of ORA Online Harms Report output.

Bibliography

- [1] Alan Abramowitz and Jennifer McCoy. United states: Racial resentment, negative partisanship, and polarization in trump’s america. *The ANNALS of the American Academy of Political and Social Science*, 681(1):137–156, 2019. 4.3.1
- [2] Dominic Abrams and Michael A Hogg. Metatheory: Lessons from social identity research. *Personality and social psychology review*, 8(2):98–106, 2004. 1.2.2
- [3] Eli Adler, Shira Hebel-Sela, Oded Adomi Leshem, Jonathan Levy, and Eran Halperin. A social virus: Intergroup dehumanization and unwillingness to aid amidst covid-19-who are the main targets? *International Journal of Intercultural Relations*, 86: 109–121, 2022. 6.2.1
- [4] Robert Agnew. A revised strain theory of delinquency. *Social forces*, 64(1):151–167, 1985. 1.2.1, 6.2.1, 6.5, 7.1.5, 7.3.3
- [5] Nuha Albadi, Maram Kurdi, and Shivakant Mishra. Hateful people or hateful bots? detection and characterization of bots spreading religious hatred in arabic social media. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–25, 2019. 5.2.3
- [6] Natalie Alkiviadou, Jacob Mchangama, and Raghav Mendiratta. Global handbook on hate speech laws. *The Future of Free Speech*, 2020. 7.3.2
- [7] Gordon W Allport. *The nature of prejudice*. Addison-Wesley, 1954. 1.2.1
- [8] Wafa Alorainy, Pete Burnap, Han Liu, Matthew Williams, and Luca Giommoni. Disrupting networks of hate: characterising hateful networks and removing critical nodes. *Social Network Analysis and Mining*, 12(1):27, 2022. 4.1, 4.2.3, 7.1.3
- [9] Neal Altman, Kathleen M Carley, and Jeffrey Reminga. ORA user’s guide 2017. Technical report, CASOS Center, Institute for Software Research, Carnegie Mellon University, 2017. 3.3.2, 4.3.2, 7.2.2
- [10] Laith Alzubaidi, Jinglan Zhang, Amjad J Humaidi, Ayad Al-Dujaili, Ye Duan, Omran Al-Shamma, José Santamaría, Mohammed A Fadhel, Muthana Al-Amidie, and Laith Farhan. Review of deep learning: concepts, cnn architectures, challenges, applications, future directions. *Journal of big Data*, 8:1–74, 2021. 2.2.3
- [11] Ida Vikøren Andersen. Hostility online: Flaming, trolling, and the public debate. *First Monday*, 2021. 5.2.2
- [12] Aymé Arango, Jorge Pérez, and Barbara Poblete. Hate speech detection is not as easy

- as you may think: A closer look at model validation (extended version). *Information Systems*, page 101584, 2020. 1.2.3, 2.2.1
- [13] Ahmer Arif, Leo Graiden Stewart, and Kate Starbird. Acting the part: Examining information operations within# blacklivesmatter discourse. *Proceedings of the ACM on Human-computer Interaction*, 2(CSCW):1–27, 2018. 5.2.1
- [14] Dennis Assenmacher, Marco Niemann, Kilian Müller, Moritz Seiler, Dennis M Riehle, and Heike Trautmann. Rp-mod&rp-crowd: Moderator-and crowd-annotated german news comment datasets. In *NeurIPS Datasets and Benchmarks*, 2021. 1.2.3, 1.3.2, 2.3.1, 2.1
- [15] Imran Awan and Irene Zempi. The affinity between online and offline anti-muslim hate crime: Dynamics and impacts. *Aggression and violent behavior*, 27:1–8, 2016. 6.2.3
- [16] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. Deep learning for hate speech detection in tweets. In *Proc. WWW Companion*, pages 759–760, 2017. 1.2.3, 2.2.1, 3.2.2
- [17] Radhika Bailurkar and Nataasha Raul. Detecting bots to distinguish hate speech on social media. In *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–5. IEEE, 2021. 5.2.3
- [18] Pablo Barberá, John T Jost, Jonathan Nagler, Joshua A Tucker, and Richard Bonneau. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological science*, 26(10):1531–1542, 2015. 1.2.4, 4.2.2, 4.5, 7.1.3
- [19] Pablo Barberá, Andreu Casas, Jonathan Nagler, Patrick J Egan, Richard Bonneau, John T Jost, and Joshua A Tucker. Who leads? who follows? measuring issue attention and agenda setting by legislators and the mass public using social media data. *American Political Science Review*, 113(4):883–901, 2019. 5.3.2
- [20] Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In Jonathan May, Ekaterina Shutova, Aurelie Herbelot, Xiaodan Zhu, Marianna Apidianaki, and Saif M. Mohammad, editors, *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/S19-2007. URL <https://aclanthology.org/S19-2007>. 2.2.1
- [21] José Alberto Benítez-Andrades, Álvaro González-Jiménez, Álvaro López-Brea, Jose Avelaira-Mata, José-Manuel Alija-Pérez, and María Teresa García-Ordás. Detecting racism and xenophobia using deep learning models on twitter data: Cnn, lstm and bert. *PeerJ Computer Science*, 8:e906, 2022. 3.2.2
- [22] Kelly Bergstrand and James M Jasper. Villains, victims, and heroes in character theory and affect control theory. *Social Psychology Quarterly*, 81(3):228–247, 2018. 1.2.2

- [23] David M Beskow and Kathleen M Carley. Bot conversations are different: leveraging network metrics for bot detection in twitter. In *Proc. ASONAM*, pages 825–832, 2018. 1.2.5, 1.4.4, 5.1, 5.2.2, 5.3.1, 6.3.1, 7.2.3
- [24] David M. Beskow and Kathleen M. Carley. Social cybersecurity: An emerging national security requirement. *Military Review*, 99(2):117, 2019. Publisher: US Army CGSC. 1.1, 1.2.5, 5.2.1, 5.2.2, 5.5, 7.1.4
- [25] David M. Beskow and Kathleen M. Carley. Characterization and comparison of Russian and Chinese disinformation campaigns. In *Disinformation, Misinformation, and Fake News in Social Media*, pages 63–81. Springer, 2020. 4.3.1
- [26] Michał Bilewicz and Wiktor Soral. Hate speech epidemic. the dynamic effects of derogatory language on intergroup relations and political radicalization. *Political Psychology*, 41:3–33, 2020. 1.2.1, 5.1
- [27] Janice T Blane, Daniele Bellutta, and Kathleen M Carley. Social-cyber maneuvers during the covid-19 vaccine initial rollout: content analysis of tweets. *Journal of Medical Internet Research*, 24(3):e34040, 2022. 1.2.5, 1.4.1, 2.3.2, 5.2.1, 5.2.2, 5.5, 7.1.4
- [28] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003. 3.3.2, 3.5
- [29] Ana-Maria Bliuc, Nicholas Faulkner, Andrew Jakubowicz, and Craig McGarty. Online networks of racial hate: A systematic review of 10 years of research on cyber-racism. *Computers in Human Behavior*, 87:75–86, 2018. 4.2.3
- [30] Vincent D Blondel, Jean-Lou Guillaume, Renaud Lambiotte, and Étienne Lefebvre. The Louvain method for community detection in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 10:P10008, 2011. 4.2.1, 4.3.3, 4.5, 7.1.2, 7.2.2
- [31] Phillip Bonacich. Some unique properties of eigenvector centrality. *Social networks*, 29(4):555–564, 2007. 4.2.1
- [32] Kevin Boyle. Hate speech—the united states versus the rest of the world. *Me. L. Rev.*, 53:487, 2001. 7.3.2
- [33] Samantha Bradshaw and Philip N Howard. The global organization of social media disinformation campaigns. *Journal of International Affairs*, 71(1.5):23–32, 2018. 1.2.5, 7.1.4, 7.2.3
- [34] William J Brady, Julian A Wills, John T Jost, Joshua A Tucker, and Jay J Van Bavel. Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, 114(28):7313–7318, 2017. 4.1, 4.2.2, 4.4.3, 7.1.3
- [35] William J Brady, Killian McLoughlin, Tuan N Doan, and Molly J Crockett. How social learning amplifies moral outrage expression in online social networks. *Science Advances*, 7(33):eabe5641, 2021. 4.1, 4.2.2, 4.4.3, 7.1.3
- [36] Marilynn B Brewer. The psychology of prejudice: Ingroup love and outgroup hate? *Journal of social issues*, 55(3):429–444, 1999. 1.2.2
- [37] Kellyton Dos Santos Brito, Rogério Luiz Cardoso Silva Filho, and Paulo Jorge Leitão

- Adeodato. A systematic review of predicting elections based on social media data: research challenges and future directions. *IEEE Transactions on Computational Social Systems*, 8(4):819–843, 2021. 6.2.2
- [38] David A Broniatowski, Amelia M Jamison, SiHua Qi, Lulwah AlKulaib, Tao Chen, Adrian Benton, Sandra C Quinn, and Mark Dredze. Weaponized health communication: Twitter bots and russian trolls amplify the vaccine debate. *American journal of public health*, 108(10):1378–1384, 2018. 5.2.3
- [39] Erin E Buckels, Paul D Trapnell, and Delroy L Paulhus. Trolls just want to have fun. *Personality and Individual Differences*, 67:97–102, 2014. 5.2.2
- [40] Carter T Butts. The complexity of social networks: theoretical and empirical findings. *Social Networks*, 23(1):31–72, 2001. 4.2.1
- [41] Manuela Caiani and Claudius Wagemann. Online networks of the italian and german extreme right: An explorative study with social network analysis. *Information, Communication & Society*, 12(1):66–109, 2009. 4.2.3
- [42] Louise Cainkar. Islamophobia and the us ideological infrastructure of white supremacy. In *The Routledge International Handbook of Islamophobia*, pages 239–251. Routledge, 2019. 3.5
- [43] Bart Cammaerts. The abnormalisation of social justice: The ‘anti-woke culture war’ discourse in the uk. *Discourse & Society*, 33(6):730–743, 2022. 3.5
- [44] Zhi Cao, Enhong Chen, Ye Huang, Shuanghong Shen, and Zhenya Huang. Learning from crowds with annotation reliability. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2103–2107, 2023. 7.1.1
- [45] Kathleen Carley. An approach for relating social structure to cognitive structure. *Journal of Mathematical Sociology*, 12(2):137–189, 1986. 3.2.3
- [46] Kathleen Carley. The value of cognitive foundations for dynamic social theory. *Journal of Mathematical Sociology*, 14(2-3):171–208, 1989. 1.2.4, 3.2.3, 4.1, 7.1.2
- [47] Kathleen Carley. A theory of group stability. *American sociological review*, pages 331–354, 1991. 1.2.4, 4.1
- [48] Kathleen Carley. Extracting culture through textual analysis. *Poetics*, 22(4):291–312, 1994. 1.2.4, 3.1, 3.2.3, 3.3.2, 3.5, 7.1.2, 7.2.2, 7.3.1
- [49] Kathleen Carley and Michael Palmquist. Extracting, representing, and analyzing mental models. *Social forces*, 70(3):601–636, 1992. 1.2.4, 3.2.3, 3.3.2, 3.5, 7.1.2, 7.2.2, 7.3.1
- [50] Kathleen M Carley. On the evolution of social and organizational networks. *Research in the Sociology of Organizations*, 16(0), 1999. 4.2.1
- [51] Kathleen M Carley. Network text analysis: The network position of concepts. In *Text analysis for the social sciences*, pages 79–100. Routledge, 2020. 1.2.4, 7.1.2
- [52] Kathleen M Carley. Social cybersecurity: an emerging science. *Computational and*

- mathematical organization theory*, 26(4):365–381, 2020. 1.1, 1.2.5, 1.4, 2.1, 2.2.3, 2.6, 3.1, 5.2.2, 5.2.3, 6.2.2, 6.5, 7, 7.1.4, 7.2.2, 7.3.1
- [53] Kathleen M Carley, Guido Cervone, Nitin Agarwal, and Huan Liu. Social cyber-security. In *Proc. SBP-BRIMS*, pages 389–394. Springer, 2018. 1.1, 1.2.5, 1.4, 2.1, 2.2.3, 2.5.1, 2.6, 5.2.1, 5.2.2, 5.2.3, 6.2.2, 6.5, 7
- [54] L Richard Carley, Jeff Reminga, and Kathleen M Carley. Ora & netmapper. In *Proc. SBP-BRIMS*, volume 3, page 7, 2018. 1.4.1, 1.4.2, 2.2.2, 2.3.2, 3.3.1, 4.2.1, 4.3.2, 4.3.3, 5.3.1, 7.2.2
- [55] Sergio Andrés Castaño-Pulgarín, Natalia Suárez-Betancur, Luz Magnolia Tilano Vega, and Harvey Mauricio Herrera López. Internet, social media and online hate speech. systematic review. *Aggression and Violent Behavior*, 58:101608, 2021. 3.2.2, 3.5
- [56] Davide Castelvechi. Can we open the black box of ai? *Nature News*, 538(7623):20, 2016. 2.2.3, 7.1.1
- [57] Chang-Feng Chen, Wen Shi, Jing Yang, and Hao-Huan Fu. Social bots’ role in climate change discussion on twitter: Measuring standpoints, topics, and interaction strategies. *Advances in Climate Change Research*, 12(6):913–923, 2021. 5.2.1
- [58] Naganna Chetty and Sreejith Alathur. Hate speech review in the context of online social networks. *Aggression and Violent Behavior*, 40:108–118, 2018. 1.1, 2.1, 4.1, 4.5, 7.1.2, 7.2.2
- [59] Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroğlu, and Marco Guerini. Conan-counter narratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proc. ACL*, pages 2819–2829, 2019. 1.2.3, 1.3.2, 2.3.1, 2.1
- [60] Mina Cikara, Vasiliki Fouka, and Marco Tabellini. Hate crime towards minoritized groups increases as they increase in sized-based rank. *Nature human behaviour*, pages 1–8, 2022. 1.2.1
- [61] Wendy Colomb and Kelly Dampousse. Examination of newspaper coverage of hate crimes: A moral panic perspective. *American Journal of Criminal Justice*, 28:147–163, 2004. 6.5, 7.1.5, 7.3.2
- [62] Matthew Costello, James Hawdon, Colin Bernatzky, and Kelly Mendes. Social group identity and perceptions of online hate. *Sociological inquiry*, 89(3):427–452, 2019. 6.2.1, 7.1.1
- [63] Kimberle Crenshaw. Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stanford Law Review*, 43:1241, 1990. 4.3.1
- [64] Stefano Cresci. A decade of social bot detection. *Communications of the ACM*, 63(10):72–83, 2020. 1.2.5, 5.2.2, 7.1.4
- [65] Molly J Crockett. Moral outrage in the digital age. *Nature human behaviour*, 1(11):769–771, 2017. 4.2.2
- [66] Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. Dealing

- with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110, 2022. 2.6
- [67] Thomas Davidson, Dana Warmley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Proc. ICWSM*, volume 11, pages 512–515, 2017. 1.2.3, 1.3.2, 2.2.1, 2.3.1, 2.1, 2.4.1
- [68] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. Racial bias in hate speech and abusive language detection datasets. In *Proc. ALW*, pages 25–35, Florence, Italy, August 2019. 1.2.3, 2.1, 2.2.1, 2.2.3, 2.4.1, 7.1.1
- [69] Ona de Gibert, Naiara Pérez, Aitor García-Pablos, and Montse Cuadros. Hate speech dataset from a white supremacy forum. In *Proc. ALW*, pages 11–20, 2018. 1.2.3, 1.3.2, 2.3.1, 2.1
- [70] Rogers Prates de Pelle and Viviane P Moreira. Offensive comments in the brazilian web: a dataset and baseline results. In *Anais do VI Brazilian Workshop on Social Network Analysis and Mining*. SBC, 2017. 1.2.3, 1.3.2, 2.3.1, 2.1
- [71] Robert F Dredrick, John M Ferron, Melinda R Hess, Kristine Y Hogarty, Jeffrey D Kromrey, Thomas R Lang, John D Niles, and Reginald S Lee. Multilevel modeling: A review of methodological issues and applications. *Review of educational research*, 79(1):69–102, 2009. 4.2.3, 7.2.2
- [72] Delan Devakumar, Sujitha Selvarajah, Geordan Shannon, Kui Muraya, Sarah Lasoye, Susanna Corona, Yin Paradies, Ibrahim Abubakar, and E Tendayi Achiume. Racism, the public health crisis we can no longer ignore. *The Lancet*, 395(10242):e112–e113, 2020. 3.2.2
- [73] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL*, pages 4171–4186, June 2019. 1.2.3, 2.2.1
- [74] Joan Donovan and Danah Boyd. Stop the presses? moving from strategic silence to strategic amplification in a networked media ecosystem. *American Behavioral Scientist*, 65(2):333–350, 2021. 6.5, 7.1.5, 7.3.2
- [75] Karen M Douglas. Psychology, discrimination and hate groups online. In *The Oxford handbook of internet psychology*, pages 155–164. Oxford University Press Oxford, 2007. 4.2.2, 4.5
- [76] Paul Ekman. Are there basic emotions? *Psychological Review*, 99(3):550–553, 1992. 1.2.1
- [77] Naomi Ellemers, Russell Spears, and Bertjan Doosje. Self and social identity. *Annual review of psychology*, 53(1):161–186, 2002. 1.2.2, 7.1.1
- [78] Mai ElSherief, Shirin Nilizadeh, Dana Nguyen, Giovanni Vigna, and Elizabeth Belding. Peer to peer hate: Hate speech instigators and their targets. In *Proc. ICWSM*, volume 12, 2018. 1.2.3, 1.3.2, 2.3.1, 2.1
- [79] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic bert sentence embedding. In *Proceedings of the 60th Annual Meet-*

- ing of the Association for Computational Linguistics (Volume 1: Long Papers), pages 878–891, 2022. 2.3.2
- [80] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. The rise of social bots. *Communications of the ACM*, 59(7):96–104, 2016. 1.2.5, 5.2.1, 5.2.2, 7.1.4, 7.1.5, 7.2.3
- [81] Emilio Ferrara, Stefano Cresci, and Luca Luceri. Misinformation, manipulation, and abuse on social media in the era of covid-19. *Journal of Computational Social Science*, 3(2):271–277, 2020. 5.5, 7.1.4, 7.2.2, 7.2.3
- [82] Agneta Fischer, Eran Halperin, Daphna Canetti, and Alba Jasini. Why we hate. *Emotion Review*, 10(4):309–320, 2018. 1.2.1, 3.1, 3.2.1, 3.2.3
- [83] Susan T Fiske. Stereotyping, prejudice, and discrimination at the seam between the centuries: Evolution, culture, mind, and brain. *European Journal of Social Psychology*, 30(3):299–322, 2000. 1.2.1
- [84] Julie Fitness and Garth JO Fletcher. Love, hate, anger, and jealousy in close relationships: A prototype and cognitive appraisal analysis. *Journal of personality and Social Psychology*, 65(5):942, 1993. 3.2.1
- [85] Cechao Foo. Protection from online falsehoods and manipulation act and the roles of internet intermediaries in regulating online falsehoods. *Singapore Academy of Law Journal*, 33(1):438–482, 2021. 7.3.2
- [86] Paula Fortuna and Sérgio Nunes. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30, 2018. 1.2.3, 2.1, 2.2.1, 2.5.1, 3.1, 3.5, 7.1.1, 7.4
- [87] Paula Fortuna, Juan Soler, and Leo Wanner. Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6786–6794, 2020. 2.3.2, 2.6, 7.1.2
- [88] Paula Fortuna, Juan Soler-Company, and Leo Wanner. How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing & Management*, 58(3):102524, 2021. 1.2.3, 2.1, 2.2.1, 2.2.3, 2.5.1, 2.6, 3.5, 7.1.1
- [89] Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. Large scale crowdsourcing and characterization of Twitter abusive behavior. In *Proc. ICWSM*, volume 12, 2018. 1.2.3, 1.3.2, 2.3.1, 2.1
- [90] Terrill L Frantz, Marcelo Cataldo, and Kathleen M Carley. Robustness of centrality measures under uncertainty: Examining the role of network topology. *Computational and Mathematical Organization Theory*, 15:303–328, 2009. 4.2.1
- [91] Markus Freitag and Nathalie Hofstetter. Pandemic threat and intergroup relations: how negative emotions associated with the threat of covid-19 shape attitudes towards immigrants. *Journal of Ethnic and Migration Studies*, 48(13):2985–3004, 2022. 6.2.1

- [92] Simona Frenda, Alessandra Teresa Cignarella, Valerio Basile, Cristina Bosco, Viviana Patti, and Paolo Rosso. The unbearable hurtfulness of sarcasm. *Expert Systems with Applications*, 193:116398, 2022. 2.6
- [93] Noah E Friedkin. *A structural theory of social influence*. Cambridge University Press, 1998. 1.2.4
- [94] Noah E Friedkin and Eugene C Johnsen. Social influence and opinions. *Journal of mathematical sociology*, 15(3-4):193–206, 1990. 1.2.4
- [95] Michael C Galgoczy, Atharva Phatak, Danielle Vinson, Vijay K Mago, and Philippe J Giabbanelli. (re) shaping online narratives: when bots promote the message of president trump during his first impeachment. *PeerJ Computer Science*, 8:e947, 2022. 5.2.3
- [96] John D Gallacher, Marc W Heerdink, and Miles Hewstone. Online engagement between opposing political protest groups via social media is linked to physical violence of offline encounters. *Social Media+ Society*, 7(1):2056305120984445, 2021. 6.2.3
- [97] Cherian George. Hate speech law and policy. *The international encyclopedia of digital communication and society*, pages 1–10, 2015. 7.3.2
- [98] Andrea Gibbons. The five refusals of white supremacy. *American Journal of Economics and Sociology*, 77(3-4):729–755, 2018. 3.5
- [99] Homero Gil de Zúñiga, Nakwon Jung, and Sebastián Valenzuela. Social media use for news and individuals’ social capital, civic engagement and political participation. *Journal of Computer-Mediated Communication*, 17(3):319–336, 2012. 6.1, 6.5
- [100] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE, 2018. 2.2.3
- [101] Edmund M Glaser. Is there a neurobiology of hate? *J. Hate Stud.*, 7:7, 2008. 1.2.1
- [102] Raul Gomez, Jaume Gibert, Lluís Gomez, and Dimosthenis Karatzas. Exploring hate speech detection in multimodal publications. In *Proc. WACV*, pages 1470–1478, 2020. 1.2.3, 1.3.2, 2.3.1, 2.1, 3.1, 3.3.1
- [103] Felipe González-Pizarro and Savvas Zannettou. Understanding and detecting hateful content using contrastive learning. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 257–268, 2023. 2.3.2
- [104] Megan Goodwin. Unmasking islamophobia: anti-muslim hostility and/as white supremacy. *Journal of the American Academy of Religion*, 88(2):354–386, 2020. 3.5
- [105] Angela R Gover, Shannon B Harper, and Lynn Langton. Anti-asian hate crime during the covid-19 pandemic: Exploring the reproduction of inequality. *American journal of criminal justice*, 45(4):647–667, 2020. 6.2.3, 7.3.3
- [106] Hedy Greijdanus, Carlos A de Matos Fernandes, Felicity Turner-Zwinkels, Ali Honari, Carla A Roos, Hannes Rosenbusch, and Tom Postmes. The psychology of online activism and social movements: Relations between online and offline collective action.

- Current Opinion in Psychology*, 35:49–54, 2020. 1.2.5, 6.1, 6.2.2, 6.5, 7.1.5
- [107] Phillip L Hammack. Narrative and the cultural psychology of identity. *Personality and social psychology review*, 12(3):222–247, 2008. 3.1
- [108] S Alexander Haslam, Niklas K Steffens, Stephen D Reicher, and Sarah V Bentley. Identity leadership in a crisis: A 5r framework for learning from responses to covid-19. *Social Issues and Policy Review*, 15(1):35–83, 2021. 6.5, 7.1.1, 7.1.5, 7.3.2
- [109] Bing He, Caleb Ziems, Sandeep Soni, Naren Ramakrishnan, Diyi Yang, and Srijan Kumar. Racism is a virus: Anti-asian hate and counterspeech in social media during the covid-19 crisis. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 90–94, 2021. 4.3.1, 4.5, 7.1.5
- [110] Xiaoxin He, Xavier Bresson, Thomas Laurent, Adam Perold, Yann LeCun, and Bryan Hooi. Harnessing explanations: Llm-to-lm interpreter for enhanced text-attributed graph representation learning. In *The Twelfth International Conference on Learning Representations*, 2023. 3.3.2, 3.5
- [111] David R Heise. Affect control theory: Concepts and model. *Journal of Mathematical Sociology*, 13(1-2):1–33, 1987. 1.2.2, 1.4.1, 2.3.2, 3.1
- [112] Cornelia Herbert, Eileen Bendig, and Roberto Rojas. My sadness–our happiness: writing about positive, negative, and neutral autobiographical life events reveals linguistic markers of self-positivity and individual well-being. *Frontiers in Psychology*, 9:423215, 2019. 2.2.2, 2.6
- [113] Daniel Hickey, Matheus Schmitz, Daniel Fessler, Paul E Smaldino, Goran Muric, and Keith Burghardt. Auditing elon musk’s impact on hate speech and bots. In *Proceedings of the international AAAI conference on web and social media*, volume 17, pages 1133–1137, 2023. 5.2.3
- [114] Omi Hodwitz and Kelley Massingale. Rhetoric and hate crimes: examining the public response to the trump narrative. *Behavioral sciences of terrorism and political aggression*, 15(3):303–320, 2023. 5.5
- [115] Joe Hoover, Mohammad Atari, Aida Mostafazadeh Davani, Brendan Kennedy, Gwenyth Portillo-Wightman, Leigh Yeh, and Morteza Dehghani. Investigating the role of group-based morality in extreme behavioral expressions of prejudice. *Nature Communications*, 12(1):1–13, 2021. 1.2.1
- [116] Binxuan Huang and Kathleen M Carley. On predicting geolocation of tweets using convolutional neural networks. In *Proc. ICWSM*, pages 281–291. Springer, 2017. 1.2.5, 1.4.6, 6.3.1
- [117] Binxuan Huang and Kathleen M Carley. Disinformation and misinformation on twitter during the novel coronavirus outbreak. *arXiv preprint arXiv:2006.04278*, 2020. 1.3.3, 1.4.6, 6.3.1
- [118] Mark Huberty. Can we vote with our tweet? on the perennial difficulty of election forecasting with social media. *International Journal of Forecasting*, 31(3):992–1007,

2015. 6.2.2

- [119] Molly E Ireland, Richard B Slatcher, Paul W Eastwick, Lauren E Scissors, Eli J Finkel, and James W Pennebaker. Language style matching predicts relationship initiation and stability. *Psychological science*, 22(1):39–44, 2011. 2.2.2, 2.6
- [120] Nancy Isenberg, David Silbersweig, Almut Engelen, Sylvia Emmerich, Kishor Malavade, Bradley Beattie, Andy C Leon, and Emily Stern. Linguistic threat activates the human amygdala. *Proceedings of the National Academy of Sciences*, 96(18):10456–10459, 1999. 1.2.1
- [121] Jolanda Jetten. The wealth paradox: Prosperity and opposition to immigration. *European Journal of Social Psychology*, 49(6):1097–1113, 2019. 3.2.2
- [122] Aiqi Jiang, Xiaohan Yang, Yang Liu, and Arkaitz Zubiaga. Swsr: A chinese dataset and lexicon for online sexism detection. *Online Social Networks and Media*, 27:100182, 2022. 1.2.3, 1.3.2, 2.3.1, 2.1, 3.2.2, 3.5
- [123] N. F. Johnson, R. Leahy, N. Johnson Restrepo, N. Velasquez, M. Zheng, P. Manrique, P. Devkota, and S. Wuchty. Hidden resilience and adaptive dynamics of the global online hate ecology. *Nature*, 573(7773):261–265, September 2019. ISSN 1476-4687. doi: 10.1038/s41586-019-1494-7. Number: 7773 Publisher: Nature Publishing Group. 4.1, 4.2.3
- [124] Ana Jordan. Conceptualizing backlash:(uk) men’s rights groups, anti-feminism, and postfeminism. *Canadian Journal of Women and the Law*, 28(1):18–44, 2016. 3.5
- [125] Kenneth Joseph, Geoffrey P Morgan, Michael K Martin, and Kathleen M Carley. On the coevolution of stereotype, culture, and social relationships: an agent-based model. *Social Science Computer Review*, 32(3):295–311, 2014. 1.2.4, 3.2.3, 7.1.2
- [126] Kenneth Joseph, Wei Wei, Matthew Benigni, and Kathleen M Carley. A social-event based approach to sentiment analysis of identities and behaviors in text. *The Journal of Mathematical Sociology*, 40(3):137–166, 2016. 2.2.2, 3.3.1, 4.3.3, 5.3.1, 7.1.1, 7.2.1
- [127] Jeffrey H Kahn, Renee M Tobin, Audra E Massey, and Jennifer A Anderson. Measuring emotional expression with the linguistic inquiry and word count. *The American journal of psychology*, 120(2):263–286, 2007. 2.2.2
- [128] Rune Karlsen, Kari Steen-Johnsen, Dag Wollebæk, and Bernard Enjolras. Echo chamber and trench warfare dynamics in online debates. *European journal of communication*, 32(3):257–273, 2017. 1.2.4, 4.2.2, 7.1.3
- [129] Anastasia Kavada. Creating the collective: social media, the occupy movement and its constitution as a collective actor. *Information, Communication & Society*, 18(8):872–886, 2015. 6.2.2
- [130] Anna Kende, Martijn van Zomeren, Adrienn Ujhelyi, and Nóra Anna Lantos. The social affirmation use of social media as a motivator of collective action. *Journal of Applied Social Psychology*, 46(8):453–469, 2016. 7.1.5
- [131] Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. Contextualizing hate speech classifiers with post-hoc explanation. In

Proc. ACL, pages 5435–5442, July 2020. 2.2.2

- [132] Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs, Shreya Havaladar, Gwenyth Portillo-Wightman, Elaine Gonzalez, et al. Introducing the gab hate corpus: defining and applying hate-based rhetoric to social media posts at scale. *Language Resources and Evaluation*, 56(1):79–108, 2022. 1.2.3, 1.3.2, 2.3.1, 2.1, 3.1, 3.3.1
- [133] Brian TaeHyuk Keum, Xu Li, and Michele J Wong. Hate as a system: Examining hate crimes and hate groups as state level moderators on the impact of online and offline racism on mental health. *International Journal of Intercultural Relations*, 91: 44–55, 2022. 1.2.1
- [134] Heena Khan and Joshua L Phillips. Language agnostic model: detecting islamophobic content on social media. In *Proceedings of the 2021 ACM Southeast conference*, pages 229–233, 2021. 3.2.2
- [135] Majid KhosraviNik and Eleonora Esposito. Online hate, digital discourse and critique: Exploring digitally-mediated discursive practices of gender-based hostility. *Lodz papers in pragmatics*, 14(1):45–68, 2018. 3.1, 3.2.1, 3.2.2
- [136] Bumsoo Kim. Effects of Social Grooming on Incivility in COVID-19. *Cyberpsychology, Behavior, and Social Networking*, April 2020. ISSN 2152-2715. doi: 10.1089/cyber.2020.0201. Publisher: Mary Ann Liebert, Inc., publishers. 4.3.3, 4.5
- [137] Dam Hee Kim and Nicole B Ellison. From observation on social media to offline political participation: The social media affordances approach. *New Media & Society*, page 1461444821998346, 2021. 1.2.5, 6.1, 6.2.2, 6.5
- [138] Yoon Kim. Convolutional neural networks for sentence classification. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1181. URL <https://aclanthology.org/D14-1181>. 2.2.1
- [139] Gueorgi Kossinets and Duncan J Watts. Origins of homophily in an evolving social network. *American journal of sociology*, 115(2):405–450, 2009. 1.2.4
- [140] David Krackhardt and Robert N Stern. Informal networks and organizational crises: An experimental simulation. *Social Psychology Quarterly*, pages 123–140, 1988. 4.2.1, 4.1, 4.3.3, 5.3.1, 5.1
- [141] Lance G Lachenicht. Aggravating language a study of abusive and insulting language. *Research on Language & Social Interaction*, 13(4):607–687, 1980. 2.2.2
- [142] Dorottya Lantos and Pascal Molenberghs. The neuroscience of intergroup threat and violence. *Neuroscience & Biobehavioral Reviews*, 131:77–87, 2021. 6.2.1
- [143] Ernesto Lee, Furqan Rustam, Patrick Bernard Washington, Fatima El Barakaz, Wajdi Aljedaani, and Imran Ashraf. Racism detection by analyzing differential opinions through sentiment analysis of tweets using stacked ensemble gcr-nn model. *IEEE*

Access, 10:9717–9728, 2022. 3.2.2

- [144] Stephan Lewandowsky and Sander Van Der Linden. Countering misinformation and fake news through inoculation and prebunking. *European Review of Social Psychology*, 32(2):348–384, 2021. 7.3.3
- [145] Qian Li, Hao Peng, Jianxin Li, Congying Xia, Renyu Yang, Lichao Sun, Philip S Yu, and Lifang He. A survey on text classification: From traditional to deep learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13(2):1–41, 2022. 2.2.1
- [146] Sun Sun Lim. Manufacturing hate 4.0: Can media studies rise to the challenge? *Television & New Media*, 21(6):602–607, 2020. 1.2.3
- [147] Yu-Ru Lin and Wen-Ting Chung. The dynamics of Twitter users’ gun narratives across major mass shooting events. *Humanities and Social Sciences Communications*, 7(1):1–16, 2020. 1.2.5, 4.2.3, 6.2.2
- [148] Xinsheng Liu, Eric Lindquist, and Arnold Vedlitz. Explaining media and congressional attention to global climate change, 1969-2005: An empirical test of agenda-setting theory. *Political Research Quarterly*, 64(2):405–419, 2011. 5.3.2
- [149] Yonatan Lupu, Richard Sear, Nicolas Velásquez, Rhys Leahy, Nicholas Johnson Restrepo, Beth Goldberg, and Neil F Johnson. Offline events and online hate. *PLoS one*, 18(1):e0278511, 2023. 6.2.3
- [150] Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. Hate speech detection: Challenges and solutions. *PloS one*, 14(8): e0221152, 2019. 1.2.3, 2.1, 2.2.1, 3.2.2
- [151] Thomas Magelinski and Kathleen M Carley. Community-based time segmentation from network snapshots: Streaming and holistic approaches for semi-static and dynamic nodesets. *Applied Network Science*, 4:1–19, 2019. 4.5
- [152] Kate Manne. *Down girl: The logic of misogyny*. Oxford University Press, 2017. 3.5
- [153] Tim Markham. Social media, protest cultures and political subjectivities of the arab spring. *Media, Culture & Society*, 36(1):89–104, 2014. 6.2.2
- [154] Michael K Martin, Juergen Pfeffer, and Kathleen M Carley. Network text analysis of conceptual overlap in interviews, newspaper articles and keywords. *Social Network Analysis and Mining*, 3:1165–1177, 2013. 3.1, 3.2.3, 7.1.5
- [155] Cristhian A Martínez, Jan-Willem van Prooijen, and Paul AM Van Lange. Hate: Toward understanding its distinctive features across interpersonal and intergroup targets. *Emotion*, 22(1):46, 2022. 1.2.1, 3.1, 3.2.1
- [156] Ggaliwango Marvin, Nakayiza Hellen, Daudi Jjingo, and Joyce Nakatumba-Nabende. Prompt engineering in large language models. In *International Conference on Data Intelligence and Cognitive Informatics*, pages 387–402. Springer, 2023. 3.3.2
- [157] Alice E Marwick. Morally motivated networked harassment as normative reinforcement. *Social Media+ Society*, 7(2):20563051211021378, 2021. 4.2.2

- [158] Alice E Marwick and Robyn Caplan. Drinking male tears: Language, the manosphere, and networked harassment. *Feminist Media Studies*, 18(4):543–559, 2018. 3.2.2
- [159] Ariadna Matamoros-Fernández and Johan Farkas. Racism, hate speech, and social media: A systematic review and critique. *Television & New Media*, 22(2):205–224, 2021. 1.2.3
- [160] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proc. AAAI*, volume 35, pages 14867–14875, 2021. 1.2.3, 1.3.2, 2.2.3, 2.3.1, 2.1
- [161] Tara Matthews, Jalal U Mahmud, Jilin Chen, Michael Muller, Eben Haber, and Hernan Badenes. They said what? exploring the relationship between language use and member satisfaction in communities. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 819–825, 2015. 2.2.2
- [162] Jon Mcauliffe and David Blei. Supervised topic models. *Advances in neural information processing systems*, 20, 2007. 3.3.2, 3.5, 7.2.2
- [163] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444, 2001. 1.2.4
- [164] Robert K Merton. Social structure and anomie. *American Sociological Review*, 3(5):672–682, 1938. 1.2.1, 6.2.1, 6.5, 7.1.5, 7.3.3
- [165] Doug Meyer. An intersectional analysis of lesbian, gay, bisexual, and transgender (lgbt) people’s evaluations of anti-queer violence. *Gender & Society*, 26(6):849–873, 2012. 3.5
- [166] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013. 1.2.3, 2.2.1
- [167] Semi Min and Juyong Park. Modeling narrative structure and dynamics with networks, sentiment analysis, and topic modeling. *PloS one*, 14(12):e0226025, 2019. 3.2.3
- [168] Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. Deep learning-based text classification: a comprehensive review. *ACM computing surveys (CSUR)*, 54(3):1–40, 2021. 2.2.1
- [169] Bojan Mohar. Isoperimetric numbers of graphs. *Journal of Combinatorial Theory, Series B*, 47(3):274–291, 1989. 4.2.1, 4.1, 4.3.3, 5.3.1
- [170] Zuzanna Aleksandra Molenda, Marta Marchlewska, Marta Rogoza, Piotr Michalski, Paulina Górska, Dagmara Szczepańska, and Aleksandra Cislak. What makes an internet troll? on the relationships between temperament (bis/bas), dark triad, and internet trolling. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, 16(5), 2022. 5.2.2

- [171] Ludwin E Molina, Linda R Tropp, and Chris Goode. Reflections on prejudice and intergroup relations. *Current Opinion in Psychology*, 11:120–124, 2016. 1.2.2
- [172] Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. Ethos: a multi-label hate speech detection dataset. *Complex & Intelligent Systems*, pages 1–16, 2022. 1.3.2, 2.3.1, 2.1, 3.1, 3.3.1
- [173] Cristina J Montiel, Erwine S Dela Paz, and Jose S Medriano. Narrative expansion and” terrorist” labeling: Discursive conflict escalation by state media. *Journal of Social and Political Psychology*, 10(2):518–535, 2022. 1.2.4, 3.2.3
- [174] Cristina Jayme Montiel and Joshua Uyheng. Mapping contentious collective emotions in a populist democracy: Duterte’s push for philippine federalism. *Political Psychology*, 41(4):737–754, 2020. 4.3.1
- [175] Cristina Jayme Montiel, Arvin Jay Boller, Joshua Uyheng, and Ervina A Espina. Narrative congruence between populist president duterte and the filipino public: Shifting global alliances from the united states to china. *Journal of Community & Applied Social Psychology*, 29(6):520–534, 2019. 4.3.1
- [176] Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen Carley. Is the sample good enough? comparing data from twitter’s streaming api with twitter’s firehose. In *Proceedings of the international AAAI conference on web and social media*, volume 7, pages 400–408, 2013. 4.5, 5.5, 6.5, 7.1.3, 7.1.5, 7.4
- [177] Mohsen Mosleh, Gordon Pennycook, and David G Rand. Field experiments on social media. *Current Directions in Psychological Science*, 31(1):69–75, 2022. 4.5, 6.5, 7.1.3, 7.1.5, 7.4
- [178] Michael Murray et al. Narrative psychology. *Qualitative psychology: A practical guide to research methods*, pages 85–107, 2015. 3.1
- [179] Seth A Myers and Jure Leskovec. The bursty dynamics of the Twitter information network. In *Proc. WWW*, pages 913–924, 2014. 1.2.5, 4.2.3
- [180] Ginés Navarro-Carrillo, Jorge Torres-Marín, and Hugo Carretero-Dios. Do trolls just want to have fun? assessing the role of humor-related traits in online trolling behavior. *Computers in Human Behavior*, 114:106551, 2021. 5.2.2
- [181] Lynnette Hui Xian Ng and Kathleen M Carley. Online coordination: methods and comparative case studies of coordinated groups across four events in the united states. In *Proceedings of the 14th ACM Web Science Conference 2022*, pages 12–21, 2022. 5.2.2
- [182] Lynnette Hui Xian Ng and Kathleen M Carley. Botbuster: Multi-platform bot detection using a mixture of experts. In *Proceedings of the international AAAI conference on web and social media*, volume 17, pages 686–697, 2023. 5.2.2
- [183] Lynnette Hui Xian Ng and Kathleen M Carley. Deflating the chinese balloon: types of twitter bots in us-china balloon incident. *EPJ Data Science*, 12(1):63, 2023. 5.2.2
- [184] Lynnette Hui Xian Ng, Dawn C Robertson, and Kathleen M Carley. Stabilizing a supervised bot detection algorithm: How much data is needed for consistent predic-

- tions? *Online Social Networks and Media*, 28:100198, 2022. 5.4.1
- [185] Debora Nozza, Claudia Volpetti, and Elisabetta Fersini. Unintended bias in misogyny detection. In *Ieee/wic/acm international conference on web intelligence*, pages 149–155, 2019. 3.2.2
- [186] Alexandra Olteanu, Carlos Castillo, Jeremy Boy, and Kush Varshney. The effect of extremist violence on hateful speech online. In *Proc. ICWSM*, volume 12, 2018. 1.2.1
- [187] Jonathan Corpus Ong and Jason Vincent Cabañes. When disinformation studies meets production studies: Social identities and moral justifications in the political trolling industry. *International Journal of Communication*, 13, 2019. 5.5
- [188] Jonathan Corpus Ong, Ross Tapsell, and Nicole Curato. Tracking digital disinformation in the 2019 Philippine midterm election. *New Mandala*, 2019. 4.3.1, 5.5
- [189] Susan Opatow. Moral exclusion and injustice: An introduction. *Journal of Social Issues*, 46(1):1–20, 1990. 1.2.1, 5.1, 6.2.1, 7.1.5
- [190] Susan Opatow and Sara I McClelland. The intensification of hating: A theory. *Social Justice Research*, 20(1):68–97, 2007. (document), 1.1, 1.2.1, 1.2.5, 2.2.2, 4.1, 5.1, 6.1, 6.2.1, 6.2.3, 6.5, 7, 7.3.3
- [191] Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. Multilingual and multi-aspect hate speech analysis. In *Proc. EMNLP-IJCNLP*, pages 4675–4684, 2019. 1.2.3, 1.3.2, 2.3.1, 2.1
- [192] Shriphani Palakodety, Ashiqur R KhudaBukhsh, and Jaime G Carbonell. Voice for the voiceless: Active sampling to detect comments supporting the rohingyas. In *Proc. AAAI*, volume 34, pages 454–462, 2020. 1.2.3, 2.3.1
- [193] Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. Misogyny detection in twitter: a multilingual and cross-domain study. *Information processing & management*, 57(6):102360, 2020. 3.2.2
- [194] James W Pennebaker, Matthias R Mehl, and Kate G Niederhoffer. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, 54(1):547–577, 2003. 2.2.2, 2.6, 7.1.1
- [195] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proc. EMNLP*, pages 1532–1543, 2014. 1.2.3, 2.2.1, 2.3.2
- [196] Frank S Pezzella, Matthew D Fetzer, and Tyler Keller. The dark figure of hate crime underreporting. *American Behavioral Scientist*, page 0002764218823844, 2019. 7.3.2
- [197] Shruti Phadke and Tanushree Mitra. Many faced hate: A cross platform study of content framing and information sharing by online hate groups. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–13, 2020. 4.2.3
- [198] Jason E Plaks, Jeffrey S Robinson, and Rachel Forbes. Anger and sadness as moral signals. *Social Psychological and Personality Science*, 13(2):362–371, 2022. 2.2.2

- [199] Robert Plutchik. A psychoevolutionary theory of emotions. *Social Science Information*, 21(4/5):529–553, 1982. 1.2.1
- [200] Matti Pohjonen and Sahana Udupa. Extreme speech online: An anthropological critique of hate speech debates. *International Journal of Communication*, 11:19, 2017. 1.1, 1.2.3, 2.1, 2.2.2, 3.1, 3.2.2, 3.5, 7.1.1, 7.3.3
- [201] Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55(2):477–523, 2021. 1.2.3, 1.3.2, 2.1, 2.3.1, 2.6, 7.2.1, 7.4
- [202] Clara Pretus, Jennifer L Ray, Yael Granot, William A Cunningham, and Jay J Van Bavel. The psychology of hate: Moral concerns differentiate hate from dislike. *European Journal of Social Psychology*, 53(2):336–353, 2023. 3.1
- [203] Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. A benchmark dataset for learning to intervene in online hate speech. In *Proc. EMNLP-IJCNLP*, pages 4757–4766, 2019. 1.2.3, 1.3.2, 2.3.1, 2.1
- [204] Carlos José Parales Quenza. On the structural approach to social representations. *Theory & Psychology*, 15(1):77–100, 2005. 1.2.4
- [205] Priyanka Ranade, Sanorita Dey, Anupam Joshi, and Tim Finin. Computational understanding of narratives: A survey. *IEEE Access*, 10:101575–101594, 2022. 3.2.3, 3.3.2
- [206] Steve Rathje, Jay J Van Bavel, and Sander Van Der Linden. Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences*, 118(26):e2024292118, 2021. 4.1, 4.2.2, 7.1.3
- [207] Madhavi Reddi, Rachel Kuo, and Daniel Kreiss. Identity propaganda: Racial narratives and disinformation. *New Media & Society*, page 14614448211029293, 2021. 3.1, 3.2.1, 3.2.2, 3.2.3, 3.5
- [208] Stephen Reicher and Clifford Stott. On order and disorder during the COVID-19 pandemic. *British Journal of Social Psychology*, 59(3):694–702, 2020. 4.3.1, 5.1, 7.1.5
- [209] Stephen Reicher, Nick Hopkins, Mark Levine, and Rakshi Rath. Entrepreneurs of hate and entrepreneurs of solidarity: Social identity as a basis for mass communication. *International review of the Red Cross*, 87(860):621–637, 2005. 4.2.2, 4.5, 5.1, 7.3.3
- [210] Stephen Reicher, S Alexander Haslam, and Rakshi Rath. Making a virtue of evil: A five-step social identity model of the development of collective hate. *Social and Personality Psychology Compass*, 2(3):1313–1344, 2008. 1.2.1, 1.2.2
- [211] Garry Robins and Yoshi Kashima. Social psychology and social networks: Individuals and social systems. *Asian Journal of Social Psychology*, 11(1):1–12, 2008. 4.2.2
- [212] Hannes Rosenbusch, Anthony M Evans, and Marcel Zeelenberg. Interregional and intraregional variability of intergroup attitudes predict online hostility. *European Journal of Personality*, 34(5):859–872, 2020. 6.2.3

- [213] Paul E Rutledge. Trump, covid-19, and the war on expertise. *The American Review of Public Administration*, 50(6-7):505–511, 2020. 4.3.1
- [214] Joni Salminen, Maximilian Hopf, Shammur A Chowdhury, Soon-gyo Jung, Hind Almerekhi, and Bernard J Jansen. Developing an online hate classifier for multiple social media platforms. *Human-centric Computing and Information Sciences*, 10: 1–34, 2020. 2.2.2, 2.2.3, 2.6, 7.1.1
- [215] Stefano Sbalchiero and Maciej Eder. Topic modeling, long texts and the best number of topics. some problems and solutions. *Quality & Quantity*, 54:1095–1108, 2020. 3.3.2
- [216] Klaus R Scherer. Theories in cognition & emotion—social functions of emotion. *Cognition and Emotion*, 36(3):385–387, 2022. 1.2.1
- [217] Anna Schmidt and Michael Wiegand. A survey on hate speech detection using natural language processing. In *Proc. SocialNLP*, pages 1–10, 2017. 1.2.3, 2.1, 2.2.1, 3.1, 3.2.2
- [218] Kayla Sergent and Alexander D Stajkovic. Women’s leadership is associated with fewer deaths during the covid-19 crisis: Quantitative and qualitative analyses of united states governors. *Journal of Applied Psychology*, 105(8):771, 2020. 6.5, 7.1.5, 7.3.2
- [219] Joan Serra, Ilias Leontiadis, Dimitris Spathis, Gianluca Stringhini, Jeremy Blackburn, and Athena Vakali. Class-based prediction errors to detect hate speech with out-of-vocabulary words. In *Proceedings of the first workshop on abusive language online*, pages 36–40, 2017. 2.6, 7.2.1
- [220] Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. The spread of low-credibility content by social bots. *Nature communications*, 9(1):1–9, 2018. 1.2.5
- [221] Eugenia Siapera and Paloma Viejo-Otero. Governing hate: Facebook and digital racism. *Television & New Media*, 22(2):112–130, 2021. 1.2.3
- [222] Krystal A Smalls, Arthur K Spears, and Jonathan Rosa. Introduction: Language and white supremacy. *Journal of Linguistic Anthropology*, 31(2):152–156, 2021. 3.5
- [223] Lynn Smith-Lovin. Affect control theory: An assessment. *Journal of Mathematical Sociology*, 13(1-2):171–192, 1987. 3.1
- [224] Wiktor Soral, Michał Bilewicz, and Mikołaj Winiewski. Exposure to hate speech increases prejudice through desensitization. *Aggressive behavior*, 44(2):136–146, 2018. 5.1
- [225] Russell Spears. Social influence and group identity. *Annual Review of Psychology*, 72(1):367–390, 2021. 1.2.2, 7.1.1
- [226] Kate Starbird. Disinformation’s spread: bots, trolls and all of us. *Nature*, 571(7766): 449–450, 2019. 1.2.5
- [227] Annika Stechemesser, Leonie Wenz, and Anders Levermann. Corona crisis fuels racially profiled hate in social media networks. *EClinicalMedicine*, 23, 2020. 4.5, 6.2.3

- [228] Danny D Steinberg and Natalia V Sciarini. *An introduction to psycholinguistics*. Routledge, 2013. 2.2.2
- [229] Robert J Sternberg. A duplex theory of hate: Development and application to terrorism, massacres, and genocide. *Review of general psychology*, 7(3):299–328, 2003. 1.1, 1.2.1, 2.2.2, 3.1, 3.2.1, 3.2.3, 3.4.2, 6.1, 7, 7.1.2
- [230] Robert J Sternberg. Flotsam: A model for the development and transmission of hate. *Journal of Theoretical Social Psychology*, 2(4):97–106, 2018. 1.2.1, 2.2.2, 3.1, 7.1.2
- [231] Leo Graiden Stewart, Ahmer Arif, A Conrad Nied, Emma S Spiro, and Kate Starbird. Drawing the lines of contention: Networked frame contests within# blacklivesmatter discourse. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):1–23, 2017. 5.2.1
- [232] David B Sugarman, Maury Nation, Nicole P Yuan, Gabriel P Kuperminc, Lama Hassoun Ayoub, and Sherry Hamby. Hate and violence: Addressing discrimination based on race, ethnicity, religion, sexual orientation, and gender identity. *Psychology of violence*, 8(6):649, 2018. 6.2.3
- [233] James Surowiecki. *The wisdom of crowds*. Anchor, 2005. 7.1.1
- [234] Henri Tajfel. Cognitive aspects of prejudice. *Journal of biosocial science*, 1(S1):173–191, 1969. 1.2.1, 6.2.1, 7.1.2
- [235] Henri Tajfel. *Human groups and social categories*. Cambridge University Press, 1981. 1.2.2
- [236] Henri Tajfel and John C Turner. Social psychology of intergroup relations. *Annual review of psychology*, 33(1):1–39, 1982. 1.2.2, 1.2.4, 2.2.2, 7.1.1
- [237] Henri Tajfel and John C Turner. The social identity theory of intergroup behavior. In *Political psychology*, pages 276–293. Psychology Press, 2004. 1.2.4, 7.1.1
- [238] Duyu Tang, Bing Qin, and Ting Liu. Deep learning for sentiment analysis: successful approaches and future challenges. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 5(6):292–303, 2015. 2.2.3
- [239] Abbas Tashakkori and Charles Teddlie. Putting the human back in “human research methodology”: The researcher in mixed methods research, 2010. 3.5
- [240] Abbas Tashakkori and Charles Teddlie. *Sage handbook of mixed methods in social & behavioral research*. SAGE publications, 2021. 3.5
- [241] Yla R Tausczik and James W Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54, 2010. 1.4.1, 2.1, 2.2.2, 2.3.2, 2.6, 3.3.1, 5.3.1, 7.1.1, 7.2.1
- [242] Hannah Tessler, Meera Choi, and Grace Kao. The anxiety of being asian american: Hate crimes and negative biases during the covid-19 pandemic. *American Journal of Criminal Justice*, 45:636–646, 2020. 6.2.3
- [243] Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. From louvain to leiden:

- guaranteeing well-connected communities. *Scientific reports*, 9(1):1–12, 2019. 1.4.2, 4.2.1, 4.2.2, 4.3.3, 4.5, 7.1.2, 7.2.2
- [244] Shu-Feng Tsao, Helen Chen, Therese Tisseverasinghe, Yang Yang, Lianghua Li, and Zahid A Butt. What social media told us in the time of covid-19: a scoping review. *The Lancet Digital Health*, 3(3):e175–e194, 2021. 4.5, 6.2.3, 7.1.5
- [245] Heidi JS Tworek. Fighting hate with speech law: media and german visions of democracy. *The Journal of Holocaust Research*, 35(2):106–122, 2021. 7.3.2
- [246] Aman Tyagi, Joshua Uyheng, and Kathleen M Carley. Heated conversations in a warming world: affective polarization in online climate change discourse follows real-world climate anomalies. *Social Network Analysis and Mining*, 11(1):1–12, 2021. 1.2.4, 5.2.1, 6.2.2
- [247] Joshua Uyheng and Kathleen M Carley. Characterizing bot networks on twitter: An empirical analysis of contentious issues in the asia-pacific. In *Proc. SBP-BRIMS*, pages 153–162. Springer, 2019. 1.1, 1.2.4, 1.3.3, 5.1, 5.2.1
- [248] Joshua Uyheng and Kathleen M Carley. Bot impacts on public sentiment and community structures: Comparative analysis of three elections in the asia-pacific. In *Proc. SBP-BRIMS*, pages 12–22, 2020. 1.1, 5.2.1
- [249] Joshua Uyheng and Kathleen M Carley. Bots and online hate during the covid-19 pandemic: case studies in the united states and the philippines. *Journal of computational social science*, 3(2):445–468, 2020. 1.2.4, 1.3.3, 1.4.2, 5.1, 5.3.1, 5.3.2, 5.5
- [250] Joshua Uyheng and Kathleen M Carley. Characterizing network dynamics of online hate communities around the covid-19 pandemic. *Applied Network Science*, 6(1):1–21, 2021. 1.1, 1.2.4, 1.4.2, 3.2.2, 4.2.2, 4.3.3, 5.3.1, 5.4.4, 5.5, 7.1.3, 7.3.1, A.6
- [251] Joshua Uyheng and Kathleen M Carley. An identity-based framework for generalizable hate speech detection. In *Proc. SBP-BRIMS*, pages 121–130. Springer, 2021. 1.2.3, 1.2.5, 1.4.1, 1.4.3, 2.1, 2.2.1, 2.2.3, 2.3.2, 4.3.2, 4.3.3, 4.3.4, 4.4.1, 5.3.1, 6.3.1, 7.2.1, 7.3.1
- [252] Joshua Uyheng, Thomas Magelinski, Ramon Villa-Cox, Christine Sowa, and Kathleen M Carley. Interoperable pipelines for social cyber-security: assessing twitter information operations during nato trident juncture 2018. *Computational and Mathematical Organization Theory*, 26(4):465–483, 2020. 1.2.5, 1.4, 2.1, 2.2.3, 2.5.1, 2.6, 5.1, 5.2.1, 6.3.1, 6.5, 7.2.3
- [253] Joshua Uyheng, Daniele Bellutta, and Kathleen M Carley. Bots amplify and redirect hate speech in online discourse about racism during the covid-19 pandemic. *Social Media+ Society*, 8(3):20563051221104749, 2022. 5.2.3, 5.4.4, 7.1.3, 7.3.1
- [254] Joshua Uyheng, Iain J Cruickshank, and Kathleen M Carley. Mapping state-sponsored information operations with multi-view modularity clustering. *EPJ Data Science*, 11(1):25, 2022. 1.2.5
- [255] Joshua Uyheng, JD Moffitt, and Kathleen M Carley. The language and targets of

- online trolling: A psycholinguistic approach for social cybersecurity. *Information Processing & Management*, 59(5):103012, 2022. 1.2.5, 1.4.1, 1.4.5, 2.3.2, 2.6, 5.1, 5.2.2, 5.3.1, 5.5, 6.3.1, 7.1.4
- [256] Joshua Uyheng, Dawn C Robertson, and Kathleen M Carley. Bridging online and offline dynamics of the face mask infodemic. *BMC Digital Health*, 1(1):27, 2023. 6.2.3, 6.3.1
- [257] Bertie Vidgen and Taha Yasseri. Detecting weak and strong islamophobic hate speech on social media. *Journal of Information Technology & Politics*, 17(1):66–78, 2020. 3.2.2
- [258] Sebastian Wachs, Ludwig Bilz, Alexander Wettstein, Michelle F Wright, Norman Krause, Cindy Ballaschk, and Julia Kansok-Dusche. The online hate speech cycle of violence: Moderating effects of moral disengagement and empathy in the victim-to-perpetrator relationship. *Cyberpsychology, Behavior, and Social Networking*, 25(4): 223–229, 2022. 6.2.1
- [259] Joseph B Walther. Social media and online hate. *Current Opinion in Psychology*, 45:101298, 2022. 6.2.3
- [260] Cindy Wang. Interpreting neural network hate speech classifiers. In *Proc. ALW*, pages 86–92, 2018. 1.2.3, 2.2.1
- [261] Zeerak Waseem. Are you a racist or am I seeing things? annotator influence on hate speech detection on Twitter. In *Proc. NLP+CSS*, pages 138–142, November 2016. 1.2.3, 2.2.1, 3.1
- [262] Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proc. NAACL SRW*, pages 88–93, 2016. 1.2.3, 1.3.2, 2.3.1, 2.1, 3.2.2, 3.3.1, 3.5, 7.1.2
- [263] Stanley Wasserman, Katherine Faust, et al. *Social network analysis: Methods and applications*. Cambridge university press, 1994. 1.4.2, 4.1, 4.2.1, 4.3.3, 7.1.2
- [264] Wei Wei and Kathleen M Carley. Measuring temporal patterns in dynamic social networks. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 10(1): 1–27, 2015. 4.2.1
- [265] Maximilian Wich, Jan Bauer, and Georg Groh. Impact of politically biased data on hate speech classification. In *Proc. WOAHP*, pages 54–64, 2020. 1.2.3, 2.2.1
- [266] Susann Wiedlitzka, Gabriele Prati, Rupert Brown, Josh Smith, and Mark A Walters. Hate in word and deed: the temporal association between online and offline islamophobia. *Journal of quantitative criminology*, 39(1):75–96, 2023. 6.2.3
- [267] B Dan Wood and Jeffrey S Peake. The dynamics of foreign policy agenda setting. *American Political Science Review*, 92(1):173–184, 1998. 5.3.2
- [268] Samuel C Woolley and Philip N Howard. Political communication, computational propaganda, and autonomous agents: Introduction. *International journal of Communication*, 10, 2016. 1.2.5
- [269] Ruochen Xu, Yiming Yang, Hanxiao Liu, and Andrew Hsi. Cross-lingual text clas-

- sification via model translation with limited dictionaries. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 95–104, 2016. 2.3.2
- [270] Anandita Yadav. Countering hate speech in india: Looking for answers beyond the law. *ILI Law Review*, 2:1–17, 2018. 7.3.2
- [271] Chao Yang and Padmini Srinivasan. Life satisfaction and the pursuit of happiness on twitter. *PloS one*, 11(3):e0150881, 2016. 2.2.2, 2.6
- [272] Kai-Cheng Yang, Emilio Ferrara, and Filippo Menczer. Botometer 101: Social bot practicum for computational social scientists. *Journal of computational social science*, 5(2):1511–1528, 2022. 5.2.2
- [273] Moran Yarchi, Christian Baden, and Neta Kligler-Vilenchik. Political polarization on the digital sphere: A cross-platform, over-time analysis of interactional, positional, and affective polarization on social media. *Political Communication*, 38(1-2):98–139, 2021. 1.2.4, 4.2.2, 4.5, 7.1.3
- [274] Lucy Yardley. Demonstrating the validity of qualitative research. *The journal of positive psychology*, 12(3):295–296, 2017. 3.5, 7.2.2
- [275] Savvas Zannettou, Tristan Caulfield, William Setzer, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. Who let the trolls out? towards understanding state-sponsored trolls. In *Proc. WebSci*, pages 353–362, 2019. 5.2.2, 7.1.4, 7.2.3
- [276] Savvas Zannettou, Mai ElSherief, Elizabeth Belding, Shirin Nilizadeh, and Gianluca Stringhini. Measuring and characterizing hate speech on news websites. In *Proceedings of the 12th ACM conference on web science*, pages 125–134, 2020. 5.2.3
- [277] Yin Zhang, Rong Jin, and Zhi-Hua Zhou. Understanding bag-of-words model: a statistical framework. *International journal of machine learning and cybernetics*, 1: 43–52, 2010. 2.2.1, 2.3.2
- [278] Steven Zimmerman, Udo Kruschwitz, and Chris Fox. Improving hate speech detection with deep learning ensembles. In *Proc. LREC.*, 2018. 1.2.3, 2.2.1