University of California
Santa Barbara

Computational Methods for Next-Generation Online Media Ecosystems

A dissertation submitted in partial satisfaction
of the requirements for the degree

Doctor of Philosophy
in
Computer Science

by

May E. ElSherif

Committee in charge:

Professor Elizabeth Belding, Chair
Professor William Wang
Professor Amr Elabaddi

December 2019

The Dissertation of May E. ElSherif is approved.

Professor William Wang

Professor Amr Elabaddi

Professor Elizabeth Belding, Committee Chair

August 2019

Computational Methods for Next-Generation Online Media Ecosystems

by

May E. ElSherif

In Memory of all Lives Lost due to Hate and Bigotry

Acknowledgements

This thesis is about responsibility to promote the good and prevent the harm. This work is the result of the best and most challenging years of my life.

I would like to start by thanking my family who always put education at the forefront of their priorities. I am really grateful for all their efforts, their endless support, and love throughout my life. They taught me work ethic, perseverance, and ambition by example. To my dad, thank you for your generosity in providing me with whatever I needed through my education journey. To my mum, thank you for lending me an ear, being there whenever I needed you, and believing in me against all the odds. To my sister, thank you for giving me positive vibes and listening to all my stories. To my brother, who is not here today, you will always be in my thoughts and I wish you were here today.

I cannot thank enough my research adviser, Elizabeth Belding. I have been fortunate to have her as my mentor, supporter, and my advocate. Not only did she believe in me as a researcher but also research line I chose to tackle. She always finds way to make research more interesting and impactful. I did not only learn how to be a good researcher from her but I also learned how to be a good mentor. I would not have been the researcher I am now if it weren't for her.

Many thanks to my mentor William Wang who introduced me to the field of Natural Language Processing and how to tie it to problems with social impact. I had the great privilege to learn a lot from his vast knowledge and applying this knowledge in my work.

To Amr Elabaddi, thank you for your insightful research questions and for always having an open door and open mind when I needed help. He always made me feel that I have a part of my Egyptian family away from home.

To my collaborators: Mogran Vigil-Hayes, Ramya Raghavendra, Shirin Nilizadeh, Vivek Kulkarni, and Jing Qian, thank you for contributing to this work in many ways

and it was a joy to work with you.

I had the privilege to work with and mentor many undergraduates who developed a passion for Computer Science research: Dana Nguyen, Xuewen Sherry Li, Barbara Korycki, Andrew Gaut, Tony Sun, Shirlyn Tan, and Yuxin Huang. Together, we engaged in tackling research problems and I enjoyed every discussion we had. Special thanks to Elizabeth Belding, William Wang, and Diba Mirza who guided me through the mentoring process and provided feedback whenever needed.

I would like to especially thank Dana Nguyen who has demonstrated unprecedented excitement and dedication to research. We co-authored together multiple publications through which she showed persistence, ownership, and hard work.

To my colleagues and friends: Morgan Vigil-Hayes, Victor Zakhary, Mohamed Wahba, Radwa Hamed, Nadah Feteih, Mahmoud Ramadan, Mai Said, Vivek Adarsh, and Esther Showalter. Your words of encouragement and our conversations kept me going through lots of challenges. Finally, I would like to specifically thank Mohamed Wahba for his support during lots of deadlines. He heard all my ideas, confusions, helped me practice so many talks, and constantly affirmed my place in the academic community.

Curriculum Vitæ

May E. ElSherif

Education

2014-2019	Ph.D. in Computer Science, University of California, Santa Barbara. Advisor: Elizabeth Belding Thesis: *Computational Methods for Next-Generation Online Media Ecosystems*
2011-2013	M.Sc. in Wireless Communications and Information Technology, Nile University, Giza, Egypt. Advisors: Tamer ElBatt, Ahmed Zahran and Ahmed Helmy Thesis: *Similarity Metrics and an Information-theoretic Framework for Opportunistic Mobile Social Networks*
2006-2011	B.Sc. in Computer Engineering, Cairo University, Giza, Egypt. Graduated with Distinction Advisor: Hatem ElBoghdady Thesis: *The Golden Retriever: A Robotic Odor Source Localizer*

Research Interests

Social computing, natural language processing, data science, applied machine learning, computer-supported cooperative work, computational social science

Awards & Honors

2019	UCSB Computer Science Outstanding Graduate Student Award
2018	UCSB Grad Slam Finalist (Public Speaking)
2017	UCSB Fiona and Michael Goodchild Graduate Mentoring Award
2011	Google Research Award (Nile University)
2011	Nile University Research Fellowship (Full academic scholarship)
2009	IEEE Best Project Award (Cairo University Student Brach)
2009	Ideal Student Award (Cairo University)
2008	Best Volunteer Team (Suzanne Mubarak Women's International Peace Movement NGO)

Publications

Refereed Papers

[1] J. Qian, M. ElSherief, E. Belding, and W. Yang Wang. Learning to Decipher Hate Symbols. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT'19).*

[2] A. Gaut, T. Sun, S. Tang, Y. Huang, M. ElSherief, J. Zhao, D. Mirza, E. Belding, K. Chang, and W. Yang Wang. Mitigating Gender Bias in *Natural Language Processing: Literature Review. In Annual Meeting of the Association for Computational Linguistics (ACL'19)*, 2019.

[3] J. Qian, M. ElSherief, E. Belding, and W. Yang Wang. Hierarchical CVAE for Fine-Grained Hate Speech Classification. In *Empirical Methods of Natural Language Processing (EMNLP'18)*, November 2018. [25% Acceptance Rate].

[4] J. Qian, M. ElSherief, E. Belding, and W. Yang Wang. Leveraging Intra-User and Inter-User Representation Learning for Automated Hate Speech Detection. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT'18)*, June 2018. [18% Acceptance Rate].

[5] M. ElSherief, V. Kulranki, D. Nguyen, W. Yang Wang, and E. Belding. Hate Lingo: A Target-based Linguistic Analysis of Hate Speech in Social Media. In *AAAI Conference on Web and Social Media (ICWSM'18)*, June 2018. [16% Acceptance Rate].

[6] M. ElSherief, S. Nilizadeh, D. Nguyen, G. Vigna, and E. Belding. Peer to Peer Hate: Hate Speech Instigators and Their Targets. In *AAAI Conference on Web and Social Media (ICWSM'18)*, June 2018. [16% Acceptance Rate].

[7] M. ElSherief, B. Alipour, M. Al Qathrady, T. ElBatt, A. Zahran, and A. Helmy. A Novel Mathematical Framework for Similarity-based Opportunistic Social Networks. *Elsevier Pervasive and Mobile Computing*, 42:134–150, December 2017. [Journal Extension].

[8] M. ElSherief, M. Vigil-Hayes, R. Raghavendra, and E. Belding. Whom to Query? Spatially-Blind Participatory Crowdsensing under Budget Constraints. In *ACM Workshop on Mobile Crowdsensing Systems and Applications (CrowdSenSys'17)*, November 2017.

[9] M. ElSherief, E. Belding, and D. Nguyen. #NotOkay: Understanding Gender-based Violence in Social Media. In *AAAI Conference on Web and Social Media (ICWSM'17)*, May 2017. [14% Acceptance Rate].

[10] M. ElSherief, T. ElBatt, A. Zahran, and A. Helmy. An Information-theoretic Model for Knowledge Sharing in Opportunistic Social Networks. In *IEEE Conference on Social Computing and Networking (SocialCom'15)*, December 2015. [25% Acceptance Rate].

[11] M. ElSherief and E. Belding. The Urban Characteristics of Street Harassment: A First Look. In *ACM Workshop on Smart Cities and Urban Analytics (UrbanGIS'15)*, November 2015.

[12] M. ElSherief, T. ElBatt, A. Zahran, and A. Helmy. The Quest for User Similarity in Mobile Societies. In *IEEE Workshop on Social and Community Intelligence (SCI'14)*, March 2014. [17% Acceptance Rate].

Other Contributions

[13] M. ElSherief. Compressed Lexicon and Currently the Largest Twitter Hate Speech Dataset. `https://github.com/mayelsherif/hate_speech_icwsm18`, June 2018.

[14] M. ElSherief, M. Vigil-Hayes, R. Raghavendra, and E. Belding. Whom to Query? Spatially-Blind Participatory Crowdsensing under Budget Constraints. In *AAAI Conference on Human Computation and Crowdsourcing (HCOMP'17)*, October 2017. [Poster].

[15] M. ElSherief, T. ElBatt, A. Zahran, and A. Helmy. O'BTW: An Opportunistic Similarity-based Mobile Recommendation System. In *ACM Conference on Mobile Systems, Applications and Services (MobiSys'13)*, June 2013. [Demo].

Experience

CS8: Introduction to Computer Science, *Instructor of Record, UCSB.* [Summer 2019]

UCSB MOMENT Lab, *Gradaute Research Assistant, Santa Barbara, CA.* [2014-2019]

Early Research Scholars Program (ERSP), *Lead TA, UCSB.* [Fall 2018-Spring 2019]

Berkman Klein Center for Internet and Society (BKC), Harvard University, *Research Intern, Cambridge, MA.* [Summer 2018]

Polaris Wireless Research, *Data Science Intern, Mountain View, CA.* [Summer 2016]

Wireless Intelligent Networks Center, *Research Assistant, Nile University, Giza, Egypt.* [2011-2013]

Center of Informatics Science, *Bioinformatics Intern, Nile University, Giza, Egypt.* [Summer 2010]

Teaching Assistant, *Computer Science Department, UCSB.* [Winter-Spring 2014]

Teaching Assistant, *Engineering Math and Physics Department, Cairo University.* [2012-2013]

Abstract

Computational Methods for Next-Generation Online Media Ecosystems

by

May E. ElSherif

Human biases have found their way into our digital footprints. Human corpora and human forms of expression mirror biases inherent in societies explicitly or implicitly. Now, information all around us is tainted with these biases. This has led to severe consequences from a technological perspective.

First, social and cultural biases have found their way into technology, and in particular into automated tools that rely on human-generated data leading to discriminatory systems [1]. Secondly, while online information ecosystems provide freedom of expression and give voice to individuals, they have also suffered a wave of disorder due to the prevalence of malevolent online misuse, manifested as hate speech and online misinformation, such as fake news. These problems are motivated by bias and present unprecedented challenges because they “cannot be solved in a traditional linear fashion, since the problem definition evolves as new possible solutions are considered and/or implemented” [2]. In this thesis, we investigate the digital representations of these prejudices including issues of gender equality and hate speech.

In the first part of the thesis, we begin by analyzing stories of women sharing their harassment experiences and show how targets of gender-based violence utilize social media to shift their cognitive states by leveraging storytelling. We then move into studying gender bias representations in Natural Language Processing. We provide a comprehensive review of current methods that attempt to debias corpora and prevent bias amplification in machine learning models. We then show that current Neural Relation Extraction systems exhibit gender bias.

In the next part of the thesis, we focus on online hate speech and its nuances on social media. In order to design automated hate speech detection systems, we must empirically study existing instances of hate speech. We present the first set of online hate speech studies that investigate hate instigators and hate targets, linguistic properties of directed and generalized hate speech, and online hate communities. As a result of this work, we make publicly available a high precision dataset of 28K tweets, currently the largest Twitter hate speech dataset available to the research community. Our work includes a one of a kind set of analyses pertaining to hate speech that have impacted the design of hate speech detection systems by improving the F-1 score of hate speech detection and classification systems in online social media by an average of 10%.

Our work enables the design of the next-generation hate speech detection systems and gender bias detection and mitigation systems. We conclude with an overview of our key findings as well as a discussion of future research directions inspired by the work in this dissertation.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Bias is defined as *prejudice in favor of or against one thing, person, or group compared with another, usually in a way considered to be unfair* [5]. Today, in this world, multiple biases are present through society in a complex manner. These biases do not only affect targets of prejudice but also the society at large.

On the societal level, actions motivated by biases cause domestic violence, crime, death, property loss, and expense to society in the form of court trials and providing psychological counseling [6]. Additionally, harmful prejudicial actions can create barriers for entire populations, such as women or minorities, seeking the benefits of participating in mainstream society [6].

Since the creation of online platforms including online social networks, people have carried their inherent prejudices into the digital world. Targets of biases and prejudice have used online platforms and social networks to report incidents of violence and create social activism movements including *Black Lives Matter* (#blacklivesmatter) for racial equality and *Love Wins* (#lovewins) for marriage equality [7, 8]. In this thesis, we investigate the online representations of these prejudices including issues of gender equality and hate speech.

In the first part of the thesis, we tackle issues related to gender equality and gender biases in the digital world including online platforms and text corpora. Gender equality, also known as sexual equality or equality of the sexes, is the *state of equal ease of access to resources and opportunities regardless of gender, including economic participation and decision-making; and the state of valuing different behaviors, aspirations and needs equally, regardless of gender.* There has been a global effort with respect to achieving gender equality especially by international agencies such as UNICEF [9] and UNFPA [10]. The objectives of the initiatives taken to achieve gender equality include dismantling gender stereotypes and sexism and preventing and combating violence against women among other objectives such as achieving equal access of women to justice and political and public decision making [11, 12]. We are particularly interested in analyzing how women share their gender-violence stories and the detection of gender stereotypes in text corpora.

In the second part of the thesis, we investigate the issue of online hate speech which constitutes *prejudiced attacks towards people based on their prominent and protected attributes such as race, religion, sexual orientation, gender among others.* We particularly analyze online hate actors, individual hate language instances and hate language with communities of hate. The current hate speech characterization and detection communities work independently from each other. This dissertation bridges the gap between the two research communities by calling for detection systems that are informed by the results of the studies in this dissertation.

1.1 Thesis, Contributions, and Impacts

This dissertation demonstrates that:

Data-driven analyses pertaining to understanding biases, including gender, linguistic, engagement, visibility, and personality analyses, lead to the discovery of characteristics that can have far-reaching computational implications such as nuanced hate speech

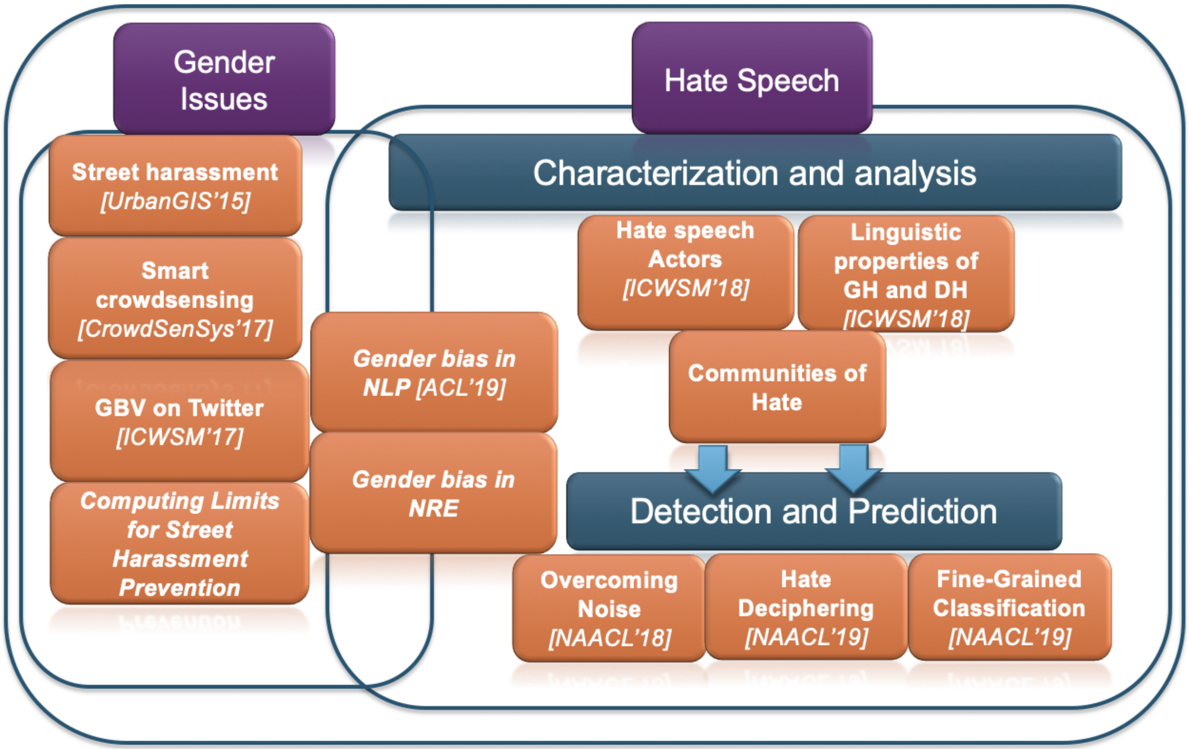

Gender Issues
Hate Speech
Street harassment [UrbanGIS'15]
Smart crowdsensing [CrowdSenSys'17]
GBV on Twitter [ICWSM'17]
Computing Limits for Street Harassment Prevention
Gender bias in NLP [ACL'19]
Gender bias in NRE
Characterization and analysis
Hate speech Actors [ICWSM'18]
Linguistic properties of GH and DH [ICWSM'18]
Communities of Hate
Detection and Prediction
Overcoming Noise [NAACL'18]
Hate Deciphering [NAACL'19]
Fine-Grained Classification [NAACL'19]

Figure 1.1: Dissertation Overview.

and biases detection and mitigation and societal implications including reducing emotional distress of victims and law and policy design for socially-harmful phenomena.

The overview of this dissertation is depicted in Figure 1.1. In this thesis, we utilize a data-centric approach to understand issues related to gender-based violence, gender biases, and hate speech. In the rest of this chapter, we provide a summary of the work associated with each category in Figure 1.1 as well as how the work contributes to the state of the art and the associated intellectual impacts.

In the first part of this thesis, we focus on two gender-equality related issues. The first issue is in respect to how online platforms are used to report street harassment and how online social networks are used in reporting and combating violence against women. The second issue pertains to combating gender stereotypes in the science of Natural Language Processing (NLP). We outline the details pertaining to the aforementioned two issues below.

Online Reporting Platforms. In this part of the thesis, we focus on how gender-based violence (GBV) victims are empowered by online reporting platforms. Specifically, we leverage a data-driven approach to analyze the stories shared by women when they report street harassment and gender-based violence. In Chapter 2, we analyze street harassment stories on the online platform Hollaback in order to improve our understanding of where harassment is likely to occur. This is the first work that investigates walkability and transit scores of 7, 800 worldwide street harassment incidents. We show that *street harassment is likely to occur in areas with high walkability scores and transit scores.* This work has the potential to enhance the urban mobility of pedestrians to avoid areas that are likely to include street harassment.

In Chapter 4, we study how online social media has been a key enabler of conversations about socially complex issues such as gender-based violence. In one of the first studies of GBV in Twitter , we collect and analyze over 300K tweets that pertain to three types of GBV: physical violence, sexual violence, and harmful practices. Through this work, we provide one of the first empirical insights into social media discourse on GBV. Our analysis show *higher user engagement with GBV tweets than with generic tweets, but that engagement is not uniform across all ages and genders.* We show that public figures used hashtags, such as #notokay, #maybehedoesnthityou, and #beenrapedneverreported, to motivate women to speak up about GBV and share their personal experiences. Our psycholinguistic analysis reveals that *anger often surfaces in GBV content.* This work shows that the pervasiveness of social media can provide platforms of community while giving voice to victims of harassment and violence. Additionally, the data we derive from our analysis can be used to complement policy design data sources.

In Chapter 5, we build onto our results from Chaptesr 2 and 4 by investigating the limits that hinder the development of information and communication technologies (ICT) for safety on the streets. We conduct an interview with the Harassmap research team, an organization that fights sexual harassment in Egypt. Based on our GBV studies and the interview, we *outline four types of limits hindering the development of technology that fights street harassment, including those imposed by platforms, society, emerging interpretations of location, and incomplete data sets.* This is the first study that not only outlines the current limits to create computational technologies that help predict and prevent gender violence but also *sheds light on three promising future directions including: the usage of social media to raise awareness and create social movements, cyber-physical systems that tie on-the-ground data to urban mobility decisions, and on the ground social change.*

Querying for Real-time Phenomenon. Motivated by the spatio-temporal nature of phenomenon like street harassment and the realistic nature of limited resources, we seek to answer the research question: Given M workers in a spatial environment and N probing resources, where $N < M$, which N workers should be queried to answer a specific question pertaining to a real-time phenomenon? In Chapter 3, we *introduce and define the task of spatially blind crowdsensing under budget constraints.* In this context, budget constraints may arise from limitations pertaining to network bandwidth, energy, user attention, time, and money. Moreover, we propose two querying algorithms: one that exploits worker feedback (DispNN) and one that does not rely on worker feedback (DispMax). Our algorithms *outperform a random selection approach by up to* 30%, *a random selection approach with feedback by up to* 35%, *a greedy heuristic by up to* 5*x times, and cover up to a median of* 96% *of the incidents.*

Gender bias in NLP. As Natural Language Processing (NLP) and Machine Learning (ML) tools rise in popularity, it becomes increasingly vital to recognize the role they play in shaping societal biases and stereotypes. Although NLP models have shown success in modeling various applications, they propagate and may even amplify gender bias found in text corpora. While the study of bias in Artificial Intelligence is not new, methods to mitigate gender bias in NLP are relatively nascent. In this thesis, we contribute three studies to the area of gender bias detection and mitigation in NLP. First, we present the first study, of the timely topic of gender bias in NLP, *that reviews and contextualizes prior works that relate to algorithmic bias in NLP under a unified framework and critiques issues with current debiasing methods.* Secondly, we analyze bias in Neural Relation Extraction (NRE) classifications in Chapter 7. Our contributions do not only encapsulate the *gender bias detection in NRE systems* but additionally, we *create Wikigender, distantly supervised dataset with a human annotated test set that has an even split of male and female sentences specifically curated to analyze gender bias in relation extraction systems.*

In the second part of this thesis we focus on issues pertaining to hate speech characterization and detection. First, we formulate novel research questions pertaining to the online representation of hate speech. Second, we collect and analyze hate speech datasets for different types of hate speech: directed, generalized, and community-based. We then show how these insights could be used to improve the detection of hate speech.

Hate Speech Actors. Little prior work has focused on the understanding or characterization of online hate speech actors. Our work represents the first comparative study of hate instigators, targets and general Twitter users in terms of profile self-presentation, Twitter visibility, and personality traits in Chapter 9. To perform this study, we first curated and made publicly available the largest Twitter hate speech dataset,[1] representing 28K tweets. To generate this dataset, we employed a novel four-step filtering process to ensure a high-fidelity dataset; this differs greatly from prior datasets, which tend to have large volumes of text incorrectly classified as hate speech. As a first step, we investigated the lexicon of Hatebase (the world's largest hate expression repository) and provided a

[1] https://github.com/mayelsherif/hate_speech_icwsm18

reduced set of 51 terms with the highest likelihood of representing hate speech content across eight different hate classes. We then outlined a method of semi-automated classification that could be used for directed explicit hate speech data curation. This work showed that *hate instigators target more visible users* and that *participating in hate commentary is associated with higher visibility.* Through use of the Big Five personality traits model, a well-studied Psychology model for modeling human engagement with the world, we also showed that *hate instigators and targets have unique personality characteristics, such as anger, depression, and immoderation, which may contribute to hate speech, and that instigators and targets could exchange roles, i.e., instigators can become targets of hate and vice versa.*

One of the main impactful findings of this chapter is that instigators share a common inherent representation. This finding have been used to improve the F-1 score of hate speech detection systems by leveraging the commonalities of inter and intra-user representation specifically to overcome the short and noisy nature of social media posts [13]. In this work, semantically similar tweets posted by other users were leveraged by implementing a reinforced bidirectional Long short-term memory network (LSTM) to interactively utilize similar tweets from a large Twitter dataset to enhance the performance of the hate speech classifier. To leverage a user's historical tweets, the tweets were collected and fed into a pre-trained model to obtain an intra-user representation. Experimentally, it was shown that leveraging these two representations can significantly improve the f-score of a strong bidirectional LSTM baseline model by as much as 10.1%.

We then note that hate speech comes in different forms as outlined in the next studies.

Linguistic Properties of Directed vs. Generalized Hate Discourse. Prior work has ignored a crucial aspect of hate speech – the target of hate speech – and only seeks to distinguish hate and non-hate speech. Such a binary distinction fails to capture the nuances of hate speech – nuances that can influence free speech policy. We performed the first linguistic and psycholinguistic analysis of these two forms of hate speech to reveal the presence of markers that distinguish these types of hate speech. We noted that hate speech can be *aimed at a specific individual (Directed) or it can be aimed at a group or class of people (Generalized).* To perform this analysis, we had to curate the first dataset that captures this distinction; we have since made this dataset publicly available. In our analyses, we trained mixed-effect topic models, and leveraged Named Entity Recognition and Frame-Semantic Parsing, to understand the nuances of Directed and Generalized hate speech. This work, discussed in Chapter 10, showed that *Directed hate speech is very personal, more informal, and angrier than Generalized hate speech*, where religious and ethnic terms dominate. Directed hate speech invokes words that suggest intentional action and explicitly uses words to hinder the action of the target. In contrast, *Generalized hate speech is dominated by quantity words* such as million, all, many; *religious words* such as Muslims, Jews, Christians; and *lethal words* such as murder, beheaded, killed, exterminate.

Hate Communities. In this work, we note that hate speech does not just take the form of individualistic posts but also could take a form of organized efforts from communities. With the growth of the number of hate groups recently and the wide reach of social media, we study the discursive practices of these communities. We collected a Twitter dataset, comprised of 4.7M tweets, for the eight types of hate ideology that constitute the largest presence in the U.S: White Nationalist, Black Nationalist, Ku Klux Klan, Anti-LGBT, Anti-Muslim, Neo- Nazi, Anti-Immigrant, and Racist Skinhead. We conduct the longest temporal linguistic analysis for these hate groups (2015-2017) and present the results in Chapter 11. Our analysis reveals the following key findings: first, unlike prior work that focuses on derogatory hate speech lexicons, *hate groups leverage formal language to disseminate their message.* Second, we find that *the majority of the hate ideologies are driven by power with the exception of the KKK which is driven* *by affiliation.* Next, we observe a *high semantic similarity (approximately 30%) between certain ideology pairs which shows the presence of common interests and could be leveraged for response prediction of certain ideologies.* Additionally, we also observed the usage of cryptic symbols that are used to encode hidden meaning among hate group members. Finally, we show the presence of questionable media sources among hate group content which opens the room for discussion to the correlation between fake news and hate speech. These findings have resulted in the following impacts:

- The introduction of a fine-grained hate speech classification task that separates tweets posted by 40 hate groups in 13 different hate group categories [14]. This is the first work on fine-grained hate speech classification that attributes individual tweets to hate groups. A novel Hierarchical Conditional Variational Autoencoder (CVAE) model was proposed for fine grained tweet hate speech classification that improves the Micro-F1 score of up to 10% over the baselines.

- The study of the application of Sequence to Sequence (Seq2Seq) models to crack the symbols used by online hate actors based on context. In this work, a novel Variational Decipher was proposed to be able to generalize to unseen hate symbols in a challenging testing setting [15]. The proposed Seq2Seq model is implemented based on an RNN Encoder-Decoder model with attention mechanism while the Variational Decipher is based on the Conditional Variational Autoencoder (CVAE) model. This work showed that the Seq2Seq model outperforms the Variational Decipher for deciphering the hate symbols with similar definitions to that in the training dataset. This means the Seq2Seq model can better explain the hate symbols when Twitter users intentionally misspell or abbreviate common slur terms. On the other hand, the Variational Decipher tends to be better at deciphering hate symbols with unseen definitions, so it can be applied to explain newly created hate symbols on Twitter.

Part I

Gender Issues

Chapter 2

The Urban Characteristics of Street Harassment

In this chapter, we seek to gain insights into the characteristics of neighborhoods in which street harassment has occurred. We analyze over $7,800$ worldwide street harassment incidents, gathered by the Hollaback project [16], to study the association of street harassment with walkability scores and the number of transit routes in the area surrounding the incident. This unveils a number of key insights. First, we show that more than 50% of the incidents occur in highly walkable areas with walkability scores ranging from 90 to 100, and that nonintuitively, as the walkability score increases, the probability of street harassment events increases. The same result is obtained for areas with high transit scores. Further, the number of transit routes within one mile of the harassment incident has a negative correlation with the number of incidents. The insights gained from our study are a step towards understanding where harassment is likely to occur, which we hope can one day be used for prevention of future incidents.

2.1 Introduction

Street harassment is a worldwide problem; not only is it a frequent occurrence in developing countries, but in many developed countries, such as the U.S., Italy and New

Zealand, women are much more likely to feel unsafe on the streets at night than men due to the potential for verbal and/or physical harassment [17]. According to one study [18], 65% of women and 25% of men have experienced street harassment in the United States. This harassment can have numerous undesirable side effects on victims, such as a reduced sense of safety, anxiety, depression, and refusal to engage in civic life [19]. Efforts to study and address street harassment from a societal point of view include [19] and [17], among others.

Hollaback [16] is a non-profit movement powered by local activists in 92 cities and 32 countries to end street harassment. The Hollaback project collects data on street harassment events worldwide. Through the Hollaback phone app and the online platform, users can report stories of street harassment to share with the Hollaback community. This empowers victims to speak out about everyday harassment and spread the word about the prevalence of these events. In some communities, local governments are informed in real-time about street harassment so that there is a system-wide level of accountability. In addition, the Hollaback app uses GPS to record a data set representing the locations of street harassment events as a means of improving the collective understanding of street harassment and how it can be prevented. As of July 2015, over $7,800$ street harassment incidents have been recorded in their dataset since February 2011. It is on this data set that our work is focused.

In this chapter, we use the Hollaback data set to study how users report street harassment stories and analyze the characteristics of the streets where the incidents occur. Our analysis of the data set results in a number of key findings, including:

- The most commonly used words reported in harassment stories are “walking”, “man/guy”, and “home”.
- Street harassment incidents occur more frequently in areas with higher walkability

scores [20].

- The most common type of harassment is verbal.
- Street harassment incidents occur frequently along streets with higher transit scores and fewer nearby transit routes (i.e. routes for buses, rail, etc.).

Through our deepened understanding of street harassment events, it is our hope that potential incidents can one day be prevented through, for instance, better route planning to avoid location and time correlations in which events are more likely to occur.

2.2 Related Work

There are several organizations that fight street harassment by building platforms where users can report incidents, share their stories and interact with others who have gone through similar experiences. Examples include Stop Street Harassment [19], and Hollaback [16], among others. These platforms aggregate user experiences and some provide a map of harassment incidents.

Our work lies in the area of urban informatics, which is an emerging field that aims to analyze data to understand how cities function and how people behave in response to different issues they face [21]. The field deals with problems related to issues ranging from traffic and morning commute to preparedness for emergencies. Urban informatics data analysis is used to enable more informed planning decisions, which results in more effective city management. For instance, street walkability can have effects on wealth [22] and health [23]. Examples include websites that can be used to learn of neighborhoods with public transit routes, better commutes and healthier lifestyles (e.g., walkscore.com [20] and walkonomics.com). Recommendation of beautiful, quiet and happy routes that can

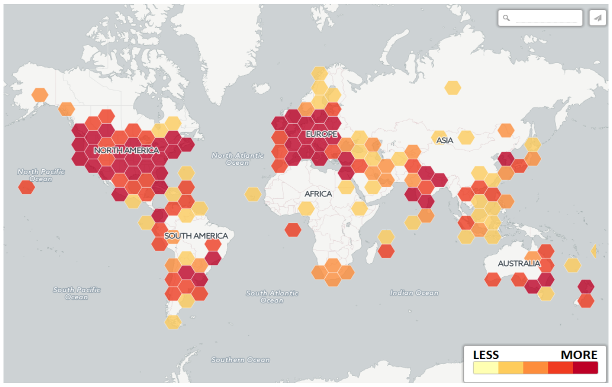


Figure 2.1: Hollaback dataset density map. Darker spots correspond to higher number of incidents.

make travel more enjoyable in cities instead of the shortest routes is explored in [24]. Automation of walkability score calculation using social media is presented in [25]. While urban planners are motivated to build walkable streets, [26] shows that adults can be dissatisfied with living on walkable streets due to the association of these streets with more aesthetics-related problems and lower safety.

2.3 Street Harassment Dataset

We analyze a dataset of $7,838$ street harassment stories provided by Hollaback [16]. Our dataset spans the period from February 2011 to July 2015. Figure 2.1 shows a heat map of reported locations during this period. Cities with the highest number of harassment incidents in this dataset include San Francisco, Los Angeles, New York, Boston, Toronto, Buenos aires, London, Berlin, Paris and Rome. Each street harassment entry is composed of a title, type of harassment, a story, report time, a latitude and a longitude. Reports can be updated after initial entry and only indicate the time of the entry or update, not the time the event actually occurred.

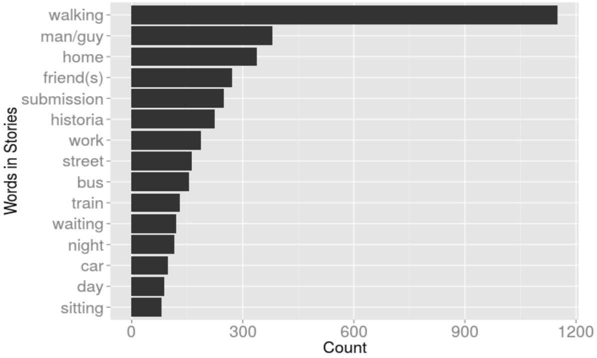


Figure 2.2: Histogram of common words in street harassment reports.

2.3.1 Preliminary Analysis

The Hollaback data set is chosen because it contains multiple components that can help us better understand street harassment. To understand how and where the harassment events take place, we examine the stories for common situational circumstances. Figure 2.2 shows the top 15 most frequent words in shared stories. We discard non-descriptive words such as “I, was, a, the, to, my, and, in”, among others. We observe that “walking” is the most frequently used word. This leads us to investigate the correlation between street harassment locations and walkability scores, which we discuss in the next subsection. The words “street” and “bus” rank roughly equivalently at eight and nine. This likely indicates that harassment occurs not only along city streets but also on buses. In the next section, we take a closer look at the urban environment surrounding the GPS locations associated with the street harassment reports.

2.3.2 Urban Analysis

The urban environment around us, whether or not we are consciously aware of it, has a number of effects, both positive and negative. To quantify these effects on human beings, urban informatics researchers have introduced the term “walkability”. In his book

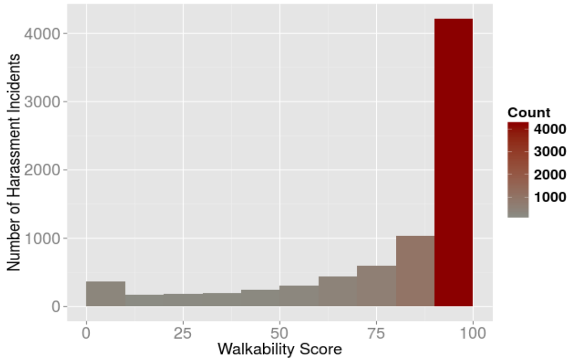


Figure 2.3: Histogram for Hollaback dataset with respect to walkability scores.

Walkable City, Jeff Speck explains that for a walk to be favorable, it has to be useful, safe, comfortable and interesting [27]. Motivated by "walking" as the most commonly used word in the harassment reports, we pose the following question: **Is street harassment related to walkability?** To answer this question, we use the GPS locations reported in our data set and submit them to the "walkscore.com" web service, which has been used by others in [28, 29]. The "walkscore.com" web service takes a GPS location and returns the walkability score computed for this location. To calculate a walkability score, "walkscore.com" computes the distance to nearby amenities and incorporates pedestrian friendliness and street dimensions.

Figure 2.3 shows a histogram of the results we obtained. We can draw two important observations from the figure. First, 53.8% of the street harassment events occurred in streets with very high walkability scores, from 90 to 100. Second, the number of street harassment occurrences increases with the increase of walkability score. This suggests that walkability scores are highly correlated with street harassment incidents. We can also observe that there is a slightly greater number of incidents associated with walkability scores from $0 - 10$, which suggests that "unwalkable" streets can be a good medium for harassers, possibly due to the lack of activity/witnesses in these areas.

The significance of this result does not only lie in the positive correlation found

Harassment Type	Severity Level
assault	5
groping	4
stalking	3
verbal	2
other	1

Table 2.1: Harassment type mapped to severity level.

between walkability and street harassment. The fact that this data is collected from different cities across multiple continents demonstrates the consistency of the results over different parts of the world. Moreover, this result agrees with [26], arguing that walkability is not necessarily positively correlated with adult satisfaction. This indeed opens room for the consideration of other dimensions in the calculation of walkability scores, including safety.

Based on the results in section 3.1, we next seek to determine **whether the degree of severity of the harassment is related to walkability scores**. To answer this question, we annotate each type of harassment with a number depending on its severity as shown in Table 2.1. The type of harassment is specified by the user reporting the incident through check boxes and he/she may choose more than one type to include. Figure 2.4 depicts a jitter plot that graphs harassment severity on the x-axis and walkability scores on the y-axis. The figure shows that the dominant harassment type is verbal, constituting approximately 52% of the entries, and occurs across virtually all walkability scores. The other types of harassment tend to occur more frequently in areas with high walkability scores. At any walkability score, the most likely type of harassment will be verbal, but surprisingly, the risk of harassment events is positively correlated with high walkability scores.

Next, we shift our attention to studying the transit properties of the environment surrounding the street harassment reports. By transit properties, we mean the number of

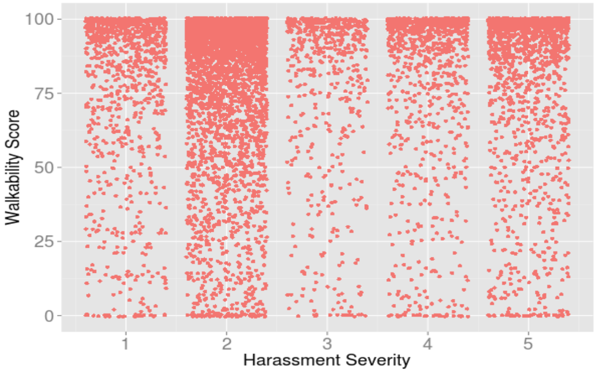


Figure 2.4: Jitter plot for walkability vs harassment scores.

transit routes in an area and the quality of service of these routes. To examine the transit properties, we investigate two metrics: the transit score and the number of transit routes. The website "walkscore.com" defines the transit score of a GPS location as a patented measure of how well a location is served by public transit on a scale from 0 to 100. The number of transit routes is a measure of the number of different routes taken by buses, trains and other transit options within one mile of the specified location. In this section, we are limited by the cities for which "walkscore.com" has transit information. Thus, our dataset is reduced to $3,289$ street harassment entries. It is worth noting that the number of transit routes and transit scores are not directly correlated. An area served by one transit route can have either a very high or low transit score depending on other characteristics such as service level/frequency and the distance to the nearest stop. This is illustrated in Figure 2.5, which shows that areas with few transit routes can have a wide spectrum of transit scores. However, in general as the number of transit routes increases, so does the transit score.

We then ask the following question: **Is street harassment correlated with transit scores and/or number of transit routes?** To answer this question, we plot a histogram of the transit scores and number of transit routes for our reduced dataset in Figures 2.6 and 2.7, respectively. From the trends in Figures 2.6 and 2.7, we note that,

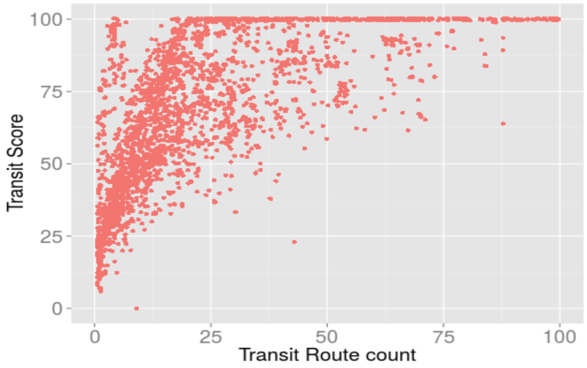


Figure 2.5: Jitter plot for number of routes vs transit scores.

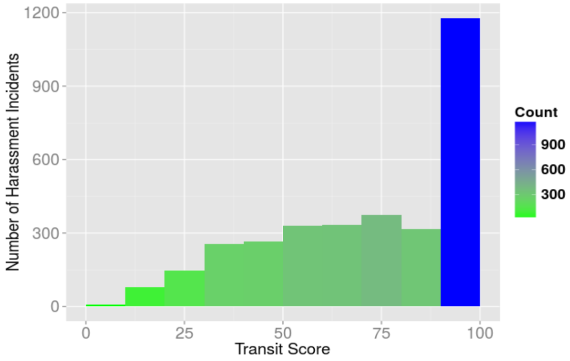


Figure 2.6: Histogram of $3,289$ incidents with respect to transit scores.

in general, the better a place is served by public transportation as measured through the transit score metric, the higher the number of street harassment events. Further, locations with fewer transit route options suffer more from harassment.

Based on these observations, we divide Figure 2.5 into four quadrants. The upper left quadrant, with high transit scores and a low number of transit routes, can be considered the most dangerous zone for street harassment. The upper right and the lower left quadrants have lower probabilities of harassment as they have either high transit scores or low route count. The lower right quadrant is considered a safe zone as the probability of experiencing harassment is very low.

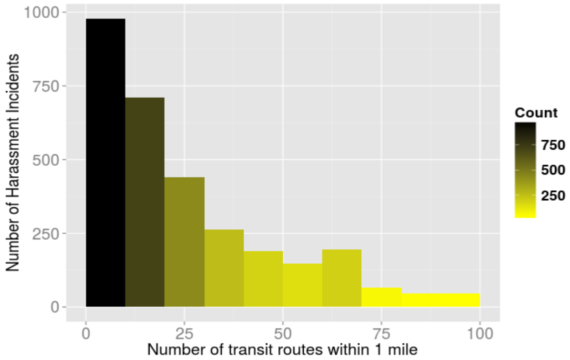


Figure 2.7: Histogram of $3,289$ harassment incidents based on local transit route availability.

2.4 Conclusions

In this chapter we sought to understand some of the urban characteristics of street harassment incidents. Our analysis shows that street harassment is more common in highly walkable areas with high transit scores and fewer nearby transit routes. On one hand, walkable streets should encourage people to walk more, but on the other hand the sexual harassment rate increases in these areas. While street harassment is considered a crime by law in some countries, other countries have laws that are more tolerant to this behavior. For the countries that criminalize street harassment, the results presented in this chapter can be utilized for better targeting of law enforcement. In all cases, we hope that tracking and analyzing street harassment datasets both spatially and temporally can lead to safer route planning for pedestrians.

2.5 Acknowledgments

We would like to thank Hollaback for sharing their collected dataset and taking the time to answer questions. We would also like to thank Morgan Vigil-Hayes for her help in obtaining access to the data.

Chapter 3

Spatially-blind Participatory Crowdsensing under Budget Constraints

The ubiquity of sensors has introduced a variety of new opportunities for data collection. In this chapter, we attempt to answer the question: Given M workers in a spatial environment and N probing resources, where $N < M$, which N workers should be queried to answer a specific question? To solve this research question, we propose two querying algorithms: one that exploits worker feedback (DispNN) and one that does not rely on worker feedback (DispMax). We evaluate DispNN and DispMax algorithms on two different event distributions: clustered and complete spatial randomness. We then apply the algorithms to a dataset of actual street harassment events provided by Hollaback. The proposed algorithms outperform a random selection approach by up to 30%, a random selection approach with feedback by up to 35%, a greedy heuristic by up to 5x times, and cover up to a median of 96% of the incidents.

3.1 Introduction

The ubiquity of mobile phones and sensors has brought participatory sensing into daily life. Participatory sensing can be defined as "*the process whereby individuals and communities use ever-more-capable mobile phones and cloud services to collect and analyze systematic data for use in discovery*" [30]. In this scenario, data can be continuously collected by leveraging user mobility and phone sensors across a range of applications including traffic monitoring [31], environmental sensors [32] and street safety [33]. Spatial crowdsourcing (SC) [34] provides a framework for the previously mentioned data collection applications where data requesters can create tasks in geographic areas of interest and workers are assigned or voluntarily choose to complete these tasks based on their spatial location. To fulfill an optimization function, such as minimizing distance traveled by workers [35], ensuring data quality [36], or maximizing task assignment [34], the task requester must accurately geolocate the location of task execution by providing geographic coordinates in the request to the SC server. But what happens when the requester is interested in sensing a geographic region, instead of a specific location, because the location of one or more events of interest is not precisely known? One solution would be to use the SC framework by modifying the request sent to the SC server to include a geographic region instead of the precise location. The SC server would then query all workers in this geographic area about the phenomenon of interest. While this solution is viable, it is impractical when it comes to a either a large-scale geographic region e.g., a city, or when the geographic region contains too many individuals to be reasonably queried. A budget constraint is a vital factor to consider in order to (i) save energy for resource constrained systems, e.g., disaster [37] and safety applications, because in an emergency, communication networks tend to fail and resources, such as bandwidth, are scarce [38]; and (ii) prevent users from becoming overwhelmed by queries and reaching a

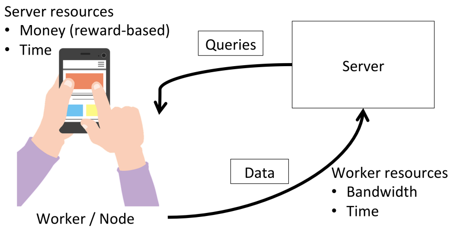


Figure 3.1: Data flow between servers and workers and the resources associated with each endpoint.

point where they cease using the crowdsensing system. Figure 3.1 depicts the resources associated with the SC server and workers which could impose budget constraints on the SC problem. In our work, we address what we term the "spatially-blind participatory crowd sensing" problem. In this problem, the SC task requester is not able to specify a precise location for a task but instead only a larger geographic region due to a lack of geographically tied distribution information about the phenomenon. In particular, our goal is to answer the following research question: *given the real-time interest of an SC requester in a specific geographic region, and a specific phenomenon of an unknown spatial distribution, who are the workers the SC server should query given a budget constraint of selecting N out of M crowd workers, where $N < M$, to maximize the probability of coverage for the phenomenon?* To answer this question, this chapter contributes the following:

(i) We define the problem of spatially-blind crowdsensing under budget constraints. To the best of our knowledge, we are the first to study this problem.

(ii) We define two types of queries under the setting of spatially-blind crowdsensing: binary and exploratory queries.

(iii) We propose two novel algorithms, one that does and one that does not rely on worker feedback (DispNN and DispMax, respectively), to select N out of M workers based on their locations, where $N < M$. We compare our algorithms to random selection and a greedy heuristic [39]. We study the performance of our proposed algorithms under two event distributions: clustered and complete spatial randomness. Our algorithms outperform random user selection by up to 30% and the greedy heuristic by up to 5x more detected incidents. We then test the algorithms on a real dataset of harassment reports in two cities and show the applicability of DispNN and DispMax in detecting incidents and locating workers close to these incidents without any prior knowledge of the incident distribution. Although we discuss the spatially-blind participatory crowdsending under budget constraints problem under the umbrella of crowdsensing, our work could be extended to other communities of artificial sensors, mobile phones or robotic sensors.

3.2 Related Work and Motivation

Since the introduction of "crowdsourcing" as a modern business term [40], a significant body of work has been dedicated to the study and implementation of crowdsourcing in real life applications. Spatial crowdsourcing (SC), where the information sought is bound to a particular geographic area, has received significant attention [34, 41, 42]. SC problems are split into two categories: problems where servers assign tasks to workers (SAT) and problems where workers select tasks (WST). Each of these two types of problems can be split further based on the worker model used for the problem; reward-based problems and self-incentivized problems. DispNN and DispMax provide a task assignment solution for reward-based SAT problems that seek to generate information about some environment (e.g., neighborhood, city, park, concert) with high coverage of the environmental area.

Beyond our contributions to the general area of SC using reward-based SAT, there is a specific SC problem that we seek to address: event-detection. Kazemi and Shahabi formally propose the *maximum task assignment* (MTA) problem as well as several solutions [34]. While solutions to the MTA problem seek to optimize task assignment given a number of spatially known tasks and workers at a specific time interval, they still require *a priori* information about the location of events and do not incorporate a notion of resource budgeting. Most similar to our work are [43, 44, 45]. In [43, 44], the goal is to maximize the system utility through a focus on task allocation under sensing capability constraints. In contrast, our goal is to maximize spatial situational awareness. In [45], To *et al.* introduce adaptive budget algorithms used to perform real-time task assignment in hyperlocal SC under budget constraints. However, the algorithms introduced require *real-time* information about the location of events of interest. In contrast, we seek to enable detection of events for which hyperlocal spatial information is not previously known. Our solution is particularly important for gathering information about small-scale, ephemeral social events.

As cities become smarter and cyberphysical systems become increasingly pervasive, there is an increasing need for SC platforms that are designed to flexibly collect quality data using methodologies that adapt to the dynamic intersections of human behavior and complex systems. One of the most critical aspects to designing city-scale SC platforms is resource scalability. To leverage the crowd for location-based data collection at a large scale, spatial crowd-sourcing platforms must be able to minimize resource consumption to harvest high quality data. For a SC task, resources may include network bandwidth, energy, user attention, time, and money. In particular, our work focuses on information queries that are best answered via human interpretations of the environment (e.g., *“Are you feeling too cold, too hot, or comfortable right now?”* vs. ‘‘*What is the temperature outside today?”* or *“On a scale from 1-10 how safe do you feel right now?”* vs. *“Is your bus stop well lit?”*).

3.3 Preliminaries

In this section, we introduce relevant definitions and offer examples of motivating queries. An **incident** is a real-time event or phenomenon that occurs at a particular location. An incident is tied with the specific geographic region around it; any worker in this region is able to sense or detect the incident. We model this region as a circular geographic space centered around the incident location with a specific radius. The larger the radius of the incident, the higher the probability that workers will be able to detect it. For example, the effect of a hurricane can be sensed over an entire city; however a street harassment incident can only be sensed if the worker is within a few meters. In the problem of "spatially-blind participatory sensing," the location of incidents is not known to the SC server or the requester. It is therefore vital to design a smart algorithm that tries to capture as many incidents as possible in the spatial area of interest. More formally, an incident i of form $< id, l, r >$ is an incident at location l and can be detected by all workers within a circular space centered at l with radius r. A **worker** is a person or device, i.e. a sensor or node, who can sense an incident in their vicinity. Formally, a worker w, of form $< id, l >$, is a mobile device carrier, or the device itself, who is a subscriber of the crowdsensing application and can report an incident of interest, in their geographic vicinity, to the SC server in real-time. A **real-time information query** is a query sent by the SC server to workers in a spatial region to inquire about one phenomenon of interest in real-time. We envision two types of queries. First, a *binary query*, which requires a yes/no response. As an example, a binary query could be *"Is your location affected by the hurricane?"* or *"Do you feel safe in your location?"*. This query is beneficial to obtain a high-level understanding of the spatial occurrence of the phenomenon of interest. A second type of query, an *exploratory query*, seeks to understand incidents at a more fine-grained level. The objective of this query is to eventually draw an approximate heat map of the phenomenon for the spatial region. Examples of this query include '*'On a scale from 1-10, how safe do you feel right now?"* and *"Is your location highly-walkable, somewhat walkable, or car-dependent?"* Finally, a **spatially blind worker selection algorithm under budget constraints** is an algorithm that runs on the SC server that aims to select workers under a specific budget of N out of M total workers without any prior knowledge of the incident spatial distribution. Since the algorithm is spatially blind to the incident spatial distribution, we cannot model the worker selection as a Maximum Task Coverage problem which is known to be strongly NP-hard [39]. Instead we have to devise a method of worker selection to maximize the spatially unpredicted incident coverage.

3.4 Problem Statement and Measures

In our system, we have a two-dimensional geographic region and a number of online workers (M) that can sense the environment around them. We investigate how to distribute queries within predefined geographic regions in the case of limited resources. To meet this constraint, we bound the system by a specific number of probes per time slot. Hence, the question becomes: *Given M workers and N resources, where $N < M$, which N workers should be queried to sufficiently answer a spatially-constrained query?* Consider the use case depicted in Figure 3.2, where we have 21 workers ($M = 21$) around the area of Central Park. The SC requester is interested in assessing harassment levels in the park but is constrained by a budget of querying only 5 workers ($N = 5$), for every real-time request. In other words, how should the SC server select these N workers?

Spatially-blind participatory crowdsensing under budget constraints.

If we tackle this question from a probabilistic point of view, then the straightforward answer is to try to select workers with the same spatial distribution as the phenomenon

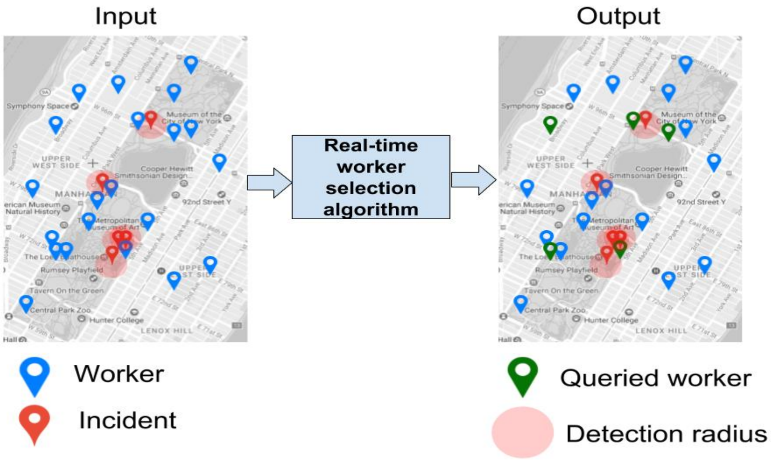


Figure 3.2: A use case for spatially-blind participatory crowdsending under budget constraints (M = 21, N = 5, 2 out of 5 harassment incidents are detected).

in the geographic region. For instance, if we know that a certain phenomenon occurs uniformly in the region, then we would have no bias in selecting the workers to query, i.e. each worker should have the same probability of selection. On the other hand, if we know the phenomenon is more prevalent in certain areas of the region, we should incorporate information when selecting the workers such that more workers are queried in the area of interest, where the phenomenon is likely to occur, and fewer workers in areas where there is a smaller probability of occurrence. The question becomes far more challenging if the distribution is not known or if it is not stationary. In this case, we ask if there is a systematic algorithm that can be used for selecting workers to spatially identify a phenomenon regardless of the probabilistic distribution or time variation.

Measures. To quantify the performance of the different approaches to solve the spatially-blind participatory crowdsensing under budget constraints problem, we propose the following three metrics for the output of the worker selection algorithm, which is the set of N workers that are queried (Queried Workers), denoted by QW. Let $QW = \{qw_1, qw_2, ..., qw_N\}$

- Coverage (COV): the number of incidents covered out of the total number of incidents that occur in the $2D$ geographic region. We define an incident as covered if the algorithm selects at least one worker in the range of the incident to be queried. Let the set of incidents that occur in the geographic region be $\{i_1, ..., i_{\mathcal{I}}\}$ and $Range(i_k)$ denote the set of workers in range of incident i_K, where a worker (w_j) is defined to be in the range of an incident if $dist(w(l)_j, i(l)) \leq i(r)$. Coverage is formally measured as:

$$COV = \sum_{k=1}^{\mathcal{I}} Coverage_k \tag{3.1}$$

where,

$$Coverage_k = \begin{cases} 1, & \text{if } (Range(i_k) \cap QW) \neq \phi \\ 0, & \text{otherwise} \end{cases}$$

- Close worker count (CWC): the absolute number of workers in the range of each incident for all incidents:

$$CWC = \sum_{k=1}^{\mathcal{I}} |(Range(i_k) \cap QW)| \tag{3.2}$$

- Redundancy (RED): the average share of workers per covered incident, defined as:

$$RED = CWC/COV \tag{3.3}$$

3.5 Algorithms and methodology

We assume that there are M online workers in a two-dimensional geographic area. The server that selects workers to query is bounded by N resources, where N and the geographic region are pre-determined by the SC requester. Each of the M workers has a specific location in the spatial area, determined by a two-dimensional system, e.g. (x, y) or a $(latitude, longitude)$. We assume that the selected workers will respond to the query. If needed, a pre-selection phase can be used to eliminate workers that are not likely to co-operate, such as requiring the installation of an app to facilitate querying. The focus of the worker selection mechanism is how to select N out of M nodes, where $N < M$, to maximize incident detection.

DispNN and DispMax algorithms. Suppose a requester wishes to identify unsafe areas in a geographic area (G) using only N worker probes. The requester provides the server with the following information: $< G,\ Q,\ N,\ ANS >$, where Q is the query related to the phenomenon of interest and ANS is the answer to the query for which the server will probe further, e.g., $ANS = No$ for $Q =$ "$Is\ it\ safe\ around\ you?''$. Since the SC server is spatially-blind with respect to the incident distribution, we can envision a solution that tries maximize the spatial variation of N worker locations so that the geographic area is covered. One measure of the degree to which points in a point set are separated from each other is spatial dispersion [46] measured as $tr(cov(P))$, where tr and cov denote the trace and covariance operations. Here, the point set is represented as a matrix P where each row represents a point p. Hence, the crowdsensing problem could be modeled as maximizing the spatial dispersion for the N workers i.e., selecting a set of N workers, $\mathcal{QW} = \{qw(j),\ j \in \{1, ..., N\}\}$, such that $\underset{\mathcal{QW}}{\operatorname{argmax}}\ tr(cov(QW(l)))$ where $QW(l)$ represents the matrix of the locations of the queried workers as follows:

$$QW(l) = \begin{bmatrix} qw_1(l) \\ \vdots \\ qw_N(l) \end{bmatrix}$$

In order to ensure a globally optimal solution, we can compute the dispersion of all $\binom{M}{N}$ worker location combinations and choose the combination with the maximum spatial dispersion as the set of queried workers. Solving $\underset{\mathcal{QW}}{\operatorname{argmax}}\ tr(cov(QW(l)))$ by generating all possible worker location combinations is of a complexity exponential in N. More generally for a fixed N, this yields a complexity of $O(M!/N!(M-N)!)$ which could become unrealistic for real-time applications as M and N increase. Instead, we propose to use Lloyd's K-means clustering algorithm [47], which tries to place the centers of the clusters as far away from each other as possible. We can then apply Lloyd's algorithm by computing the N-means clusters and choosing the workers with the closest locations to the centroid of each of the N clusters as a way of **maximizing the dispersion** of the N workers. Using Lloyd's algorithm yields a complexity of $O(MN)$, assuming constancy of point dimensions and number of iterations needed until convergence [47]. This method represents the core of DispMax and the first stage of DispNN. Another concept that can be applied to this problem is Thompson sampling, which is a heuristic for choosing actions that address the **exploration-exploitation** dilemma in the multi-armed bandit problem [48]. In our problem, we can design an algorithm that combines the concepts of exploration and exploitation. We define exploration as the process of maximizing the dispersion of worker location so that we can explore the geographic region. On the other hand, the concept of exploitation relates to making use of worker feedback about the incidents in the selection of other workers. For instance, using exploitation, if a worker (w_5) indicates that it is not safe around them by answering "No" to the query "Is it safe around you?", the server could exploit that answer and dedicate a subset of the N probes to some workers close to w_5. Querying the neighboring nodes can provide the requester with information related to spatial correlations and can help the requester bound the region in which the phenomenon occurs. Based on this prior work, our algorithm, DispNN, selects N of M workers in a geographic region by dividing the

incident count: number of incidents distributed across the cells of the spatial matrix.
incident range: the radius of an incident where, if a worker is present within the radius, he/she will be able to detect the incident.
crowd count: the M workers from which N will be chosen to query, where $N < M$.
N*:* the number of workers the SC server is limited by to query.
first stage percentage (FSP): the percentage of workers of the N resources that will be selected to query in the Disp stage.

Table 3.1: Parameters used in experiments.

selection into two phases: (1) Disp: the dispersion maximization phase (Exploration) and (2) NN: the use of worker feedback to query the nearest neighbors (Exploitation). These two phases work under the total budget constraint N; a percentage (FSP) of N is dedicated to the Disp phase and the percentage $(1 - FSP)$ of N is used to query the nearest neighbors of workers of interest based on the initial query response. If there is not sufficient feedback to locate nearest neighbors, we use the remaining resources towards another round of exploration. A variation of DispNN would be to not rely on user feedback; in this case the algorithm will dedicate all N probes towards the first phase, Disp. We call this algorithm DispMax.

3.6 Experiments and results

Experiment setup. Real world phenomenon rarely follow complete spatial randomness [49]. Hence, we study the performance of DispNN and DispMax under three different event distributions: clustered, random, and real-world datasets. There are multiple variables that can be controlled to test the behavior of DispNN and DispMax. Table 3.1 summarizes the most important experimental parameters. In all of our experiments, except the case study on real-world data, we use a 10x10 spatial grid and the Euclidean distance to measure the straight line distance between locations. Since it is unrealistic to assume that workers are uniformly distributed across the spatial area, we model the worker location distribution as a mixture of a Poisson point process [49] with $\lambda = \dfrac{crowd\ count}{2}$ and a cluster process where the other half of the crowd is distributed across a number of clusters that varies between [1, 10] and is chosen randomly. We compare our algorithms for worker selection to three alternative approaches and one optimal approach as follows:

- Random worker selection (Rand): we select N workers randomly based on a uniform distribution, i.e., each worker has the same probability of being selected.
- Greedy worker selection (Greedy): we apply the greedy heuristic proposed in [39] to solve the Maximum Task Coverage problem. At each iteration, the heuristic selects the worker that covers the maximum number of uncovered tasks; however, because the incident distribution is not known, we modify the heuristic. We choose the worker that *is likely* to cover part of the geographic space that is not covered. We start by selecting a worker randomly and then iterating through the rest of the workers and select the worker that will maximize the spatial dispersion. We continue iterating until we have N workers.
- Random with feedback worker selection (Randf): we use random worker selection in the first phase then apply the feedback process similar to the DispNN methodology.
- Optimal coverage (OptCov): we assume full knowledge of incident and worker locations and select N worker locations that maximize the number of incidents covered. We use this as a reference for the maximum coverage obtained if the server was aware of the incident distribution.

Generic observations. There are many variables that can affect the output of the experiments. For instance, we found that as the range of incidents increases, all approaches tend towards the same performance. The same observation is true as N approaches M.

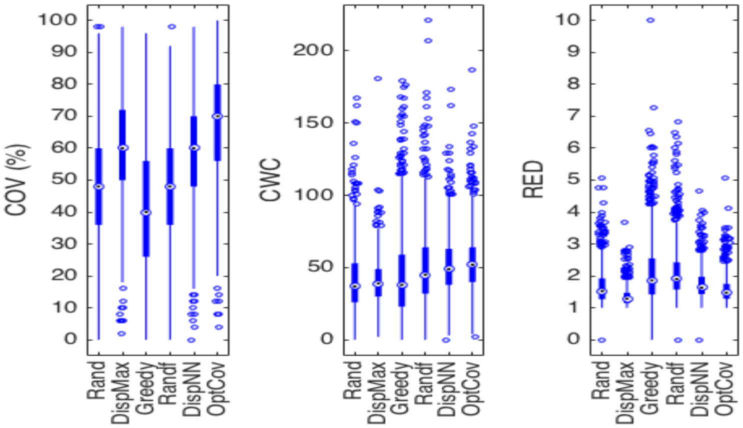


Figure 3.3: COV, CWC, and RED for distributions of clustered incidents.

As a result, we stress the different approaches by modeling incidents with smaller ranges. An interesting trade-off for DispNN is related to the choice of FSP. As FSP increases, COV increases but CWC tends to decrease and vice versa, for clustered distributions. We find that a good choice for FSP, that strikes a balance between COV and CWC, is 0.8, i.e., 80% of the probes allocated to the Disp phase and 20% for the NN phase.

3.6.1 Clustered incident experiments

Geographer Waldo R. Tobler stated in the first law of Geography: "Everything is related to everything else, but near things are more related than distant things" [50]. In this set of experiments, we assume that the incidents are related to each other, i.e. they form clusters across the $2D$ spatial region. Our goal in these experiments is to study the performance of the different query algorithms when the incidents are clustered. We vary the number of clusters in our $2D$ spatial area from one to ten while fixing the incident count to be 50 with a range of 1 unit distance. We set crowd count to 60 and $N = 30$. To enforce data variability, we model the size of each cluster as a random variable while ensuring that the aggregated size of all the clusters is equal to crowd count. For each number of clusters, we average results over 100 different random configurations.

Figure 3.3 illustrates COV, CWC and RED aggregated over all random configurations of clustered incidents. DispNN and DispMax outperform Rand, Greedy and Randf. DispNN and DispMax achieve a median coverage of 60% while the median OptCov is 70%. Rand ($\mu = 47.8, \sigma = 18.4$) and Randf ($\mu = 47.7, \sigma = 17.9$) provide a median coverage of 48%, while Greedy ($\mu = 41.9, \sigma = 20$) results in a median of 40% coverage. With respect to coverage, DispNN outperforms Rand, Greedy, and Randf by an average of 22.5%, 39.8% and 22.7%, respectively, and it comes within 13.3% of OptCov. Similarly, DispMax outperforms Rand, Greedy, and Randf by an average of 23.8%, 41.3%, and 24.1%, respectively, and comes within 12.4% of OptCov. Randf and DispNN achieve a higher CWC than Rand and DispMax since they rely on worker feedback; their NN selection phase selects workers that uncover other incidents because of the clustered nature of the incidents. DispMax achieves the lowest median RED of 1.3, since it maximizes the location dispersion of workers without relying on any feedback.

In the next set of experiments, the probability of occurrence of an incident is uniform across the spatial region. Incident occurrence in the spatial area follows a Poisson point process with $\mu = \lambda = incident\_count$. We randomly generate 100 different spatial region incident configurations. On average, the spatial matrix contains *incident_count* incidents. We operate under the same settings where $incident\_count = 50$ with a range of 1 unit distance and $M = 60$, and $N = 30$.

3.6.2 Complete spatial randomness experiments

Figure 3.4 shows that DispNN outperforms Rand, Greedy, and Randf in terms of coverage by 18.4%, 62%, and 26.2%, respectively, and comes within 11.7% of OptCov. Similarly, DispMax outperforms Rand, Greedy, and Randf by $26.9\%, 73.6\%, 35.2\%$, respec-

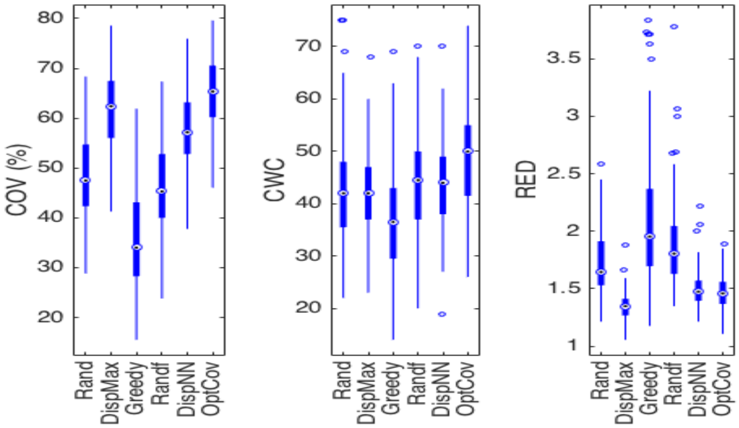


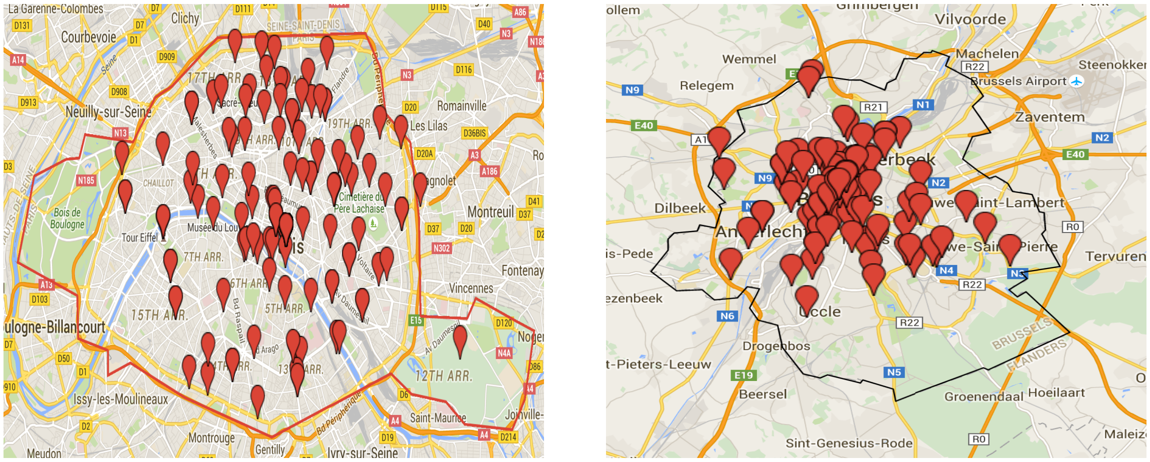
Figure 3.4: COV, CWC, and RED for incidents that follow complete spatial randomness.

tively, and comes within 5.4% of OptCov. We note that DispMax consistently performs closer to OptCov than DispNN. Because of the random distribution of incidents, there are no spatial correlations, unlike in the previous clustered distribution. Hence, there are fewer workers for DispNN to exploit in the NN phase. For the same reason, Randf performs slightly worse than Rand in terms of coverage. Apart from OptCov, DispNN and DispMax achieve the lowest RED since they focus on maximizing the dispersion. The result is higher incident coverage, on average, with workers more geographically dispersed.

3.6.3 Case study: Hollaback harassment dataset

After applying DispNN and DispMax to the previous two distributions, we wish to examine the algorithms under real incident distributions. To do so, we test our algorithm on a global street harassment dataset provided by Hollaback [16].

Data overview. We leverage the GPS coordinates from the street harassment dataset recorded by Hollaback in different cities. We refer the reader to Chapter 2 for an overview of Hollaback efforts. As of January 2016, over 8000 street harassment incidents have been



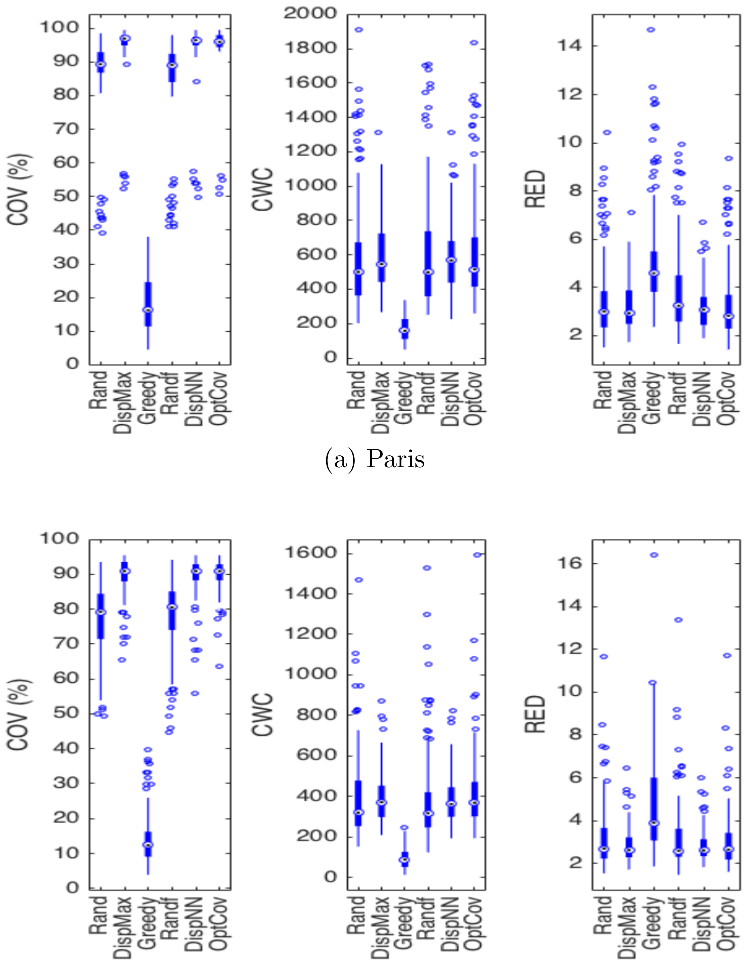
(a) Paris (b) Brussels

Figure 3.5: Distribution of harassment incidents across representative city datasets.

recorded in the dataset since February 2011. It is on this data set that we test DispNN and DispMax.

Analysis. From the Hollaback dataset, we select two cities (Paris, and Brussels) for which we have enough harassment samples for statistical significance (i.e. more than 30 samples). We test the performance of the six querying approaches on these cities. As a first step, we parse the Hollaback dataset such that incident reports are grouped by city. To do so, we use bounding box coordinates and shape files for each city to determine incidents bounded by the city borders and we remove any outliers. Figure 3.5 shows the resulting distribution of events for the two cities. The Paris dataset contains 197 harassment incidents and covers an area of 28.2 mi^2, while the Brussels dataset contains 154 incidents covering a geographic area of 28.4 mi^2.

For each of the cities, we generate 100 different variations of crowd locations ($M = 1000$) and set $N = 500$. In this analysis, *incident_count* is taken directly from the Hollaback dataset. We update the distance metric and use the Haversine formula to



(a) Paris

(b) Brussels

Figure 3.6: COV, CWC, and RED for Paris, and Brussels.

calculate the great-circle distance between two points as follows:

$$d = 2R * atan2(\sqrt{a}, \sqrt{1-a}) \tag{3.4}$$

where a is calculated as $\sin^2((\Delta\phi)/2) + \cos(\phi_1)\cos(\phi_2) * \sin^2((\Delta\lambda)/2)$; $\Delta\phi$ and $\Delta\lambda$ are calculated as the radian difference between the latitudes and longitudes, respectively; and R is the Earth's radius. Since a harassment incident cannot be witnessed unless a worker is very close, we adjust the incident range to 5 meters. We measure COV, CWC and RED aggregated over all random configurations of worker distributions for all six querying approaches and plot the results in Figure 3.6. DispNN and DispMax achieve close to optimal coverage in the case of Paris and Brussels. The median coverage using DispNN and DispMax for Paris and Brussels was found to be 96.4 and 90.1, respectively. We note that Greedy performs poorly for all cities. The reason is that at each step, Greedy chooses the point that maximizes the dispersion. The result is it selects the majority of the workers around the borders of the geographic region where the number of harassment incidents are minimal.

3.7 Conclusion

This chapter proposes DispNN and DispMax, spatial querying algorithms that select workers to discover randomly placed events within a $2D$ spatial environment through intelligent probing of worker resources. While the experimental evaluation confirms the applicability of proposed approaches, the algorithms could be adjusted to accommodate prior information about the nature of the events. If an approximate spatial distribution is known, we can use weights to reflect the probability of occurrence in each spatial sub-region and then apply DispNN and DispMax on each of the sub-regions. On the other hand, knowledge of spatial correlations and event stationarity could be used to manipulate worker selection. Our work is applicable in numerous scenarios, particularly when resource preservation is important and when querying all nodes will cause too large a disturbance or a response implosion.

3.8 Acknowledgments

The content of this chapter is the result of a collaboration with Morgan Vigil-Hayes Hayes and Ramya Raghevendra and has previously appeared in ACM CrowdSenSys 2017.

Chapter 4

Understanding Gender-based Violence in Social Media

Gender-based violence (GBV) is a global epidemic that is powered, in part, by a culture of silence and denial of the seriousness of its repercussions. In this chapter, we present one of the first investigations of GBV in social media. Considering Twitter as an open pervasive platform that provides means for open discourse and community engagement, we study user engagement with GBV related posts, and age and gender dynamics of users who post GBV content. We also study the specific language nuances of GBV-related posts. We find evidence for increased engagement with GBV-related tweets in comparison to other non-GBV tweets. Our hashtag-based topical analysis shows that users engage online in commentary and discussion about political, social movement-based, and common-place GBV incidents. Finally, with the rise of public figures encouraging women to speak up, we observe a unique blended experience of non-anonymous self-reported assault stories and an online community of support around victims of GBV. We discuss the role of social media and online anti-GBV campaigns in enabling an open conversation about GBV topics and how these conversations provide a lens into a socially complex and vulnerable issue like GBV.

4.1 Introduction

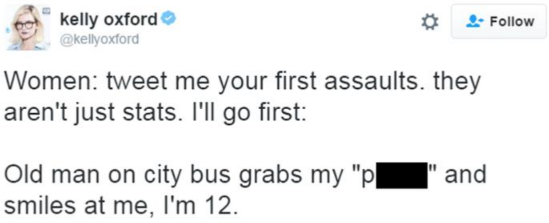
Gender-based violence (GBV) is one of the most prevalent human rights violations in the world. GBV is commonly defined as "*any form of violence that results in, or is likely to result in, physical, sexual or psychological harm or suffering to women, including threats of such acts, coercion or arbitrary deprivations of liberty, whether occurring in public or in private life*" [51]. According to the United Nations Population Fund (UNFPA), worldwide, one in three women will experience physical or sexual abuse in her lifetime [52]. Collected data reveal that GBV is pervasive across all social, economic and national strata [53, 54].

Vital to the design of social and economic policies that target GBV at its roots is the availability of data. The analysis of GBV through data is not only crucial to understanding GBV patterns, it is critical to measuring community-wide engagement, public opinion, and expression sensing as well as designing data-driven policies for raising awareness [55]. Despite the significant on-going effort into gathering GBV data [56], many specifics of GBV remain a grey zone due to a variety of reasons including victim blaming, and shamefulness [57], among others.

Within the past decade, social media has become a platform for social activism movements including *Black Lives Matter* (#blacklivesmatter) for racial equality and *Love Wins* (#lovewins) for marriage equality; the same can be said for GBV-based context. With 313 million active users, 1 billion monthly visits to sites with embedded tweets, and 79% of accounts outside the US, Twitter[1] is a pervasive open platform that facilitates a unique lens into GBV, both in terms of victims sharing their stories as well as the promotion of GBV, and subsequent reactions, both positive and negative.

We are driven by Twitter as an infrastructure for social activism to study characteris-

[1] `https://about.twitter.com/company`



kelly oxford
@kellyoxford

Women: tweet me your first assaults. they aren't just stats. I'll go first:

Old man on city bus grabs my "p[redacted]" and smiles at me, I'm 12.

Figure 4.1: Kelly Oxford invites women to share sexual assault stories on Twitter.

tics of online GBV. We are also inspired by recent events across the globe that led to the movement of GBV victims sharing their stories on the Twitter platform. For instance, as recently as October 2016, Canadian author and social media blogger Kelly Oxford started a conversation on Twitter encouraging women to share their first assault experiences, as shown in Figure 4.1. The response was overwhelming; she reported that she received 1 million tweets in one night with a minimum rate of 50 tweets per minute [58].

Through social media, we can thus study aspects related to self-reported stories, GBV news shares and user participation in the discussion. Our research seeks to understand user engagement with GBV posts, how users shape their GBV stories and the role of age and gender in online GBV contexts. To do so, we mine approximately 300,000 Twitter tweets, between April and November 2016.

Specifically, we seek to answer the following research questions using our datasets:

- *RQ 1:* What are the characteristics of user engagement with GBV stories?
- *RQ 2:* How do GBV tweet characteristics and content vary based on user demographics such as age and gender?
- *RQ 3:* How do authors present GBV stories?

Previous work has explored the use of misogynistic language in Twitter [59] and investigated the correlation between misogynistic content in Twitter [60] and the FBI

Uniform Crime Reports[2] for rape statistics in 2012. While the work in [61] and [62] examines GBV properties across geographic locations and anti-GBV campaigns in Twitter, respectively, online GBV computational studies through social media is still in its initial forms. Our work represents a first attempt to characterize user engagement, author story representation, and author demographics in the context of GBV in social media.

Our results show that social media is a key enabler for people to discuss GBV issues – this is apparent by the large number of self-reported stories and the sharing of news domains that host GBV-related stories. We also find, on average, higher engagement associated with GBV posts in comparison with generic tweets and that female participation is higher for ages less than 30 while male participation is higher for ages above 30. Finally, we show that GBV hashtags inspire self-expression and communal coping through sharing and support.

4.2 Background and Related Work

Our work is best understood in the realm of the following theories:

4.2.1 Social Movements

Looking at GBV as a global crisis, anti-GBV campaigns can be viewed as social movements to increase awareness against GBV and provide venues for people from different backgrounds to participate in the conversation. In 2016, the US White House's #StateOfWomen[3] summit deliberated violence against women under the umbrella of gender equality issues. UN Women aimed to increase the engagement of males through the

[2] https://ucr.fbi.gov/crime-in-the-u.s/2012/crime-in-the-u.s.-2012

[3] The United State of Women:
http://www.theunitedstateofwomen.org/

#HeforShe[4] campaign against inequalities faced by women. These anti-GBV campaigns, as well as others such as ItsOnUs[5], used hashtags on social media websites to spread the word globally. Use of social media by individuals and organizations to promote collective action and engagement is not new. Prior work that has extensively studied social movements in social media includes studies of the "Black Lives Matter" movement for racial equality [8, 7] and revolutions that helped shape the Arab Spring [63, 64]. While these studies focus on issues other than GBV, some of the research questions regarding users' topical engagement, demographics and attitude remain equivocal in the context of GBV.

Close to our work is the work of [61, 62]. Purohit et al. [61] used a key phrase based approach to gather GBV tweets over a period of 10 months and used a mixed methods approach [65] to focus on analyses such as volume, gender and language indicators. Our work differs in that we scrutinize user and tweet related key aspects such as common demographics, tweet visibility, and GBV story representations. The work in [62] examines the communities of three anti-GBV campaigns: #ItsOnUS, #StateOfWomen and #HeForShe and their community overlap. While we take these hashtags into consideration, our analysis complements this work by covering a broader set of hashtags. Our diverse hashtag set includes, among campaign related hashtags, ones that involve sharing personal experiences such as #NotOkay, #WhyIStayed and #BeenRapedNeverReported and others that aim to discuss and answer GBV related reality issues such as #WhyWomenDontReport, #MaybeHeDoesntHitYou, and #IBelieveSurvivors.

4.2.2 Influence

Social contacts in the physical world [66, 67] or in social media [68, 69] can have a strong influence on the attitude of individuals. An extensive body of literature has

[4] UN Women HeForShe campaign:
http://www.heforshe.org/en

[5] http://itsonus.org/

GBV category:	**Physical Violence (PhysViol)**	**Sexual Violence (SexViol)**	**Harmful Practices (HarmPrac)**
Key phrases	woman/women/girl/female beat up	sexual assault	child/children/underage/forced marriage
	woman/women/girl/female acid attack	sexual violence	sex/child/children trafficking
	woman/women/girl/female violence	woman/women/girl/female harass	woman/women/girl/female trafficking
	woman/women/girl/female punched	woman/women/girl/female attacked	*child molestation/bride/sex/*
	woman/women/girl/female attacked	boyfriend/boy-friend assault	*child violence/abuse/bullying/beat*
	gender/domestic violence	stalking woman/women/girl/female	*spouse abuse*
	intimate partner violence	groping woman/women/girl/female	*sex/women/forced slave*
	physical abuse/violence	*sexual/rape victim*	*female genital mutilation (fgm)*
		gang rape	*early marriage*
		victim blam	*pedophilia*
		sex predator	*human trafficking*
		woman/women/girl/female forced	*woman abuse*

Table 4.1: Key phrases used to identify GBV tweets. Newly identified key phrases are italicized.

#notokay: author Kelly Oxford invites women to share assault stories on Twitter (2016)

#whyistayed: users discuss their experience of domestic violence in the wake of the Ray Rice abuse incident (2014)

#yesallwomen: users share stories of misogyny and violence against women following the Isla Vista killings (2014)

#whywomendontreport: Vox correspondent Elizabeth Plank asked her Twitter followers why women do not report sexual assault (2016)

#beenrapedneverreported: Montgomery and Zerbisias co-created the hashtag to tweet support for the women who alleged they were assaulted by former CBC radio host Jian Ghomeshi (2014)

#ibelievesurvivors: brings up the issues around victim shaming and women reporting sexual assault allegations to police (2016)

#itsonus: hashtag associated with movement dedicated to changing the culture around campus sexual assault (2014)

#stateofwomen: hashtag associated with the White House summit discussing challenges that face women (2016)

#heforshe: UN's women campaign for gender equality aiming to engage men and boys as agents for change (2014)

#maybehedoesnthityou: writer and artist Zahira Kelly used Twitter to publicly share her emotional abuse experience (2016)

Table 4.2: Hashtags used in the context of GBV.

studied how social media's exposure can influence an individual's psychological states [70, 71, 72, 73]. Other work has explored the influence of content creation on social media attitude, such as retweeting, replying or favoriting, for example in the context of the Twitter platform. For instance, Levitt et al. [74] classified user's influence into two types: content-based and conversation-based. This work concluded that influential people such as celebrities were better at starting conversations on social media while news outlets content resulted in more retweets.

The ultimate form of influence is to promote collective action via social networks; this was visible in the Black Lives Matter (BLM) movement [8] and the Arab Spring [63]. On the theoretical end of studying influence and factors that promote users to endorse certain campaigns, points of view or products, lie the theories of Influence Maximization and Contagion. Influence Maximization is the problem of finding a set of nodes in a network that maximizes the spread of an idea or campaign. Greedy algorithms and heuristics that were proposed to solve this problem were studied in [75, 76, 77]. The Contagion theory aims to explain how ideas spread across human social networks. Granovetter et al. [78] explains that people will engage in a certain behavior by contagion if the number of people in the group who adopt that behavior exceeds a certain threshold. In the context of social networks, [79] found that political and idiom tags had a higher rate of contagion growth than other random topics on Twitter. Other work supported the contagion theory for petition virality [80] and showing support for same-sex marriage by overlaying profile pictures in Facebook [81]. To understand how to maximize GBV visibility, we explore how users engage in the context of GBV in Twitter. We examine both favorite rate and retweet rate of original GBV content on Twitter and compare these metrics for different forms of GBV as well as comparing GBV tweets with generic tweets.

4.3 Data and Methods

4.3.1 Social Media Data

We collected data from Twitter via two methodologies:

*(1) Key phrase-based dataset (GBV-KP-*1%*)*: For this dataset, we used Twitter's Streaming API to procure a 1% sample of Twitter's public stream. We then applied our own filtering process by using the key phrases in Table 4.1 to identify relevant GBV tweets. Specifically, we first used the key phrases identified by UNFPA domain experts in [61]. Purohit et al. analyze a dataset of 13.9 million tweets from Jan 1st to Oct 31st, 2014 in non-uniform time slices and differentiate between three categories of GBV: Physical Violence, Sexual Violence, and Harmful Practices. In our work, we adopt the same categorization scheme. Upon examining the results of our initial crawling attempt, we excluded a set of key phrases that resulted in irrelevant content. These key phrases contained keywords that were used colloquially in discourse and contexts that were extraneous to GBV. For the Physical Violence category, we excluded the key phrases containing the words dragged, kicked, beaten, and burn. For the Sexual Violence category, we excluded the word "rape" but replaced it with the more specific "rape victim" and "gang rape" key phrases. In the analysis conducted in [59], it has been shown that the word "rape" appeared in serious/news contexts 40% of the time and 60% in other types of discourse including casual and metaphor categories. Following a snowball approach and multiple crawling phases, we were able to identify 35 unique key phrases. Table 4.1 encompasses both UNFPA key phrases and our newly-identified key phrases.

(2) Hashtag-based datasets: For a more detailed study of recent events, we include two other datasets based on the 10 hashtags specified in Table 4.2. For the first dataset (*GBV-HT*-1%), we filtered the 6-month 1% sample of Twitter's public stream using these hashtags. Table 4.2 depicts the used hashtags, the initial incidents that sparked their

Dataset	Time Range	Tweets	Users	Content creators
GBV-KP-1%:	04/13/16-10/13/16			
PhysViol		34,380	31,085	8,574
SexViol		93,567	82,132	18,160
HarmPrac		108,822	92,499	21,925
GBV-HT-1%	04/13/16-10/13/16	6,454	5,999	1,602
GBV-HT-Comp	10/26/16-11/26/16	58,908	34,450	35,490
General-1%	10/26/16-11/26/16	33,055,294	11,394,125	2,572,617

Table 4.3: Descriptive statistics of GBV Twitter datasets.

creation, and the year they first appeared. For the hashtag #notokay, we only include tweets that also contain the mention @kellyoxford in order to exclude tweets that mention the hashtag but discuss issues other than GBV.

For a more comprehensive[6] hashtag-based dataset (*GBV-HT-Comp*), we use Twitter's public streaming API[7] to collect tweets from October 26th to November 26th, 2016 that contain the indicated hashtags. Because Twitter's Streaming API cannot be used to track certain hashtags, we specify the hashtags as keywords (e.g. notokay), then we apply a string matching approach to identify the # symbol followed by the hashtag string (e.g. #notokay).

To provide a larger context for interpretation within our experiments, we compare the *GBV-HT-Comp* dataset with a 1% sample of all tweets (including non-GBV tweets) using a 1-month dataset (*General*-1%) spanning the same time period (10/26/16 - 11/26/16). We filter all non-English tweets from our datasets. We also apply preprocessing to eliminate repeated tweets and tweets from authors with zero followers. Table 4.3 constitutes an overview of the time-span covered by each dataset, the number of tweets, and number of unique users and content creators.

[6] as opposed to the 1% sample

[7] Twitter's public Streaming API
`https://dev.twitter.com/streaming/public`

4.3.2 Measures

In our investigation, we adopt several measures based on prior work in order to answer the proposed research questions. For content sharing and engagement, we examine multiple metrics including favorite rate, retweet rate, and number of tweets containing links and media. To identify influential topics, we look at the prevalence of hashtags. Focusing on the actual nature of GBV stories and how authors represent GBV, we use the psycholinguistic lexicon LIWC [82] to measure interpersonal awareness, affect, and emotional expressions. In our analysis, we differentiate between the notions of *perceived vs actual user characteristics*. When we look at account characteristics of content creators or consumers, we study the perceived account characteristics (e.g. gender and age) that are visible in their account. Nilizadeh et al. [83] studied the association between perceived gender and measures of online visibility. Recent work that investigates the inference of actual user characteristics from online content in social networks, aka *user profiling*, include age, gender, and occupation estimation in [84, 85, 86, 87]. We specifically study user perceived age and gender using an automatic facial feature recognition service "Face++" [88].

4.4 Analysis

4.4.1 RQ1: User engagement with GBV tweets

To answer RQ 1, we begin by exploring the engagement of users with GBV content on Twitter. In particular, we examine metrics related to favoriting and re-sharing a tweet (retweeting). In Twitter, once a user favorites a tweet, that tweet is automatically archived in the user's profile for the user and their network to read later. Retweeting is the act of resharing content with followers of the user. Retweets do not necessarily indicate content endorsement but suggest content to be viewed by the retweeter's network. Retweets provide a powerful tool for tweets to be shared beyond the content creators' network of followers [89]. As a user's follower network grows, so does the visibility of their content on Twitter. To incorporate this effect, we normalize favorite and retweet counts by the size of a user's follower network. We, therefore, compute two metrics for each tweet, favorite rate and retweet rate, which are defined as follows:

$$\begin{aligned} Favorite\ rate\ (FR) &= \frac{Favorite\ count}{Followers\ count} \\ Retweet\ rate\ (RR) &= \frac{Retweet\ count}{Followers\ count} \end{aligned} \tag{4.1}$$

where favorite count and retweet count indicate how many times a tweet is favorited and retweeted, respectively. We note that favorite count and retweet count are a function of the tweet while the followers count depends on the user's network. The content captured in our datasets falls into one of three categories: original, retweet, and reply. If a retweet exists, this suggests that the retweet count for the original tweet reflects the resharing accordingly. For this analysis, we thus consider only original tweets in our datasets. Since the datasets used in our analyses were gathered using the Twitter streaming API at the time of their creation, the corresponding favorite and retweet counts associated with each tweet's body of information were zero-valued. In order to accurately capture the eventual favorite and retweet counts, we queried the Twitter API again at a later time[8] to allow user engagement with tweets. Table 4.4 depicts the number of original tweets investigated for each dataset, favorite count and rate, and retweet count and rate descriptive statistics.

We are particularly interested in exploring two questions. First, *do different types of GBV tweets exhibit different engagement patterns?* and second, *how does engagement*

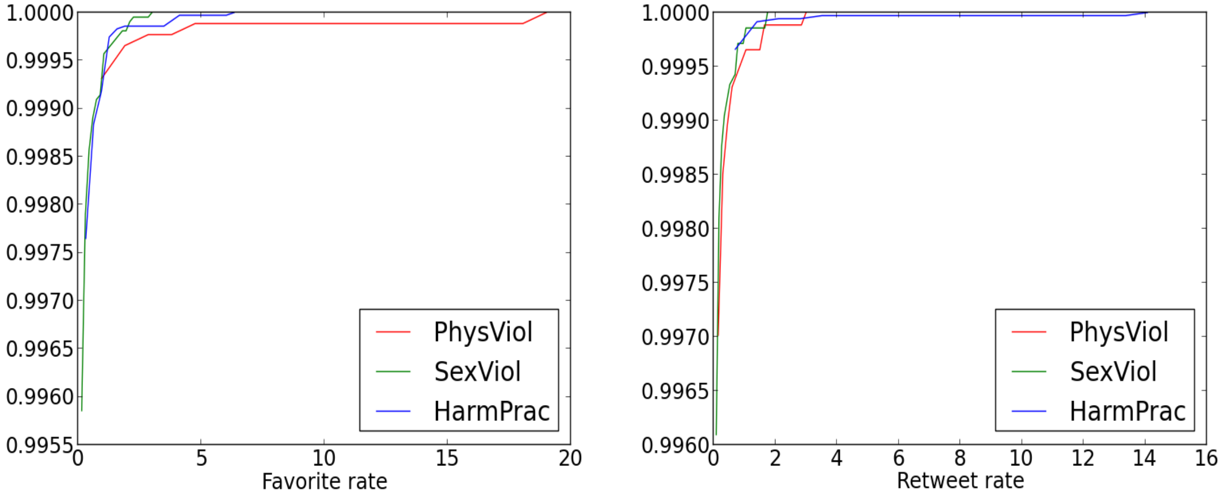
[8] in December 2016, resulting in a minimum of one month and a maximum of eight months of interaction

Dataset	Engagement stats
GBV-KP-1%: *PhysViol*	**Original tweets:** 8,711 **Favorite count:** Min = 0, Max = 1394, $\mu = 2.39$, $\sigma = 28.16$ **Favorite rate:** Min = 0, Max = 19, $\mu = 0.0067$, $\sigma = 0.21$ **Retweet count:** Min = 0, Max = 2282, $\mu = 2.14$, $\sigma = 31.5$ **Retweet rate:** Min = 0, Max = 3, $\mu = 0.0030$, $\sigma = 0.046$
SexViol	**Original tweets:** 20,999 **Favorite count:** Min = 0, Max = 4844, $\mu = 3.74$, $\sigma = 64.02$ **Favorite rate:** Min = 0, Max = 3, $\mu = 0.0042$, $\sigma = 0.0486$ **Retweet count:** Min = 0, Max = 2816, $\mu = 2.79$, $\sigma = 42.37$ **Retweet rate:** Min = 0, Max = 1.75 , $\mu = 0.0021$, $\sigma = 0.0302$
HarmPrac	**Original tweets:** 35,315 **Favorite count:** Min = 0, Max = 1497 , $\mu = 1.45$, $\sigma = 15.48$ **Favorite rate:** Min = 0, Max = 6.33, $\mu = 0.0043$, $\sigma = 0.0652$ **Retweet count:** Min = 0, Max = 1168, $\mu = 1.09$, $\sigma = 11.64$ **Retweet rate:** Min = 0, Max = 14.07, $\mu = 0.0022$, $\sigma = 0.08$
GBV-HT-Comp	**Original tweets:** 13,871 **Favorite count:** Min = 0, Max = 3447, $\mu = 6.16$, $\sigma = 59.65$ **Favorite rate:** Min = 0, Max = 21.75, $\mu = 0.0191$, $\sigma = 0.2466$ **Retweet count:** Min = 0, Max = 1351, $\mu = 2.81$, $\sigma = 23.44$ **Retweet rate:** Min = 0, Max = 17.75, $\mu = 0.0071$, $\sigma = 0.1581$
General-1%	**Original tweets:** 82,083 **Favorite count:** Min = 0, Max =29819, $\mu = 2.62$, $\sigma = 111.24$ **Favorite rate:** Min = 0, Max = 12, $\mu = 0.0056$, $\sigma = 0.0786$ **Retweet count:** Min = 0, Max = 7055, $\mu = 1.25$, $\sigma = 37.4$ **Retweet rate:** Min = 0, Max = 8.69 , $\mu = 0.0021$, $\sigma = 0.05432$

Table 4.4: Descriptive statistics for engagement with GBV posts.

with a GBV tweet differ from a generic non-GBV tweet?

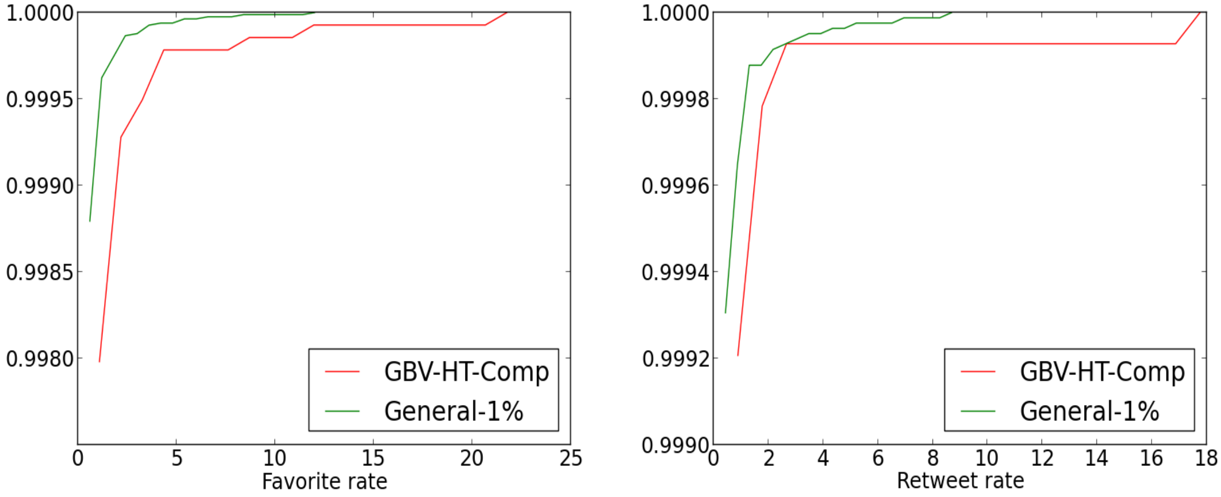
Engagement based on tweet GBV category. To answer the first question, we study PhysViol, SexViol, and HarmPrac categories in the *GBV-KP*-1% dataset. All three categories were collected over the same six-month duration, from April 13th to October 16th, 2016. We compute the Favorite rate and Retweet rate for the three categories of GBV tweets and plot the corresponding Cumulative Distribution Functions (CDF) in Figure 4.2. To determine whether there are significant differences between the three datasets, we used Kruskal-Wallis H test for the Favorite rate and the Retweet rate. The test statistic for the Favorite rate was $H = 73.8$ with p-value $< .001$ and for the Retweet rate was $H = 75.4$ with p-value $< .001$. On average, PhysViol tweets were favorited approximately 1.6× more than SexViol and HarmPrac tweets ($\mu_{FR-PhysViol} = 0.0067$ vs $\mu_{FR-SexViol} = 0.0042$ and $\mu_{FR-HarmPrac} = 0.0043$). We also noted that the probability of a tweet's favorite count extending beyond network size (i.e. $P(FR > 1)$) is larger for PhysViol tweets and approximately the same for SexViol and HarmPrac tweets. Following the same pattern, PhysViol tweets were retweeted on average 1.4× more than SexViol and HarmPrac tweets ($\mu_{RR-PhysViol} = 0.0030$ vs $\mu_{RR-SexViol} = 0.0021$ and $\mu_{RR-HarmPrac} = 0.0022$); $P(RR > 1)$ is larger for PhysViol tweets and approximately the same for SexViol and HarmPrac tweets.



(a)

(b)

Figure 4.2: Cumulative distribution associated with (a) Favorite rate and (b) Retweet rate for three categories of GBV.



(a) (b)

Figure 4.3: Cumulative distribution associated with (a) Favorite rate and (b) Retweet rate for GBV tweets versus General tweets.

Engagement with GBV tweets vs General tweets. To answer the second question, we study the *GBV-HT-Comp* dataset and compare it with a random sample of $82,083$ original tweets from the *General*-1% dataset from the same time period. We plot the CDF for Favorite rate and Retweet rate for both datasets in Figure 4.3. To determine if there are significant differences between the two distributions, we conduct the Wilcoxon-Mann-Whitney test for the Favorite rate and Retweet rate. The test statistic for the Favorite rate was $U = 43.7$ with p-value $< .001$ and for the Retweet rate was $U = 47.3$ with p-value $< .001$. On average, a GBV tweet was favorited 3.41× more than a General tweet ($\mu_{FR-GBV-HT-Comp} = 0.0191 > \mu_{FR-General-1\%} = 0.0056$). We also noted that the probability of a tweet's favorite count extending beyond network size (i.e. $P(FR > 1)$) is larger for GBV tweets than General tweets. A similar result was found for the Retweet rate ($\mu_{RR-GBV-HT-Comp} = 0.0071 > \mu_{FR-General-1\%} = 0.0021$) and $P(RR > 1)$ is larger for GBV tweets.

4.4.2 RQ2: Age and gender variables for users in the GBV context

We utilize descriptive statistical analysis to discover relationships among tweets, gender, and age collected from the Twitter REST API and the Face++ API. In this experiment, we combine the hashtag-based datasets, GBV-HT-1% and GBV-HT-Comp, into one dataset (HT) since the emphasis of the experiment is to identify demographic variables for users regardless of time span. For all types of tweets (original, reply, or a retweet), we identify Twitter user IDs associated with each tweet and query the Twitter REST API to extract the user's profile picture url. We then feed the picture's url to the Face++ API, which predicts the demographic information of a given photo (e.g. age, gender, and race). Upon compiling the demographic information of each user, Face++ returns a confidence level for its detection. We omit any results with a confidence level below 95% (21.7% of the total queries). This results in data for 9,837 users for PhysViol, 7,373 users for SexViol, 10,591 users for HarmPrac and 12,996 for HT.

We plot the age-gender distribution for the combined datasets PhysViol, SexViol, and HarmPrac in Figure 4.4. Figure 4.4 shows that highest participation is in the age range 20-29, followed by 30-39, and then 10-19. We list the percentages of female vs male participation across age ranges in Table 4.5. We note that female participation is dominant across age ranges ≤ 9, 10-19 and 20-29, and decreases as age increases, while male participation dominates above 30, increasing with age[9]. The same observations were consistent across individual datasets: PhysViol, SexViol, HarmPrac, and HT; we omit the results due to space limitations.

Dominant female participation in the range of [55.76%-100%] was also observed for

[9] We show the results for ages 0-9 despite the fact that the majority of the actual users are not likely to be in this age range. Upon investigation, we found this age range to include users that have cartoon pictures as profile pictures or photos of their children as their account profile photo.

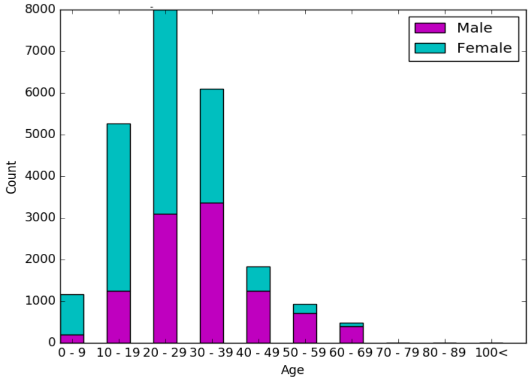


Figure 4.4: Breakdown of users by perceived age and gender for PhysViol, SexViol and HarmPrac datasets.

all the hashtags in Table 4.2. We study gender participation for different types of tweets (original, retweet and reply) across all datasets and note that female participation with original content is dominant across all datasets, ranging from [55%-68.21%], and for retweeting ranging from [56%-69.5%]. In the case of replies, male responses dominate in the HarmPrac dataset with 55% while females dominate in PhysViol, SexViol, and HT in the range [53.5%-64.2%]. Higher female participation was also noted in [62] in the context of anti-GBV activism.

Despite the previous results, there remains a need to provide a more comprehensive gender breakdown with respect to the specific context of a GBV tweet. For instance, do women provide more content focusing on raising awareness of GBV? Do women provide more content that reports GBV events on behalf of themselves or others? Are men and women equally likely to tweet support for GBV victims? Our future work will more deeply correlate content type with content creator demographics.

4.4.3 RQ3: GBV story representation on Twitter

In order to understand GBV story representation, we examine three different parameters: the use of embedded urls, topics of interest by looking at viral hashtags, and the

Age range	Female (%)	Male (%)
≤ 9	82.53	17.47
10-19	76.32	23.68
20-29	61.19	38.81
30-39	44.95	55.05
40-49	31.44	68.56
50-59	22.57	77.43
60-69	16.32	83.68

Table 4.5: Percentage of female and male participaton across age ranges in PhysViol, SexViol and HarmPrac.

common linguistic properties in tweets.

Shared content via url usage.

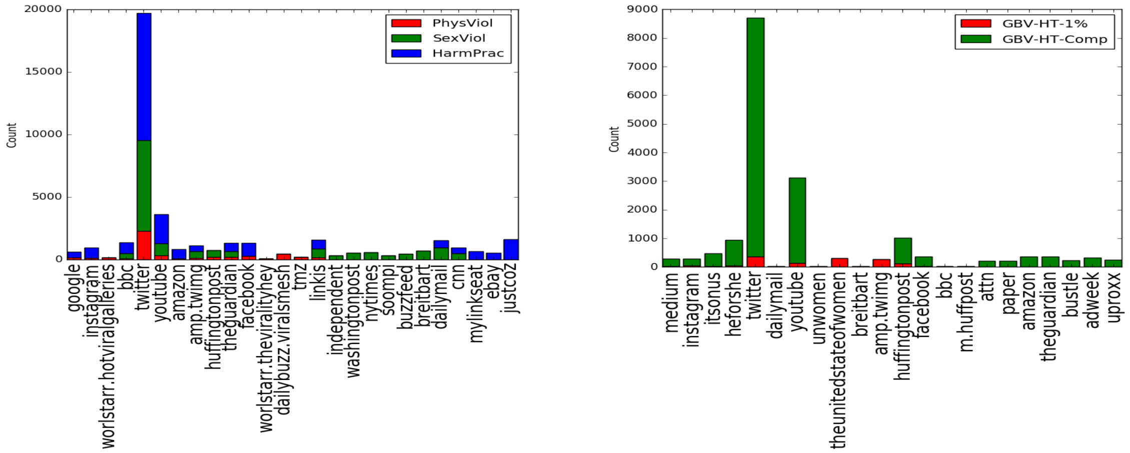
Since a tweet is bound to a maximum of 140 characters, Twitter users commonly embed urls that redirect readers to relevant content. To more deeply understand GBV tweets, we quantify the usage of urls and examine the top visited domains in our datasets. We parse the tweet text to extract urls and perform a GET request with as many redirections on each url as needed until the last destination is hit. Upon reaching the target url, we capture subdomains and domains. Table 4.6 depicts the percentage of tweets containing one or more embedded urls and the number of unique domains for each dataset. The dataset with the lowest percentage of tweets containing urls was GBV-HT-1% with a percentage of 34.75%; the dataset with the highest number of urls was HarmPrac with 58.21%. On average, 43.7% of tweets across all datasets contained one or more url.

Next, we examine the top 15 domains for each dataset. We group the results for PhysViol, SexViol and HarmPrac datasets in Figure 4.5(a) and the results for GBV-HT-1% and GBV-HT-Comp, since they cover the same set of hashtags, in Figure 4.5(b). Figure 4.5 shows a large presence of social media websites (e.g. Twitter, Instagram, Facebook, and Youtube). Upon inspection of the tweets, we found out that users often reference other GBV content, such as a status on Twitter, a Facebook post, a Youtube

Dataset	One or more url (%)	# Unique domains
GBV-KP-1%:		
PhysViol	41.04	3,247
SexViol	44.15	5,219
HarmPrac	58.21	12,491
GBV-HT-1%	34.75	385
GBV-HT-Comp	37.36	990

Table 4.6: Descriptive statistics of url usage in GBV tweets.

video or an Instagram picture. We also note the huge presence of news and blog websites that share full GBV stories. Examples of news domains include BBC, Independent, Washington Post, New York Post, CNN, Daily Mail, and The Huffington Post. Across the blog websites, the most frequently occurring were medium, bustle (offering online content for women and by women) and adweek. Since some hashtags were related to anti-GBV campaigns, domains referencing these initiatives, were also encountered e.g., heforshe.org, itsonus.org, and theunitedstateofwomen.org.



(a)

(b)

Figure 4.5: Distribution of top 15 urls used across (a) PhysViol, SexViol and HarmPrac datasets and (b) GBV-HT-1% and GBV-HT-Comp.

Relationship to on the ground realities.

To identify current on the ground topics related to GBV, we investigate the trending hashtags for each dataset which act as topical labels to their tweets. Table 4.7 depicts the top 10 hashtags for each dataset. We discern four types of hashtags: *social-movement*, *political*, *violence incidents*, and *generic* hashtags. Social-movement hashtags include #ghanaendsdomesticviolence, #youoksis, #mcug16, #internationalmensday, #heforshe and the hashtags #shiftyourperspective and #turnstwo which are specifically associated with #heforshe. The hashtag #ghanaendsdomesticviolence discusses the launch of the Government of Ghana's National Survey on Domestic Violence as a mean of advancing its gender equality agenda. The goal of the #youoksis movement is to inspire people of both genders to intervene with street harassment situations by engaging with the victim of said harassment. The hashtag #turnstwo celebrates the second anniversary of the launch of the HeForShe movement, while #shiftyourperspective is also associated with tweets asking males and boys to change their perspective as a part of the HeForShe campaign. Political hashtags are also observed in our datasets since the time duration of our datasets coincided with the 2016 US Presidential elections. These include the hashtags #trump, #tcot, #trumptapes, #hillary, #iamwithher, #trump2016, and #americafirst. These hashtags were typically used to talk about sexual assault allegations in the political context. Hashtags concerning violence include #ripamy, #justice4cindy, #terencecrutcher, #brockturner. These incidents cover a range of types of violence including physical violence, domestic abuse, and rape. Other hashtags encountered cover the broader scope of GBV by opening a discussion about #domesticviolence, #violenceagainstwomen, and #womenrights, among others.

GBV-KP-1%

- PhysViol: ripamy, maybehedoesnthityou, ghanaendsdomesticviolence, youoksis, violence, justice4cindy, terencecrutcher, violenceagainstwomen, domesticviolence, news
- SexViol: trump, tcot, trumptapes, brockturner, sexualassault, hillary, sexualviolence, imwithher, trump2016, news
- HarmPrac: child, endviolence, endfgm, endchildmarriage, sex, fgm, abuse, nsfw, childabuse, humantrafficking

GBV-HT-1%*:* americafirst, tcot, itsonus, notokay, whatweshare, heforshe, shiftyourperspective, turnstwo, rape, yesallwomen

GBV-HT-Comp: globalgoals, itsonus, notokay, womensrights, internationalmensday, heforshe, whywomendontreport, imwithher, mcug16, genderequality

Table 4.7: Top 10 hashtags for GBV datasets.

Linguistic properties for GBV tweets.

Next, we examine different language attributes associated with the set of hashtags under investigation. In particular we wish to *examine different interpersonal awareness and affect patterns of GBV hashtags.* As a preprocessing step, we remove retweet headers, screen names, and urls. We use the LIWC 2015 software [82] for our linguistic analysis. First, we measure interpersonal awareness based on **linguistic dimensions** including the frequency of usage of *1st person singular* (1st p. singular), *1st person plural* (1st p. plural), *2nd person* (2nd pp.) and *3rd person singular* (3rd p. singular) pronouns. We investigate **temporal references** based on the usage of past, present and future tenses. We consider two measures of affect: *positive affect* (PA), and *negative affect* (NA). Under the umbrella of NA, we examine three measures of emotional expression: *anxiety, anger*, and *sadness.* The average percentage of usage of linguistic pronoun dimensions and temporal references are depicted in Table 4.8 and the average corresponding affective attributes in Table 4.9. We note the following observations.

Observation 1: *GBV hashtags inspire both self-expression and communal attachment.*

Higher usage of 1st p. singular (e.g. I, me, mine) is associated with hashtags #notokay

Hashtag	1st p. singular	1st p. plural	2nd pp.	3rd p. singular	past tense	present tense	future tense
#beenrapedneverreported	0.64	0.68	0.27	0.19	2.94	3.23	0.14
#heforshe	0.94	3.16	2.19	0.4	1.08	9.41	0.40
#ibelievesurvivors	2.16	0.27	5.58	0.33	2.81	8.84	1.60
#itsonus	0.79	2.77	3.06	0.11	1.51	10.23	0.51
#maybehedoesnthityou	0.71	0.08	9.15	6.45	1.36	11.24	1.30
#notokay	5.42	5.11	3.85	0.31	9.53	9.28	0.21
#stateofwomen	1.3	4.3	1.42	0.17	0.89	10.38	1.67
#whyistayed	5.42	0.56	0.88	2.85	4.55	6.74	0.66
#whywomendontreport	2.36	0.84	1.4	1.16	2.66	10	0.56
#yesallwomen	1.79	0.52	1.18	0.44	1.57	7.96	0.45

Table 4.8: Average linguistic dimensions and temporal references percentages associated with GBV hashtags.

Hashtag	PA	NA:	anxiety	anger	sadness
#beenrapedneverreported	6.36	1.73	0	1.73	0
#heforshe	6.46	1.07	0.09	0.5	0.12
#ibelievesurvivors	3.48	5.06	0	2.4	1.6
#itsonus	3.25	4.13	0.09	3.09	0.12
#maybehedoesnthityou	2.63	6.74	0.88	2.68	1.14
#notokay	4.81	4.07	0.13	3.46	0.24
#stateofwomen	4.70	1.23	0.05	0.76	0.22
#whyistayed	3.24	5.69	1.4	2.55	0.43
#whywomendontreport	2.53	7.4	1.15	4.92	0.6
#yesallwomen	5.28	3.59	0.37	2.23	0.46

Table 4.9: Average affective attributes' percentages associated with GBV hashtags.

and #whyistayed. Moreover, #notokay, #whyistayed, and #beenrapedneverreported exhibit focus on past and present temporal forms. This indicates a recall of self-relevant information including current and previous GBV experiences. Examples include the following tweets:

"The first time I was harassed I was 5 yo and a boy looked up my dress and commented on my ass #notokay @kellyoxford"

"#WhyIStayed because he made me distance myself from everyone and he always told me If i left I would be alone..." Higher usage of 1st p. plural (e.g. we, us, our) is associated with hashtags #notokay, #stateofwomen and #heforshe. This indicates a sense of greater social awareness and support within the anti-GBV community. This is anticipated in the context of anti-GBV campaigns (State of Women and HeForShe) where individuals provide support for each other. On the other hand, #notokay provided a virtual space for both self-reported GBV incidents and mutual support. Examples include the following tweets:

"@kellyoxford I want to thank you for starting #notokay ...It is one of the reasons I had the courage to write this http://ndsmcobserver.com/2016/11/remembering-my-racist/"

"MT @FLOTUS Together, we are stronger. Together we can change tomorrow. Stand with us: http://www.theunitedstateofwomen.org #StateOfWomen @USWomen2016"

With the higher usage of 2nd pp. (e.g. you, your), the hashtags #maybehedoesnthityou and #ibelievesurvivors were primarily used to provide greater social awareness in the context of GBV. #Maybehedoesnthityou was used to bring attention to other forms of non-physical relationship abuse and #ibelievesurvivors was used to shed light on sexual assault victims speaking up but not being believed.

"#MaybeHeDoesntHitYou but he's isolated you from and turned you against everyone who you care about"

"When your role models fail you, become the role model you wish they were. #ubcac-

countable #ibelievesurvivors"

From a temporal perspective, we observe that the hashtags #itsonus, #heforshe, #stateofwomen, #whywomendontreport, #yesallwomen #ibelievesurvivors, and #maybehedoesnthityou focus more on present issues than past and future.

Observation 2: *Mixed positive and negative emotions present in anti-GBV posts.*

Hashtags with the highest PA include #heforshe, #beenrapedneverreported, #notokay, #stateofwomen. The tweets associated with #heforshe and #stateofwomen encourage men to take solidarity with women and the unity of women, respectively, hence the higher PA scores. Upon inspection of the #beenrapedneverreported tweets, we discover that the captured tweets in 2016 discuss the spread of GBV underreporting and urge others to spread the word; these tweets rarely contain self-reported stories. On the other hand, hashtags with highest NA include #whywomendontreport, #maybehedoesnthityou, #whyistayed and #ibelievesurvivors. Most interesting are hashtags that combine both higher levels of PA and NA at the same time. These include #ibelievesurvivors (PA = 3.48, NA = 5.06), #itsonus (PA = 3.25, NA = 4.13), #notokay (PA = 4.81, NA = 4.07), #whyistayed (PA = 3.24, NA = 5.69) and #yesallwomen (PA = 5.28, NA = 3.59). The tweets associated with these hashtags, in some cases, contain both PA and NA simultaneously as indicated in Table 4.9. In these tweets, users exhibit NA due to the nature of GBV reported issues but at the same time, they express optimism about either the notion of women speaking up and sharing their personal experiences or hope for a change in their partners or the overall GBV situation. An example is the tweet: *"RT: KellyOxford: I am in such horrendous shock and yet so proud of the women sharing their assaults. #notokay is trending in US. Not our shame anymore"*.

Observation 3: *Anger is more prevalent than anxiety and sadness across all GBV hashtags.*

Among the negative affect attributes, we examine anxiety, anger and sadness attributes as computed by the LIWC software. Hashtags with the highest score of anger included #whywomendontreport, #notokay, and #itsonus. We also observe that the average anger scores are greater than anxiety according to Wilcoxon-Mann-Whitney test (U = 6.0 and p-value < .001) and the same for sadness scores (U = 5.0 and p-value < .001). Examples of tweets with high anger scores include:

"Is there One woman out there that has not been violated? *#YouOkSis #WhyWomenDontReport #WhyILeft #WhyIStayed #RapeCulture"*

"Under no circumstance is assaulting a woman acceptable. Abuse is abuse. Rape is rape. No means no. https://amp.twimg.com/v/1cf894c7-e211-4415-a809-d6ae71cd6ded . . . #ItsOnUs"

4.5 Discussion

Digital Storytelling. In our investigations, we did encounter tweets of women sharing their personal assault stories as a part of the #notokay and #whyistayed hashtags, among others. This gives a new perspective on the role of digital storytelling in the context of GBV. Narrative and storytelling have played a huge role in the contexts of social justice [90] and social movements [91]. Until recently, online platforms have been used to encourage users to anonymously share their harassment stories, resulting in shifting their cognitive and emotional orientation towards their experience [92]. What was intriguing in this case was the rise of non-anonymous self-reported stories, which can be viewed as a social movement by women expressing anger about the occurrence of GBV.

Public Figures and Digital Activism. Public figures played a vital role in encouraging people to take a stand against GBV. Four of our GBV-related hashtags (#notokay, #whywomendontreport, #maybehedoesnthityou, and #beenrapedneverreported) were inspired by public figures. We also note that public figures use Twitter as a channel for *digital activism* and promoting collective action in the GBV context, as in: "*In October I asked if we could all share our stories of sexual assaults. #notokay was born. Can you March on Washington JAN 21 with me?*", written by Kelly Oxford.

Limitations and Critique of Methodology. There are limitations to our methodology and findings. Recent studies [93, 94] discuss common issues associated with social media analysis and sample quality of the Twitter's Streaming API. We cannot claim to have captured a complete representation of GBV on Twitter or in the physical world, as we highly depended on the set of GBV key-phrases provided by UNFPA domain experts in [61] as a starting point to our analysis. Our primary objectives were to investigate engagement patterns with GBV content and analyze gender and age demographics. The realm of GBV-related social interactions is clearly greater than what can be captured by a single platform; however, Twitter enables public visibility for user-generated content and the platform has played a key role in enabling women to share. Hence, Twitter is an excellent starting point in our attempt to understand GBV nuances as they take place over a single platform.

4.6 Conclusion

We provide some of the first empirical insights into social media discourse on the sensitive topic of GBV. In our analysis, Twitter has provided a powerful reflection of multiple aspects of GBV. While our analysis shows more engagement with GBV tweets in comparison to generic tweets, the engagement is not uniform across all ages and genders. Although Twitter has been an open platform for all sorts of discussions, it is only recently that public figures have encouraged people to share their personal stories. The data derived from our analysis can be used to complement policy design data sources. Our results show the need for more policies and programs that work to combat GBV.

We also note that anger often surfaces in GBV content. It is our hope that this anger will lead to further progress towards raising awareness and eventually eradicating GBV.

4.7 Acknowledgments

The content of this chapter is the result of a collaboration with Dana Nguyen and has previously appeared in the AAAI Conference on Web and Social Media in 2017.

Chapter 5

On the Limits of Computing for Street Harassment Prevention

Street harassment in public places is a global epidemic that can be magnified by cultures of silence and denial. While different organizations are involved in the fight against street harassment (e.g., non-governmental organizations and law-enforcement), we, as computer scientists, can contribute to this fight by designing technologies that seek to understand and combat these events. In this chapter, we discuss the status quo of information and communication technologies (ICT) for safety on the streets. We also shed light on four distinct types of limits, including those imposed by platforms, society, emerging interpretations of location, and incomplete data sets. These limits stem from the social complexity and on-the-ground realities of this topic, which strongly hinder the progress of ICT for harassment prevention. The limits we discuss are derived from two studies we conducted that pertain to gender-based violence on the streets and in social media. We complement our discussion by an interview we conducted with the Harassmap research team, an organization that fights sexual harassment in Egypt.

5.1 Introduction

The statistics are sobering. One in three women worldwide has experienced physical or sexual violence in their lifetime[1]. In the United States, 65% of women and 25% of men have experienced street harassment (a higher percentage of LGBT-identified men than heterosexual men reported this, and their most common form of harassment was homophobic or transphobic slurs)[2]. In 2014, Cornell University and Hollaback [92] conducted a global study[3] of street harassment[4]. They reached 16,607 respondents across 42 sites worldwide. In the United States, a sample size of 4,872 women reported a number of alarming statistics: nearly every respondent reported experiencing verbal and/or non-verbal harassment in the past year. Half of respondents under 40 reported being groped or fondled in the past year, and 77% of these respondents reported that they had been followed by a man or group of men in a way that made them feel unsafe during the past year.

These events happen in a variety of locations. For example, 55% of respondents under 40 reported that it occurred on the street; 31% reported it on public transit; 30% on the way to work; and 40% in a well-lit area. 46% reported harassment during the day, while 35% reported it late at night. As a result, over 85% of these women reported taking a different route home or to their destination in order to avoid harassment. 73% took a different mode of transportation, while 72% avoided a city or area and 67% changed the time they left an event or location in order to prevent harassment. Unfortunately, a quick glance at the sample of results of this study in other countries reveals that women

[1] U.N. Women: `http://www.unwomen.org/en/news/in-focus/end-violence-against-women`

[2] National Street Harassment Report: `http://www.stopstreetharassment.org/our-work/`

[3] Cornell International Survey on Street Harassment: `http://www.ihollaback.org/cornell-international-survey-on-street-harassment/`

[4] In this work, we base our definition of street harassment on that provided by Hollaback. Street harassment is a form of sexual and gender-based harassment that takes place in public spaces. It can be sexist, racist, transphobic, ableist, sizeist and/or classist. It includes verbal attacks, as well as groping, flashing and assault. http://www.ihollaback.org/street-harassment

around the world experience similar levels of harassment. In some cases, it is much worse: in Egypt, an astounding 99.3% of women studied by the United Nations reported having been sexually harassed[5],. These events leave victims with a variety of reactions, most notably including anger, fear and anxiety, which can lead to depression and low self esteem.

Despite the significant on-going effort to give a voice to victims of harassment, a variety of limits hinder the progress of ICT in this area. In this chapter, we discuss boundaries of ICT for the street safety research agenda. Our discussion stems from (i) previous studies of gender-based violence [61, 62, 92, 95]; (ii) our own analysis of platforms that seek to raise awareness about street harassment and gender-based violence [96, 97]; and (iii) an interview with a non-governmental organization (NGO) whose mission is to raise societal awareness around issues of gender-based violence and street harassment. Overall, we discern four different types of limits: platform-imposed, societal, emergent interpretations of location, and data completeness.

The rest of this chapter is organized as follows. We discuss the status quo of the area of ICT and street safety in Section 5.2, and we shed light on the current platforms used in the context of safety. Section 5.3 identifies the limits of current platforms and users' misconceptions and preferences for safety platforms. In Section 5.4, we discuss challenges associated with emergent interpretations of location. Section 5.5 explores aspects related to cultural norms, gender limitations, and societal impacts. In Section 5.6, we discuss limits on data collection methodologies and the impact that incomplete data has on accurate modeling and prediction. We conclude by discussing future directions in Section 5.7 and emphasize key steps needed to change the dynamics of the ICT for safety agenda.

[5] UNFPA Egypt Sexual Harassment Report: `http://www.dailynewsegypt.com/2013/04/28/99-3-of-egyptian-women-experienced-sexual-harassment-report/`

5.2 ICT and Safety on the Streets

What can be done to address street harassment? Clearly a multi-pronged approach is needed, and fortunately, any quick Internet search will reveal innumerable on-going efforts and programs, many of which have been in place for decades, from education in schools, to community support groups for victims, to workplace trainings. Most recently, technology has played an increasingly larger role in giving a voice to victims of verbal and physical assaults.

Reporting platforms. There are several existing platforms [92, 95] that are geared towards providing a mechanism to report harassment incidents, providing support resources to victims and advice on how to react to different scenarios of harassment. Hollaback [92] is a non-profit activist movement working to end harassment in public spaces. Hollaback provides a blog for people to report harassment incidents through web-based and mobile app interfaces. Victims can anonymously report a description of the harassment incident, location where the incident occurred, and harassment type (e.g., stalking, verbal, homophobic, groping, assault). Through the sharing of harassment stories and experience of community support, victims go through a shift in their cognitive and emotional orientation towards their experience [92]. Harassmap is a volunteer-based initiative rooted in Egypt. The organization works to spread awareness of the epidemic of street harassment. They provide a mobile friendly map of Egypt where harassment incidents are displayed. Victims can report harassment incidents using four channels: web-based, SMS, email and Twitter. Protibadi [95], a similar web and mobile app-based system, was developed to report incidents in Bangladesh. The mobile app is equipped with a “Save Me” button that generates a loud sound that can draw attention of nearby people. The app also sends emergency text messages to indicated contacts. In 2017, mobile phones in India were mandated by the Indian government to include a panic button.When pressed, the button routes to India's new emergency number [98].

Connectivity apps. Beyond the previously mentioned platforms, there is an array of safety apps designed to connect victims to either surrounding bystanders, friends or emergency contacts to seek immediate help. Designed as a digital companion for users walking alone, the Circleof6 app[6] enables a user to place six people into their contact circle. Using a simple set of icons, these six friends can be notified to come pick up the user, or can be texted with a request to call the user to interrupt an uncomfortable situation. The Companion app[7], on the other hand, provides users with a digital companion that monitors the user as they walk or travel. It provides an "I feel nervous" button and an "Emergency" button that can be clicked to alert the digital companion of any dangers around the user. Along with notifying an unlimited number of contacts, the bSafe app[8] provides a time and location-stamped video recording of the incident, upon pressing the alarm button, to potentially be used as evidence against perpetrators. Guardly[9] leverages geofencing to detect whether the victim is within a university campus and, if so connects the victim with campus security phone numbers. The Safetipin app and web platform[10] goes a step further by rendering safe zones on maps based on attributes such as and public transport.

Victim response. In general, victims tend to react differently to various types of harassment, in part based on the degree severity of the incident. Victims may ignore a verbal comment while they are more likely to speak up if the incident included stalking or groping. Other types of responses include reporting the harasser to the police, or

[6] Circle of 6: `https://www.circleof6app.com/`

[7] Companion: `http://www.companionapp.io/`

[8] bSafe: `http://getbsafe.com/`

[9] Guardly: `https://appsagainstabuse.devpost.com/submissions/4899-guardly`

[10] Safetipin: `http://www.safetipin.com/`

taking a picture of the harasser[11]. In Toronto, a mobile app coined "Not Your Baby"[12] provides an interface where the user can choose different combinations of *"who"* (e.g., family, stranger, teacher) is committing the act of harassment and *"where"* the incident is taking place (e.g., home, school, work). The app then generates an appropriate type of response or action for the victim to take in the selected situation. Users can also contribute by adding their stories in the corresponding situations.

While each of the aforementioned projects provides a valuable contribution to the promotion of physical safety and the fight against street harassment, the majority of these projects have one thing in common: they are used after or while the harassment occurs. Moving forward with the research agenda for safety, *the next set of goals will revolve around developing technologies that prevent harassment before it occurs.* To achieve this goal, we need a fundamental understanding of existing limits of current platforms and methodologies, as we discuss in the next sections.

5.3 Platform-imposed Limits

To understand user experiences with current platforms beyond basic reporting mechanisms, we conducted a semi-structured interview with the Harassmap[13] research team. The research team has conducted numerous studies[14] based on interviews with harassment victims.Our discussions with Harassmap identified the following issues about user participation that contribute to underreporting of harassment using existing platforms.

Multi-organization integrated platforms. While Harassmap users emphasized the

[11] Why One Woman Is Photographing Her Catcallers: http://www.huffingtonpost.com/2014/09/11/street-harassment-hey-baby-photo-project_n_5799296.html

[12] Not Your Baby app: http://www.metrac.org/resources/not-your-baby-app/

[13] Harassmap sexual harassment reporting tool: http://www.harassmap.org/en/

[14] Sexual Harassment in Greater Cairo: Effectiveness of Crowdsourced Data - Towards a safer city: http://harassmap.org/en/wp-content/uploads/2013/03/Towards-A-Safer-City_full-report_EN-.pdf

importance of user-friendly interfaces, they were reluctant to be referred to other organizations for legal aid and psychological counseling. The users mentioned that they would be more willing to report harassment online if Harassmap offered the help and support services as a strongly tied part of their organization. To improve the reporting system, Harassmap also suggested that a highly available online chatting service for harassment victims would encourage people to report. Users expressed concern about the relationship of law-enforcement with online reported harassment incidents. Again in this case, users felt that they would be motivated to report harassment to Harassmap if their reports were linked to police and law-enforcement agencies. Users also emphasized the importance of the presence of harassment reporting platforms on pervasive social media websites (e.g., Twitter and Facebook) and not only using the social media pages as gateways to the original platform website, which was implemented later by Harassmap on their Facebook page (under the Report Sexual Harassment button)[15].

Reporting incentives. Similar to Protibadi [95], the Harassmap team noticed that the simple sharing of harassment stories was not a strong motivation for users to use the platform. Users expected some sort of return from the reporting platform. Examples of questions asked by users after reporting include *"What is next after reporting?"*, ''*What will you do?"*, and *"Will you be able catch the harasser?"*

This raises an interesting question: *How can we design incentives that motivate users to report harassment while preserving user anonymity?* While providing multi-organization integrated platforms, as mentioned earlier, could enhance the aggregate level of motivation for users, there is still a need for per-user incentives. Examples of these incentives include informing the users of the impact of their reports.

Simple interfaces. The need for simple interfaces for victims of harassment imposes a significant limitation on the amount of information that can be gathered. As an example,

[15] Harassmap Facebook page: `https://www.facebook.com/HarassMapEgypt/`

the Harassmap research team wanted to include the following two questions related to harassment incidents on their reporting interface: *"What was the reaction of bystanders to the incident?"* (seeking more detailed description of bystander reaction)and *"What was the personal appearance of the victim?"* While adding these two questions to the current interface is simplistic in terms of implementation, they were not added in order to avoid overwhelming the user with required inputs. Overwhelming users with too many input parameters can cause users to pull away from reporting [99]. Designing simple interfaces not only encourages victims to report but also makes it easier for users to understand the mechanism of reporting. To this end, the Harassmap team discussed how users with lower level of education/literacy may struggle with the more complex reporting mechanisms.

Despite the emergence of numerous harassment and safety incident reporting platforms and the increase in the ability of smart communities to engage in environmental monitoring through the implementation of cyber-physical systems, these platforms and systems fail to flawlessly capture and report all harassment and safety incidents due to lack of coverage, infrastructure availability, and data completeness (discussed in Section 5.6). Infrastructure availability has a major impact on the timeliness with which reports can be made. For example, only 29% of people living in rural areas of the world have access to 3G or better coverage[16]. Particularly in developing areas, the cost of access may be prohibitive for people trying to report harassment and safety concerns in real-time; rather, users may need to wait until they have access to a less-expensive per-byte connection to the Internet or forgo reporting at all.

[16] ICT Facts & Figures: `http://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2015.pdf`

5.4 Interpretations of Location

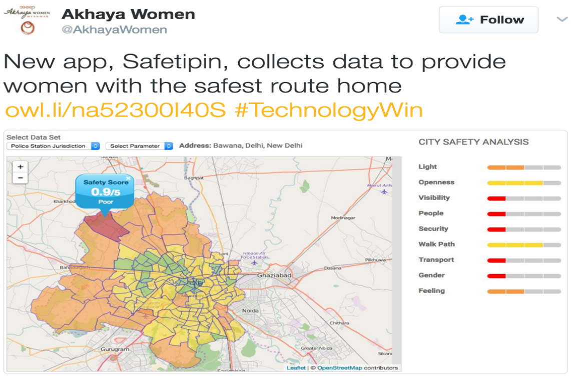
Street harassment characterization requires a nuanced understanding of interpretations of location [100, 101]. Here, we identify some spatial characteristics that effect the detection and prevention of street harassment.

Location takes on varying tenors for individuals through the course of time in both short-term and long-term scales. For instance, a bus stop on an urban street corner may be perceived as safe and secure for a woman on her commute to work at 8 AM on a weekday, but may not hold the same sense of security (and in fact, may actually be perceived as unsafe) when she goes for a walk at 8 PM that evening. There are also long-term trends in levels of harassment that must be considered. For example, in a neighborhood where lighting is installed and parking is metered, the number of harassment incidents may decrease from one month to the next.

In our urban analysis [96] of street harassment incidents reported to Hollaback[17], we analyzed the characteristics of neighborhoods where harassment occurred. Our objective was to try to understand where harassment is likely to occur. Motivated by the term "walking" as the most common word in the reports as well as a high frequency of other terms related to transportation (e.g, "bus" and "train"), we examined walkability and transit scores[18] of GPS locations associated with harassment reports. The definition of walkability states that for a walk to be favorable, it has to be "useful, safe, comfortable and interesting" [27]. Contrary to this definition, and non-intuitively, we found a positive correlation between the walkability scores of a location and the number of harassment incidents. While preliminary, this result sheds light on an inherent tradeoff in walkable streets; they encourage more people to walk at the cost of increasing the probability of

[17] Hollaback. Read and Share Stories. When it comes to street harassment, you are not alone: //www.ihollaback.org/share/

[18] Walk Score Methodology: https://www.walkscore.com/methodology.shtml



Akhaya Women
@AkhayaWomen

New app, Safetipin, collects data to provide women with the safest route home owl.li/na52300I40S #TechnologyWin

Select Data Set
Police Station Jurisdiction
Select Parameter
Address: Bawana, Delhi, New Delhi

CITY SAFETY ANALYSIS

Light
Openness
Visibility
People
Security
Walk Path
Transport
Gender
Feeling

Figure 5.1: A tweet depicting parameters used by Safetipin to calculate safety scores.

occurrence of harassment. We obtained a similar result for transit scores. In essence, these scores are not adequate to capture street harassment safety. There is a need for another score that accounts for street characteristics and place safety. The work done by Safetipin considers a number of characteristics including light, walk paths, transport, people, and visibility when calculating safety scores and represents an excellent start to more nuanced safety metrics as shown in Figure 5.1.

Studies of gender-based violence and street harassment suggest that a significant portion of street harassment incidents take place in the context of mass transit [102, 100] and are less likely to take place in malls and upscale neighborhoods due to their attraction to people with higher socioeconomic status [98]. With respect to sensing, reporting, and predicting instances of street harassment, mobility adds complexity as it represents fluidity in user context over space and time. This contextual fluidity is easily observable in the case of public transportation. For example, if a harassment incident occurs on a bus in transit, recording the location of the incident is non-trivial. If the victim reports the incident via a reporting platform, the representation of the incident in space is ambiguous. Should the incident be recorded as taking place at the exact geolocation recorded at the time the report is created? Should the incident be recorded as taking place on a bus route, and if so, is it possible to identify that route (e.g., "the red line" or "28 express bus") automatically? Similarly, the time of the event is also important. In any scenario, but particularly in mobile scenarios, the difference between the time an incident occurs, the time an incident report is created, and the time an incident report is transmitted can have significant implications for the accurate prediction and effective prevention of street harassment. Both geospatial and temporal information are critical for the prediction and prevention of future incidents, but depending on the area through which a person is traveling, it may be difficult to automatically identify a route (e.g., in an urban setting where multiple bus lines overlap during the same time window) or an exact location (e.g., in areas with dense foliage that prevent precise geolocation). Thus, even when temporal and geographical information is provided, there are limits to guarantees of perfect integration of those data points with the actual incident experienced by a reporting user.

5.5 Cultural and Societal Limits

In order to prevent street harassment, we need to understand underlying forces that enable its existence. Cultural and societal perspectives are key forces that affect the perception of street harassment.

Gender-imposed limits. Inherently, the sexual harassment problem demonstrates a gender power gap. From the technology design perspective, female participation is hindered by the biases women face in science-related disciplines [103, 104]. From the perspective of technology access, prior work on the digital gender divide attributed the divide to factors of employment, income, and education [105]. In their 2011 study, Hilbert demonstrated that when controlling for variables of employment, income, and education, women turn out to be more active users of digital tools than men. The studied variables are inherently gender-biased [106] especially in developing countries where gender roles and stereotypes are emphasized [107]. Ahmed et al. discussed gender inequality and analyzed the impact of gender inequality on the stages of user research, design, deployment, and use of computing technologies which is presented in previous literature [108]. In this work, the authors outline the notions of reflective design [109] and Feminist HCI [92] as key approaches to start mitigating the effect of the gender gap in the design of technologies pertaining to women.

Privacy assurances. Since the discussion of gender-based violence was considered taboo for most participants in the Harassmap interview process, privacy concerns were mentioned multiple times. Through interviews, the Harassmap team reported the need to reassure women multiple times that their feedback would remain anonymous. We can only imagine that a significant number of women do not report online due to anonymity concerns as well.

5.6 Limited Data

Computational approaches to reporting and predicting street harassment and safety issues depend on the consistency and completeness of data provided to computational algorithms. A limitation with respect to data used to characterize, predict, and prevent incidents of street harassment is a lack of completeness. In addition to issues of under-reporting, it is also difficult to know which factors might be critical for predicting street harassment. While some factors may be easy to record automatically (e.g., time of day and geolocation), others may require more complex sensory information (e.g., number of people nearby and lighting) or direct user input (e.g., current activity and personal appearance[19]).

Even when there is a well-understood set of factors that may be predictive of incidents

[19] Personal appearance, especially with respect to clothing style, is a data point requested in numerous studies on gender-based violence and street harassment to understand the impact (if any) appearance has on experiences of harassment [100, 102].

of street harassment, there are energy and connectivity costs associated with constantly collecting environmental data about the aforementioned factors. For instance, assume lighting and the number of surrounding people were predictive of whether or not a person would encounter street harassment. An application seeking to route a user safely through space would rely not only on information about the user's current environment with respect to crowd size and lighting, but there would need to be accurate information about the crowd sizes and lighting around all areas between the user's current location and target destination. This would require constant environmental sensing performed either in a crowd-sourced manner (where individuals' devices are constantly collecting environmental data about their current location and transmitting it to the cloud for broader access) or via cyber-physical systems (CPS) where there are pervasive lighting and infrared sensors and data is frequently transmitted to a centralized entity for processing. While CPS or crowd-sensing approaches are feasible, there are potential limits to scale, particularly in rural and developing areas where underlying connectivity between sensors and processing clouds is limited. Indeed, even in areas where connectivity is not currently an issue, previous literature indicates that a confluence of information-generating sensors and applications and increasing demand will invariably run up against infrastructural limits, specifically energy and bandwidth limits [110, 111, 112].

5.7 Future Directions

Based on our analysis, we envision three areas that will serve as a vital basis for future street harassment prevention research.

Social media usage. In our study [97] pertaining to gender-based violence in social media, it was shown that public figures used hashtags to motivate women to speak up about gender-based violence and share their personal experiences. Examples of the studied hashtags include *#beenrapedneverreported*, which was co-created by Montgomery and Zerbisias to tweet support for the women who alleged they were assaulted by former CBC radio host, Jian Ghomeshi, in 2014. The hashtag *#maybehedoesnthityou* was used by writer and artist, Zahira Kelly, to publicly share her emotional abuse experience in 2016; and *#notokay* started in October 2016 by Kelly Oxford, Canadian author and social media blogger, to encourage women to share their first assault experiences. In this latter case, Kelly Oxford reported that she received 1 million tweets in one night with a minimum rate of 50 tweets per minute[20]. The phrase "Me Too" by social activist Tarana Burke was first used in 2006 to raise awareness for women of color who experienced sexual abuse on MySpace and was later popularized by actress Alyssa Milano on October 2017 who coined the phrase into #MeToo, encouraging women to share their sexual abuse stories to understand the magnitude of the problem. In less than 24 hours, Facebook soared with responses of 4.7 million people using #MeToo in 12 million posts[21]. On Twitter, the hashtag was tweeted over 825,000 times. With the rise of public figures, it is clear that the pervasiveness of social media can provide platforms of community while giving voice to victims of harassment and violence.

Cyber-physical systems. As communities become smarter and more connected, CPS present a realistic and holistic method in which to report, predict, and prevent street harassment. CPS enables characteristics and phenomena occurring in the "real-world" to be translated to the cyber-world by combining wireless sensors, Internet of things paradigms, machine-to-machine communication, wireless networking, and crowd-sensing.One of the promising aspects of CPS for the the prevention of gender-based violence and street safety is that they allow for individual and organizational feedback to close information

[20] Sexual assault and the Trump tape: 1 million women say it's #notokay: `http://money.cnn.com/2016/10/08/technology/notokay-twitter-donald-trump/`

[21] A Little Girl and the Heartbreaking Origin of "Me too": `http://www.cnn.com/2017/10/17/us/me-too-tarana-burke-origin-trnd/index.html`

loops caused by emergent characteristics that can be difficult for computational systems to detect, identify, and interpret. Future work investigating ICT for street safety will look to CPS that integrate pervasive sensors, crowd-sensing, law enforcement feedback, and pedestrian feedback. In designing these systems with respect to computational limits, research must address challenges associated with bandwidth and energy constraints imposed by constant sensing and data collection [113, 114].

Social Change. There is no doubt that street harassment and the associated problem of under-reporting of victims is multi-pronged. Technology alone will not cause social change [115], thus significant work needs to be done on multiple levels. Encouraging victims to speak up through social movements and on-the ground campaigns is one approach that should include tackling the sociocultural norms that encourage victim blaming, shamefulness and the taboo culture that revolves around the topic of sexual harassment.

5.8 Acknowledgments

This chapter is a result of a collaboration with Morgan Vigil-Hayes. We would like to also thank Enas Hamdy from Harassmap for taking the time to answer questions as part of our semi-structured interview.

Chapter 6

Mitigating Gender Bias in Natural Language Processing

In this chapter, we review contemporary studies on recognizing and mitigating gender bias in NLP. We discuss gender bias based on four forms of representation bias and analyze methods recognizing gender bias. Furthermore, we discuss the advantages and drawbacks of existing gender debiasing methods. Finally, we discuss future studies for recognizing and mitigating gender bias in NLP.

6.1 Introduction

Gender bias is the preference or prejudice toward one gender over the other [116]. Gender bias is exhibited in multiple parts of a Natural Language Processing (NLP) system, including the training data, resources, pre-trained models (e.g. word embeddings), and algorithms themselves [117, 118, 119, 120]. NLP systems that contain bias in any of these parts can produce gender biased predictions and sometimes even amplify biases present in the training sets [121].

The propagation of gender bias in NLP algorithms poses the danger of reinforcing damaging stereotypes in downstream applications. This has real-world consequences; for example, concerns have been raised about automatic resume filtering systems giv-

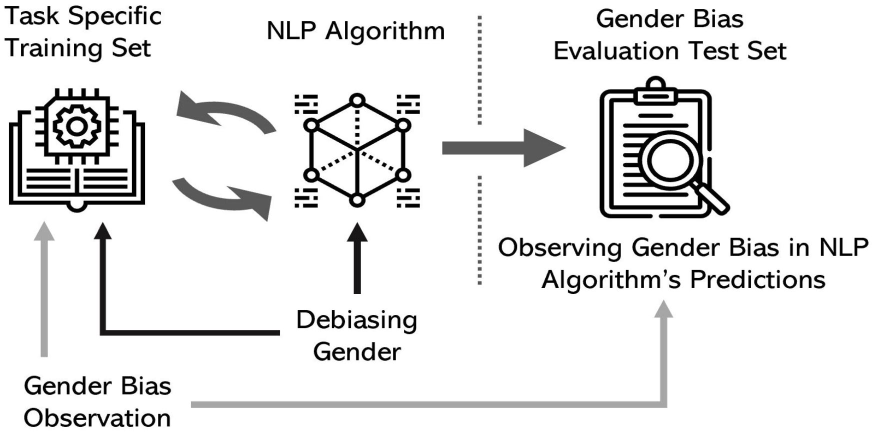


Figure 6.1: Observation and evaluation of gender bias in NLP. Bias observation occurs in both the training sets and the test sets specifically for evaluating the gender bias of a given algorithm's predictions. Debiasing gender occurs in both the training set and within the algorithm itself.

ing preference to male applicants when the only distinguishing factor is the applicants' gender.

One way to categorize bias is in terms of allocation and representation bias [4]. Allocation bias can be framed as an economic issue in which a system unfairly allocates resources to certain groups over others, while representation bias occurs when systems detract from the social identity and representation of certain groups [4]. In terms of NLP applications, allocation bias is reflected when models often perform better on data associated with majority gender, and representation bias is reflected when associations between gender with certain concepts are captured in word embedding and model parameters. In Table 6.1, we categorize common examples of gender bias in NLP following [4]. Briefly, denigration refers to the use of culturally or historically derogatory terms; stereotyping reinforces existing societal stereotypes; recognition bias involves a given algorithm's inaccuracy in recognition tasks; and under-representation bias is the disproportionately low representation of a specific group. We identify that both allocative and representational harms often arise in NLP systems due to statistical patterns in the training corpora,

Task	**Example of Representation Bias in the Context of Gender**	**D**	**S**	**R**	**U**
Machine Translation	Translating “He is a nurse. She is a doctor.” to Hungarian and back to English results in “She is a nurse. He is a doctor.” [122]		✓	✓	
Caption Generation	An image captioning model incorrectly predicts the agent to be male because there is a computer nearby [123].		✓	✓	
Speech Recognition	Automatic speech detection works better with male voices than female voices [124].			✓	✓
Sentiment Analysis	Sentiment Analysis Systems rank sentences containing female noun phrases to be indicative of anger more often than sentences containing male noun phrases [125].		✓		
Language Model	“He is doctor” has a higher conditional likelihood than “She is doctor” [126].		✓	✓	✓
Word Embedding	Analogies such as “man : woman :: computer programmer : homemaker” are automatically generated by models trained on biased word embeddings [118].	✓	✓	✓	✓

Table 6.1: Following the talk by [4], we categorize representation bias in NLP tasks into the following four categories: (D)enigration, (S)tereotyping, (R)ecognition, (U)nder-representation.

which are then embedded in semantic representations and the model.

Gender bias in NLP is a complex and compound issue, requiring interdisciplinary communication. As NLP systems have been increasingly integrated with our daily life thanks to modern AI developments, we need both immediate solutions to patch current systems as well as fundamental approaches to debias. In this chapter, we provide a comprehensive literature review to summarize recent attempts for recognizing and mitigating bias in NLP systems. Most debiasing methods can be depicted as a special case in Figure 6.1.

We make two primary contributions. (1) We summarize recent studies of algorithmic bias in NLP under a unified framework for the ease of future discussion. (2) We critically discuss issues with current debiasing methods with the purpose of identifying optimizations, knowledge gaps, and directions for future research.

6.2 Observing Gender Bias

Recent work in analyzing gender bias in NLP has focused on quantifying bias through psychological tests, performance differences between genders for various tasks, and the geometry of vector spaces. We provide an overview of gender bias evaluation methods and discuss types of representation bias each method identifies.

6.2.1 Adopting Psychological Tests

In psychology, the Implicit Association Test (IAT) is used to measure subconscious gender bias in humans, which can be quantified as the difference in time and accuracy for humans to categorize words as relating to two concepts they find similar versus two concepts they find different [127, 119]. For instance, to measure subconscious associations of genders with arts and sciences, participants are asked to categorize words as pertaining to (males or the sciences) or (females or the arts) [128]. The participants are then asked to categorize words as pertaining to (males or the arts) or (females or the sciences). If participants answered faster and more accurately in the former setting, it indicates that humans subconsciously associate males with the sciences and females with the arts.

[119] adopt the IAT's core concept, measuring gender bias through the difference in strength of association of concepts, to measure bias in word embeddings using the Word Embedding Association Test (WEAT) [119]. The authors confirm that human biases found through IAT tests exist in GloVe and Word2Vec embeddings. Finally, the authors demonstrate a positive correlation between the strength of association of an occupation word embedding with the female gender and the percentage of females in that occupation in United States, with the percentages taken from Bureau of Labor Statistics labor force participation data. Notably, [120] show that bias in word embeddings can be used to track social changes such as increased or decreased female participation in the workforce.

[129] extend WEAT to create the Sentence Encoder Association Test (SEAT), capable of testing sentence encoders (e.g., ELMo [130]) for human biases found in IAT tests.

6.2.2 Analyzing Gender Sub-space in Embeddings

[118] define gender bias as the correlation between the magnitude of the projection onto the gender subspace of a word embedding representing a gender-neutral word and that word's bias rating, as rated by crowd workers. To identify the gender subspace, they first build a linear support vector machine to classify words into a set of gender-specific and a set of gender-neutral words based on a training set of hand-selected gender-specific words. The authors then identify a gender direction by aggregating ten gender pairs (e.g. she-he, her-his, woman-man, etc.) and using principal component analysis to find a single eigenvector that exhibits significantly greater variance than the rest. [131] extend this method and their approach can be used to find non-binary gender bias by aggregating n-tuples instead of gender pairs.

However, [132] note that the above method fails to capture the full picture of gender bias in vector spaces. Specifically, even after the projections of word embeddings representing gender-neutral words onto the gender subspace have been removed, word embeddings representing words with similar biases still cluster together. They further introduce the notion of cluster bias. Cluster bias of a word w can be measured as the percentage of male or female stereotypical words among the k nearest neighbors of w's embedding where the male or female stereotypical words are obtained through human annotation.

6.2.3 Measuring Performance Differences Across Genders

In most NLP tasks, a model's prediction should not be heavily influenced by the gender of the entity mentions or contexts in the input. To evaluate whether or not this is the case, consider two sentences that act as the inputs to a model for which the only differences are the words that correspond to gender, such as "*He* went to the park" vs "*She* went to the park". We refer to changing the gender of the gendered nouns as *gender-swapping*. Gender-swapping can be generalized to sentences by swapping each male-definitional word with its respective female equivalent and vice-versa [117, 126, 133]. If the model does not make decisions based on genders, it should perform equally for both sentences. Otherwise, the difference in evaluation scores reflects the extent of gender bias found in the system.

For example, [134] introduce two metrics to measure these performance differences – False Positive Equality Difference (FPED) and False Negative Equality Difference (FNED) – that have been used to measure gender bias in abusive language detection [125]. These are defined as the differences in the false positive and false negative rates, respectively, of predictions of a model between original and gender-swapped inputs. We note that these measurements can generalize to tasks aside from abusive language detection.

By designing test sets, measuring performance differences between genders reveals representational gender bias in the context of recognition, stereotyping, and under-representation. If, for instance, an image captioning model is worse at recognizing a *woman* than a *man* when they are each sitting in front of a computer [123], that is a clear indicator of recognition bias. If this prediction inaccuracy arises as a consequence of the algorithm's association between *man* and *computer*, then this example also reveals stereotyping in the image captioning model. One can also imagine that if the model is

Data Set	**Task**	**Probing Concept**	**Size**
Winogender Schemas [135]	Coreference Resolution	Occupation	720 English Sentences
WinoBias [117]	Coreference Resolution	Occupation	3,160 English Sentences
GAP [136]	Coreference Resolution	Names	4,454 English Contexts
EEC [133]	Sentiment Analysis	Emotion	8,640 English Sentences

Table 6.2: Summary of GBETs. GBETs evaluate models trained for specific tasks for gender bias. GBETs use differences in values of the probing concept or prediction accuracies relating to the probing concept between gender-swapped data points to measure bias.

not debiased and these errors propagate over a large sample of images, then the model may further contribute to the under-representation of minority.

Standard evaluation data sets in NLP are inadequate for measuring gender bias. For one, these data sets often also contain biases (such as containing more male references than female references), so evaluation on them might not reveal gender bias. Furthermore, predictions made by systems performing complex NLP tasks depend on many factors; we must carefully design data sets to isolate the effect of gender of the output in order to be able to probe gender bias. We name these data sets Gender Bias Evaluation Testsets (GBETs).

The goal of designing GBETs is to provide check that NLP systems avoid making mistakes due to gender bias. Some may argue that the artificial design of GBETs does not reflect the true distribution of the data, implying that these evaluations are artificial. We argue that if humans can avoid making mistakes due to gender bias, then machines should as well. Additionally, systems that make biased predictions may discourage minorities from using those systems and having their data collected, thus worsening the disparity in the data sets [137]. We provide an overview of publicly available GBETs in Table 6.2.

Gender-swapped GBETs: In the following, we review GBETs in coreference resolution and sentiment analysis applications.

For coreference resolution, [135] and [138] independently designed GBETs based on Winograd Schemas. The corpus consists of sentences which contain a gender-neutral occupation (e.g., doctor), a secondary participant (e.g., patient), and a gendered pronoun that refers either the occupation or the participant. The coreference resolution system requires the identification of the antecedent of the pronoun. For each sentence, [135] consider three types of pronouns (female, male, or neutral), and [138] consider male and female pronouns. The two datasets have a few notable differences (see the discussion in [135]).

Note that simply measuring a global difference in accuracies of a model between inputs with different gendered pronouns is insufficient. For example, a model could predict females and males to be coreferent to "secretary" with 60% and 20% accuracy, respectively. If that same model predicts females and males coreferent to "doctor" with 20% and 60% accuracy, respectively, then the global average accuracy for each gender is equivalent, yet the model exhibits bias.[1] Therefore, [138] and [135] design metrics to analyze gender bias by examining how the performance difference between genders with respect to each occupation correlate with the occupational gender statistics from the U.S Bureau of Labor Statistics.

Another GBET for coreference resolution named GAP contains sentences mined from Wikipedia and thus can perform an evaluation with sentences taken from real contexts as opposed to artificially generated ones [136]. GAP does not include stereotypical nouns; instead, pronouns refer to names only. Gender bias can be measured as the ratio of F_1 scores on inputs for which the pronoun is female to inputs for which the pronoun is male. Notably, sentences are not gender-swapped, so there may be differences in difficulty between sentences in male and female test sets.

For sentiment analysis, a GBET dataset named *Equity Evaluation Corpus* (EEC) [133] is designed. Each EEC sentence contains an emotional word (e.g., anger, fear, joy, sad-

[1] For the sake of simplicity, we illustrate the motivation in accuracy. The coreference resolution systems may be evaluated using a different metric.

ness), with one of five intensities for each emotion and a gender-specific word. Gender bias is measured as the difference in emotional intensity predictions between gender-swapped sentences.

6.3 Debiasing Methods Using Data Manipulation

Several approaches have been proposed for debiasing gender stereotypes in NLP by working on two tangents: (1) text corpora and their representations and (2) prediction algorithms. In this section, we will discuss the techniques to debias text corpora and word embeddings. We do the same for techniques to mitigate gender bias in algorithms in Section 6.4.

We note that debiasing methods can be categorized as retraining and inference (see Table 6.3). Retraining methods require that the model is trained again, while inference methods reduce bias without requiring the existence of the original training set. Retraining methods tend to address gender bias in its early stages or even at its source. However, retraining a model on a new data set can be costly in terms of resources and time. Inference methods, on the other hand, do not require models to be retrained; instead, they patch existing models to adjust their outputs providing a testing-time debiasing. We will discuss different debiasing methods from these two perspectives.

6.3.1 Debiasing Training Corpora

We review three approaches for debiasing gender in the literature.

Data Augmentation

Oftentimes a data set has a disproportionate number of references to one gender (e.g. OntoNotes 5.0) [117]. To mitigate this, [117] proposed to create an augmented data set identical to the original data set but biased towards the opposite gender and to train on the union of the original and data-swapped sets. The augmented data set is created using gender-swapping. This is similar to the method used to create GBETs; however, the goal of data augmentation is to debias predictions by training the model on a gender-balanced data set, while GBETs are created specifically to evaluate the gender bias of those predictions both before and after debiasing.

Data augmentation works as follows: for every sentence in the original data set, create that sentence's gender-swapped equivalent using the procedure described in 6.2.3. Next, apply name-anonymization to every original sentence and its gender-swapped equivalent. Name anonymization consists of replacing all named entities with anonymized entities, such as "E1". For instance, *Mary likes her mother Jan* becomes *E1 likes his father E2* after applying gender-swapping and name anonymization for data augmentation. This removes gender associations with named entities in sentences. The model is then trained on the union of the original data set with name-anonymization and the augmented data set. The identification of gender-specific words and their equivalent opposite gender word requires lists typically created by crowd workers.

Data augmentation has been shown to be flexible; it can mitigate gender bias in several different models in many different tasks. When applied to a neural network based coreference resolution model [139, 140] originally trained on OntoNotes 5.0 which was tested on WinoBias, gender augmentation lowered the difference between F_1 scores on pro-stereotypical and anti-stereotypical test sets significantly, which indicates the model was less inclined to make gender-biased predictions [117, **?**]. In hate speech detection, data augmentation reduced FNED and FPED differences between male and female predictions of a Convolutional Neural Network by a wide margin [125]. Data augmentation without name-anonymization has also been used to debias knowledge graphs built from Bollywood movie scripts [141] by swapping the nodes for the lead actor and actress, but

Methods	**Method Type**
Data Augmentation by Gender-Swapping	Retraining
Gender Tagging	Retraining
Bias Fine-Tuning	Retraining
Hard Debiasing	Inference
Learning Gender-Neutral Embeddings	Retraining
Constraining Predictions	Inference
Adjusting Adversarial Discriminator	Retraining

Table 6.3: Debiasing methods can be categorized according to how they affect the model. Some debiasing methods require the model to be retrained after debiasing (Retraining). Others modify existing models' predictions or representations (Inference).

metrics evaluating the success of gender-swapping were not provided.

Data augmentation is easy to implement, but creating the annotated list can be expensive if there is high variability in the data or if the data set is large since more annotations will be required. Furthermore, data augmentation doubles the size of the training set, which can increase training time by a factor specific to the task at hand. Lastly, blindly gender-swapping can create nonsensical sentences – for example, gender-swapping "*she* gave birth" to "*he* gave birth" [141].

Gender Tagging

In some tasks, like Machine Translation (MT), confounding the gender of the source of a data point can lead to inaccurate predictions. Current MT models predict the source to be male a disproportionate amount of time [142, 143]. This happens because training sets are dominated by male-sourced data points, so the models learn skewed statistical relationships and are thus more likely to predict the speaker to be male when the gender of the source is ambiguous [143].

Gender tagging mitigates this by adding a tag indicating the gender of the source of the data point to the beginning of every data point. For instance, "I'm happy" would change to "MALE I'm happy." In theory, encoding gender information in sentences could improve translations in which the gender of the speaker affects the translation (i.e. "I am happy" could translate to "?Je suis heureux" [M] or "Je suis heureuse" [F]), since English does not mark the gender of the speaker in this case. The tag is then parsed separately from the rest of the data by the model. The goal is to preserve the gender of the source so the model can create more accurate translations [143].

Gender tagging is effective: a Sequence-to-Sequence Neural Network trained on Europarl increased BLEU scores significantly for machine translations from English to French in which the first-person speaker was female [143]. Sentences with male first-person speakers had accuracy increases by a sizeable margin. However, gender-tagging can be expensive: knowing the gender of the source of a data point requires meta-information, and obtaining this could be costly in terms of memory usage and time. Furthermore, MT models may need to be redesigned to correctly parse the gender tags.

Bias Fine-Tuning

Unbiased data sets for a given task may be scarce, but there may exist unbiased data sets for a related task. Bias fine-tuning incorporates transfer learning from an unbiased data set to ensure that a model contains minimal bias before fine-tuning the model on a more biased data set used to train for the target task directly [125]. This allows models to avoid learning biases from training sets while still being adequately trained to perform a task.

Bias fine-tuning has been shown to be relatively effective. [125] use transfer learning from a gender unbiased abusive tweets data set [144] and fine-tuning on a gender-biased sexist tweets data set [145] to train a Convolutional Neural Network (CNN). They evaluate the CNN using a GBET evaluation set (which is private, so not mentioned in 6.2.3). They tested the same model after training it on gender-swapped data sets as well. [125] find that gender-swapping was more effective at both removing bias and retaining performance than bias fine-tuning. However, transfer learning may have been ineffective in this case because abusive language detection data sets and sexist language detection data sets have significant differences. For more similar data sets, bias fine-tuning may be more effective; further testing is necessary.

6.3.2 Debiasing Gender in Word Embeddings

Word embeddings represent words in a vector space. These embeddings have been demonstrated to reflect societal biases and changing views during social movements in the United States [120].

As the word embedding model is a fundamental component in many NLP systems, mitigating bias in embeddings plays a key role in the reduction of bias that is propagated to downstream tasks (e.g., [117]). However, it is debatable if debiasing word embeddings is a philosophically right step towards mitigating bias in NLP. [119] argue that debiasing word embeddings blinds an AI agent's perception rather than teaching it to perform fair actions. We refer readers to the discussion in [119].

It is also important to recognize that removing gender bias from the embedding space entirely is difficult. While existing methods successfully mitigate bias with respect to projection onto the gender subspace in some degrees, [132] show that gender bias based on more subtle metrics such as cluster bias still exist.

In the following we review two families of approaches to debias gender in word embeddings. One difference between these two types of methods is that the former does not require retraining embeddings, whereas the latter does.

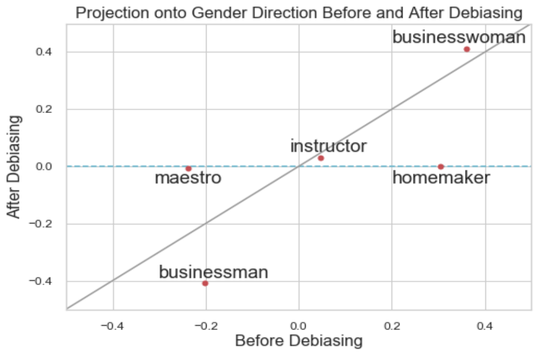


Figure 6.2: We project five word2vec embeddings onto the 'he' - 'she' direction before and after neutralizing the gender-neutral words *maestro, instructor*, and *homemaker* and equalizing the gender-specific pair *businessman* and *businesswoman* [3]. For both x and y-axes, negative values represent male gender bias and positive values represent female gender bias.

Removing Gender Subspace in Word Embeddings

[146] first removed similarity to the gender subspace in word embeddings by building a genderless framework using cosine similarity and orthogonal vectors [146]. Removing the gender component, though, pushes the word *he* to become the 6th closest word to *she* when it was the 1,826th closest previously. The genderless framework may be flawed because the semantic definition of a given word may be closely tied to its gender component. However, a case can also be made that a word's gender component should play a key role in its semantic definition. We encourage future work to collaborate with social scientists for further discussion on this topic.

[118] build upon [146] and propose to surgerically alter the embedding space by removing the gender component only from gender-neutral words. Instead of removing gender altogether, debiasing involves making gender-neutral words orthogonal to the gender direction (see Figure 6.2). Ultimately, word embeddings with reduced bias performed just as well as unaltered embeddings on coherence and analogy-solving tasks [118].

Learning Gender-Neutral Word Embeddings

[138] propose a new method called GN-GloVe that does not use a classifier to create a set of gender-specific words. The authors train the word embeddings by isolating gender information in specific dimensions and maintaining gender-neutral information in the other dimensions. They do this by (1) minimizing the negative difference (i.e. maximizing the difference) between the gender dimension in male and female definitional word embeddings and (2) maximizing the difference between the gender direction and the other neutral dimensions in the word embeddings. This allows for greater flexibility; the gender dimensions can be used or neglected.

Finally, we note that both aforementioned approaches [118, 138] used to debias word embeddings may not work with embeddings in a non-Euclidean space, such as Poincare embeddings [147], because the notion of cosine similarity would no longer apply. Also, it is unclear if these approaches can be extended to other languages beyond English, especially for languages with grammatical genders.

6.4 Debiasing by Adjusting Algorithms

Some gender debiasing methods in NLP adjust predictions in NLP systems. We call these algorithm adjustment methods. In this section, we discuss two such approaches.

6.4.1 Constraining Predictions

Zhao et al.[121] show that an NLP model risks amplifying bias by making predictions which exacerbate biases present in the training set. For instance, if 80% of coreferents of "secretary" are female in a training set and a model trained on that set predicts 90% of coreferents of "secretary" in a test set to be female, then that model amplifies bias.

Zhao et al.[121] proposed Reducing Bias Amplification (RBA) based on a constrained conditional model [148], which takes an existing model's optimization function and constrains that function to ensure its predictions fit defined conditions. For example, when RBA was applied to the visual semantic role labelling [149], it restricted the ratio of males to females predicted to be doing particular activities to prevent the model from amplifying bias through predictions. The approximate inference can be efficient solved by Lagrangian relaxation [150].

6.4.2 Adversarial Learning: Adjusting the Discriminator

Zhang et al.[151] propose a variation on the traditional generative adversarial network [152] by having the generator learn with respect to a protected gender attribute. In other words, the generator attempts to prevent the discriminator from identifying the gender in a given task such as analogy completion. This method has the potential to be generalizable: it can be used to debias any model that uses gradient-based learning.

6.5 Conclusion and Future Directions

In this chapter, we summarize recent literature about recognizing and mitigating gender bias in NLP. We acknowledge that the scope of this chapter is limited. There is a long history of gender stereotype study in law, psychology, media study, and many other disciplines which we do not discuss. Similar issues of algorithmic bias have also been discussed extensively in artificial intelligence, machine learning, data mining, and several other application fields (e.g., [153, 154, 155, 156, 157, 158, 159, 156]). Other important aspects such as model/data transparency [160, 161] and privacy preservation [162, 163, 164] are also not covered in this literature survey. Besides, we refer the readers to [165] for a more general discussion of ethical concern in NLP.

The study of gender bias in NLP is still relatively nascent and consequently lacks unified metrics and benchmarks for evaluation. We urge researchers in related fields to work together to create standardized metrics that rigorously measure the gender bias in NLP applications. However, we recognize that different applications may require different metrics and there are trade-offs between different notions of biases [166, 167].

Gender debiasing methods in NLP are not sufficient to debias models end-to-end for many applications. We note the following limitations of current approaches. First, the majority of debiasing techniques focus on a single, modular process of an end-to-end NLP system. It remains to be discovered how these individual parts harmonize together to form an ideally unbiased system. Second, most gender debiasing methods have only been empirically verified in limited applications [151, 121], and it is not clear that these methods can generalize to other tasks or models. Third, we note that certain debiasing techniques may introduce noise into a NLP model, causing performance degradation. Finally, hand-craft debiasing approaches may unintentionally encode the implicit bias of the developers.

Below, we identify a few future directions.

Mitigating Gender Bias in Languages Beyond English. With few exceptions [143, 142], prior work has focused on mitigating gender bias in the English language. Future work can look to apply existing methods or devise new techniques towards mitigating gender bias in other languages as well. However, such a task is not trivial. Methods such as gender-swapping are relatively easy in English because English does not distinguish gender linguistically. However, in languages such as Spanish, each noun has its own gender and corresponding modifiers of the noun need to align with the gender of the noun. To perform gender-swapping in such languages, besides swapping those gendered nouns, we also need to change the modifiers.

Non-Binary Gender Bias. With few exceptions [131], work on debiasing in NLP has assumed that the protected attribute being discriminated against is binary. Non-binary genders [168] as well as racial biases have largely been ignored in NLP and should be considered in future work.

Interdisciplinary Collaboration. As mentioned in Section 6.1, gender bias is not a problem that is unique to NLP; other fields in computer science such as data mining, machine learning, and security also study gender bias [153, 154, 155, 156, 157, 158, 159, 169]. Many of these technical methods could be applicable to NLP yet to our knowledge have not been studied.

Additionally, mitigating gender bias in NLP is both a sociological and an engineering problem. To completely debias effectively, it is important to understand how machine learning methods encode biases and how humans perceive biases. A few interdisciplinary studies [170, 171, 172, 173] have emerged, and we urge more interdisciplinary discussions in terms of gender bias. Approaches from other technical fields may improve current debiasing methods in NLP or inspire the development of new, more effective methods even if the properties of the data or problem are different across fields. Discussions between computer scientists and sociologists may improve understanding of latent gender bias found in machine learning data sets and model predictions.

6.6 Acknowledgements

The content of this chapter is the result of a collaboration with Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Jieyu Zhao, Diba Mirza, and Kai-Wei Chang. The chapter appeared in ACL 2019.

Chapter 7

Understanding Gender Bias in Neural Relation Extraction

Recent developments in Neural Relation Extraction (NRE) have made significant strides towards increasingly reliable classifications. While much attention has been dedicated towards improvements in accuracy, there have been no attempts in the literature to evaluate social biases in NRE systems. In this chapter, we analyze gender bias in NRE systems. We create and publicly release Wikigender, a distantly supervised dataset with a human annotated test set that has an even split of male and female sentences specifically curated to analyze gender bias in relation extraction systems. We use Wikigender to evaluate systems for bias and find that NRE systems exhibit gender biased predictions and lay groundwork for future evaluation of bias in NRE.

7.1 Introduction

With the wealth of information being posted online daily, Information Extraction (IE), the task of extracting information from unstructured text, has become increasingly important. One sizable sub-domain of IE called Relation Extraction aims specifically to extract relations from raw sentences and represent them as succinct relation tuples of the form *(head, relation, tail)*. An example is *(Barack Obama, spouse, Michelle Obama)*.

The concise representations provided by Neural Relation Extraction (NRE) models are used to extend Knowledge Bases (KBs), that are in turn heavily used to understand the meaning of sentences in downstream NLP tasks like search and QA [?]. In recent years, much focus in the NRE community has been centered on improvements in model precision and the reduction of noise [?, ?, ?, ?, ?]. Yet, little attention has been devoted towards the fairness of such systems.

In this paper, we take the first step at understanding and evaluating gender bias in NRE systems. Gender bias in NRE models takes the form of differences in model performance in extractions that have a gendered relation, provided other dimensions or variables are fixed. High differences in performance metrics, like accuracy, between genders could diminish the fairness of systems and distort outcomes in applications that use them. For example, if a model better predicts the *occupation* relation for with higher accuracy for male entities, this could lead to KBs having more occupation information for males. Downstream search tasks using that KB could produce biased predictions, such as ranking articles about female computer scientists below articles about their male peers.

We provide the first evaluation of social bias in NRE models; specifically, we evaluate gender bias in English language predictions of a collection of popularly used and open source NRE models[1] [?, ?, ?, ?]. We propose to measure gender bias in NRE by measuring the difference in accuracy with which NRE models extract relations for sentences from articles written about female entities and articles written about male entities.

However, carrying out such an evaluation is difficult with existing NRE datasets, such as the NYT dataset from [?], because there is no reliable way to obtain gender information about the entities. Thus, we create a new dataset specifically aimed at evaluating gender bias for NRE, just as prior work has done for other tasks like Coreference Resolution [138,

[1] `https://github.com/thunlp/OpenNRE/`

135]. We call our dataset Wikigender, and we make it publicly available for others who would like to evaluate gender bias in NRE models. Our contributions are as such:

- Wikigender is the first dataset aimed at training and evaluating NRE systems for gender bias. It contains ground truth labels for the test set and about 45,000 sentences in total.
- We provide the first evaluation of OpenNRE for gender bias and find that it exhibit gender bias.
- We find that the performance of OpenNRE does not differ significantly when trained using default word embeddings vs debiased word embeddings.

7.2 Methodology

To compare the performance of NRE models when extracting relations from articles written about male entities and articles written about female entities, we need articles with gender information about the entity being written about. We use Wikipedia articles, since most Wikipedia articles about people contain gender information. We have dataset statistics in Table 7.1, as well as splits for the training and development set, which have 30,456 male sentences and 10,532 female sentences. To make sure that test data is unseen in training and development, each entity and all its corresponding sentences may only be present in one of the training, development, and test datasets. We then train and evaluate the NRE models on Wikigender.

7.2.1 Generation of Datasets

The models we evaluate are supervised RE models; as such, they require labeled data. Obtaining labeled data to train RE models is tedious and expensive to scale up

	Head Entities	**Number of Sentences**	**Distantly Supervised?**
Train	4118	36365	Yes
Dev	528	4560	Yes
Test (Male)	255	2320	Ground Truth
Test (Female)	268	2284	Ground Truth
Test (Total)	523	4604	Ground Truth

Table 7.1: Wikigender's Dataset Splits.

[?]. Hence, to generate training and development data for supervised NRE models, we use the distant supervision assumption: we obtain *(head, relation, tail)* pairs from the cross-domain KB DBPedia and assume that all sentences we encounter which contain both the *head* and *tail* entities expresses the relation given from DBPedia [?]. We use the distant supervision assumption because it provides a scalable means to find data for supervised NRE models.

Although using distant supervision is an effective way of generating a large dataset, it introduces noise in our data in cases where the assumption does not hold. However, there exist various effective ways to mitigate the effect of this noise [?, ?, ?, ?, ?].

We compose Wikigender such that, for each relation corresponding to a sentence, the sentence is taken from the article written about the head entity and the tail entity is found by querying the head entity's DBPedia page for the aforementioned relations. We only take head entities which have: (1) corresponding tail entities for *spouse*, *hypernym*, *birthDate*, and *birthPlace*, and *gender* on DBPedia and (2) at least one sentence mentioning both the head entity and the tail entity in the head entity's article for each of the *spouse*, *hypernym*, *birthPlace*, and *birthDate* relations.

We find extractions for the relations *hypernym*, *spouse*, *birthDate*, and *birthPlace* and compare the prediction accuracies for each. We choose these four relations because it is expected that *birthDate* and *birthPlace* are relatively gender-neutral relations compared to spouse and hypernym. Consequently, we expect that if OpenNRE implicitly contains gender bias, it will have skewed predictions for the *spouse* and *hypernym* relations and predictions that are not skewed for *birthDate* and *birthPlace*.

7.2.2 Test Set

We partition the test set into two subsets: one with sentences from female articles, and one with sentences from male articles (see Table 7.1). To generate test data, we collect distantly supervised data. However, as noted earlier, some sentences will be noisy. Evaluating models on noisy data is unfair since the model could be penalized for correctly predicting the relation is not expressed in the sentence. Thus, we had to obtain ground truth labels.

To find the ground truth, we collect annotations from AMT workers. We asked these workers to determine whether or not a given sentence expressed a given relation. If the majority answer was no, then we labeled that sentence as expressing no relation. (We denote no relation as NA in our publicly released dataset.) Each sentence was annotated by three different workers. Each worker was paid 15 cents per annotation. We only accepted workers from England, the US or Australia and with HIT Approval Rate $> 95\%$ and Number of HITs > 100. We found the pairwise inter-annotator agreement as measured by Fleiss' Kappa [?] κ to be 0.44, which is consistent across both genders and signals moderate agreement. We note that our κ value is affected by asking workers to make binary classifications, which leads to a relatively lower degree of agreement due to a strong baseline. We also found the pairwise inter-annotator agreement to be 84%.

	Spouse			Birth Date			Birth Place			Hypernym		
	M	F	**Diff**	M	F	Diff	M	F	Diff	M	F	Diff
BiRNN +ATT	.613	.326	**.286**	.788	.755	.033	.473	.496	-.022	.223	.340	-.117
+AVG	.620	.349	**.271**	.802	.746	.055	.461	.496	-.034	.226	.327	-.101
PCNN +ATT	.592	.324	**.267**	.765	.755	.009	.441	.454	-.012	.215	.326	-.110
+AVG	.602	.331	**.271**	.765	.743	.021	.453	.473	-.019	.211	.326	-.115
CNN +ATT	.585	.324	**.260**	.718	.703	.014	.449	.450	-.000	.203	.311	-.107
+AVG	.613	.341	**.271**	.768	.755	.013	.409	.412	-.002	.214	.331	-.116
RNN +ATT	.613	.349	**.264**	.791	.755	.036	.469	.492	-.022	.223	.332	-.109
+AVG	.613	.349	**.264**	.798	.755	.043	.465	.477	-.011	.229	.334	-.105

Table 7.2: Results from running combinations of encoders and selectors of the OpenNRE model for the male and female genders of each relation. Columns labeled M give $P(\hat{Y} = 1|A = male, Y = 1)$, and columns labeled F give $P(\hat{Y} = 1|A = female, Y = 1)$. Columns labeled Diff give $P(\hat{Y} = 1|A = male, Y = 1) - P(\hat{Y} = 1|A = female, Y = 1)$, where a positive difference means a higher prediction accuracy for male entities.

7.2.3 Model Evaluation

We evaluate NRE models from a popular open-source code repository called OpenNRE. OpenNRE models combine methods including usage of selective attention to reduce the weight given to noisy sentences at prediction time [?] as well as methods to reduce noise at an entity-pair level [?] and innovations in adversarial training of NRE models [?]. We evaluate models with every combination of four encoders (PCNN, CNN, RNN, and Bi-RNN) and two selectors (Attention and Average) for a total of 8 models.[2] It should be noted that a PCNN is simply a CNN which has a piecewise max-pooling operation [?].

To evaluate gender bias in the models, we use a metric proposed for measuring bias in the literature of bias and social responsibility but not often used in NLP called Equality of Opportunity (EOP) in Supervised Learning [155]. We use EOP to ensure that NRE models have similar accuracies on male and female article sentences. Equality of Opportunity is defined in terms of the joint distribution of (X, A, Y), where X is the input, A is

[2] We performed Grid Search to determine the optimal hyperparameters. We set $\eta = 0.5$, *batch size*$= 160$ and *sliding window size*$= 3$ (for CNN and PCNN).

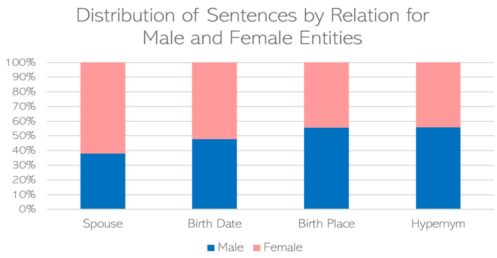


Figure 7.1: Proportion of sentences corresponding to a given relation over total sentences extracted to Wikigender for each entity.

a protected attribute that should not influence the prediction, and Y is the true label. In our case $A \in \{male, female\}$, because gender is our protected attribute and we assume it to be binary. We evaluate EOP on a per-relation, one-versus-rest basis. Thus, we calculate one EOP where *spouse* is the positive class and all other classes are negative; in this case, $Y = 1$ corresponds to the true-label being spouse and $Y = 0$ corresponds to the true label being *hypernym*, *birthDate*, *birthPlace*, or *NA*. We then do another calculation for each relation where $Y = 1$ corresponds to that relation being expressed and $Y = 0$ corresponds to any other relation being expressed. Then, a predictor satisfies EOP if and only if the following probability is satisfied for each relation:

$P(\hat{Y} = 1 | A = male, Y = 1) = P(\hat{Y} = 1 | A = female, Y = 1)$

We expand this probability equation to:

$$\frac{P(\hat{Y} = 1, A = a, Y = 1)}{P(A = a, Y = 1)}$$

, where $a \in \{male, female\}$.

Note that this equates to the number of true positives for a given gender and a given relation divided by the total number of sentences that express that relation for that gender, so we calculate the above probabilities using these statistics.

7.3 Results

7.3.1 Wikigender

In our creation of Wikigender, we performed some statistical analysis on the Wikipedia data we obtained. We build on the work of [?], who discover that a higher proportion of Wikipedia Infoboxes on Wikipedia pages of female entities have spouse information than Wikipedia Infoboxes on Wikipedia pages of male entities. However, Figure 7.1 demonstrates a further discrepancy: that amongst articles for female and men which contain *spouse* information, articles written about females mention females' spouses far more often than articles written about men. Additionally, we show that amongst female and male articles we sampled, *hypernyms* are mentioned far more often in male than female articles.

That female articles mention the females' spouses more often than male articles indicates gender bias in Wikipedia's composition; authors do not write about the two genders equally.

7.3.2 Model Evaluation

We find that not all combinations of encoders and selectors satisfy Equality of Opportunity, especially for the *spouse* or *hypernym* relations, as seen in Table 7.2. There are negligible differences in accuracy for *birthPlace* and *birthDate*, which we expect to be gender-neutral relations. However, there are significant discrepancies in prediction accuracy between sentences for male and female entities for the *spouse* and *hypernym*, which may suggest that the OpenNRE model picks up on subtle gender biases found in the training data. From Table 7.2, it is clear that the probability that a prediction for spouse is correct when the head entity is female is much higher than that same probability when the head entity is male. Since the sentence input for the model is from the head entity's article, this means that NRE models were much more likely to predict the spouse relation correctly when predicting on sentences taken from female articles versus sentences taken from male articles.

We also find that there is little discrepancy in gender bias in predictions produced by models with attention and average selectors (Table 7.2). We do find that the BiRNN and RNN encoders exhibit, on average, lower gender bias than the PCNN and CNN encoders. Furthermore, all models have the highest prediction accuracy for the *birthDate* and have the lowest for the *hypernym*.

7.3.3 Model Evaluation with Debiased Embeddings

Prior work has noted that word embeddings can contain gender biases [118, 119, 120] and also that Wikipedia articles themselves contain some biases in that female entities are often written about in a more sexualized way than male entities, among other things [?].

In order to account for the possibility that gender bias arises as a consequence of stereotypes found in Wikipedia articles rather than OpenNRE, we also train OpenNRE using hard-debiased embeddings [118] and give the results in Appendix B. However, we find that the performance of the model does not differ significantly whether or not we use debiased embeddings. This suggests two possiblitites: 1) OpenNRE makes biased predictions in spite of gender-neutral word embeddings, and we encourage future work to delve deeper into causes. 2) The debiasing technique used [3] does not actually remove gender bias from word embeddings [132], and it is inconclusive as to whether or not gender bias arises from the word embeddings or the OpenNRE model.

7.4 Conclusion

In our study, we create Wikigender: the largest dataset for gender bias evaluation to date across all NLP tasks to our knowledge. We train OpenNRE models on the Wikigender dataset and test them on gender-separated test sets. We notice that there is a significant difference in accuracy for the spouse relation between the male and female genders. Our results indicate that OpenNRE models may implicitly contain gender bias, although our study is preliminary and meant to be a first step towards understanding bias in NRE models. We encourage future work to build on our results.

7.5 Acknowledgments

The content of this chapter is the result of a collaboration with Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, and Diba Mirza.

Part II

Hate Speech

Chapter 8

Research Background

Social media has become a ubiquitous, powerful communication tool. On one hand, it enables people to share information, provides a framework for support during a crisis [174], aids law enforcement agencies [175] and more generally facilitates insight into society at large. However, it has also facilitated anti-social behavior including online harassment, trolling, cyberbullying, and hate speech. In a 2017 Pew Research Center study[1], 66% of Internet users had witnessed some form of online harassment, with 39% revealing that they had seen someone targeted by aggressive behavior such as sustained harassment, physical threatening, or stalking.

In this part of the thesis, we focus on speech that denigrates a person because of their innate and protected characteristics, which is also known as *hate speech*. While there is no consensus on the definition of hate speech, prior work has shown that people are primarily bullied for their *perceived or actual* ethnicity, behavior, physical characteristics, sexual orientation, class or gender [176]. Targeting a community or individual because of their immutable or prominent characteristics slowly eradicates feelings of safety and security [177, 178].

Prior work has presented evidence that social media can be used to obtain valuable data that incorporates facets of the virtual and physical worlds of bullying [179]. We

[1] http://www.pewinternet.org/2017/07/11/online-harassment-2017/

choose Twitter because it provides a platform for open discourse and a cross-section of the general public, with 328 million monthly active users in 2017 [180].

The current literature that pertains to hate speech can be classified into three main areas:

Anti-social behavior. As early as 1997, Spertus [181] introduced some classes of expression to classify online flames including second-person rules, profanity, condescension, epithets, insult rules and polite and praise rules. Cyberbullying has been studied on numerous social media platforms, *e.g.,* Twitter [176] and YouTube [182]. In 2012, Warner and Hirschberg detected anti-semitic speech on Yahoo News and adopted a definition of hate speech as “abusive speech targeting specific group characteristics, such as ethnic origin, religion, gender, or sexual orientation” [183]. Other work has focused on detecting personal insults, profanity, and offensive language [184]. Sood *et al.* show that users can circumvent profanity detection systems by using misspellings and abbreviations of insults and that profanity contexts can affect community tolerance to such insults [185]. To detect offensive language, Razavi *et al.* leverage statistical models and rule-based patterns [186], while Xiang *et al.* exploit linguistic regularities via statistical topic modeling [187], and Burnap and Williams use text parsing to extract typed dependencies, which represent syntactic and grammatical relationships between words [188]. Vulgar language and profanity are modeled as a linguistic style in Twitter using a bootstrapping approach [187] while Xu *et al.* study teasing in messages that represent (possibly less severe) bullying incidents [189]. On the other hand, othering language, which divides people into us and them in racist comments, is investigated as part of hate speech classification in [188].

Hate speech detection. A proposed solution for mitigating hate speech is to design automated detection tools with social content moderation. A recent survey outlined eight categories of features used in hate speech detection [190]: simple surface [145, 188, 191], word generalization [192, 193], sentiment analysis [194, 195], lexical resources [196, 191, 188], linguistic features [188, 191, 192], knowledge-based features [182], meta-information [145, 197], and multi-modal information [197, 192]. Despite the body of work on hate speech detection, it is still a difficult, unsolved problem.

While any deployed classifier may use different types of features, the classification approach mainly entails supervised learning. More recently, Waseem and Hovey analyze the impact of extra-linguistic features such as gender and geographic location in conjunction with character n-grams for hate speech detection [145]. Along with capturing othering language in their classifier, Burnap and Williams improve classification by using dependency relationships and inspecting multiple attacked characteristics in the same content, e.g., hate speech that could fall into both categories of race and sexual orientation, which mirrors intersectionality of hate crime [188]. Nobata et al. develop an NLP-based classifier, incorporating n-grams, linguistic, syntactic and distributional semantics, that outperforms deep learning hate speech classification approaches [191]. On the other hand, features including topics determined from image captions and visual features are leveraged to identify instances of images that could trigger hate speech in [192].

Hate speech characterization. The characterization and correlation of hate speech with contributing factors has recently received attention. Factors include on-the-ground "trigger" events, *e.g.,* terrorist attacks [198], crime [199], and news [200].

In this part of the thesis, we aim to bridge the gap between the hate speech characterization and detection research communities. By characterizing and understanding the nature of hate speech, the design of automated hate speech detection systems is driven by the resultant insights and becomes more impactful. We begin by adopting the definition of hate speech along the same lines of prior literature [200, 201] and inspired by social networking community standards and hateful conduct policy [202, 203] as "*direct* *and serious attacks on any protected category of people based on their race, ethnicity, national origin, religion, sex, gender, sexual orientation, disability or disease*". The rest of this thesis is organized as follows. Chapter 9 investigates the distinctive characteristics of hate instigators and targets in terms of their profile self-presentation, activities, and online visibility. In Chapter 10, we present the first extensive study that explores different forms of hate speech based on the target of hate. Finally, we note that hate speech does not just represent individualistic efforts but also can take a form of organized behavior manifested as hate groups. We discuss the specific nature of this community efforts in Chapter 11.

Chapter 9

Hate Speech Instigators and Their Targets

In this chapter, we present the first comparative study of hate speech instigators and target users on Twitter. Through a multi-step classification process, we curate a comprehensive hate speech dataset capturing various types of hate. We study the distinctive characteristics of hate instigators and targets in terms of their profile self-presentation, activities, and online visibility. We find that hate instigators target more popular and high profile Twitter users, and that participating in hate speech can result in greater online visibility. We conduct a personality analysis of hate instigators and targets and show that both groups have eccentric personality facets that differ from the general Twitter population. Our results advance the state of the art of understanding online hate speech engagement.

9.1 Introduction

Prior studies have focused on online hate speech detection [190] and characterization, e.g., effect of banning hate speech [204]; on-the-ground events that are triggered by hate speech [198, 199, 205]; and semi-organized raids by instigators to cripple hate speech detection technology [200]. Despite this work, little is known about online hate speech actors, including hate speech instigators and targets.

We present the first comparative study of online hate speech instigators and targets. We curate a dataset of 27,330 hate speech Twitter tweets and extract 25,278 instigator and 22,287 target accounts. Our work seeks to answer the following research questions:
RQ1: How do hate instigator and target account characteristics, online visibility, and perceived demographics differ from each other and from generic Twitter account holders?
RQ2: Are there key personality differences between hate speech instigators, targets and general Twitter users?

Due to the lack of public hate speech datasets that include labeled roles of instigators and targets, we curate our own dataset for what we coin "*Peer to peer*" hate speech. This chapter presents the following contributions:

- We present the first comparison of hate instigators, targets and general Twitter users in terms of profile self-presentation, Twitter visibility, and personality traits.

- We provide a compressed lexicon of Hatebase (the world's largest hate expression repository) for hate speech researchers, comprised of 51 terms likely to result in hate speech content across eight different hate classes. We outline a method of semi-automated classification that could be used for directed explicit hate speech data curation. We curate a dataset of 27,330 hate speech tweets, which we make publicly available for other researchers.[1]

- We examine the visibility of Twitter users through multi-variant regression models and controlling for variables that can impact visibility measures. Because visibility measures exhibit skewed distributions, we adopt quartile regression technique to analyze the data in quartiles.

[1] The lexicon and the dataset are available here: `https://github.com/mayelsherif/hate_speech_icwsm18`

Our study yields multiple important findings. First, hate targets often have older accounts while instigators often have younger accounts. Compared to general users, both instigators and targets are more active in terms of becoming friends with others, posting tweets, and populating profile content. Targets include 60% and 40% more verified accounts than instigators and general users, respectively. Even when controlling for variables that can impact visibility measures, we find that higher visibility and participation in hate are correlated. More visible Twitter users (with more followers, retweets and lists) are more likely to become targets of hate. Interactions between instigators and targets perceived as *male* are more likely between younger instigators and older targets, while the interactions between instigators and targets perceived as *female* are more likely between older instigators and younger targets. Users perceived as female are less engaged in hate discussions and male to male hate is predominant. Finally, the personality traits of instigators and targets span both the physical and digital worlds: both hate speech instigators and targets share some personality traits such as suspiciousness, low emotional awareness, and high anger and immoderation, which differ from personality traits of the general Twitter user population.

9.2 Related Work

Most closely related to our work are [206, 207, 208, 176, 145]. Chatzakou *et al.* [207] study the users of tweets with the #Gamergate hashtag. Similar to our results, they found that these users tend to have more friends and followers, and are generally more engaged than random users. Chatzakou *et al.* [206] study the properties of bullies and aggressors and employ supervised machine learning to classify Twitter users into four classes: bully, aggressive, spam, and normal. In contrast to their dataset, our dataset is more diverse and not biased towards specific types of hate speech. Moreover, we compare the characteristics of hate instigators and the targets of hate from multiple perspectives and show that, even when controlling for features that capture the activity level of the users, both hate instigator and target users are more likely to get attention on Twitter, *i.e.,* they obtain more followers, are retweeted and listed more.

Alternatively, [208] find that prior negative mood and the context of the discussion can combine to double participants' baseline engagement in trolling behavior. While the authors only used sentiment analysis to investigate mood, we incorporate a full analysis of the Big Five personality traits. In addition, we study the personality traits of both instigators and targets and compare results to a random sample of general Twitter users. Silva *et al.* [176] identified hate speech and hate targets on Twitter and Whisper by searching for sentence structures similar to "I <intensity> hate <targeted group>" and differentiate hate based on the innate characteristic of targets, *e.g.,* class and ethnicity. They find that the top targeted groups are primarily bullied for their ethnicity, behavior, physical characteristics, sexual orientation, class, or gender. However, when we analyze targets, we do not extract target groups using sentence structures. We identify the actual accounts of hate targets on Twitter, *i.e.,* those that are explicitly mentioned by hate instigators, thus the tweets are considered a personal hateful attack (Directed hate). Therefore, our analysis provides a unique lens to analyze characteristics of target accounts. The distinction between hate aimed at a specific individual or entity (Directed hate) and a group of people sharing a protected characteristic (Generalized hate) can have implications related to free speech policy and has been discussed in depth in [209].

9.3 Preliminaries

We define the following entities:

- A **hate tweet** is an explicit directed tweet that contains one or more hate speech terms used against a Twitter account holder. An example from our dataset is: "*@usr n*gger f*ck u igger n*gger n*gger n*gger.*"[2] This tweet is explicit because of the word "n*gger;" it is directed because it targets a specific account (@usr).[3]

- A **hate instigator (HI)** is a Twitter account that posts one or more hate tweets.
- A **hate target (HT)** is a Twitter account targeted by a hate tweet and explicitly mentioned in the tweet using the mention sign (@), e.g., *usr* in our example. We note that role labels are not mutually exclusive in our dataset; a HI account may be a HT in another hate tweet.

Xu et al. discuss the different roles accounts can exhibit in the cyberbullying context such as a bully, victim, assistant, defender, bystander, reinforcer, reporter, and accuser [179] and the challenges associated with role labeling. The role of intermediaries and points of intervention such as law enforcement, public figures, media, bystanders, NGOs, and educators is discussed in [210]. We note that the definition for what constitutes a HI and a HT is by nature contextual depending on a specific online conversation. It is worth mentioning that prior literature [208] has discussed the causes of trolling behavior in online conversations and showed that prior mood and discussion context increases the likelihood of trolling. Additionally, these insights were used to develop a predictive model that suggests that ordinary people can, under the right circumstances, behave like trolls.

9.4 Data and Methods

Despite the existence of a body of work dedicated to detecting hate speech [190], accurate hate speech detection is still extremely challenging [211]. A key problem is the

[2] We replace select vowels with the star (*) character in obscene language.

[3] We anonymize all user mentions by replacing them with *@usr*.

lack of a commonly accepted benchmark corpus for the task. Each classifier is tested on a corpus of labeled comments ranging from a hundred to several thousand [182, 195, 193]. Despite the presence of public crowdsourced slur databases [212, 213], filters and classifiers based on specific hate terms have proven to be unreliable since (i) malicious users often use misspellings and abbreviations to avoid classifiers [185]; (ii) many keywords can be used in different contexts, both benign and hateful; and (iii) the interpretation or severity of hate terms can vary based on community tolerance and contextual attributes, and (iv) online harmful behavior is often implicit in nature or exhibited superficially or ambiguously and may require additional contextual information to be detected. Another option for collecting a dataset is filtering comments based on hate terms and annotating them. This is challenging because (i) annotation is time consuming and the percentage of hate tweets is very small relative to the total; and (ii) there is no consensus on the definition of hate speech [214]. Some work has distinguished between profanity, insults and hate speech [201], while other work has considered any insult based on the intrinsic characteristics of the person (e.g. ethnicity, sexual orientation, gender) to be hate speech related [183].

This annotation process can become even harder for role labeling, i.e., annotating actors as instigators, targets, bystanders [179]. This is particularly challenging for social networking APIs that do not provide the whole thread of the conversation but only a random sample of comments, as in the case of the Twitter Streaming API. In this work, we adopt a definition of hate speech inspired by Facebook's community standards [202] and Twitter's hateful conduct policy [203] as "*direct and serious attacks on any protected category of people based on their race, ethnicity, national origin, religion, sex, gender, sexual orientation, disability or disease.*" To mitigate the aforementioned challenges, we collect our own explicit Twitter hate speech dataset. We describe our semi-automated detection approach for directed explicit hate speech in the following subsections.

9.4.1 Data Collection

(1) Key phrase-based dataset (HS-1%): We adopt a multi-step classification approach. First, we use Twitter's Streaming API[4] to procure a 1% sample of Twitter's public stream from January 1st, 2016 to July 31st, 2017. Due to the sheer volume of Twitter data, our main focus is to curate a relevant and accurate hate speech dataset with minimal amount of noise. We began by inspecting hate speech keyphrases in the Hatebase repository[5], the world's largest online repository of structured, multilingual, usage-based hate speech[6], which has been widely used as a tool to collect hate speech keywords by other researchers such as [176, 201]. Online users can contribute to Hatebase by adding new derogatory words or phrases, their meaning, and language. Hatebase asks users who add terms to classify the term under one or more of the following hate categories: archaic, class, disability, ethnicity, gender, nationality, religion, and sexual orientation (SexOrient). We use Hatebase as a lexical resource to retrieve English hate keyterms, broken down as: 42 archaic terms, 57 class, 7 disability, 427 ethnicity, 13 gender, 147 nationality-related, 38 religion, and 9 related to sexual orientation. After careful inspection and five iterations of keyword scrutiny by human experts, we removed keyphrases that resulted in tweets with uses distinct from hate speech or key phrases that were extremely context sensitive. For example, the word "pancake" appears in Hatebase, but is more commonly used in benign contexts. Since our goal was a high quality dataset, we only included key phrases that were highly likely to indicate hate speech. The result is 8, 8, 2, 12, 4, 11, 4, and 2 keyphrases for the above, respective, hate speech classes.

Despite the qualitative inspection of the keyphrases, when we used the resultant keyphrases to filter tweets from the 1% public stream, non-hate speech tweets remained

[4] Twitter Streaming APIs: https://dev.twitter.com/streaming/overview

[5] Hatebase: https://www.hatebase.org/

[6] We refer to hate speech terms as keyphrases, keywords, hate terms and hate expressions, interchangeably.

in our dataset. To mitigate the effects of obscure contexts and stance on the filtering process, we were in need of a hate speech classifier that could remove non-hate speech tweets. Consider the following two tweets:

(a): "*@usr_1 i'll tear your limbs apart and feed them to the f*cking sharks you n*gger*"

(b): "*@usr_2 what influence?? that you can say n*gger and get away with it if you say sorry??*.

While both of these tweets contain the word "n*gger", the first tweet (a) is pro-hate speech where the hate instigator is attacking *usr_1*; the second tweet (b) is anti-hate speech in which the tweet author denounces the comments of *usr_2*. Thus stance detection is vital to consider when classifying hate speech tweets. To mitigate the effects of obscure contexts and stance with respect to hate speech on the filtering process, we used the Perspective API[7] developed by Jigsaw and the Google Counter-Abuse technology team, the model for which is comprehensively discussed in [215].[8]

The Perspective API contains different models of classification including: toxicity, attack of commenter, inflammatory, and obscene, among others. When a request is sent to the API with specific model parameters, a probability value [0, 1] is returned for each model type. For our datasets, we focus on two models: `toxicity` and `attack_on_commenter`. The `toxicity` model is a convolutional neural network trained with word-vector inputs. It measures how likely a comment will make people leave a discussion. The `attack_on_commenter` model measures the probability a comment is an attack on a fellow commenter and is trained on a New York Times dataset tagged by their moderation team. After inspecting the `toxicity` and `attack_on_commenter` scores for the tweets filtered by the Hatebase phrases, we found that a threshold of 0.8 for `toxicity` scores and 0.5 for `attack_on_commenter` scores yielded a high quality dataset.

[7] Conversation AI source code: https://conversationai.github.io/

[8] We also experimented with classifiers including [201] but found Perspective API to be empirically better.

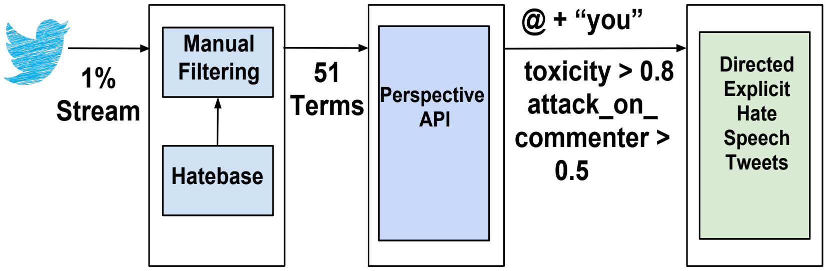


Figure 9.1: Flowchart of the filtering process used to obtain our dataset.

As a final step to ensure that the resultant tweets attacked a specific Twitter user, we took the remaining tweets in our hate dataset and retained only those tweets that both mention another account (@) and that contain second person pronouns (e.g., "you", "your", "u", "ur"). The use of second person pronouns has been found to occur with high prevalence in directed hostile messages [181]. The result of applying these filters is a high precision hate speech dataset of 27,330 tweets in which HIs use explicit Hatebase expressions against HTs. Figure 9.1 depicts the filtering process along with our workflow.

(2) General dataset (Gen-1%): To provide a larger context for interpretation of our analyses, we compare data from the HS-1% dataset with a random sample of all general Twitter accounts. To create this dataset, we use the Twitter Streaming API to obtain a 1% sample of tweets posted per day within the same 18 month collection window and extract the union set of users who posted them. We then remove accounts appearing in the HS-1% dataset, and randomly sample 60K of the remaining users. To mitigate the bias towards more active users, we sample from the union set of users to ensure equiprobable selection of all users, regardless of activity level. While we try our best to remove all the bias, we acknowledge the possibility that this set might include some HIs and HTs. However, later our results show that this bias is likely to have have little impact because we observe significant differences between characteristics of HIs and HTs compared to the general dataset.

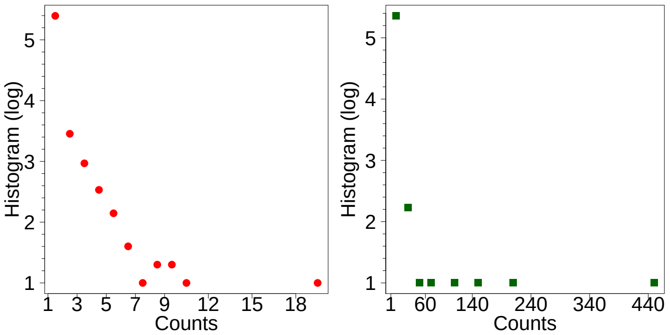
	Total Unique Users		Suspended		Deleted	
HS Type	HI	HT	HI (%)	HT (%)	HI (%)	HT (%)
Archaic	169	169	8.3	11.2	4.1	4.1
Class	849	837	10.0	7.3	4.9	4.4
Disability	8,044	7,930	11.8	6.7	5.7	4.3
Ethnicity	2,073	2,045	18.8	11.3	6.6	5.2
Gender	13,195	13,340	9.4	5.7	5.6	4.7
Nationality	78	79	9.0	11.4	6.4	3.8
Religion	45	47	13.3	19.1	13.3	2.1
SexOrient	3,638	3,584	15.3	9.0	6.9	6.0
HS-1%	25,278	22,857	12.8	8.3	6.5	5.7
Gen-1%	60,000		5.2		3.2	

Table 9.1: Suspended and deleted accounts for all datasets.

Table 9.1 shows the number of users in each of our datasets. In total, our dataset includes 25,278 hate instigators and 22,857 targets. The table shows the quantity of hate tweets for different hate classes.

The number of keywords used for identifying each class of hate can have an impact on the number of detected HIs and HTs. However, we observe that some classes with fewer keywords, such as *gender*, *disability* and *sexual orientation*, with 4, 2 and 2 keywords, have a higher contribution to our dataset, with 52%, 32% and 14% of HIs. This shows the prevalence of these hate keywords on Twitter.

Table 9.1 also shows the percentages of suspended and deleted accounts. The Twitter API returns an error message when the user account is suspended or the user is not found. According to Twitter, account suspensions occur when the account is spam, its security is at risk, or it is engaged in abusive tweets or behaviors. Twitter accounts that are not found (deleted) occur when the user does not exist. This error could arise for a variety of reasons: the user deactivated their account, the account was permanently deleted after thirty days of deactivation, etc. We label users that no longer exist as *deleted*. On average, suspended accounts comprise 12.8% of instigators, 8.3% of targets, and 5.2%



(a) Tweets by instigators (b) Tweets against targets

Figure 9.2: Frequency of hate tweets in HS-1%.

of general Twitter users. Additionally, on average, deleted accounts comprise 6.5% of instigators, 5.7% of targets, and 3.2% of general Twitter accounts. Our findings show that instigators and targets are more likely to have their accounts suspended or deleted than general Twitter users, with instigators as the most likely.

Across each hate class, approximately 5% of accounts are deleted. The only exception is the *Religion* class, where 13% of hate instigator accounts are deleted. However, this may be the result of the small sample from this class. Interestingly, it seems Twitter is more successful in detecting hate related to *Ethnicity*, *SexOrient* and *Religion* as these categories have the highest number of suspended instigator accounts, with about 19%, 15% and 13% of the instigators in these classes being suspended.

Many account holders in HS-1% either post more than one hateful tweet, or are hate targets more than once. Further, we identify 2,077 (approximately 5%) accounts that are both hate instigators and targets. Figure 9.2a illustrates the logarithmic histogram for the number of hate tweets posted by each instigator account. In our HS-1% dataset, about 10% of instigator accounts have posted more than one hate tweet. In particular, 2,014 accounts posted two, 285 posted three, and one account posted 20 hateful tweets. Figure 9.2b illustrates the histogram representing the number of hate tweets against other accounts. Approximately 11% of accounts are mentioned in more than two tweets, while two specific accounts are mentioned in 449 and 210 hate tweets.

Human-centered dataset evaluation.

We evaluate the quality of our final dataset by incorporating human judgment using Crowdflower. We provided annotators with a class balanced random sample of 1000 tweets.[9] To aid annotation, all annotators were provided a set of precise instructions. This included the definition of hate speech according to the social media community (Facebook and Twitter) and examples of hate tweets selected from each of our eight hate speech categories. Then, for each tweet, we asked annotators two questions: (1) whether the tweet is hate speech, and (2) whether the tweet is a direct attack towards the account mentioned in the tweet. To improve the quality of responses, before assigning a task to annotators, we asked them five test questions with already known responses. If they could not answer at least 80% of these questions correctly, we identified them as unreliable annotators and removed them from the task. Each tweet was labeled by at least three independent Crowdflower annotators.

Using MACE [216][10], an unsupervised Bayesian annotation model, we found that annotators were predicted to have labeled 89.1% of the tweets as hate speech and 71.6% of tweets as an attack towards the mentioned account. We then evaluated the inter-annotator reliability by measuring the agreement percentage of annotators for each of the questions. We found that the agreement percentage for the first question is 92.8%, and for second question is 82.6%. These results shows that our hate speech dataset is reliable with minimal noise.

[9] We used a random sample of 1000 tweets to keep the monetary cost manageable.

[10] We experimented with different parameter settings, and found that the predictions remained constant. Thus, we report our results from MACE with default parameter settings.

9.4.2 Measures

We adopt several measures based on prior work to answer our research questions. To compare the account characteristics of HIs and HTs, we investigate whether users have a profile image, set a geo-location and a timezone, whether the account is verified, and the length of the profile description. We study the number of tweets and retweets, friends, followers, and whether the account is enlisted. Similar to Nilizadeh *et al.* [217], we differentiate accounts by *perceived*, as opposed to *actual*, user characteristics. This is because we can only study how an account holder chooses to represent him/herself, i.e., through a profile photo, and cannot determine their actual characteristics. When we look at account characteristics of instigators and targets, we study the perceived account characteristics (e.g., gender and age) that are visible in their account. Nilizadeh *et al.* [217] studied the association between perceived gender and measures of online visibility. Recent work that investigates the inference of actual user characteristics from online content in social networks, a.k.a. *user profiling*, include age, gender, and occupation estimation [84, 85, 86, 87]. Specifically, we study perceived user age and gender using an automatic facial feature recognition service "Face++" [88].

We predict user gender by extracting first names and comparing them with those listed in the 1900 – 2013 U.S. Census [218, 217]. We leverage the IBM Watson Personality Insights API [219] to quantify the Big Five personality traits for HIs and HTs. The API has been used in prior studies to correlate personality traits with information-spreading [220] and targeted advertising [221].

9.5 Analysis

9.5.1 RQ1: Account Characteristics

Our first objective is to understand the differences of self presentation through profile configurations, activity level, and interaction with other users. To study profile presentation, we analyze whether profile image, location, and timezone are provided by the user; whether the user has enabled the geo-location to be posted along with their tweets; whether the account is verified by Twitter; and the length of profile description in characters. For user activity level, we analyze number of tweets, friends, followers, lists, and retweets. The last three of these indicate how Twitter users interact with an account and are used as visibility measures [217, 222].

All characteristics can be extracted from the meta-data provided with the tweets, except the retweet count. For every user, we count the number of times the user's tweets are reposted in our 1% dataset. Although the obtained retweet counts only represent a subset of the actual retweets, they provide useful insight when comparing different samples. We determine the gender of users by extracting first names and comparing them with first names listed in the U.S. Census dataset obtained from 1900 – 2013 [218]. Some first names are gender-neutral, such as "Pat," which based on the U.S. Census dataset, 40% and 60% of the time has been used for females and males. Similar to other work [218], if a name has a female-to-male ratio larger than 0.95 or smaller than 0.05, we label it as female or male; other names are labeled as 'gender ambiguous'. We are able to extract first names for 53% of HIs, 55% of HTs and 56% of general users. HIs use pseudonyms more than others, which can be an indication of desire to hide their identities. 25%, 23% and 8% of users in the Gen-1% dataset; 35%, 10% and 8% of users in the instigator dataset; and 35%, 12% and 8% of users in the instigator dataset are male, female and gender ambiguous, respectively. Instigator and target datasets include

	Gen-1% users				HIs				HTs			
Statistic	Mean	Min	Max	Median	Mean	Min	Max	Median	Mean	Min	Max	Median
Followers count	932	0	4,589,177	93	1,358	0	1,006,790	259	229,676	0	102,008,153	857
Friends count	408	0	243,937	160	663	0	1,012,412	239	1,897	0	1,698,640	396
Tweets count	4,384	0	570,550	545	14,160	0	4,321,652	3,266	29,559	1	3,644,240	10,902
Listed count	8	0	10,118	0	13	0	7,855	2	755	0	616,271	9
Retweet counts	3	0	13,220	0	30	0	27,390	2	623	0	304,900	10
Account age (years)	3.73	0.09	10.99	3.33	3.67	0.09	10.66	3.22	4.40	0.09	11.37	4.16
len. description (chars)	45	0	164	28	53	0	164	37	63	0	164	49
Profile image	0.95	0	1	NA	0.97	0	1	NA	0.99	0	1	NA
Profile URL	0.23	0	1	NA	0.24	0	1	NA	0.40	0	1	NA
Geo location	0.33	0	1	NA	0.39	0	1	NA	0.51	0	1	NA
Location	0.53	0	1	NA	0.61	0	1	NA	0.69	0	1	NA
Timezone	0.40	0	1	NA	0.52	0	1	NA	0.68	0	1	NA
Verified	0.003	0	1	NA	0.002	0	1	NA	0.12	0	1	NA
	$N = 60,000$				$N = 25,278$				$N = 22,857$			

Table 9.2: Descriptive statistics of our datasets.

10% more male and 13% fewer female users than the Gen-1% dataset, which implies that *users with female account names are less engaged in hate discussions.*

Table 9.2 statistically describes the users in our Gen-1% and HS-1% datasets. Since the distribution of most characteristics is skewed, in addition to mean, the table also shows the min, max and median of values. The table illustrates multiple differences between user types. The t-tests for account age (by year) suggest that, on average, the accounts for HTs are older than those of HIs ($\mu = 4.40$, vs. $\mu = 3.67$) ($t = 32.18$, $p < 0.001$) and generic random users ($\mu = 4.40$, vs. $\mu = 3.73$) ($t = 32.91$, $p < 0.001$). Also, the accounts for HIs are younger than those of general random users ($\mu = 3.67$ vs. $\mu = 3.73$) ($t = 3.33$, $p < 0.001$). We observe that compared to random users, HIs and HTs are more active in becoming friends with others, posting tweets, and providing more content on their profiles.

The t-tests for profile description length (in characters) show that, on average, the descriptions provided by HTs are longer than those for HIs ($\mu = 63$, vs. $\mu = 53$) ($t = 20.14$, $p < 0.001$). The descriptions provided by hate targets and instigators are longer than those of generic random users ($\mu = 63$, vs. $\mu = 45$) ($t = 40.04$, $p < 0.001$), ($\mu = 53$, vs. $\mu = 45$) ($t = 19.56$, $p < 0.001$). These results may suggest that both HIs and HTs are more willing to present themselves.

		HT vs. HI		Gen-1% vs. HT		Gen-1% vs. HI	
	df	X^2	p	X^2	p	X^2	p
Profile image	1	7633	***	672	***	4901	***
Profile URL	1	325	***	1858	***	3546	***
Geo location	1	3.53	0.06	1937	***	1801	***
Location	1	1606	***	1389	***	66	***
Timezone	1	1389	***	4444	***	797	***
Verified	1	99	***	6226	***	4789	***
Gender (name)	1	1318	***	1230	***	21	***
Invalid image	1	2,088,900	***	1,221	***	4,827,400	***
Detected face	1	1,138,200	***	505	***	1,821,700	***
Multiple faces	1	282,530	***	127	***	368,000	***
One face (Male)	1	289,160	***	24,493	***	224,900	***
One face (Female)	1	270,580	***	197,900	***	933,780	***

Note: *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

Table 9.3: Pearson's Chi square tests.

Table 9.3 shows the results of Chi-square tests for the binary variables. In general, HTs reveal more information on their profiles; they are more likely to add image, URL, location and timezone to their profiles compared to both HIs and general Twitter users. There is only one exception where the difference between the distribution of geo-location for HIs and that of HTs is not significant ($p = 0.06$).

Twitter verifies accounts that are of public interest. When accounts are verified, a blue badge appears next to the user's name on their profile.[11] Interestingly, when comparing HIs and HTs, we observe that HTs include significantly more high profile and established users; 12% belong to verified accounts. However, HIs themselves are less likely to have verified accounts, even compared to random general users.

Next, we examine the activity and visibility levels of account holders. We compare these variables by using Mann-Whitney U tests, because they do not follow a normal distribution. These results are provided in Table 9.4. Interestingly, HTs have more

[11] Request to verify an account:
https://support.twitter.com/articles/119135#

	U (HT vs. HI)	U (Gen-1% vs. HT)	U (Gen-1% vs. HI)	p
Followers	321,900K	183,400K	504,620K	***
Tweets	294,930K	190,920K	445,380K	***
Friends	278,670K	316,970K	586,540K	***
Lists	305,450K	221,840K	503,890K	***
Retweets	304,560K	139,270K	369,650K	***

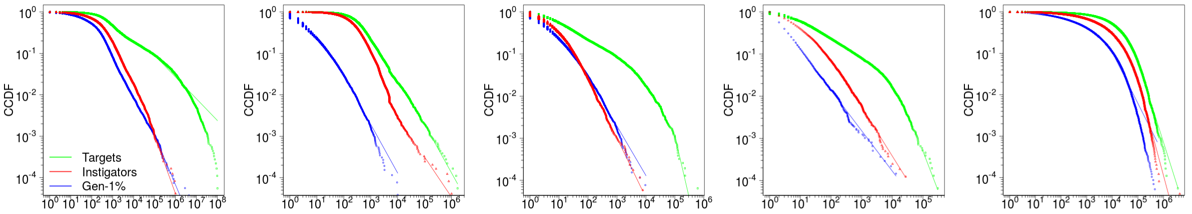
Table 9.4: Mann-Whitney U tests.

friends and post more tweets than both HIs and general users. They also have higher visibility and influence; their median numbers of followers and retweets are larger than those of both HIs and general users.

Twitter's 'List' feature allows users to organize others by creating topical user lists. If some users are known for something, *e.g.,* are computer scientists, then they might be listed by others in "Computer Scientists" list. Organizing Twitter users into lists helps track tweets from those in the list. These lists have been used to accurately identify domain experts [223]. Our results show that targets of hate are listed more often.

Figure 9.3 compares the distribution of the activity and visibility characteristics of HIs and HTs with those from the Gen-1% dataset. This figure shows CCDF plots for variables that exhibit heavy-tailed distributions. Figure 9.3a shows that HTs on average have more followers than both HIs and general Twitter users, while the distribution of followers count for HIs is more similar to that of general Twitter users. Specifically, the difference between HTs and others is more significant for visibility measures including followers, lists and retweet counts.

Visibility: We next examine the visibility of HIs and HTs by controlling for variables that can have an impact on the visibility measures. For example, older accounts have had more time to accumulate followers; following many others usually yields more followers by sheer reciprocity; and posting many tweets can increase the chances to be noticed. Thus, we incorporate the following control variables in our models: account age, number of



(a) Followers count (b) Friends count (c) Listed count (d) Retweets count (e) Tweets count

Figure 9.3: Comparison of account characteristics of HIs, HTs, and general users.

tweets, number of friends, and profile characteristics such as URL, location, image, length of user description, timezone and verified, as well as perceived user gender. We control for profile characteristics and gender because user self-presentation can affect the way people perceive them, and therefore, can have an impact on visibility measures [224, 217].

We select three dependent variables as the main measures of online visibility on Twitter: 'number of followers', 'retweets,' and 'lists.' We apply multiple multivariate regression models and present the results from our Poisson regression model. Linear and negative binomial regression models show qualitatively consistent results, although a couple did not converge.

Since our dependent variables exhibit a skewed distribution, examining the whole population may not capture more nuanced patterns [225]. For example, in Table 9.2, we observe that a hate target account holder has more than 100M followers and this user alone can impact the overall and average statistical results. Thus, we tested multiple statistical methods to account for the skewed distributions. First, we adopt the quartile regression technique to analyze our dataset in each quartile. We divide the data into quartiles based on each dependent variable and apply multivariate regression models. Second, we log transform the dependent variables, considering only those observations with dependent variables not equal to zero, and then run OLS regression. This dataset includes 37,437 observations. Third, we remove the large outliers from our dataset, and re-run the Poisson models. In particular, we removed users with zero followers, lists, and retweets, as well as users with more than 1M followers (571), 100K lists (17), and 100K retweets. This dataset includes 36,860 observations.

Although we include control variables in all models, for brevity, we omit them from the result tables; full tables are available upon request. We add followers count as a control for the retweets and lists count models because more followers may result in being retweeted and listed more. We add lists count as a control for the retweets count model because being listed by many people may result in being retweeted more. We report Incident Rate Ratios (IRRs), the exponentiated coefficients of Poisson regressions, which allow us to compare the rates of variables between HIs, HTs, and general users.

	Hate Targets vs. All				
	Followers count				
	Poisson	0.25 Qrt.	0.5 Qrt.	0.75 Qrt.	1.00 Qrt.
HT	2.68***	0.41***	0.10***	0.05***	2.36***
IRRs	14.64	1.51	1.11	1.05	10.60
	Lists count				
HT	1.93***	0.04	0.08***	0.06***	1.59***
IRRs	6.92	1.036	1.08	1.06	4.92
	Retweet count				
HT	4.06***	2.57***	4.18***	3.76***	3.35***
IRRs	57.94	13.00	65.01	42.98	28.53
Observations	100,346	25,084	25,088	25,086	25,088
Note:			*p<0.05; **p<0.01; ***p<0.001		

Table 9.5: HTs vs. All Poisson Regressions.

Table 9.5 shows the results of Poisson regression for the followers, lists and retweets

	Hate Targets vs. All		
	Log(Followers count)	*Log(Lists count)*	*Log(Retweet count)*
HT	0.764*** (0.016)	0.574*** (0.015)	1.617*** (0.021)
IRRs	2.15	1.78	5.04
Observations	37,436		

Table 9.6: HTs vs. All, OLS Regression on log transfomation of dependent variables.

	Hate Targets vs. All		
	Followers count	*Lists count*	*Retweet count*
HT	1.40*** (0.0002)	0.96 *** (0.002)	3.0*** (0.001)
IRRs	4.068	2.614	20.24
Observations	36,859		

Table 9.7: HTs vs. All, Poisson Regression on the dataset without outliers.

counts comparing HTs vs. the union of HIs and general users. The first column shows the result for the entire sample such that HTs have significantly more followers, are listed and retweeted more than all other users ($p < 0.001$). Particularly, for followers, lists and retweet counts, the HTs have IRRs 14.64, 6.92 and 57.94 times of those of the union of HIs and general users. Tables 9.6 and 9.7 illustrate that using other methods, such as removing the outliers and using the Possion model or using OLS regression on the log transformation of the dependent variables, we still obtain the same findings – HTs have significantly more followers, and are listed and retweeted more than all other users ($p < 0.001$).

	Hate Targets vs. Hate Instigators				
	Followers count				
	Poisson	0.25 Qrt.	0.5 Qrt.	0.75 Qrt.	1.00 Qrt.
HT	2.03***	0.21***	0.08***	0.02***	1.82***
IRRs	7.65	1.23	1.08	1.02	6.14
	Lists count				
HT	1.59***	−0.08	−0.01	−0.05***	1.42***
IRRs	4.90	0.93	0.99	0.95	4.15
	Retweet count				
HT	3.15***	1.27***	2.9***	3.12***	2.91***
IRRs	23.32	3.57	18.20	22.76	18.35
Observations	100,346	25,084	25,088	25,086	25,088
Note:					*p<0.05; **p<0.01; ***p<0.001

Table 9.8: HTs vs. HIs Poisson Regressions.

Table 9.8 illustrates that these findings hold even when HTs are compared only with

	Hate Targets vs. Hate Instigators		
	Log(Followers count)	*Log(Lists count)*	*Log(Retweet count)*
HT	0.692*** (0.018)	0.478*** (0.016)	1.181*** (0.025)
IRRs	1.997	1.612	3.256
Observations	37,436		

Table 9.9: HTs vs. Instigators, OLS Regression on log transfomation of dependent varaibles.

	Hate Targets vs. Instigators		
	Followers count	*Lists count*	*Retweet count*
HT	1.420*** (0.0002)	1.036*** (0.002)	2.641*** (0.001)
IRRs	4.136	2.817	14.029
Observations	36,859		

Table 9.10: HTs vs. Instigators, Poisson Regression on the dataset without outliers.

HIs ($p < 0.001$). For example, for followers, lists, and retweet counts, the hate targets are expected to have incidence rates 7.65, 4.90 and 23.32 times those of the hate instigators. Results in Tables 9.9 and 9.10 confirm that these findings are not the artifacts of the used method.

These results suggest that regardless of user activity level, profile self-presentation, and gender, more visible Twitter users (with more followers, lists, and retweets) are more likely to become target of hate.

Table 9.11 demonstrates the results of models for HIs vs. general users. The coefficients for both overall and quartiles models are positive and larger than one, which indicate that HIs are positively associated with being visible. There is one exception when the dependent variable is followers count. While the least visible HIs (quartile one) are more likely to be followed, in quartile four the HIs have a lower chance to be followed by others. The results in Table 9.12 confirm findings obtained from Table 9.11. Interestingly, results in Table 9.13 are more consistent with the results obtained from the forth quartile in Table 9.11. Investigating more, we found that almost all the removed outliers (567) in terms of followers count are among hate targets. Also, all the removed

Hate Instigators vs. Gen-1%					
	Followers count				
	Poisson	0.25 Qrt.	0.5 Qrt.	0.75 Qrt.	1.00 Qrt.
HI	0.46^{***}	0.23^{***}	0.04^{***}	0.03^{***}	-0.04^{***}
IRRs	1.59	1.26	1.04	1.03	0.96
	Lists count				
HI	0.49^{***}	0.11^{***}	0.12^{***}	0.12^{***}	0.08^{***}
IRRs	1.62	1.11	1.12	1.13	1.09
	Retweet count				
HI	1.98^{***}	2.96^{***}	2.65^{***}	1.37^{***}	1.43^{***}
IRRs	7.26	19.3	14.13	3.92	4.17
Observations	100,346	25,084	25,088	25,086	25,088
Note:			*p<0.05; **p<0.01; ***p<0.001		

Table 9.11: HIs vs. Gen-1% Poisson Regressions.

Hate Instigators vs. Gen-1%			
	Log(Followers count)	*Log(Lists count)*	*Log(Retweet count)*
HI	0.071^{***} (0.017)	0.202^{***} (0.016)	1.054^{***} (0.020)
IRRs	1.07	1.224	2.869
Observations	37,436		

Table 9.12: Hate Instigators vs. Gen-1%, OLS Regression on log transfomation of dependent varaibles.

Hate Instigators vs. Gen-1%			
	Followers count	*Lists count*	*Retweet count*
HI	-0.072^{***} (0.0003)	-0.035^{***} (0.003)	1.274^{***} (0.003)
IRRs	0.93	0.96	3.575
Observations	36,859		

Table 9.13: Hate Instigators vs. Gen-1%, Poisson Regression on the dataset without outliers.

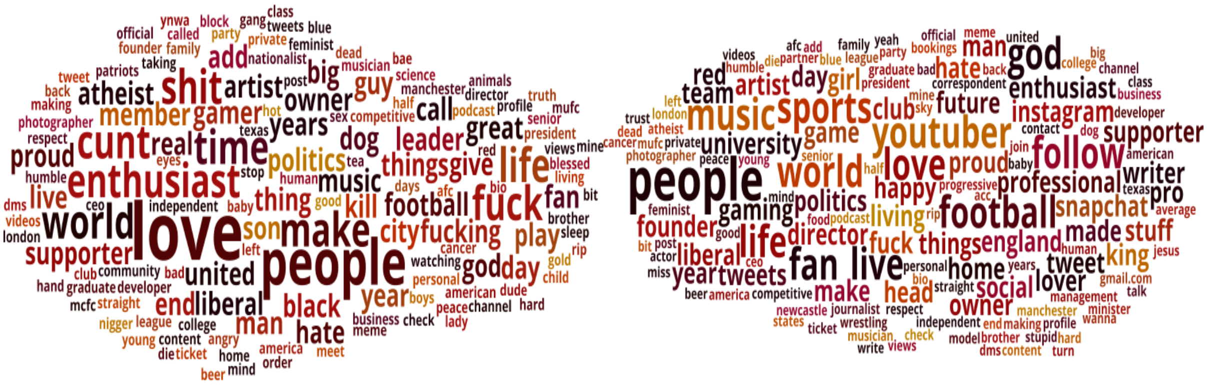
outliers for list counts are from hate targets. This shows that there are still outliers in the instigator and general datasets. While one can pick different thresholds for each distribution, it makes it hard to measure and compare the results. Thus, we argue that the quartile regression method can better show how outliers impact the overall results.

In Table 9.5, quartile regression reveals that the overall and average results are not just the effects of most visible users, and in each quartile, the HTs are more visible than

HIs and general users. Although the effect of HTs (IRR) increases as one moves from the least visible to most visible users, in almost all quartiles values are larger than one. There is an exception in Table 9.8 when the dependent variable is lists count. While the most visible HTs (quartile four) are more likely to be listed, in other quartiles the HIs are listed more.

Comparing the IRR results with those in Tables 9.5 and 9.8 shows that the differences between the HTs and HIs are significantly higher than those of HIs and general users. *These results also suggest that participating in hate speech and being more visible and popular are related; even when controlling for all mentioned independent variables, both HIs and HTs are more popular and visible than general users.*

Instigator and Target Profile Description: In order to get a sense of who HIs and HTs are and how they tend to describe themselves, we fetch the Profile Description from the meta-data. Out of the 25,278 instigators and 22,857 targets, 80% and 76% had a non-empty profile description, respectively. We train an LDA topic model with 25 topics on the instigator and target profile descriptions and 20 words per topic. Figure 9.4 depicts the results of the LDA topic models for HI and HT profile descriptions. A qualitative analysis of the results indicate the following observations. HIs tend to use the word “love” in benign contexts such as “Music lover” and “love food”. We also note the presence of more profane words in HI profile descriptions such as c*nt, f*ck, and n*gger. Example of profile descriptions containing profane words include “F*ck off you c*nt” and “F*ck Muslims”. We also note the presence of common political terms for HIs and HTs profile descriptions such as “trump”, “maga”, “america”, “proud”, “christian”. For HIs, unique political terms include “conservative”, “supporter”, “altright”, and “patriot” and for HTs, unique political terms include “liberal”, “pro”, and “government”. Another observation is that HTs tend to include more social media info in their profile description by including words such as “email”, “snapchat”, and “inquiries”. Additionally, we



(a) HI profile description

(b) HT profile description

Figure 9.4: Words mentioned in HIs and HTs profile description. Note the presence of more profane words in HI profile descriptions such as c*nt, f*ck, and n*gger and the usage of occupations and interests in HTs profile descriptions such as artists, sports, director, journalist, and enthusiast.

found that HTs tend to be more descriptive in terms of listing their occupations and interests. They use words such as “designer”, “artist”, “founder”, “feminist”, “activist”, “footballer”, “journalist”, “youtuber”, and “author”.

Perceived Demographics: We extract profile image URLs from the meta-data and examine how each group represents themselves through their profile image. After careful considerations of various facial recognition tools[12], we use the Face++ API [88] with a similar approach to [227, 228]. Face++ has the functionality to predict demographic information of a given photo including age, gender, and race. The API returns results when faces are detected, and also indicates when no faces are detected. Faces detected by Face++ can be from images that contain multiple faces, one face, or unidentified faces (faces that have unknown gender and age). We exclude default profile URLs and invalid

[12] In an evaluation of facial recognition tools conducted by [226], Face++ achieved high accuracy for predicting gender with a rate of 92%, and performed on par with other widely used facial recognition with regards to detecting age.

profile URLs and present the image analysis results in Table 9.14. Table 9.3 shows the results of Chi-square tests for the variables discussed in Table 9.14.

	Invalid URLs (%)	Default Image (%)	Detected Face (%)	Multiple Faces (%)	One Face Male (%)	One Face Female (%)
Instigators	70.2	3.0	15.9	5	7.9	2.7
Targets	0.8	1.3	66.3	22	27.2	17.1
Gen-1%	1.1	4.8	67.2	22.4	21.6	34.5

Table 9.14: Profile image URLs categorization.

To evaluate the gender inferences made by the Face++ API, we compare our Face++ gender results to our results for gender by name from the US Census. Of the users that we were able to infer gender by image urls from Face++, we were able to obtain the user's gender by first name from the US Census for 1,433 HIs and 5,309 HTs. We find that the percentage of gender that matched was 82% and 83% for HIs and HTs respectively, which is similar to the accuracies of other studies that utilize Face++ [227, 228].

We observe that HIs are much more likely to have default profile pictures (3%) compared to HTs (1.3%). There is a noticeable difference between the percentage of invalid image URLs for HIs (70%) and HTs (only 0.8%). The URLs are invalid when the accounts are suspended, deleted or the images have changed since our data collection. This may suggest that HIs are more likely to change their profile images than HTs. Table 9.14 also shows that HTs are significantly more likely to have profile images of faces (66%) compared to instigators (about 16%). This may suggest the desire of HIs to remain anonymous. This is consistent with our findings that HIs are less likely to provide names on their profiles.

Table 9.14 also presents the gender of HIs and HTs determined by Face++, for those images with only one detected face. The total number of profile photos with only one face detected by Face++ is 4,023 instigators and 15,161 targets. Consistent with our findings when detecting gender by name, the percentage of male participation in hate

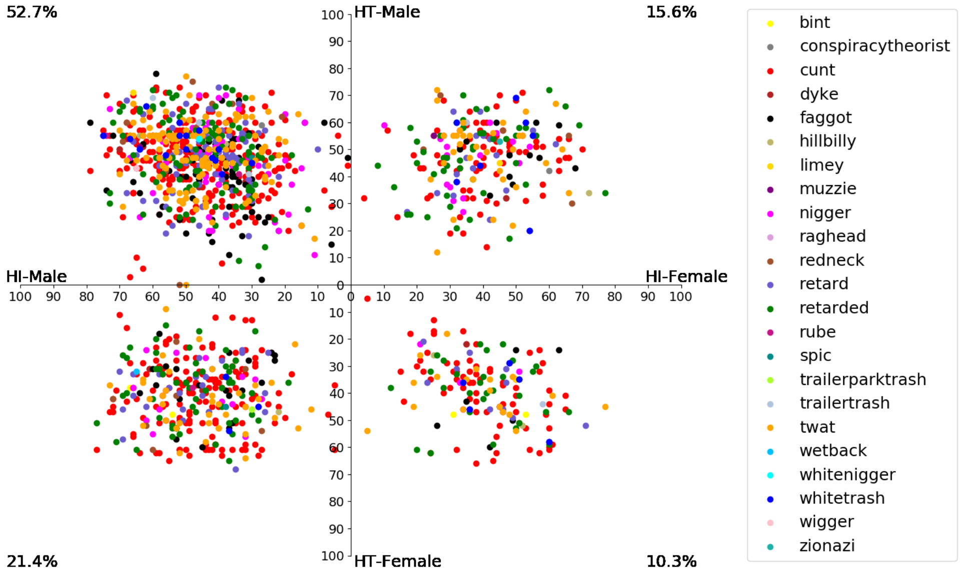


Figure 9.5: Perceived age and gender across hate phrases.

discussions is significantly higher than females for both HIs and HTs.

We then examine the perceived gender and age characteristics of HIs and HTs and the hate speech keyword used in their tweet.[13] We depict the data in Figure 9.5, and indicate the top five keywords used in each quadrant in Table 9.15. Overall, we observe 1,546 points where both HIs and HTs have faces detected, and display the percentages of points in each quadrant. The majority of hate speech (52.7%) occurs between male HIs and male HTs, and the minority (10.3%) between female HIs and female HTs. In Table 9.16, we compare the ages for HIs and HTs in each quadrant. Our findings show that interaction between HIs of both genders and male HTs are more likely between younger HIs (38%) and older male HTs (57%). Contrastingly, interaction between HIs of both genders and female HTs occurs more with older HIs (57%) and younger female HTs (37%). We further visualize this in Figure 9.6, where the majority of hate terms is depicted to be used with HIs older than female HTs and HIs younger than male HTs.

[13] Because we use Face++ for gender detection, we also use it for age categorization in this section to have consistent mappings between age and gender.

HI-M to HT-M		HI-M to HT-F		HI-F to HT-M		HI-F to HT-F	
keyword	(%)	keyword	(%)	keyword	(%)	keyword	(%)
cunt	35.6	cunt	45.3	cunt	34.0	cunt	45.9
twat	15.7	retarded	16.6	retarded	18.7	retarded	18.2
retarded	12.8	twat	11.5	twat	15.4	twat	11.3
faggot	10.1	retard	7.6	nigger	7.1	retard	6.9
retard	9.9	faggot	6.6	white trash	5.8	white trash	6.3

Table 9.15: Top five keywords by quadrant.

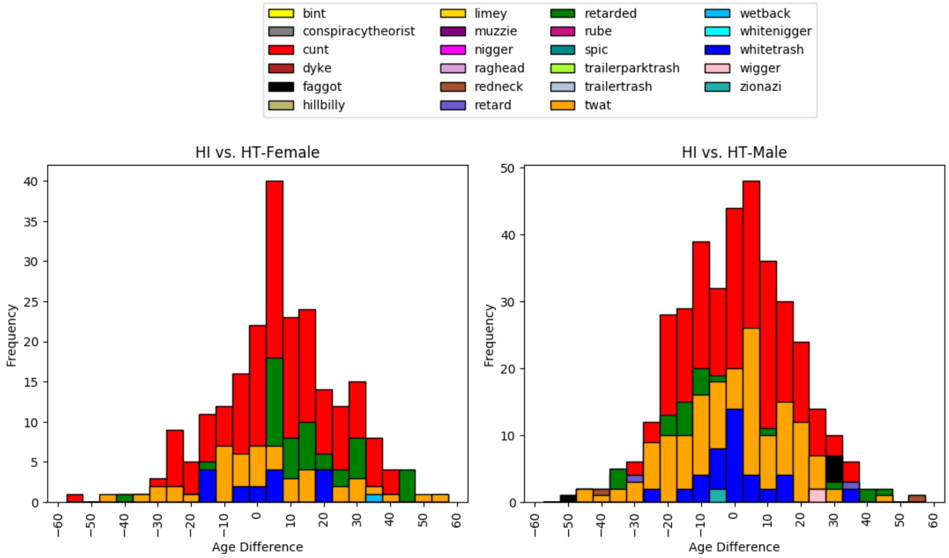


Figure 9.6: Frequency of hate keyterms among HIs and respective genders of HTs across age differences. Positive age differences denote the HIs are older than HTs. Negative age differences denote that HTs are older than HTs.

Quadrant	Total Users	HI > HT (%)	HI < *HT* (%)	HI =HT (%)
HI-M to HT-M	815	41.1	53.7	5.2
HI-F to HT-M	241	29.5	67.6	2.9
HI-M to HT-F	331	61.6	35.6	2.7
HI-F to HT-F	159	47.2	39.0	13.8

Table 9.16: Age Percentages of HIs and HTs for each quadrant.

To study those images with no faces detected by Face++ for instigators and their corresponding targets, we use the IBM Watson Visual Recognition API, which analyzes and classifies content in a given image. We filter out those classes that have less than 75% confidence, and categorize the classes into clusters that contain their semantic sim-

Instigators		
(%)	Exemplars	Clusterings
30.8	dolphin	eel, lemur, pigeon, tiger, swine
14.2	earplug	can, coil, defibrillator, router
13.1	jabot	tapestry, dolls, ring, hosiery, eyeliner
9.8	scarf	cloak, flag, pistol, toupee, tongue
6.4	vehicle	aircraft, artillery, boat, highway, tricycle

Table 9.17: Top five exemplars and clusterings of profile images for instigators.

Targets		
(%)	Exemplars	Clusterings
14.0	owl	bearcat, flatfish, wolf, dolphin, larva
13.0	truck	aircraft, boat, bridge, machinery, wheel
12.9	hatbox	android, beanbag, dictionary, magnifier, utensils
11.7	skylight	bedroom, skyscraper, hall, grotto, ridge
10.0	game	basketball, crowd, rusher, batter, pit

Table 9.18: Top five exemplars and clusterings of profile images for targets.

ilarities. We observe 5,448 classes that were detected for our instigator's profile images and 7,816 classes for targets. We next compute the semantic similarities of the classes by generating word vectors. Since words that occur in the same contexts have similar meanings (as stated by the Distributional Hypothesis), semantically similar words are embedded close to one another in the vector space. To cluster our words, we use Affinity Propagation [229], which takes in a set of pairwise similarities between data points and maximizes the similarity between their exemplar (members of the input set that represent the clusters) and their data points to form clusters. The results of the top five exemplars and their first five clusterings for instigators and targets are presented in Tables 9.17 and 9.18. We explore the clusters in more detail, and find that targets tend to have more images of scenery and entertainment than instigators. On the other hand, instigators have more images of modes of transportation and animals. Furthermore, we conducted a qualitative investigation of the those profile images that were not detected by Face++, but were classified as either ”people”, ”person”, ”man”, ”woman”, ”men”, ”women”, ”male”, or ”female”. We found that those profile images were not representative of the

Twitter account user. For example, some of the images contained basketball players, mannequins/masks, or figurines.

9.5.2 RQ2: Personality Traits

Personality traits have long been shown to affect various human behaviors [230, 231] including health [232], career adaptability [233], risky decision making [234], and consumer preferences [235, 236]. To study the key similarities and differences between the personalities of HIs, HTs, and the general population, we use the Twitter REST API to fetch tweet traces of users. A Twitter user can share content on their profile in three different ways: an original tweet, a reply to a tweet written by another user, or a redistribution of a tweet written by another account (retweeting). Retweets do not necessarily indicate content endorsement but suggest content to be viewed by the retweeter's network. Since retweeting content might not reflect the author's point of view, we only include original tweets and replies as part of our personality analysis. We attempt to fetch the most recent 2000 tweets (excluding retweets) for each account. We use IBM Watson Personality Insights API[14] for our personality analysis.The models reported by the IBM personality service are based on research in the fields of psychology, psycholinguistics, and marketing.[15] The IBM model infers personality characteristics from textual information based on an open-vocabulary approach. This method reflects the latest trend in the research about personality inference [237, 238, 239]. A validation study has been conducted to understand the accuracy of the service's approach to inferring a personality profile. IBM collected survey responses and Twitter feeds from between 1500 and 2000 participants for all characteristics and languages. To establish ground truth, participants took four sets of standard psychometric tests. The average Mean Absolute Error reported for the

[14] https://www.ibm.com/watson/services/personality-insights/

[15] IBM Bluemix. The Science Behind the Service. https://console.bluemix.net/docs/services/personality-insights/science.html#science

	Medians			HI vs. HT		HI vs. Gen-1%		HT vs. Gen-1%		Hellinger distances		
Personality facet	HI	HT	Gen-1%	U	p	U	p	U	p	HI-HT	HI-Gen-1%	HT-Gen-1%
Agreeableness	0.06	0.1	0.4	134,790K	***	47,512K	***	61,130K	***	0.11	0.37	0.27
Openness	0.49	0.51	0.5	152,400K	***	114,760K	0.18	115,840K	***	0.03	0.03	0.04
Emotional range	0.18	0.22	0.38	142,360K	***	77,917K	***	87,490K	***	0.08	0.22	0.15
Conscientiousness	0.02	0.05	0.31	128,370K	***	35,667K	***	55,020K	***	0.18	0.46	0.31
Extraversion	0.23	0.31	0.47	149,410K	***	83,693K	***	88,067K	***	0.04	0.17	0.13

Note: *$p < 0.05$ **< 0.01 ***< 0.001

Table 9.19: Scores and Hellinger distances for the Big Five personality traits of HIs, HTs and general users.

English language was found to be 0.12. The IBM personality traits model is placed at the forefront of personality inference from textual data as indicated by [238, 239].

Since the Personality Insights API requires a minimum of 600 words to obtain statistically significant result estimates, we discard any accounts that do not satisfy this requirement. After discarding suspended and deleted accounts, accounts with statistical insignificance, and accounts with languages other than English, we were able to fetch tweets for a total of 17,951 unique HIs, 17,553 unique HTs, and 12,900 unique general users (pulled from Gen-1%).[16] We use the general users personality results as a means of account sample representation on Twitter. The word count distribution is $(\mu = 11,045.6, \sigma = 7,230.5)$ for HI accounts, $(\mu = 12,316.1, \sigma = 7,308.7)$ for HT accounts, and $(\mu = 8,108.2, \sigma = 7,288.7)$ for accounts in Gen-1%.

The IBM Watson Personality API infers personality characteristics from textual information based on an open-vocabulary approach [219]. The API's machine learning algorithm is trained using scores obtained from surveys conducted among thousands of users along with data from their Twitter feeds. The API provides scores [0, 1] that reflect the normalized percentile score for the characteristic. We analyze the results of the *Big Five* personality model, the most widely used model for generally describing how a person engages with the world. The model includes five primary dimensions: Agreeableness, Conscientiousness, Extraversion, Emotional range, and Openness. Each of these

[16] All sampling errors in our results are less than 0.1.

	Medians			HI vs. HT		HI vs. Gen-1%		HT vs. Gen-1%		Hellinger distances		
Personality facet	HI	HT	Gen-1%	U	p	U	p	U	p	HI-HT	HI-Gen-1%	HT-Gen-1%
Agreeableness - Modesty	0.14	0.19	0.5	137,190K	***	59,460K	***	69,780K	***	0.08	0.3	0.23
Agreeableness - Trust	0.1	0.14	0.39	132,170K	***	50,520K	***	65,800K	***	0.14	0.4	0.27
Agreeableness - Sympathy	0.48	0.53	0.6	147,850K	***	103,070K	***	108,540K	***	0.04	0.08	0.09
Agreeableness - Cooperation	0.04	0.07	0.37	122,750K	***	31,220K	***	49,120K	***	0.17	0.5	0.35
Agreeableness - Altruism	0.2	0.26	0.5	138,030K	***	63,060K	***	76,730K	***	0.1	0.29	0.19
Agreeableness - Morality	0.05	0.09	0.36	130,340K	***	4,040K	***	57,780K	***	0.14	0.41	0.28
Openness-Emotionality	0.31	0.36	0.56	144,620K	***	75,760K	***	82,580K	***	0.06	0.23	0.18
Openness-Adventurousness	0.24	0.27	0.38	143,900K	***	83,990K	***	91,660K	***	0.07	0.2	0.13
Openness-Imagination	0.87	0.81	0.62	182,700K	***	170,950K	***	150,280K	***	0.11	0.3	0.2
Openness-Artistic interests	0.43	0.47	0.56	151,370K	0.24	93,430K	***	96,070K	***	0.02	0.12	0.1
Openness-Intellect	0.56	0.56	0.5	156,410K	***	124,590K	***	122,300K	***	0.02	0.08	0.09
Openness-Liberalism	0.66	0.66	0.57	157,570K	0.98	131,850K	***	128,430K	***	0.03	0.12	0.1
Emotional range-Anger	0.95	0.91	0.66	190,100K	***	195,420K	***	171,840K	***	0.16	0.46	0.32
Emotional range-Anxiety	0.81	0.77	0.61	172,270K	***	153,680K	***	140,030K	***	0.08	0.24	0.17
Emotional range-Depression	0.91	0.88	0.68	175,070K	***	167,530K	***	151,670K	***	0.1	0.31	0.22
Emotional range-Immoderation	0.69	0.64	0.53	173,840K	***	145,150K	***	130,160K	***	0.08	0.16	0.09
Emotional range-Self-consciousness	0.77	0.75	0.56	167,090K	***	154,830K	***	143,590K	***	0.08	0.23	0.17
Emotional range-Vulnerability	0.73	0.7	0.62	165,010K	***	133,430K	***	124,830K	***	0.06	0.13	0.08
Consciousness-Achievement-striving	0.06	0.09	0.36	135,160K	***	47,640K	***	63,470K	***	0.14	0.39	0.26
Consciousness-Self-efficacy	0.38	0.38	0.46	154,750K	**	99,890K	***	100,160K	***	0.04	0.1	0.08
Consciousness-Dutifulness	0.05	0.1	0.41	125,280K	***	36,430K	***	55,630K	***	0.17	0.48	0.31
Consciousness-Self-discipline	0.03	0.05	0.3	132,480K	***	42,070K	***	58,490K	***	0.15	0.43	0.29
Consciousness-Orderliness	0.18	0.2	0.35	149,970K	***	81,820K	***	85,040K	***	0.03	0.2	0.17
Consciousness-Cautiousness	0.04	0.1	0.34	126,030K	***	46,140K	***	67,830K	***	0.17	0.38	0.22
Extraversion-Assertiveness	0.48	0.49	0.5	153,200K	***	109,290K	***	109,980K	***	0.04	0.09	0.05
Extraversion-Gregariousness	0.34	0.33	0.51	160,450K	**	95,880K	***	92,370K	***	0.02	0.12	0.12
Extraversion-Activity level	0.13	0.17	0.39	133,910K	***	46,940K	***	61,710K	***	0.13	0.41	0.28
Extraversion-Excitement-seeking	0.75	0.69	0.61	180,590K	***	151,180K	***	131,000K	***	0.11	0.21	0.11
Extraversion-Friendliness	0.19	0.22	0.48	147,420K	***	69,910K	***	76,340K	***	0.06	0.26	0.21
Extraversion-Cheerfulness	0.17	0.2	0.5	150,300K	***	68,920K	***	73,730K	***	0.04	0.25	0.21
Note: *$p < 0.05$ **< 0.01 ***< 0.001												

Table 9.20: Personality facet scores and Hellinger distances for the lower level facets of the Big Five personality traits of HIs, HTs and general users.

top-level dimensions has six facets that further characterize an individual. For example, Emotional range is broken down into Anger, Anxiety, Depression, Immoderation, Self-consciousness, and Vulnerability. The Big Five personality traits, their associated facets, and how to interpret them are defined in detail in [240].

ﬄ

To quantify the difference between the continuous distributions of different personality aspects, we compute the Hellinger distance [241]. The Hellinger distance between two measures P and Q represented by two distributions $f(x)$ and $g(x)$, respectively, is defined as:

$$H(P,Q) = \sqrt{\frac{1}{2}\int (\sqrt{f(x)} - \sqrt{g(x)}\,)^2\,dx}\,, \tag{9.1}$$

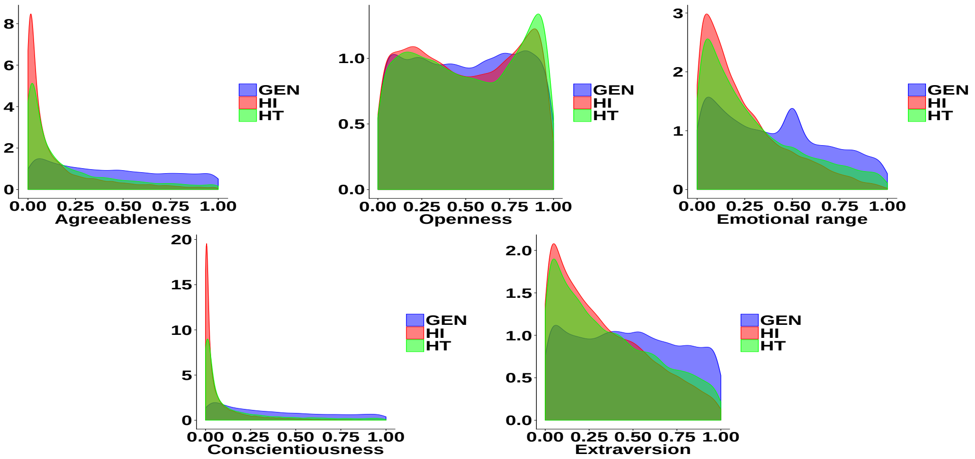


Figure 9.7: Distribution of scores for the Big Five personality traits.

where $H(P, Q) \in [0, 1]$. The minimum distance of 0 is achieved when P and Q exhibit exactly the same distributions; the maximum distance of 1 is achieved when P assigns probability zero to every set to which Q assigns a positive probability, and vice versa. Figure 9.7 shows the probability density functions for the Big Five personality traits for HIs, HTs, and the general population while Tables 9.19- 9.20 depict the pairwise distribution distances between HIs and HTs (HI-HT), and the distance between the HI and HT distributions and the general users, (HI-Gen-1%) and (HT-Gen-1%), respectively. We also report the results of the Mann-Whitney U tests.

HIs and HTs personalities differ from general users: For all the personality traits depicted in Table 9.19, the Hellinger distance of (HI-HT) is always less than or equal to (HI-Gen-1%) and (HT-Gen-1%). This indicates that HIs and HTs have more similar personalities to each other than general users. This is also shown for each personality trait's median. With the exception of Openness, the median for HIs personality facets is closer to the median of HTs than Gen-1%.

Both HIs and HTs exhibit lower Agreeableness than general users. This is also true across facets under the Agreeableness trait. Figure 9.8 shows that the gap stems from the underlying facets of Modesty, Trust, Cooperation, Altruism, and Morality (both

HIs and HTs have lower scores than the Gen-1%). Lower Agreeableness scores are often associated with suspicious and antagonistic behaviors [242]. Our results indicate that HIs and HTs are more self-focused, contrary, proud, cautious of others, and can compromise morality.

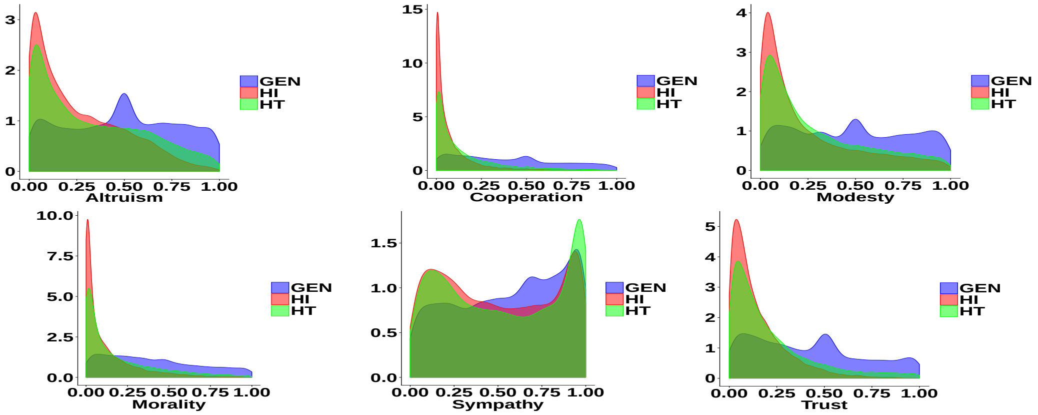


Figure 9.8: Distribution of Agreeableness scores.

While Figure 9.7 shows that the distributions for HIs, HTs, and general users are close (with a median of approximately 0.5), when we investigate Openness, we find discrepancies in the lower level facets: Adventurousness, Emotionality, and Imagination as shown in Figure 9.9. Both HIs and HTs exhibit lower scores for Emotionality and Adventurousness, and higher Imagination scores, in comparison to the general users. Moreover, HIs and HTs have similar distributions for Artistic Interests ($p = 0.24$) and Liberalism ($p = 0.98$). These results indicate that HIs and HTs are less emotionally aware and less adventurous with a wild imagination (lower preference to facts), and more authority challenging behavior, in comparison to the general users.

For Emotional range, HIs and HTs have lower scores than general users as shown in Figure 9.7. HIs have slightly lower scores, but still statistically significant, than HTs. Low Emotional range scores are correlated with high scores for Anger, Anxiety, Depression, Immoderation, and Self-consciousness as depicted in Figure 9.10. This indicates that HIs

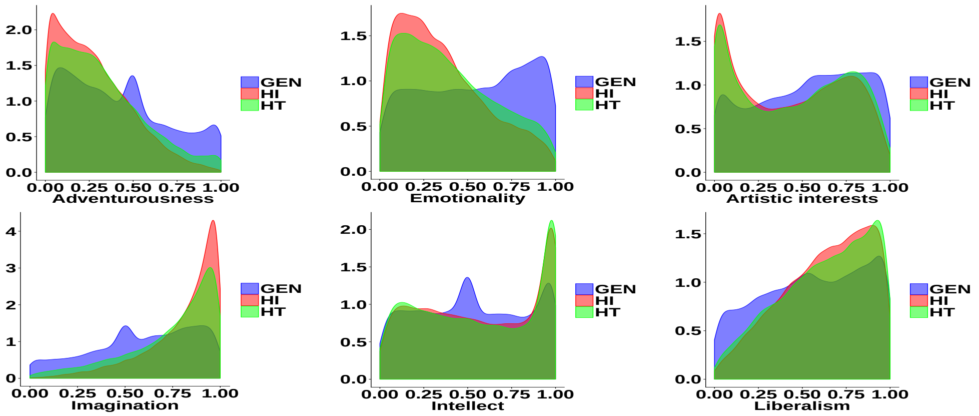


Figure 9.9: Distribution of Openness scores.

and HTs are more fiery, prone-to worry, melancholy, hedonistic, and susceptible to stress. Could all of the aforementioned traits contribute to engaging in hate speech? Cheng *et al.* observe that negative mood increased a user's probability to engage in trolling, and that anger begets more anger [208]. It seems that Emotional range facets such as Anxiety, Depression, Immoderation, and Self-consciousness are embodied more in the tweets of HIs and HTs but further work is needed to directly correlate these parameters with hate speech and online trolling.

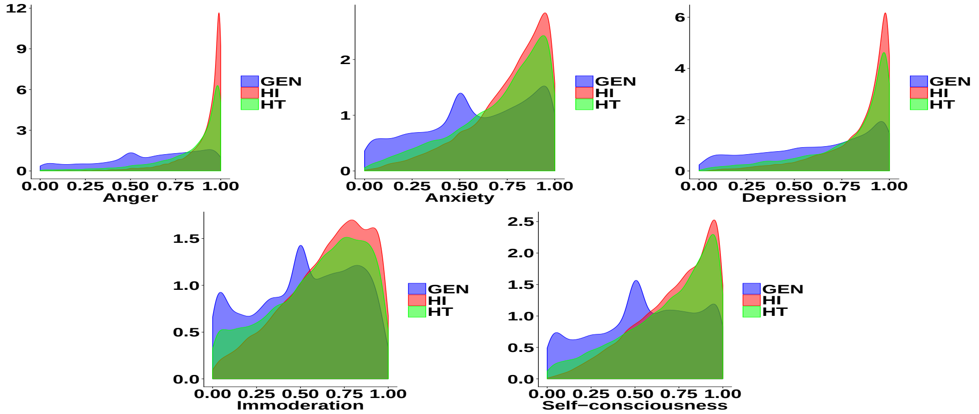


Figure 9.10: Distribution of Emotional range scores.

For Conscientiousness, HIs and HTs generally have lower scores than general users. The gap is particularly large for Achievement-striving, Dutifulness, Self-discipline, and

Cautiousness as shown in Figure 9.11. Consistently, HTs score slightly higher, but still statistically significant, than HIs. Our results suggest that HIs and HTs show lower drive, persistence, and structure. Moreover, HIs and HTs tend to disregard rules and obligations, as indicated by low dutifulness scores, and would rather take action immediately than spend time deliberating a decision, as indicated by low Cautiousness scores.

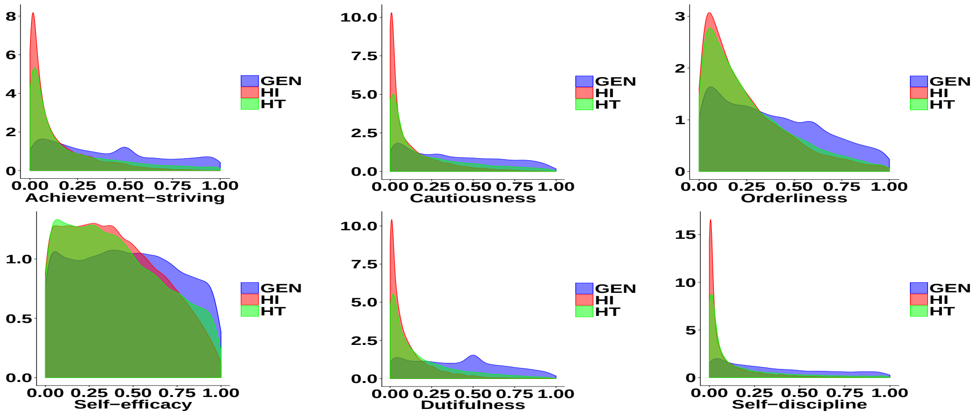


Figure 9.11: Distribution of Conscientiousness scores.

As for Extraversion, HIs and HTs tend to have lower scores of Activity-level, Friendliness, and Cheerfulness but higher scores for Excitement seeking, in comparison to general users as depicted in Figure 9.12. Our results indicate that HIs and HTs tend to have a less energetic life style. Moreover, they are inclined to be less sociable, less assertive, and more solemn. Additionally, HIs and HTs tend be more excited by taking risks in comparison to the general users; hence the higher Excitement-seeking scores.

HIs and HTs tend to share personality facets: It is possible that the personality facets for HIs and HTs could contribute to the problem of hate speech. Our results show that indeed the personalities of HIs and HTs are much closer to each other than to the general users. The very similar personality trait results for HIs and HTs could be attributed to the exchange of roles for the HIs and HTs throughout an online toxic conversation. As discussed in Section 9.4, 5% of the total accounts appear in both lists of

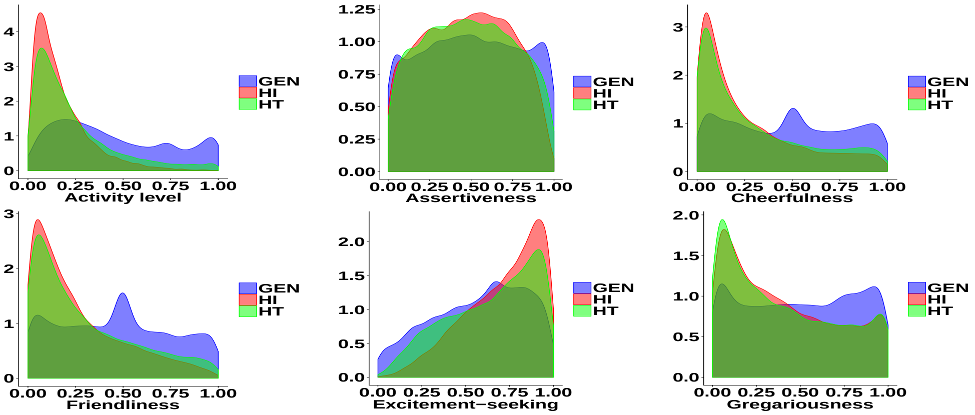


Figure 9.12: Distribution of Extraversion scores.

HIs and HTs. To verify this hypothesis, there would be a need to collect whole Twitter conversations to track the roles of different participants in hate speech conversations. This functionality is not currently supported by the Twitter API. It is also worth noting that automatic role labeling for actors in toxic converstaions, e.g., instigator, target, assistant, and defender is still an open research problem [179].

Despite that limitation, our results agree with prior work conducted for victims of bullying. Prior studies, in workplaces and schools, have shown that bullying victims tend to show depression and helplessness as a result of bullying [243]. Moreover victims are described as lacking social skills, tending to show emotions, e.g., crying easily [244], and are likely to experience anxiety, loneliness, and hyperactivity [245, 246]. Our work also agrees with studies that show that bullies and victims share a wide range of bully-typifying personality traits such as machiavellianism, narcissism, psychoticism, and aggression, and that bullies and victims could exchange roles [247]. Interestingly, in this work we have shown that these personality signals have been mirrored from the physical world and now have a presence in the digital world as well.

9.6 Discussion and Conclusion

Hate mitigation and counter speech. Successful counter speech is a direct response to hateful comments aimed at influencing discourse and behavior [205, 248]. Recently, Munger showed that counter speech using automated bots can reduce instances of racist speech if instigators are sanctioned by a high-follower white male [249]. If AI-powered counter speech bots are widely deployed [250], a research challenge would then be how we can design these bots to achieve maximum impact. Prior work has shown that people respond more positively to messages tailored to their personality [235]. For instance, Myszkowski and Storme correlated Openness with product design and found that individuals with low openness scores respond to product appearance and, conversely, high openness individuals tend to focus on product aspects, leading them to disregard aesthetic characteristics [236]. Our personality analyses could be used to design next generation counter speech bots of increased effectiveness. Moreover, our personality results show that 50% of HIs and HTs score above 0.53 for the Openness to change personality facet, which may imply that counter speech could be successfully used to decrease hate speech.

Profile-based data collection. Most common methods of data collection use hate terms and trained classifiers to classify new content as hateful or benign. Another method employs bootstrapping, which is used in [187] to obtain training data by classifying Twitter accounts as either "good" or "bad" based on usage of offensive terms. All tweets from "bad accounts" are marked as hate speech instances. Our results could be incorporated through the use of personality scores as features to classify users. Alternatively, a user could be represented as a vector of personality facets and then compared to values for hate speech accounts. This could be especially useful for content curation for cases when the instigator is likely to engage in hate speech more than once [187, 206] or as features for early instigator identification [251] and implicit hate speech detection.

Critique of methodology and limitations. There are limitations to our methodology and findings. Recent studies [93, 94] discuss common issues associated with social media analysis and the sample quality of the Twitter Streaming API. Our analysis focused on explicit hate speech and relied on keyword-based methods, which have been shown to miss instances of hateful speech [252]. However, while we cannot claim to have captured a complete representation of hate speech on Twitter, as our starting point for tweet filtering was based on a set of hate terms from Hatebase, our primary objective was to investigate hate speech instigator and target accounts with a high precision dataset. We believe that our careful curation methodology achieved this end goal.

Conclusion. We have presented the first comparative study of hate speech instigators, targets, and general Twitter users. We have outlined a semi-automated classification approach for curation of directed explicit hate speech. Our analysis yields a number of interesting and unexpected findings about actors of hate speech. For example, we found that hate instigators target more visible users and that participating in hate commentary is associated with higher visibility. We also showed that hate instigators and targets have unique personality characteristics that may contribute to hate speech such as anger, depression, and immoderation. We hope that our results can be used as meta-information to improve hate speech classification, detection and mitigation to combat this increasingly pervasive problem.

9.7 Appendix

Archaic key phrases: boojie, surrender monkey, chinaman, hillbilly, whigger, white nigger, wigger, wigerette

Class key phrases: bitter clinger, conspiracy theorist, redneck, rube, trailer park trash, trailer trash, white trash, yobbo

Disability key phrases: retard, retarded

Ethnicity key phrases: nigger, white trash, trailer trash, coonass, trailer park trash, raghead, house nigger, white nigger, camel fucker, moon cricket, wetback, spic

Gender key phrases: bint, cunt, dyke, twat

Nationality key phrases: bamboo coon, camel fucker, chinaman, limey, plastic paddy, sideways pussy, surrender monkey, whigger, white nigger, wigger, zionazi

Religion key phrases: camel fucker, muzzie, soup taker, zionazi

Sexual Orientation key phrases: dyke, faggot

9.8 Acknowledgments

The content of this chapter is the result of a collaboration with Shirin Nilizadeh, Dana Nguyen, and Giovanni Vigna and has previously appeared in the AAAI International Conference on Web and Social Media in 2018.

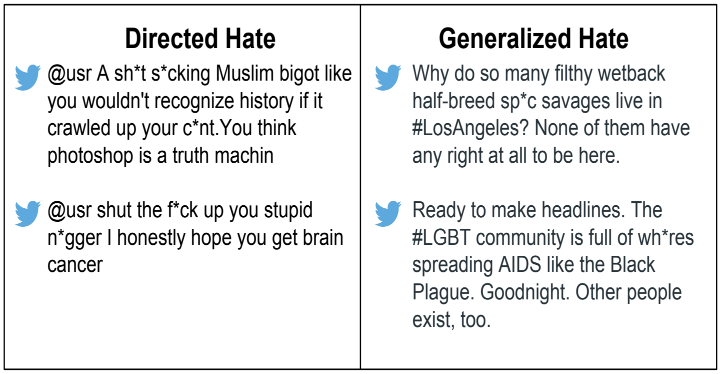
Chapter 10

A Target-based Linguistic Analysis of Hate Speech in Social Media

In this chapter, we deepen our understanding of online hate speech by focusing on a largely neglected but crucial aspect of hate speech – its *target*: either *directed* towards a specific person or entity, or *generalized* towards a group of people sharing a common protected characteristic. We perform the first linguistic and psycholinguistic analysis of these two forms of hate speech and reveal the presence of interesting markers that distinguish these types of hate speech. Our analysis reveals that Directed hate speech, in addition to being more personal and directed, is more informal, angrier, and often explicitly attacks the target (via name calling) with fewer analytic words and more words suggesting authority and influence. Generalized hate speech, on the other hand, is dominated by religious hate, is characterized by the use of lethal words such as murder, exterminate, and kill; and quantity words such as million and many. Altogether, our work provides a data-driven analysis of the nuances of online-hate speech that enables not only a deepened understanding of hate speech and its social implications, but also its detection.

10.1 Introduction

Prior work ignores a crucial aspect of hate speech – the *target of hate speech* – and only seeks to distinguish *hate* and *non-hate speech.* Such a binary distinction fails to capture the nuances of hate speech – nuances that can influence free speech policy. First, hate speech can be directed at a specific individual (**Directed**) or it can be directed at a group or class of people (**Generalized**). Figure 10.1 provides an example of each hate speech type. Second, the target of hate speech can have legal implications with regards to right to free speech (the First Amendment).[1]



Directed Hate	Generalized Hate
@usr A sh*t s*cking Muslim bigot like you wouldn't recognize history if it crawled up your c*nt.You think photoshop is a truth machin	Why do so many filthy wetback half-breed sp*c savages live in #LosAngeles? None of them have any right at all to be here.
@usr shut the f*ck up you stupid n*gger I honestly hope you get brain cancer	Ready to make headlines. The #LGBT community is full of wh*res spreading AIDS like the Black Plague. Goodnight. Other people exist, too.

Figure 10.1: Examples of two different types of hate speech. Directed hate speech is explicitly directed at an individual entity while Generalized hate speech targets a particular community or group. Note that throughout the chapter, explicit text has been modified to include a star (*).

In this chapter, we bridge the gaps identified above by analyzing Directed and Generalized hate speech to provide a thorough characterization. Our analysis reveals several differences between **Directed** and **Generalized** hate speech. First, we observe that Directed hate speech is very personal, in contrast to Generalized hate speech, where religious and ethnic terms dominate. Further, we observe that generalized hate speech is dominated by hate towards religions as opposed to other categories, such as Nationality, Gender or Sexual Orientation. We also observe key differences in the linguistic patterns,

[1] We refer the reader to [253] for a detailed discussion of one such case and its implications.

such as the semantic frames, evoked in these two types. More specifically, we note that **Directed** hate speech invokes words that suggest *intentional action*, *make statements* and explicitly uses words to hinder the action of the target (e.g. calling the target a `retard`). In contrast, **Generalized** hate speech is dominated by *quantity words* such as `million, all, many`, *religious words* such as `Muslims, Jews, Christians` and *lethal words* such as `murder, beheaded, killed, exterminate`. Finally, our psycholinguistic analysis reveals language markers suggesting differences between the two categories. One key implication of our analysis suggests that **Directed** hate speech is more informal, angrier and indicates higher clout than **Generalized** hate speech. Altogether, our analysis sheds light on the types of digital hate speech, and their distinguishing characteristics, and paves the way for future research seeking to improve our understanding of hate speech, its detection and its larger implication to society. This chapter presents the following contributions:

- We present the first extensive study that explores different forms of hate speech based on the target of hate.
- We study the lexical and semantic properties characterizing both **Directed** and **Generalized** hate speech and reveal key linguistic and psycholinguistic patterns that distinguish these two types of hate speech.
- We curate and contribute a dataset of 28,318 Directed hate speech tweets and 331 Generalized hate speech tweets to the existing public hate speech corpus.[2]

[2] The datasets are available here: `https://github.com/mayelsherif/hate_speech_icwsm18`

10.2 Related Work

Hate speech detection. Hate speech detection has been supplemented by a variety of features including lexical properties such as n-gram features [191], character n-gram features [254], average word embeddings, and paragraph embeddings [191, 193]. Other work has leveraged sentiment markers, specifically negative polarity and sentiment strength in preprocessing [182, 255, 194] and as features for hate speech classification [195, 256]. In contrast, our work reveals novel linguistic, psychological, and affective features inferred using an open vocabulary approach to characterize Directed and Generalized hate speech.

Hate speech targets. Silva *et al.* study the targets of online speech by searching for sentence structures similar to "I <intensity> hate <targeted group>". They find that the top targeted groups are primarily bullied for their ethnicity, behavior, physical characteristics, sexual orientation, class, or gender. Similar to [176], we differentiate between hate speech based on the innate characteristic of targets, *e.g.,* class and ethnicity. However, when we collect our datasets, we use a set of diverse techniques and do not limit our curation to a specific sentence structure.

10.3 Data, Definitions and Measures

Waseem et al. [257] outline a typology of abuse language and differentiate between Directed and Generalized language. We adopt the same typology and define the following in the context of hate speech:

- **Directed hate**: hate language towards a specific individual or entity. An example is: *"@usr[3] your a f*cking queer f*gg*t b*tch".*
- **Generalized hate**: hate language towards a general group of individuals who

[3] *Note that we anonymize all user mentions by replacing them with @usr.*

Category	**Key phrase-based**	**Hashtag-based**	**Davidson *et al.***	**Waseem *et al.***	**NHSM**	**Generalized**	**Directed**	**Gen-1%**
Archaic	169	0	7	0	0	5	171	-
Class	917	0	138	0	0	107	948	-
Disability	8,059	0	63	0	0	35	8,087	-
Ethnicity	2,083	220	617	0	16	648	2,288	-
Gender	13,272	0	58	0	2	43	13,289	-
Nationality	81	0	4	0	5	8	83	-
Religion	48	70	46	1,651	9	1444	380	-
Sexorient	3,689	0	394	0	9	253	3,840	-
Total	28,318	290	1,327	1,651	41	2,543	29,086	85,000

Table 10.1: Categorization of all collected datasets.

share a common protected characteristic, e.g., ethnicity or sexual orientation. An example is: *"— was born a racist and — will die a racist! — will not rest until every worthless n*gger is rounded up and hung, n*ggers are the scum of the earth!! wPww WHITE America"*.

10.3.1 Data and Methods

To mitigate the challenges associated with identifying hate speech as discussed in Chapter 9, we adopt several strategies including a comprehensive human evaluation. We describe the construction of our datasets below in detail. The datasets themselves are summarized in Table 11.1.

(1) Key phrase-based dataset: We adopt the dataset curated through a multi-step classification approach discussed in Chapter 9.

Using the aforementioned classification method, we obtain a high precision hate speech dataset of 28,318 tweets in which hate instigators use explicit Hatebase expressions against hate target accounts.

(2) Hashtag-based dataset: In addition to keyphrases, we also incorporated hashtags. We examined a set of hashtags that are used heavily in the context of hate speech. We started with 13 hashtags that are likely to result in hate speech such as #killallniggers, #internationaloffendafeministday, #getbackinkitchen. As we filtered the 1% sample of Twitter's public stream from January 1st, 2016 to July 31st, 2017 for these hashtags;

we eliminated hashtags with no significant presence. We include in our datasets the four hashtags that had the most hateful usage by Twitter users: #istandwithhatespeech, #whitepower, #blackpeoplesuck, #nomuslimrefugees. Finally, we obtained 597 tweets for #istandwithhatespeech, 195 for #whitepower, 25 for #blackpeoplesuck, and 70 for #nomuslimrefugees. We include #istandwithhatespeech in our lexical analysis but omit it from subsequent analyses because while these tweets discuss hate speech, they are not actually hate speech themselves.

(3) Public datasets: To expand our hate speech corpus, we evaluate publicly available hate speech datasets and add tweet content from these datasets into our keyphrase and hashtag datasets, as appropriate. We start with datasets obtained by Waseem and Hovy [145] and Davidson *et al.* [201]. We examine these existing datasets and eliminate tweets that contain foul and offensive language but that do not fit our definition of hate speech (for example, *"RT @usr: I can't even sit down and watch a period of women's hockey let alone a 3 hour class on it...#notsexist just not exciting"*). We then inspect the remaining tweets and assign each to its most appropriate hate speech category using a combination of our Hatebase keyword filter and manual annotations. Tweets that were not filtered by our Hatebase keyword approach were carefully examined and annotated manually. We obtain a total of 1,651 tweets from [145] and 1,327 tweets from [201].

Finally, we also examine hate speech reports on the No Hate Speech Movement (NHSM) website[4]. The campaign allows online users to contribute instances of hate speech on different social media platforms. We retrieve a total of 41 English hate tweets.

(4) General dataset (Gen-1%): To provide a larger context for interpretation of our analyses, we compare data from our collection of hate speech datasets with a random sample of all general Twitter tweets. To create this dataset, we use the Twitter Streaming API to obtain a 1% sample of tweets within the same 18 month collection window. From

[4] No Hate Speech Movement Campaign: https://www.nohatespeechmovement.org/

this random 1% sample, we randomly select 85,000 English tweets.

Human-centered dataset evaluation. We evaluate the quality of our final datasets by incorporating human judgment using Crowdflower. We provided annotators with a class balanced random sample of 2000 tweets and asked them to annotate whether or not the tweet was hate speech or not, and whether the tweet was directed towards a group of people (Generalized hate speech) or directed towards an individual (Directed hate speech). To aid annotation, all annotators were provided a set of precise instructions. This included the definition of hate speech according to the social media community (Facebook and Twitter) and examples of hate tweets selected from each of our eight hate speech categories. Each tweet was labeled by at least three independent Crowdflower annotators, and all annotators were required to maintain at least an 80% accuracy based on their performance of five test questions - falling below this accuracy resulted in automatic removal from the task. We then measured the inter-annotator reliability to assess the quality of our dataset. For the representative sample from our Generalized hate speech dataset, annotators labeled 95.6% of the tweets as hate speech and 87.5% of tweets as hate speech directed towards a group of people. For the representative sample from our Directed hate speech dataset, annotators labeled 97.8% of the tweets as hate speech and 94.3% of tweets as hate speech directed towards an individual. Our dataset obtained a Krippendorf's alpha of 0.622, which is 38% higher than other crowd-sourced studies that observed online harmful behavior [215].

10.3.2 Measures

In our investigation, we adopt several measures based on prior work in order to study linguistic features that differentiate between Directed and Generalized hate speech. To alleviate the effects of domain shift in our choice of models, we use tools that are developed and trained using Twitter data when available and fall back to state of the art models that were trained on English data in the event of unavailability of Twitter-specific tools. To analyze the salient words for each category of hate speech keywords (e.g., ethnicity, class, gender) and specific language semantics associated with hashtags, we use SAGE [258], a mixed-effect topic model that implements the L1-regularized version of sparse additive generative models of text. SAGE has been used in several Natural Language Processing (NLP) applications including [259] that provides a joint probabilistic model of who cites whom in computational linguistics, and [260] which aims to understand how opinions change temporally around the topic of slavery-related United States property law judgments. To extract entities from the collected tweets, we leverage T-NER, a system developed specifically to perform the task of Named Entity Recognition on tweets [261]. To understand the linguistic dimension and psychological processes identified among Directed hate, Generalized hate, and general Twitter tweets, we use the psycholinguistic lexicon software LIWC2015 [262], a text analysis tool that measures psychological dimensions, such as affection and cognition. To analyze frame semantics of hate speech, we use SemaFor [263], which annotates text with their evoked frames as defined by FrameNet [264, 265]. While we acknowledge that Semafor is not trained on Twitter, it has been found that it is actually more robust to domain-shift and its performance on Twitter is comparable to that on Newswire [266].

10.4 Analysis

10.4.1 Lexical Analysis

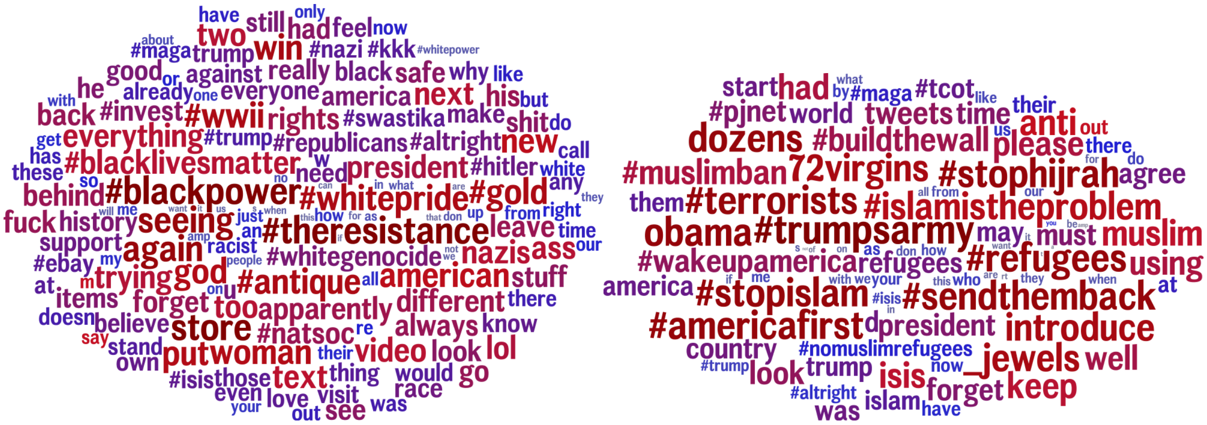
To analyze salient words that characterize different hate speech types, we use SAGE [258]. SAGE offers the advantages of being supervised, building relatively clean topic models by

Archaic Generalized	**Archaic Directed**	**Class Generalized**	**Class Directed**
Anti	hillbilly	Catholics	Rube
wigger	chinaman	hollering	#redneck
hillbilly	verbally	#racist	ALABAMA
bitch	prostitute	Cracker	batshit
white	vegetables	#Virginia	DRINKS
Disability Generalized	**Disability Directed**	**Ethnicity Generalized**	**Ethnicity Directed**
retards	#Retard	Anglo	coons
legit	sniping	spics	Redskins
Only	#retarded	breeds	Rhodes
yo	Asshole	hollering	#wifebeater
phone	upbringing	actin	plantation
Gender Generalized	**Gender Directed**	**Nationality Generalized**	**Nationality Directed**
dyke(s)	#CUNT	Anti	chinaman
chick	judgemental	wigger	Zionazi(s)
cunts	aitercation	bitch	#BoycottIsrael
hoes	Scouse	white	prostitute
bitches	traitorous		#BDS
Religion Generalized	**Religion Directed**	**SexOrient Generalized**	**SexOrient Directed**
Algebra	catapults	meh	pansy
Israelis	Muzzie	#faggot(s)	Cuck
extermination	Zionazi	queers	CHILDREN
Jihadi	#BoycottIsrael	hipster	FOH
lunatics	rationalize	NFL	wrists

Table 10.2: Top five keywords learned by SAGE for each hate speech class. Note the presence of distinctive words related to each class (both for Generalized and Directed hate).

taking into account additive effects and combining multiple generative facets, including background, topic and perspective distributions of words. In our analysis, each tweet is treated as a document and we only include words that appear at least five times in the entire corpus. This step is crucial to ensure that SAGE's supervised learning model will find salient words that not only identify each hate speech type or hashtag, but also are well-represented in our datasets.

What are the salient words characterizing different hate speech categories? Table 10.2 shows the top five salient words learned by SAGE for each hate speech type. We note that there is minimal intersection of salient words between different hate speech categories, e.g., ethnicity, archaic, and SexOrient, and between the generalized and directed versions of each hate speech type. Although a tweet could contain several key-



(a) #whitepower

(b) #nomuslimrefugees

Figure 10.2: The salient words for tweets associated with #whitepower and #nomuslimrefugees learned by the sparse additive generative model of text. A larger font corresponds to a higher score output by the model.

words pertaining to different types of hate speech, the top salient words indicate that hate speech categories have distinct topic domains with minimal overlap. For example, note the presence of words `retards, #Retard` used in hate speech related to disability. Similarly, note the presence of religion related words like `Jihadis, extermination, Zionazi, Muzzie` for religion-related hate speech.

We show the results of SAGE for the hashtags #whitepower (categorized as ethnicity-based hate) and #nomuslimrefugees (categorized as religion-based hate) in Figure 10.2. Among the salient words for the hashtag #whitepower are #whitepride, #whitegenocide, the resistance, #wwii, nazi, #kkk, #altright, and republicans. For the hashtag #nomuslimrefugees, salient words include #stopislam, #islamistheproblem, #trumpsarmy, #terrorists, #muslimban, #sendthemback, and #americafirst.

What are the prevalent themes in hate speech participation? We examine the salient words for #istandwithhatespeech to gain insight into why people participate in hate speech. The top five salient words for #istandwithhatespeech are *banned, allowed, opinion, #1a,* and *violence*. Further inspection of tweets for these keywords revealed the following themes: **(a)** hate and other offensive speech should be allowed on the Internet;

(b) not participating in hate speech implies the inability to handle different opinions; **(c)** hate speech is truth telling; and **(d)** the First Amendment (#1a) grants the right to participate in hate speech. Some example tweets representing these viewpoints include: *@usr: people should be allowed to tell the truth no matter how it affects other people. #istandwithhatespeech*; *@usr: #istandwithhatespeech because the eu shouldn't dictate what is allowed on the internet, a global communication system*; and *#istandwithhatespeech b/c if you really can't hear an opinion different from your own you need f*cking therapy.*

How are named entities represented across Directed and Generalized hate? Named Entity Recognition seeks to identify names of persons, organizations, locations, expressions of times, brands, and companies among other categories within selected text. For example, consider the following tweet: *"@usr Obama and Hillary ain't gone protect you when trump is president. btw you need some braces you f*ckin dyke."* The task of Named Entity Recognition would identify *Obama, Hillary*, and *trump* as person entities.

Figure 10.3 shows a breakdown of entities identified by T-NER for Directed hate, Generalized hate and Gen-1% tweets. We first note that Directed hate tends to have a higher percentage of person entities (55.8%) as opposed to Generalized hate (42.1%), and Gen-1% (46.4%). This is expected since Directed hate speech is often a personal attack on specific person(s). We find that tweets have other entities that do not belong to persons, brands, companies, facilities, geo-locations, movies, products, sports teams or TV shows. These include Islam and Jews; we separate these tweets into an "other" category.

We inspect all the entities recognized by T-NER and represent them in Figure 10.4. We note that some entities are universally present in different categories. These include *Trump, Hillary, Islam, Mohammed, Google, ISIS,* and *America.* Additionally, we find that Directed hate contains more common names such as *Scott, Sam, Andrew, Katie, Ben, Ryan, Jamie,* and *Lucy.* Generalized hate tends to contain religious-based entities

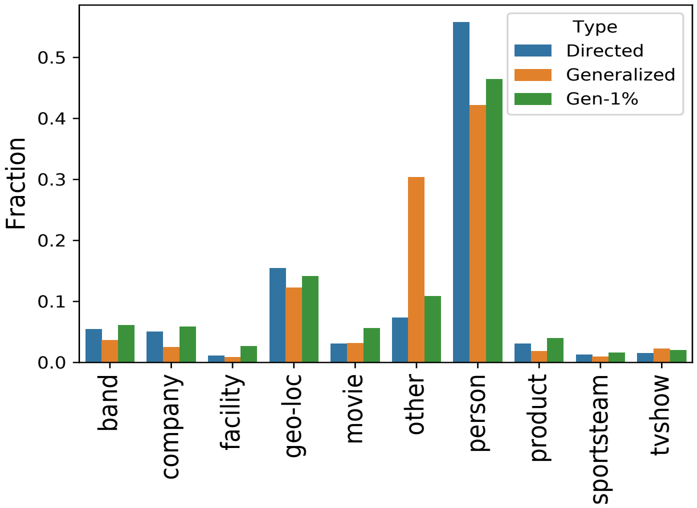
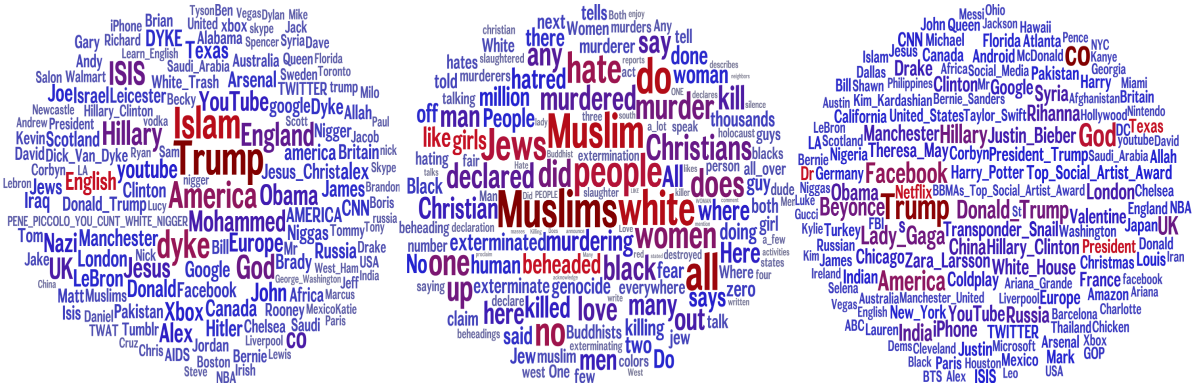


Figure 10.3: Proportion of entity types in hate speech. Note the much higher proportion of Person mentions in Directed hate speech, suggesting direct attacks. In contrast, there is a higher proportion of Other in Generalized hate speech, which are primarily religious entities (i.e. `Islam, Muslim, Jews, Christians`).

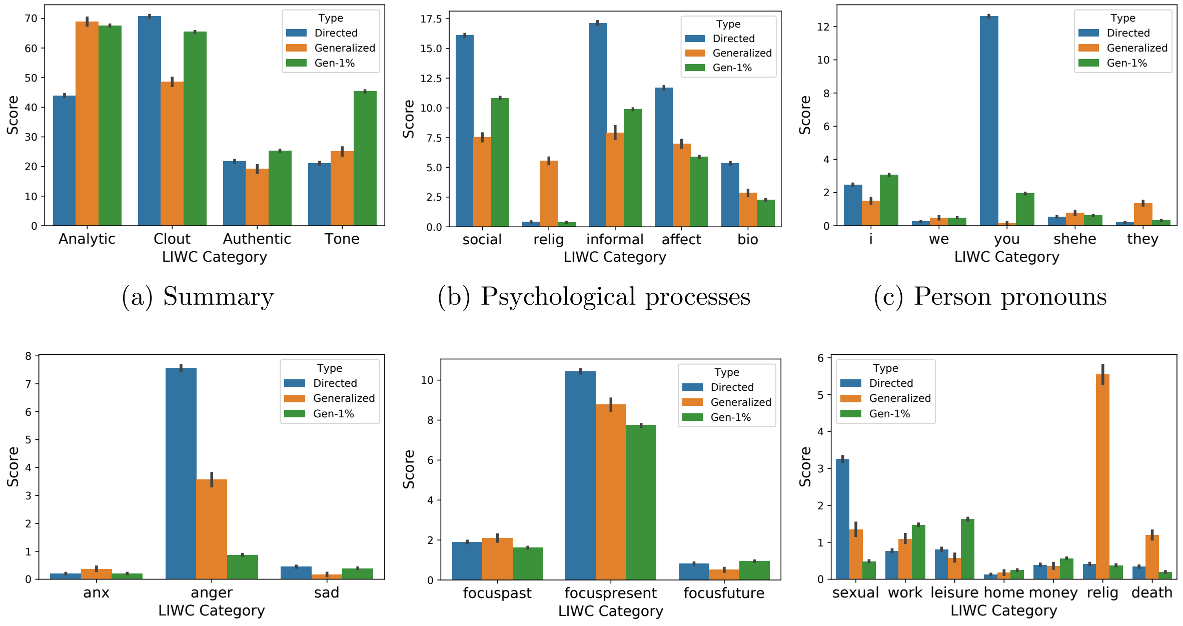
such as *Jews, Muslims, Christians, Hindus, Shia, Madina,* and *Hammas*, and entities involved in political and religious disputes and conflicts such as *Hamas, Palestine,* and *Israel*. This is also consistent with our observation that the majority of the Generalized hate speech tweets happen to be related to Religion (although no specific filtering for religion was done in the data collection step). On the other hand, we observe that certain popular individuals, such as *Theresa May, Beyonce, Justin Bieber, Lady Gaga, Taylor Swift, Tom Brady,* and *Katy Perry*, exist only in Gen-1%, suggesting that these categories differ in their focus.

In summary, our lexical analysis highlights salient features and entities that distinguish between Directed and Generalized hate speech while also revealing evident themes that indicate why people choose to participate in hate speech.



(a) Directed hate (b) Generalized hate (c) General-1%

Figure 10.4: Top entity mentions in Directed, Generalized and Gen-1% sample. Note the presence of many more person names in Directed hate speech. Generalized hate speech is dominated by religious and ethnicity words, while the Gen-1% is dominated by celebrity names.



(a) Summary (b) Psychological processes (c) Person pronouns

(d) Negative emotions (e) Temporal focus (f) Personal concerns

Figure 10.5: Mean scores for LIWC categories. Several differences exist between Directed hate speech and Generalized hate speech. For example, Directed hate speech exhibits more anger than Generalized hate speech, and Generalized hate speech is primarily associated with religion. Error bars show 95% confidence intervals of the mean.

10.4.2 Psycholinguistic Analysis

For a full psycho-linguistic analysis, we use LIWC [262]. Specifically, we focus on the following dimensions: summary scores, psychological processes, and linguistic dimensions. A detailed description of these dimensions and their attributes can be found in the LIWC2015 language manual [262]. Figure 10.5 shows the mean scores for our key LIWC attributes. Our analysis yields the following observations.

Directed hate speech exhibits the highest clout and the least analytical thinking, while general tweets exhibit the highest authenticity and emotional tone. Figure 10.5(a) shows the key summary language values obtained from LIWC2015 averaged over all tweets for Directed hate, Generalized hate, and Gen-1%. We show that Directed hate has the lowest mean for analytical thinking scores ($\mu = 43.9$, $p < 0.001$) in comparison to Generalized hate ($\mu = 68.9$) and Gen-1% ($\mu = 67.6$). We also note that Directed hate demonstrates higher mean clout (influence and power) values ($\mu = 70.7$, $p < 0.001$) than Generalized hate ($\mu = 48.5$) and Gen-1% ($\mu = 65.4$). This result resonates with the nature of personal directed hate attacks, in which persons exhibit dominance and power over others. Moreover, Figure 10.5 (a) indicates that tweets in the Gen-1% dataset have the highest mean value of authenticity (Authentic) ($\mu = 25.3$, $p < 0.001$) in comparison to hate tweets: directed ($\mu = 21.7$) and generalized ($\mu = 19.2$). Additionally, we note that Gen-1% ($\mu = 41.4$, $p < 0.001$) has the highest mean score of emotional tone (Tone) followed by Generalized ($\mu = 25.1$) and Directed hate ($\mu = 21.1$). This indicates that general tweets are associated with a more positive tone, while Generalized and Directed hate language reveal greater hostility.

Directed hate speech is more informal and social than generalized hate and general tweets. Figure 10.5(b) shows that Directed hate has a much higher mean informal score ($\mu = 17.1$, $p < 0.001$) in comparison to generalized hate ($\mu = 7.9$) and

Gen-1% ($\mu = 9.9$). Informality includes the usage of swear words and abbreviations, e.g., btw, thx. Additionally, Directed hate tends to have higher social components ($\mu = 16.1$ vs. 7.5 for generalized hate and 10.9 for general tweets, $p < 0.001$) inherent in its linguistic style, which manifests in greater usage of language related to family, friends, and male and female references.

Generalized hate speech emphasizes "they" and not "we". Figure 10.5(c) shows that generalized hate speech has higher usage of third personal plural pronouns (they) than first personal plural pronouns (we). The mean score for third person pronoun usage is 1.4, in comparison to 0.5; 2.8x higher ($p < 0.001$). An example tweet is: "*Muslims are not a race, idiot, they are a cult of murder and terrorism.*"

Directed hate speech is angrier than generalized hate speech, which in turn is angrier than general tweets. We show that anger manifests differently across Generalized and Directed hate speech. Figure 10.5(d) shows that Directed hate contains the angriest voices ($\mu = 7.6$, $p < 0.001$) followed by Generalized hate ($\mu = 3.6$); general tweets are the least angry ($\mu = 0.9$). In [208], the authors observe that negative mood increased a user's probability to engage in trolling, and that anger begets more anger. Our results complement this observation by differentiating between levels of anger for Directed and Generalized hate. Example tweets include: "*@usr F*ckin muzzie c*nts, should all be deported, savages*" and "*f*ck n*ggers, faggots, chinks, sand n*ggers and everyone who isnt white.*"

Both categories of hate speech are more focused on the present than general tweets. Figure 10.5(e) shows that hate speech ($\mu = 10.4$ and $= 8.7$ for Directed and Generalized hate, respectively, $p < 0.001$) more commonly emphasizes the present than general tweets ($\mu = 7.7$). Examples include: "*How the f*ck does a foreigner win miss America? She is Arab! #idiots*" and "*@usr Those n*ggers disgust me. They should have dealt with 100 years ago, we wouldn't be having these problems now*".

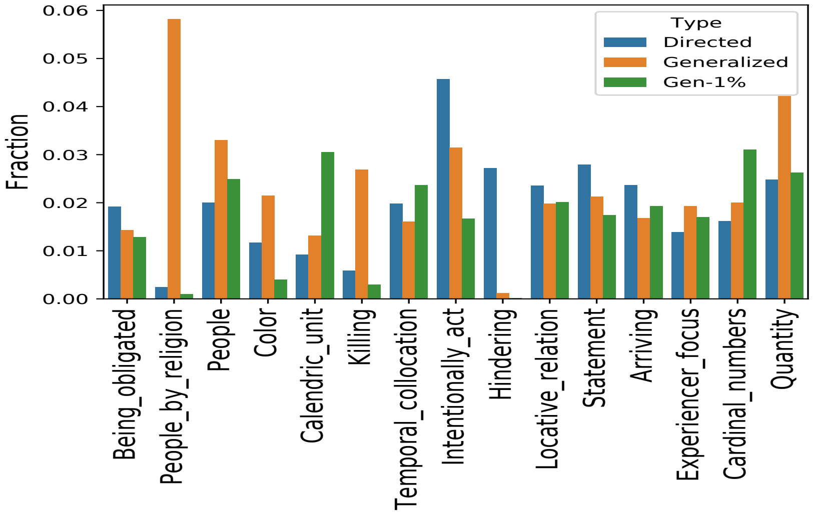
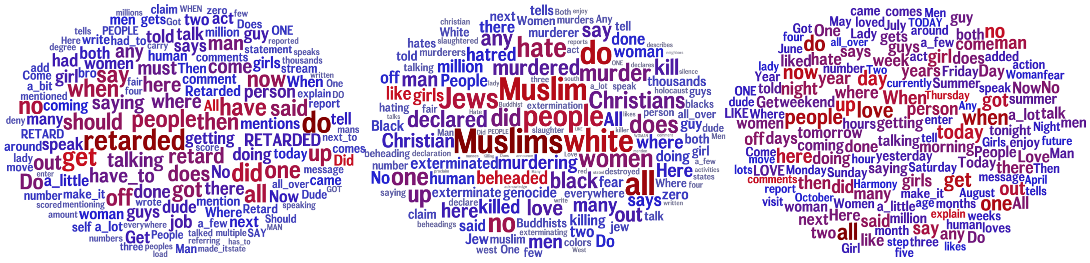


Figure 10.6: Proportion of frames in different types. Note the much higher proportion of PEOPLE_BY_RELIGION frame mentions in Generalized hate speech. In contrast, Directed hate speech evokes frames such as INTENTIONALLY_ACT and HINDERING.

General tweets have the fewest sexual references while generalized hate has the most death references. Figure 10.5(e) shows that general tweets have the lowest mean score for sexual references (μ = 0.5, $p < 0.001$) in comparison to Directed hate (μ = 3.3) and Generalized hate (μ = 1.3). Moreover, our analysis shows that, compared to general tweets ($\mu = 0.2$), hate tweets are more likely to incorporate death language (μ = 1.2, $p = 0.1$ for Generalized hate and = 0.34 for Directed hate, $p < 0.001$).

10.4.3 Semantic Analysis

In this section, we turn our attention to the frame-semantics of the hate speech categories. Using frame-semantics, we can analyze higher-level rich structures called *frames* that represent real world concepts (or stereotypical situations) that are evoked by words. For example, the frame ATTACK would represent the concept of a `person` being attacked by an `attacker` with perhaps a `weapon` situated at some point in `space` and `time`.



(a) Directed hate (b) Generalized hate (c) Gen-1%

Figure 10.7: Words evoked by the top 10 semantic frames in each hate class. In Directed hate speech, note the presence of action words such as `do, did, now, saying, must, done` and words that condemn actions (`retard, retarded`). In sharp contrast, Generalized hate speech evokes words related to Killing, Religion and Quantity such as `Muslim, Muslims, Jews, Christian, murder, killed, kill, exterminated,` and `million`.

After annotating Directed and Generalized hate speech tweets using SemaFor, we compute the distribution over evoked frames for each type of hate speech. Figure 10.6 shows proportions for 15 frame types (top 5 from each type) for Directed hate, Generalized hate and Gen-1%. We make the following observations.

Directed hate speech evokes *intentional acts, statements* and *hindering.* Our analysis reveals that the Directed hate speech has a higher proportion of intentionally_act frames (0.05, $p < 0.01$) than generalized hate (0.03) and general tweets (0.016). An example of a tweet with an intentionally_act frame is: "*@usr if you **don't**[5] choose @usr you're the biggest f*ggot to ever touch the face of the earth*". Moreover, Directed hate has the highest proportion of statement frames and hindering frames (0.03 and 0.03, respectively, $p < 0.01$) when compared to generalized hate (0.02 and 0.001) and general tweets (0.017 and 0.0001). Examples of tweets with statement and hindering frames are: "*I do not like **talking** to you f*ggot and I did but in a nicely way f*g*" and "*Your Son is a **Retarded** f*ggot like his Cowardly Daddy*", respectively. Additionally, Directed hate speech has the highest proportions of being_obligated frames (0.02, $p < 0.01$) in compar-

[5] *Bold font indicates words that evoked the corresponding frames.*

ison to generalized hate (0.014) and general tweets (0.013). A tweet that demonstrates this is *"@usr your a f*ggot and should suck my tiny c*ck block me pls"*.

Generalized hate speech evokes concepts such as *People by religion, Killing, Color, People,* and *Quantity*. Figure 10.6 shows that generalized hate has the highest proportion of frames related to People (0.033 vs 0.02 for Directed hate and 0.025 for Gen-1%, $p < 0.01$), People_by_religion (0.06 vs 0.002 for Directed hate and 0.001 for Gen-1%, $p < 0.01$), Killing (0.03 vs 0.006 for Directed hate and 0.003 for Gen-1%, $p < 0.01$), Color (0.02 vs 0.012 for Directed hate vs 0.004 for Gen-1%, $p < 0.01$), and Quantity (0.042 vs 0.025 for Directed hate and 0.026 for Gen-1%, $p < 0.01$). Example tweets include: "*@usr @usr @usr Anything to trash this **black** President!!*"; "*Why **people** think gay marriage is okay is beyond me. Sorry I don't want my future son seeing 2 f*gs walking down the street holding hands*"; and "*@usr how **many** f*ckin fags did a even get? Shouldnt be allowed into my wallet whilst under the influence haha*".

General tweets (Gen-1%) primarily evoke concepts related to the *Cardinal Numbers* and *Calendric Units.* General tweets have been found to have the highest proportion of cardinal numbers (0.03 vs 0.016 for Directed hate and 0.02 for Generalized hate, $p < 0.01$) and calendric units (0.031 vs 0.01 for Directed hate and 0.013 for Generalized hate, $p < 0.01$). Examples include: "*I LOVE you usr! xxx **February 20, 2017** at **05:45AM** #AlwaysSuperCute*" and "*Women's Basketball trails Fitchburg at the half **39-32**. Chelsea Johnson leads the Bulldogs with **12**. Live stats link: https://t.co/uRRZosr7Cl.*"

As a final step, we analyze the top words that evoked the top 10 frames in each type. We summarize these results in Figure 10.7. In Directed hate speech, we observe the presence of words like `do, doing, does, did, get, mentions, says`, which evoke the concept of Intentional Acts. This suggests that Directed hate speech directly and explicitly calls out the action of or toward the target. We also note the presence of hindering words like `retard, retarded`, which are explicitly used to attack the target entity. In contrast, Generalized hate speech is dominated by words that evoke Killing (`kill, murder, exterminate`), words that categorize people by religion (`jews, christians, muslims, islam`) and words that refer to a Quantity (`million, several, many`). This suggests the broad and general nature of Generalized hate speech, which seeks to associate hate with a general large community or group of people.

10.5 Discussion and Conclusion

Social Implications. The distinction between Directed and Generalized hate speech has important implications to law, public policy and the society. [253] raises the intriguing question of whether one needs to distinguish between emotional harm imposed on private individuals from emotional harm imposed on public political figures or from racist/hateful remarks targeted at a general community and no specific individual in particular [253]. One position is that according to the First Amendment, one needs to provide adequate opportunities to express differing opinions and engage in public political debate. However, [253] also notes that in the case of private individuals, the focus shifts towards emotional health and therefore directed/personal attacks or hate speech aimed at a particular individual must be prohibited. According to this position, hate speech directed at a public political figure or a community or no one in particular might be protected. On the other hand, one might argue that hate speech directed at a community has the potential to mobilize a large number of people by enabling a wider reach and can have devastating consequences to society. However, prohibiting all kinds of offensive/hate speech – Directed or Generalized opens up a slew of other questions with regards to censorship and the role of the government. In summary, this distinction between Generalized and Directed hate speech has widespread and far-reaching societal implications ranging from the role of the government to the framing of laws and policies.

Hate Speech Detection and Counter Speech. Current hate speech detection systems primarily focus on distinguishing between hate speech and non-hate speech. However as our analysis reveals, hate speech is far more nuanced. We argue that modeling these nuances is critical for effectively combating online hate speech. Our research points towards a richer view of hate speech that not only focuses on language but on the people generating it. For example, we show that Generalized hate exhibits the presence of the "Us Vs. Them" mentality [267] by emphasizing the usage of third person plural pronouns. Moreover, our results distinguish the different roles intermediaries could develop to deal with digital hate – one is educating communities to advance digital citizenship and facilitating counter speech [268]. Our study opens the door to research investigating whether different strategies should be designed to combat Directed and Generalized hate.

Conclusion. In this work, we shed light on an important aspect of hate speech – its target. We analyzed two different kinds of hate speech based on the target of hate: **Directed** and **Generalized**. By focusing on the target of hate speech, we demonstrated that online hate speech exhibits nuances that are not captured by a monolithic view of hate speech - nuances that have social bearing. Our work revealed key differences in linguistic and psycholinguistic properties of these two types of hate speech, sometimes revealing subtle nuances between directed and generalized hate speech. Additionally, our work highlights present challenges in the hate speech domain. One key challenge is the variety of platforms that incubate hate speech other than Twitter. Other challenges include overcoming sample quality issues and other issues associated with Twitter Streaming API as discussed by [93, 94], and the need to move beyond keyword-based methods that have been shown to miss many instances of hateful speech [252]. Despite these challenges, our approach has enabled us to amass a large dataset, which led us to a number of novel and important understandings about hate speech and its usage. We hope that our findings enable additional progress within counter speech research.

10.6 Acknowledgments

The content of this chapter is the result of a collaboration with Vivek Kulkarni and Dana Nguyen. The chapter has previously appeared in the AAAI International Conference on Web and Social Media in 2018.

Chapter 11

A Temporal Linguistic Study of the Most Prevalent Hate Ideologies in the U.S. on Twitter

Since 2014, there has been a steady rise in the number of online hate groups in the U.S.; hate groups are now present in each of the 50 states for the very first time in eight years. Given the wide reach of social media, many hate groups leverage social networks to not only propagate hate messages but also grow their base. While traditionally there have been efforts to track the evolution of hate groups via surveys and questionnaires, detecting and tracking the evolving dynamics of hate groups remains a significant challenge. On one hand, some hate groups might choose to operate in a clandestine manner, obfuscating their presence. On the other hand, other hate groups may grossly over-represent their user base. In this chapter, we present the first linguistic analysis of the dynamics of the most prevalent hate ideologies in the U.S. based on their Twitter footprints. Our analysis reveals that the strongest drive for all hate ideologies is power except Ku Klux Klan (KKK) hate groups which are mostly driven by affiliation. Additionally, our future-tense analysis reveals call for actions focusing on deportation for Anti-Immigration hate groups and pro-life arguments for Anti-LGBT hate groups. Our semantic similarity analysis yielded multiple observations including that Neo-Nazi and Anti-Muslim hate groups were consistently in the top two of the most semantically similar ideology pairs across 2015, 2016, and 2017. Anti-LGBT hate groups were also frequently found to be in the top five pairs across the three years, sharing common semantics with Anti-Muslim, White Nationalist, and Anti-Immigration hate groups. Anti-LGBT's ability to pair with these primarily race-oriented hate groups stems from its ability to transcend beyond its typical discourse of conservative sexuality and gender topics to topics of white supremacy and racism. We note the usage of cryptic emojis that are used to encode hidden meaning among hate group members. Finally, we show the presence of questionable media sources among hate group content which opens the room for discussion to the correlation between fake news and hate speech. Altogether, our work provides an unprecedented lens into the temporal language of hate present in online hate communities. We conclude by reflecting on our results discussing implications for hate speech detection.

11.1 Introduction

Human beings are social creatures that have the fundamental need to belong [269]; a lack of social connectedness has been shown to be detrimental to an individual's well being [270, 271]. Social media has gained immense popularity in part because it can provide ways for humans to feel socially connected [272]. However, it serves other roles as well, such as a platform for opinion formation. Social media has had a particularly profound effect as a political and social activism arena [273, 8, 274]. Unfortunately, social media has recently witnessed a wave of dysfunction due to the spread of challenges such as hate speech, online harassment, fake news, and manipulation by artificial entities such as bots.[1]

[1] Pew Research Center. The Future of Free Speech, Trolls, Anonymity and Fake News Online. https://www.pewinternet.org/2017/03/29/the-future-of-free-speech-trolls-anonymity-and-fake-news-online/

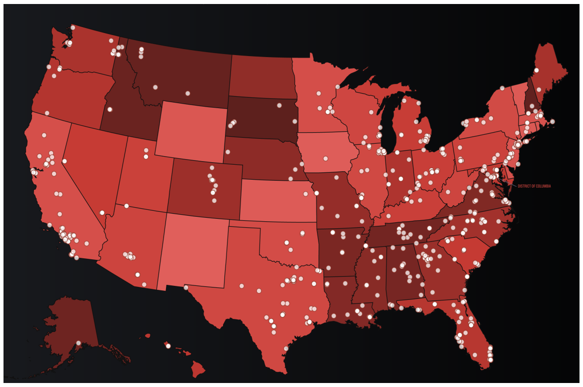


Figure 11.1: The SPLC's map for tracking hate groups across the U.S. in 2017. A darker color corresponds to a higher number of hate groups per capita.

One of the challenges of online social media is the growth of "*echo chambers*". The term echo chamber refers to the overall phenomenon by which individuals are exposed only to information from like-minded individuals [275]. As an example, one purveyor of information makes a claim, which many like-minded people then repeat, overhear, and repeat again until most people assume that some variation of the story is true [276]. In the most generic sense, echo chambers are considered dangerous because they lead to undermining opinion plurality and principles of diversity and democracy, narrowing of political worldviews, and can potentially lead to polarization and extremism [277, 276].

One dangerous combination of hate speech and echo chambers is manifested in the form of online hate groups. Hate group numbers and sizes are surging in the United States (U.S.) as shown in Figure 11.1. In February 2018, the Southern Poverty Law Center (SPLC) announced that for the first time in eight years, hate groups existed in all 50 states. [2] According to the SPLC, there has been 30% increase in U.S. hate groups over the past four years and a 7% increase in hate groups in 2018 alone.[3] In addition to their

[2] CNN. Number of neo-Nazi and black nationalist hate groups grew in 2017, SPLC says. `https://www.cnn.com/2018/02/21/us/splc-hate-group-report-2017/index.html`

[3] NPR. U.S. Hate Groups Rose 30 Percent In Recent Years, Watchdog Group Reports. https://www.npr.org/2019/02/20/696217158/u-s-hate-groups-rose-sharply-in-recent-years-watchdog-group-reports

physical presence, hate communities have embraced new online platforms to promote their ideologies and to recruit and expand their base to include younger audiences [278, 279]. For example, according to [280] over 22K white nationalists opened Twitter accounts since 2012, a 600% growth rate from 3,542 users in 2012 to 25,406 users in 2016 [280].

According to the United States Federal Bureau of Investigation (FBI), a hate group is a social group whose primary purpose is to *"promote animosity, hostility, and malice against persons belonging to a race, religion, disability, sexual orientation, or ethnicity/national origin which differs from that of the members of the organization"* [281]. Online hate groups typically work towards one or more of the following goals: to educate group members and the public about their viewpoints, to encourage participation, to promote a divine calling and privilege, and to cast out-groups or members as the enemy – "othering" [282]. While much of the content promoted by hate groups is explicitly violent or hateful, other content may appear patriotic or benign; the latter method of portrayal may contribute to the appeal of the groups [282]. The active efforts of these groups to expand their base have resulted in hate speech in particular to become a major concern. Moreover, the promotion of hate linked to these hate groups is not only an issue bounded by online communication; it can also be linked to offline societal issues and even acts of physical violence and hate crimes [283].

While there is a significant amount of prior work investigating the characterization and detection of online hate speech, the focus is primarily on individualistic posts on online platforms such as Twitter [284, 209], Reddit [204, 285], and 4chan [200]. However, the nature of hate speech is not just individual interactions, but also an organized effort of communities. In this chapter, we provide the first large scale temporal linguistic analysis of online hate groups, with a goal of understanding online hate community discourse. Specifically, this chapter seeks to answer the following research questions:

- **RQ1**: What are the pyscholinguistic characteristics of the language used by hate groups and how do they evolve over time?
- **RQ2**: Which hate ideologies are more semantically similar to other ideologies? How does inter-ideology similarity evolve over time? What are the key topics representations that evolved the course of three years?
- **RQ3**: How do hate groups leverage the power of emojis, news sites, and hashtags as forms of expression in their discourse?

The goal of **RQ1** is to derive an in-depth characterization of the temporal psycholinguistic trends for analytic and emotional content as well as investigating the primary drives, needs and future plans for different hate ideologies. On the other hand, **RQ2** focuses on the semantics of hate group discourse specifically how topics discussed within different hate ideologies are intertwined. While **RQ3** sheds light on new methods leveraged by hate group members to encode hateful content by using symbolic emojis and persistent hashtags.

Due to the lack of public hate speech datasets that include online hate communities, we curate a large scale dataset, comprised of approximately 4.7M tweets, that captures the Twitter presence of 24 hate groups belonging to eight of the most prominent hate ideologies in the U.S. from January 2015 to December 2017. This chapter presents the following contributions:

- This is the first study that investigates the discursive practices of 24 hate groups spanning the most prevalent eight hate ideologies across the U.S. and a time duration of three years (2015-2017).
- We conduct the first temporal linguistic analysis for the different hate ideologies that focuses on temporal psycholinguistic properties, inter-ideology semantics, and

hate symbols.

11.2 Related Work

Next, we survey two lines of research related to the work presented in this chapter: *online abusive language and hate speech* and *online white identity groups*.

11.2.1 Online Abusive Language and Hate Speech

Internet studies is the study of the Internet, and the Internet is a complicated assemblage of people, institution, and technologies designed to allow for the transmission of information between devices [286]. One part of the field of Internet Studies explores the nature of problematic interactions between people, institutions, and groups on the internet such as harassment, bullying, threats, and criminal acts. Critical Internet research has a responsibility to engage with the discourses, ideologies at all levels of the Internet and Internet research. There has been a long history of work in anti-social behavior detection on the Internet. Automated detection of hostile/offensive messages using machine learning models marks back to the work of [181] who proposed the use of decision trees to detect classes of abusive messages. More recently there has been a flurry of work on detecting personal insults and offensive language (e.g. [189, 196]). With the evolution of social media platforms, cyberbullying on such platforms has also been studied extensively, *e.g.,* Twitter [196, 176] and YouTube [182]. Successes in automatic detection of offensive messages have led to the development of automated methods for identifying and detecting hate speech as well [14, 287].

Most closely related to our work is the linguistic analysis done by [209] to study the subtle nuances in the language of Directed hate (hate speech aimed at a specific person or entity) and Generalized hate (hate speech aimed at a community sharing a protected attribute) [209]. In [209], the authors conduct a linguistic analysis aimed studying the lexical, psycholinguistic, and semantic differences between Directed hate and Generalized hate. Similar to [209], we leverage a linguistic-centric methodology to conduct our analysis. However, our focus is not on analyzing isolated instances of hate but rather on understanding language of communities of online hate groups. Additionally, and as opposed to [209], our study incorporates a temporal dimension for the linguistic analysis that allows us to study the evolution of hate content for the duration 2015-2017. Moreover, we adopt different methods such as Skip-Thought vectors, time series decomposition, emoji detection, and measuring prevalence and persistence of hashtags in order to understand the nuances of hate language in these communities.

11.2.2 White Identity Groups

The Internet has been a safe space for early white supremacist forums such as Stormfront.org by providing users the feeling of belonging to an online community [288]. Prior work that focused on content analysis revealed that white supremacists use forums primarily for information provision, recruitment and networking [278]. Further, white identity subgroups tend to link each other online [289].

Research on white identity groups on social media has primarily focused on Twitter [290, 291, 292]. [290] showed that words used by extremist groups are not isolated from terms used in the mainstream political discourse [290]. Similarly, white extremists try to move their racial ideology into mainstream political discourse through mixing hyperlinks of extremist web pages with hyperlinks from mainstream pages, referred to as "information laundering" [291]. Others have shown that the number of offensive and hateful tweets from members of the alt-right have increased during the 2016 election campaign [292].

Most closely related to our work in this area is [293]. [293] investigate comments and video content in a set of right-wing Youtube channels and compare it to a set of baseline channels [293]. Similar to our work, they conduct lexical and topic analyses for video comments. However, their analysis spans the period September to October 2017 and focuses only on right-wing YouTube channels. In contrast, our dataset captures eight different hate ideologies over a longitudinal duration of three years. Additionally, we conduct a detailed psycholinguistic analysis and pinpoint the exact topics that trigger the psycholinguistic categories. Moreover, we investigate temporal semantic similarities between the different hate ideologies and investigate hate symbols used in their discourse.

11.3 Theoretical Frameworks

In this section, we explore and situate our work within theoretical-based frameworks that have guided the design of our research questions and the interpretation of our results.

11.3.1 Hate Models

Shafer and Navarro [294] found that hate groups develop in multiple stages. They proposed a seven stage hate group model: (i) the haters gather; (ii) the hate group defines itself; (iii) the hate group disparages the target; (iv) the hate group taunts the target; (v) the hate group attacks the target without weapons; (vi) the hate group attacks the target with weapons; (vii) the hate group destroys the target. The authors note a transition occurring during the stages when the hate group changes from verbal to physical attacks, which differentiates ”hard-core haters” from ”rhetorical haters” [294]. The study emphasizes the importance of understanding and identifying hate groups language, especially before the transition from verbal to physical abuse.

Building onto this hate model, the Power Devaluation Model could further explain the emergence of hate groups. The Power Devaluation Model argues that right-wing extremist groups arise when their power is threatened in one or more of the following areas: economics, politics, and cultural status [295]. McVeigh uses this model to explain the dynamics and rising of the Ku Klux Klan (KKK) movement. Additionally, the appearance of hate groups could be explained by Group Position Theory. Group Position Theory suggests that when an outgroup's increasing size threatens the ingroup, this results in the ingroup reacting, oftentimes negatively, towards the perceived threat [296].

Using Twitter, a social media platform, we integrate the first four stages of Shafer and Navarro's hate model in our study to observe the vocalization of hate groups through their use of public tweets. During data interpretation of our collected tweets, we incorporate the Power Devaluation Model and Group Position Theory to our analysis.

11.3.2 Symbolic Group Convergence

Symbolic Convergence Theory (SCT) is a general theory of persuasive language in which groups create and share goals about the group and outside groups and therefore build a shared identity. The theory provides an explanation accounting for the creation, raising and maintenance of group consciousness through communication. Through stories and rituals the members of a group create a common consciousness – a shared perception of the group and what it means to be a member [297, 298, 299]. We adopt SCT in our work by incorporating hate group language which can form and organize meaning, manipulate emotions, and motivate people to act either peacefully, verbally, or physically.

11.4 Datasets and Methods

11.4.1 Definitions

The FBI does not officially keep a record of hate groups. Hence, we utilize data collected by the Southern Poverty Law Center (SPLC). The SPLC, an organization founded on the basis of protecting the rights of minorities and impoverished individuals, tracks hate and other extremist groups across the United States [300]. Their data on hate groups has arguably become the most widely known and accepted in the United States, in part due to their exhaustive list of hate groups and the limitations of other available data [301]. Their data has been used in prior hate research such as [302, 303].

In our investigation, we focus on studying eight types of hate ideology: *White Nationalist* (WN), *Black Nationalist* (BN), *Ku Klux Klan* (KKK), *Anti-LGBT* (A-LGBT), *Anti-Muslim* (A-MUS), *Neo-Nazi*, *Anti-Immigrant* (A-Immgr), and *Racist Skinhead* (Rac-Skin). We select these ideologies because they constitute the largest presence in the U.S. with respect to the total number of hate groups[4] with the following percentages: Black Nationalist (24.4%), Ku Klux Klan (7%), Neo-Nazi (12.7%), Racist Skinhead (7.4%), White Nationalist (10.5%), Anti-Immigrant (2.3%), Anti-Muslim (11.9%), and Anti-LGBT (5.3%).[5] The detailed themes and core values behind each hate ideology are discussed in the SPLC ideology section[6] and in greater detail in literature of discrimination and racism [296, 294, 295].

11.4.2 Data Curation

Our study and curation methods are motivated by three pivotal points:

[4] https://www.cnn.com/2017/08/17/us/hate-groups-us-map-trnd/index.html

[5] These percentages are calculated based on the map published by the SPLC in April, 2018. Source: https://www.splcenter.org/hate-map

[6] https://www.splcenter.org/fighting-hate/extremist-files/ideology

- The evolution of the Internet as a place where both active and passive radicalization of marginalized individuals occur.

- The emergence of social media content as being tied to a desire to be informed and educated [304]. A Pew survey administered in 2017 revealed that 67% of Americans use social media to access news (a 5% increase since 2016). This trend cuts across age groups and platforms; however, the social media platform most associated with use for news is Twitter, with 74% of surveyed Twitter users reporting their use of the media for news access [304].

- The prominence of online hate groups, the dissemination of their ideology, and the growth of their online audiences [305].

We use Crimson Hexagon (CH)[7], a real-time web-based library of social media posts, as an interface to the Twitter Firehose[8], which guarantees delivery of 100% of the tweets that match a certain criterion. For each of the eight hate ideologies, we collect a set of Twitter handles based on hate groups identified by the SPLC.[9] We select the three hate groups with the highest number of followers for each hate ideology. We then proceed to use CH to collect all tweets from each selected hate group from January 1, 2015 to December 31, 2017 – a longitudinal time duration that allows us to analyze the temporal evolution of hate groups over a three year span. Note that due to the sensitive nature of the data, we anonymize references to the three Twitter handles for each hate ideology by using the acronyms (hg_1, hg_2, hg_3) in Table 11.1.[10]

Note that during the course of analysis, we discovered that as of April 2018, two of three WN groups for which we collected tweets, one of three groups for each of RacSkin,

[7] https://www.crimsonhexagon.com/

[8] https://developer.twitter.com/en/docs/tweets/sample-realtime/overview/decahose

[9] https://www.splcenter.org/fighting-hate/extremist-files/groups

[10] We include the names and the Twitter handles of the hate groups in the Appendix for the reviewers.

Neo-Nazi, and BN, and all three KKK hate groups were removed from Twitter. As a result, our dataset provides a unique, unreproducible lens for studying these hate groups from a social computational perspective. Table 11.1 provides an overview of the number of tweets collected per year for each hate ideology and for the top three hate groups in each category.

Hate ideology	hg_1	hg_2	hg_3	**2015**	**2016**	**2017**
White Nationalist	91,610	228,504	112,241	16,373	73,510	342,472
Black Nationalist	30,919	2,144	11,542	5,008	16,456	23,141
Ku Klux Klan	1,311	12,155	1,642	9,248	4,839	1,021
Anti-LGBT	176,332	388,173	268,820	134,814	134,824	563,687
Anti-Muslim	348,272	890,800	891,320	502,946	582,628	1,045,814
Neo-Nazi	28,554	87,986	53,978	60,144	64,570	45,804
Anti-Immigrant	34,168	35,197	1,051,084	159,024	397,104	664,321
Racist Skinhead	550	396	542	508	577	403
Total	-	-	-	788,065	1,273,498	2,686,663

Table 11.1: Overview of the number of tweets collected for each hate group and the total tweets per ideology broken down by year.

11.5 Analysis

11.5.1 Temporal Psycholinguistic Patterns

To analyze temporal patterns, we use the additive model of Time Series Decomposition to deconstruct a time series into three components: trend, seasonality, and noise [306]. This statistical technique has been widely deployed to understand the complex nature of time series. This approach is optimal for our analysis since our main focus is to inspect trends in hate ideology language while decoupling seasonal and noise factors.

Language provides us with information individuals select from their internal dialogue. To understand the linguistic dimensions and psychological processes identified among hate groups, we use the psycholinguistic lexicon LIWC2015 [262]. LIWC2015 analyzes language by comparing each word in some text (target words) against the LIWC2015 dictionary, which consists of almost 6400 words, each corresponding to one or more psychological dimensions. It then computes the frequency of words in the text that are associated with each category. In particular we focus on the following categories: Analytical Thinking, Cognitive Processes, Emotional Tone, Drives, and Time Orientations. A detailed description of LIWC dimensions can be found in the LIWC2015 language manual [262]

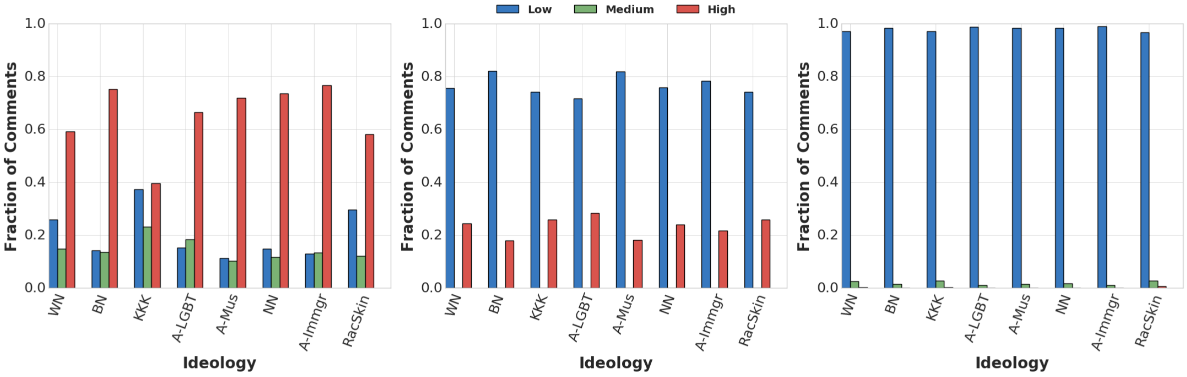
RQ1a: How do analytical thoughts or emotional experiences prevail in hate group content?

Prior studies of hate group websites have concluded that hate groups use cognitive skills as a way of logically convincing audiences of their ideologies [307]. For example, Lacy studied white supremacists' "othering" behavior and found that the behavior relies on a language that emphasizes causality [308].

To determine whether the premise of previous literature holds in online social networks, we examine the analytical thinking (Analytic) dimension from LIWC, which captures the degree to which people use words that suggest formal patterns. Additionally, we investigate language indicative of cognitive processes (CogProc). Higher cognitive processes have been correlated with more complex language that invokes insight, causation, and discrepancies. To inspect emotional tone, we examine the Tone metric, in which a high score is correlated with a more positive upbeat style; a low score reveals greater anxiety, sadness, or hostility.

Figure 11.2 depicts the fraction of total comments[11] in our dataset for each hate ideology that exhibit Analytic, Tone, and CogProc LIWC scores. We designate labels to LIWC scores for each of the three LIWC categories as follows: comments that had a LIWC score under 33.0 were labeled 'Low', comments that had a LIWC score between

[11] We use the terms "comments" and "tweets" interchangeably.



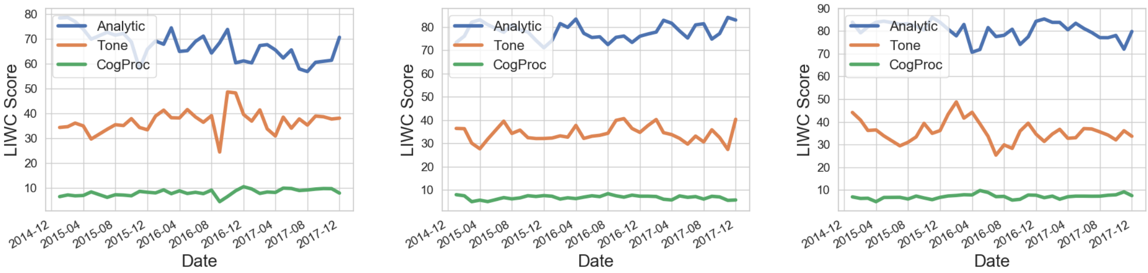
(a) Analytic (b) Tone (c) CogProc

Figure 11.2: Analytic, Tone, and CogProc score ranges for all hate ideologies. Note that more than 60% of comments have high Analytic scores and more than 90% of comments have low CogProc scores in all hate ideologies.

	CogProc	**Tone**	**Analytic**
CogProc	1.0	-0.075	-0.87
Tone	-0.075	1.0	-0.43
Analytic	-0.87	-0.43	1.0

Table 11.2: Pearson's correlation coefficients between Analytic, Tone, and CogProc scores. Positive scores denote positive correlation: as one variable increases, the other variable tends to increase. Negative scores denote negative correlations: as one variable increases, the other variable tends to decrease.

33.0 to 66.0 were labeled 'Medium', and comments that had a LIWC score above 66.0 were labeled 'High'. We observe that all hate ideologies harbor content with higher analytical structure and lower emotional tone and cognitive processes. We calculate the inter-correlations among Analytic, Tone, and CogProc scores of all hate ideologies and present the data in Table 11.2 using Pearson's correlation coefficient. Table 11.2 shows the relationship between analytical thought and cognitive processes to be highly negatively correlated (-0.87), suggesting that as hate groups tend to express their ideas with very formal language that exhibits lower tendencies for insight and causation. The Table also shows that there is a medium negative correlation (-0.43) between Analytic and Tone which indicates that oftentimes the formal language is intertwined with negative



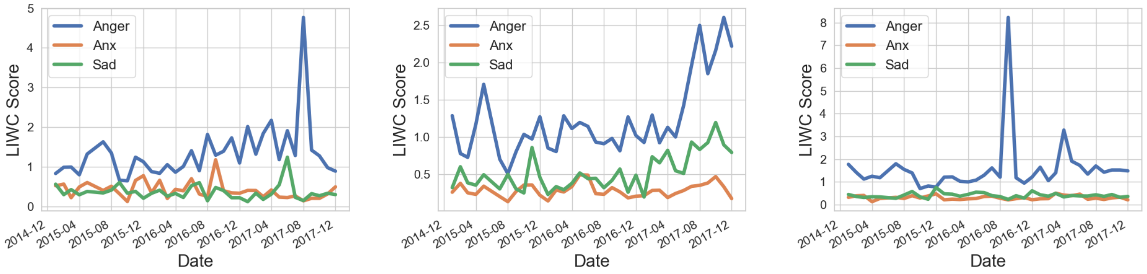
(a) WN (b) NN (c) A-Immgr

Figure 11.3: Trends of Analytic vs. Tone vs. CogProc scores for WN, Neo-Nazi, and A-Immgr ideologies. Note the higher scores for Analytic and the lower scores for Tone and CogProc.

emotions.

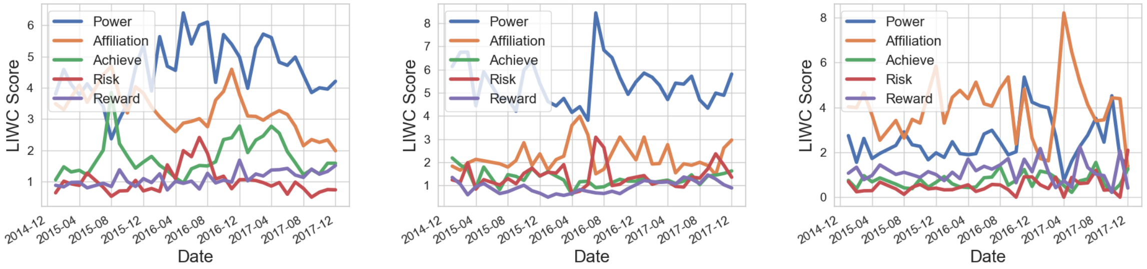
Furthermore, we analyze changes of these scores over time and present the weekly Analytic, Tone, and CogProc mean score trends for WN, NN, and A-Immgr in Figure 11.3. The figure shows that across 2015-2017, the same trends persist across all hate ideologies: hate groups tend to focus on content that contains analytical language and negative emotional tone. An example of Analytical Thinking for WN is: *”Hah! Hacked, infected, is always a possibility. Again, I’ll not debate the various levels of intellectual approach taken by bitter people.”* An example of content depicting Cognitive Processes: *”I think you’re scared to venture out of your safe space and confront uncomfortable truths about the world around you.”*

Across the spectrum of negative emotions, anger was found to be more prevalent than anxiety (Anx) and sadness (Sad) across all hate ideologies. For most of the hate ideologies, the average anger score was found to be quasi-constant with the exception of a general increase in A-LGBT anger as well as occasional spikes in other hate ideology anger. Figure 11.4 demonstrates the emotional trends of the A-LGBT ideology, as well as those of the WN and A-Immgr ideologies for comparison. Examples of anger found in tweets include: *”SCR*W THE KORAN”*, *”No. F*ck you! It’s your turn b*tch! How does*



(a) A-Immgr (b) A-LGBT (c) WN

Figure 11.4: Trends for a negative emotion components (anxiety, anger, and sadness) for A-Immgr, A-LGBT, and WN ideologies. Notice the increase in anger of the second ideology, and the spikes in anger for the A-Immgr and WN ideologies.



(a) A-LGBT (b) A-Immgr (c) KKK

Figure 11.5: Trends for hate ideology drives (affiliation, achievement, power, reward, and risk) for A-LGBT, A-Immgr and KKK ideologies. Note that power motivates A-LGBT and A-Immgr ideologies the most, while KKK groups are most motivated by affiliation.

it feel to know you are being hunted? Scared? Good.”, ”*No time for sentimentality the deportations must be cold calculating and ruthless!”*, and ”*send them back”*. The angry nature of these groups is a natural symptom of their hateful philosophies and goals.

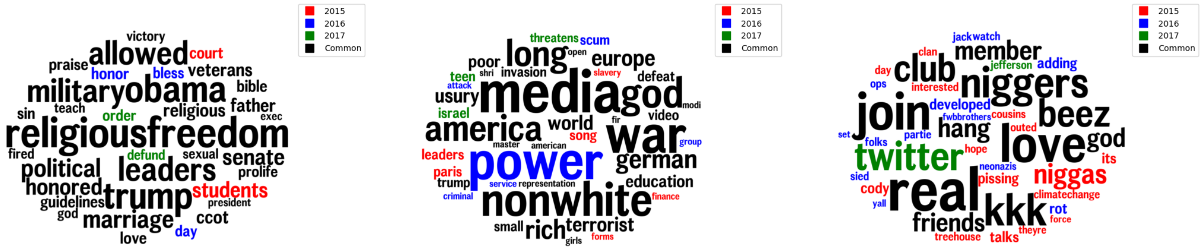
Many of the observed spikes in anger are triggered by offline events. For example, there is a WN spike that occurs in September 2016. The primary trigger for this spike lies in the large number of retweets of the tweet *“Hispanic activist @TonyYapias, who criticized Trump’s comments about Mexican rapists, has been charged with rape.”* The use of logic in this tweet is inline the observation that hate groups tend to employ analytical language. The A-Immgr hate ideology experiences a large spike in anger in August 2017. This increase can be attributed to the shooting in Fresno, California in which a black man, who had previously expressed anti-white sentiments, killed three white men. The anger originates from the ideology's perceived lack of public response: *"Police chief: #Fresno killings a hate crime 3 innocent white males murdered in the streets. Where is the outcry?"*. Once again, this peak in anger correlates with an analytic argument. It is also important to note that the triggering event is related to race rather than immigration-specific issues. This reveals the A-Immgr ideology's ability to extend their content from their main immigration-related discourse to more generally race-oriented messages.

RQ1b: What are the primary drives and needs that affect different hate ideologies?

Humans are motivated to use language and engage in actions based on drives such as affiliation, achievement, power, reward, and risk. For almost every group we analyzed, power emerged as the strongest motivator in their natural language, as shown in Figure 11.5. The KKK was an outlier in this respect because its strongest motivator was affiliation. The KKK's differing motivation could be explained by the fact that this hate ideology is a uniquely structured and official organization, in contrast to the other groups in our study. The notorious hierarchical ranking system of the KKK, from Klokard to Grand Wizard, could justify the tendency for KKK members to be motivated by social affiliation rather than power. Figure 11.6c displays the word cloud of the KKK's most affiliation-driven topics and reveals the group-oriented nature of the hate ideology. All word clouds in Figure 11.6 were obtained by training an LDA model per month per ideology on 30 topics and 30 keywords on the top 200 tweets that month over a threshold of LIWC score 10. The topics that indicate affiliation drivers consist of inclusive terminology such as "join", "club", "friends", and "love". This inviting vocabulary demonstrates how the cult attempts to promote a friendly and social image, paradoxical to their hateful message. The social nature of the KKK is contrasted by the A-LGBT ideology's motivation of power. The word cloud of A-LGBT topics that are most motivated by power are very politically charged, as demonstrated in Figure 11.6a. Power-driven A-LGBT discourse is dominated by controversial political topics such as "trump," "obama," "marriage," "religiousfreedom," and "freespeech." The fact that A-LGBT's most power-motivated topics are so politically oriented reveals that this group's sense of power lies in politics. The primary keyword in Figure 11.6b, "media", originates from Neo-Nazi's distrust of media. For example, the tweet *"#Jews admit control of the #media - #msm #Israel #WakeUp"* suggests that Neo-Nazis derive power from the perceived knowledge they maintain about media and its control.

RQ1c: How is the future-tense used by different hate ideologies?

To analyze the future calls for the different hate ideologies, we leverage LIWC's Time Orientation category and investigate the future focus scores. These future focus scores are correlated with the usage of words such as "will" and "soon" [262]. We then extract



(a) A-LGBT

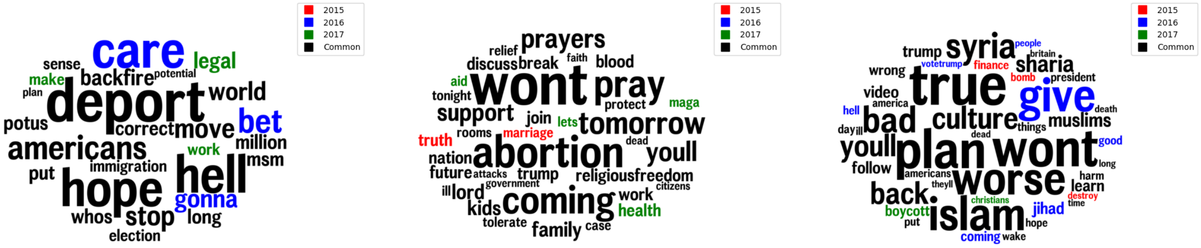
(b) NN

(c) KKK

Figure 11.6: Word clouds of the topics that represent the strongest drivers of the KKK, Neo-Nazi, and A-LGBT hate ideologies. The A-LGBT and Neo-Nazi ideologies are most motivated by power, whereas the KKK is most motivated by affiliation. The black "common" terms represent words that triggered the motivation in more than one year.

the tweets with a focus on the future based on the LIWC's future focused scores and train an LDA topic model to extract the themes related to the ideologies' future calls. Figure 11.7 shows the word clouds of the topics associated with future-focused tweets for several ideologies.

Because these word clouds demonstrate the topics discussed in a future tense, they provide insight into the groups' calls for action: what do they see happening in the future and what do they want to accomplish in the future? As shown in Figure 11.7a, "deport" dominates the Anti-Immigration ideology's future oriented tweets. For example, the A-Immgr tweet "Don't worry. Criminal aliens will and are going to be deported. Stay strong with the President. Look at the alternative for God's sake!!" explicitly demonstrates the goals of this hate group. For A-LGBT, "Pray" and "abortion" are two of the largest key words in the ideolgy's future-focus word cloud (pictured in Figure 11.7b), revealing the prevalent religious themes in the hate ideology's future-focused discourse. Examples of these religious-oriented and future focused A-LGBT tweets include *"Churches should never be forced to cover elective abortion in their insurance plans"* and *"@AllianceDefends #PPMURDERSBabies #ArrestPPAbortionists & staff! Triple your fasts, prayers, marches. Teach your children to respect life!"*. A-LGBT's discussion of abortion (which is most prominent in their future-focused discourse, as seen in Figure 11.7b), reveals the ideology's ability to extend their topics of discourse from typical issues of gender and sexuality to religious issues like the abortion debate. Religion is also seen as a major player in the ideology's use of the word "Prayer," which serves as a common call to action for the group. An example of this appears in the tweet *"The #March4Marriage happens today! Pray for the marchers, Supreme Court Justices, and the future of #marriage! #1m1w."*. These religious themes prevail



(a) A-Immgr (b) A-LGBT (c) A-MUS

Figure 11.7: A-LGBT, A-Immgr, and A-MUS topic word clouds that correlate with the strongest future-focused tweets. Words are color-coded based on the year in which the topic occurred, and are black if they appear in more than one year.

11.5.2 Semantic Similarity and Dynamic Topics

Our research includes an analysis of a variety of hate ideologies, each promoting a different message. One way of identifying overlap between the ideologies is by investigating users who engage in and follow these hate groups. Since some of these groups were taken down by Twitter, it was impossible for us to retrieve the list of users in a given group

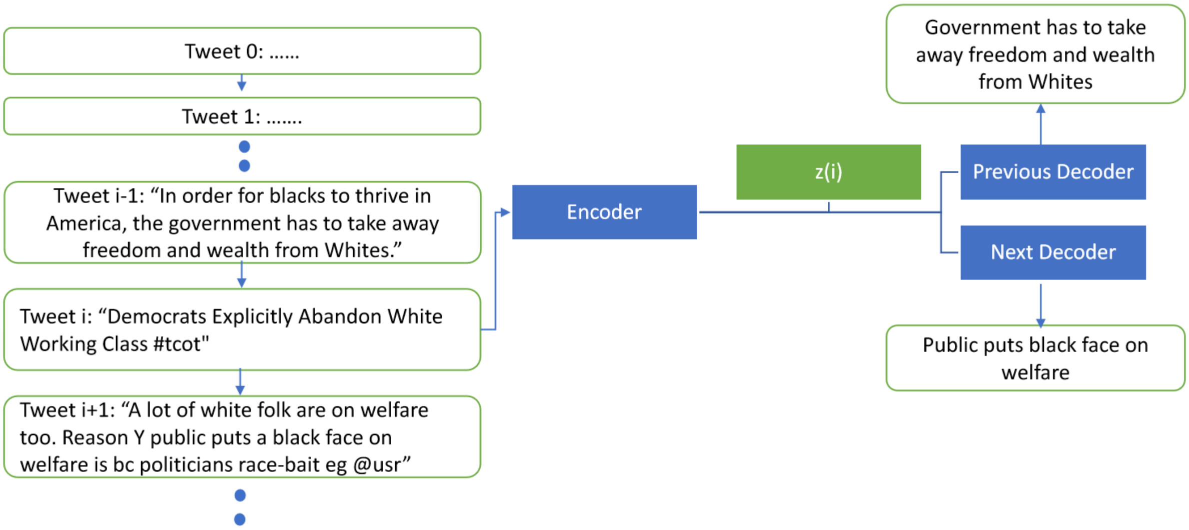

Tweet 0: ......
Tweet 1: ....
Tweet i-1: “In order for blacks to thrive in America, the government has to take away freedom and wealth from Whites.”
Tweet i: “Democrats Explicitly Abandon White Working Class #tcot"
Tweet i+1: “A lot of white folk are on welfare too. Reason Y public puts a black face on welfare is bc politicians race-bait eg @usr”
Encoder
z(i)
Previous Decoder
Next Decoder
Government has to take away freedom and wealth from Whites
Public puts black face on welfare

Figure 11.8: Architecture of the Skip-Thought RNN. During training, the Previous and Next Decoders attempt to predict the previous and next tweets. The result is capturing a vector representation of a tweet (i) in $z(i)$.

which deemed this approach infeasible. In light of the differences in ideologies and the implications of Symbolic Convergence Theory, it is interesting to study how the hate ideologies overlap semantically. Are their semantics as different as their philosophies? Are there any discussed topics that build a sense of shared identity among these ideologies? To answer this research question we use two approaches: A deep learning method via Skip-Thought Vector similarity and a lexical method using a topic-level Jaccard Index (JI) similarity.

For the deep learning method, Skip-Thought vectors proved to be an effective way to measure the similarity between two sentences, or in our case, between two tweets [309]. We train a Skip-Thought encoder, a bidirectional Recurrent Neural Network (RNN) trained on the BookCorpus dataset. The model is composed of two 1200-dimension encoders that encode sentences, with one receiving the sentence in the correct order and the other receiving the sentence backwards. Training lasted two weeks and involved using two decoder networks, that when given a sentence attempt to recreate the previous and next sentence while minimizing reconstruction error, which motivates the encoder to optimize the information contained in the sentence representations via back-propagation. By forcing the two decoders to predict the previous and next tweet, the meaning of the tweet is then captured in a vector representation $z(i)$ for a given tweet i.[12]. The model's architecture is shown in Figure 11.8. After using this model to generate encodings for every tweet in our dataset, we compute the cosine similarity between the Skip-Thought Vectors of every pair of tweets belonging to different hate ideologies. This computation was done on a monthly basis, and the final average distance between two ideologies was measured as the yearly average (for 2015, 2016, and 2017) of the monthly averages of vector distances.

[12] The model was pretrained by orthogonally initializing recurrent matricies and initializing non-recurrent weights with a uniform distribution. To optimize, the Adam algorithm was used [310]

For the lexical approach, we trained an LDA topic model [311] with a distribution of 30 topics and 15 keywords per topic for each hate ideology per month, using their stemmed and lemmatized tweets. This results in a list of 450 words representing the prominent themes of discourse for these ideologies per month. To measure the pair-wise similarity between pairs of different ideologies, we words the Jaccard Index (Similarity Coefficient) for the resultant topics from the previous step for each month so that we can capture timely linguistic similarities. The Jaccard Index (JI) of two sets A and B is calculated as follows

$$\frac{|A \cap B|}{|A \cup B|}$$

. We computed the Jaccard Index for pairs of hate ideologies by dividing the size of the intersection of the sets of topic keywords of the two ideologies by the size of the union for each month, and then taking the yearly average.

Kendall's tau, a statistic measuring correspondence between two sets of rankings, was used to measure the correspondence between the rankings produced by the deep learning and lexical methods. These values are shown in Table 11.3. The Kendall's tau value of these two approaches decreased between years over the three year time period. The average delta, however, remained small in all three years, and even did not change from 2016 to 2017, when the Kendall's tau value dropped by a difference of 0.244. The differences in method analysis show that while Skip-Thought Vectors and JI produced different rankings of pair-wise similarities, the actual similarity values that were produced varied very little. This indicates that both methods are really close in assessing the similarity between different hate ideologies.

RQ2a: Which hate ideologies are more semantically similar to other ideologies? How does the inter-ideology similarity evolve over time?

Tables 11.4, 11.5, and 11.6 enumerate the top and bottom semantic similarity pairings

Year	Kendall's Tau, p-value	Average Delta
2015	0.624, 0.000003	0.038
2016	0.450, 0.0008	0.052
2017	0.206, 0.123	0.052

Table 11.3: Measurements of how much the similarity rankings produced by Skip-Thought Vector similarity and topic-level Jaccard Index correspond. A Kendall's tau value of 1 demonstrates strong agreement, and a value of -1 demonstrates strong disagreement. The average delta value measures the average difference between each similarity value produced by the two methods.

I1	I2	JI	I1 Keywords	I2 Keywords	Topic Theme
A-LGBT	A-MUS	0.324	jihadists, islam, religion	taliban, sharialaw, nationalsecurity	Islamic Terrorism, Immigration
A-MUS	Neo-Nazi	0.310	nazis, nukes, terrorists	brotherhood, taliban, jihad	White Supremacy, Islamic Terrorism
Neo-Nazi	WN	0.300	whitegenocide, socialism, whitelivesmatter	thuglivesdontmatter, parisattacks, nazism	White Supremacy
A-LGBT	A-Immgr	0.288	refugee, muslim, terrorism	isis, christian, amnesty	Islamic Terrorism, Immigration
A-LGBT	RacSkin	0.275	anti-gay, blacks, terrorists	ppsellsbabyparts, police, isis	LGBTQ,Planned Parenthood, Islamic Terrorism
RacSkin	A-MUS	0.081	israel, irandeal, hillary	jihadists, islamism, benghazi	Islamic Terrorism, Hillary Clinton
RacSkin	WN	0.078	whitegenocide, aryan, charliehebdo	whitelivesmatter, isis, onepeopleonenation	White Supremacy, Islamic Terrorism
RacSkin	A-Immgr	0.076	secureourborders, hispanic, refugees	illegals, assimilation, daca	Immigration
RacSkin	KKK	0.064	terrorists, isis, patriot	refugees, whiteisright, whitesupremacist	Islamic Terrorism, White Supremacy
RacSkin	BN	0.058	nazi, racism, gays	alt-right, hatecrime, segregation	White Supremacy

Table 11.4: The Jaccard Index of the semantic similarities of the five most semantically similar and five least semantically similar pairs of hate ideologies in 2015 as well as their commonly discussed topics.

according to JI for 2015, 2016, and 2017, respectively. To provide context for each pairing's semantic similarity, we included keywords indicative of common discourse, which are also listed in the tables. These keywords were sampled from the intersection of the two ideologies' topics that were produced during our lexical semantic similarity analysis.

The top and bottom five semantic similarity pairings for 2015 are shown in Table 11.4. In 2015, the Jaccard Index for semantic similarities ranged from 0.058 to 0.324, with the most semantically similar pairing being between the A-LGBT and A-MUS ideologies. The semantic similarity in this pair lies in A-LGBT's ability to expand from its core message to topics commonly discusses by other ideologies, while the A-MUS ideology did not deviate from its mainstream discourses. For example, the A-LGBT ideology shares

I1	I2	JI	I1 Keywords	I2 Keywords	Topic Theme
A-MUS	Neo-Nazi	0.234	trump, makeamericagreatagain, presidentialelection	wikileaks, buildthewall, hillary	Presidential Election
WN	Neo-Nazi	0.225	banislam, rapefugees, whitegenocide	jihad, deportation, whitepride	Immigration, Islam, White Supremacy
Neo-Nazi	A-Immgr	0.195	rapefugee, ice, sanders	illegalimmigrants, securetheborders, maga	Immigration, Presidential Election
A-LGBT	A-MUS	0.191	jihad, electionday, benghazi	isis, presidentialelection, debatenight	Islamic Terrorism, Presidential Election
A-MUS	WN	0.189	nazis, terrorism, trumpstrong	whitegenocide, isis, hillary	White Supremacy, Islamic Terrorism, Presidential Election
RacSkin	A-MUS	0.085	neverhillary, killary, whitelivesmatter	crookedhillary, blm, russia	Presidential Election, Black Issues
RacSkin	A-LGBT	0.075	transngender, bathroom, womenfortrump	gays, media, trump	LGBTQ, Presidential Election
RacSkin	A-Immgr	0.074	trumptrain, makeamericagreatagain, muslims	treason, russia, isis	Islam, Presidential Election
RacSkin	BN	0.068	crookedhillary, maga, #islamisevil	protestors, wakeupamerica, neverhillary	Islam, Presidential Election
RacSkin	KKK	0.065	xenophobic, ferguson, trumptrain	maga, illegals, treason	Black Issues, Immigration, Presidential Election

Table 11.5: The Jaccard Index of the semantic similarities of the five most semantically similar and five least semantically similar pairs hate ideologies in 2016 as well as their commonly discussed topics.

I1	I2	JI	I1 Keywords	I2 Keywords	Topic Theme
Neo-Nazi	WN	0.214	whitegenocide, anti-white, russia	multiculturalism, putin, russia	White Supremacy, Russia
A-MUS	Neo-Nazi	0.192	marchagainstsharia, jihad, makeamericasafeagain	sharia, terrorists, nucelear	Islamic Terrorism
A-MUS	WN	0.191	terrorism, illegals, daca	stopsanctuarycities, noamnestytrump, terrorists	Terrorism, Immigration
A-MUS	A-Immgr	0.185	travelban, deportations, daca	deportdaca, muslims, maga	Immigration
A-LGBT	WN	0.185	altright, antifa, blm	pro-white, supremacy, blm	White Supremacy, Black Issues
A-MUS	KKK	0.078	russiagate, brotherhood, marchagainstsharia	jeffsessions, genocide, isis	White Supremacy, Islamic Terrorism
A-Immgr	KKK	0.076	wall, trump, nodaca	migration, murders, deportation	Immigration
KKK	BN	0.069	oppression, blacklivesmatter, charlottesville	domesticterrorists, blm, thebluekluxklan	White Supremacy, Black Issues
RacSkin	BN	0.060	blacklivesmatter, genocide, racism	police, racism, policebrutality	Black Issues
RacSkin	KKK	0.056	wall, immigrants, anti-antifa	refugees, nazis, confederatestatues	Immigration, White Supremacy

Table 11.6: The Jaccard Index of the semantic similarities of the five most semantically similar and five least semantically similar pairs of hate ideologies in 2017 as well as their commonly discussed topics.

in the A-MUS ideology's discourse related to Islam, as demonstrated in the example A-LGBT tweet *"Islam is a death cult not a religion. Besides we'll shoot to kill ISIS. #tcot #ctot #ccot #pjnet #2A #1"*. The A-LGBT ideology also shared a relatively high semantic similarity with the RacSkin hate ideology, even though RacSkin had the lowest semantic similarity scores with every other hate ideology. This pairing dropped to the lowest ten scores in 2016 and 2017. In 2015, these two ideologies shared discourse on a number of topics, including typical A-LGBT content like planned parenthood and homosexuality, as well as black allegations of police brutality, the refugee crisis and islamic terrorism. One odd semantically similar pairing lies in the relationship between WN and BN. These opposing ideologies scored high on the Jaccard Index – in the top 10 for 2015. This similarity could be attributed to their quantitatively equal online discussion of race.

The semantic similarities between ideologies in 2016 decreased in range from 2015 by a value of 0.95, varying from 0.065 to 0.236, as shown in Table 11.5. A-Immgr and Neo-Nazi ideologies rose up five rankings in their semantic similarity, landing in the top three for 2016. The 2016 U.S. Presidential election could be attributed to bringing these two groups together in the topics they discussed, as demonstrated by their shared topics. The election was a common topic among all pairs, with "maga" or "makeamericagreatagain" a shared topic among every pairing. Several anti-Hillary Clinton topic keywords became popular, such as "killary," "neverhillary," and "crookedhillary," also relating to the 2016 Presidential election.

Finally, Table 11.6 shows that in 2017 the average semantic similarity scores dropped yet again, with the top score being 0.214, a whole tenth of a point lower than the top score in 2015. The pair with the greatest topic overlap was the Neo-Nazi and WN pairing. Their greatest point of agreement was in their perceived threat of destruction of the white race. Neo-Nazi tweets such as *"Another (((rabbi))) celebrates the destruction of the white race. It is time to wake up! #WhiteGenocide"* and White Nationalist tweets like *"One people. One nation. End immigration. No #WhiteGenocide on our watch."* demonstrate these feelings. Note the usage of the hate symbol "((()))" to mark a Jewish person.[13]

[13] According to the Anti-Defamation League's hate symbols database, "((()))" is used to refer to Jewish

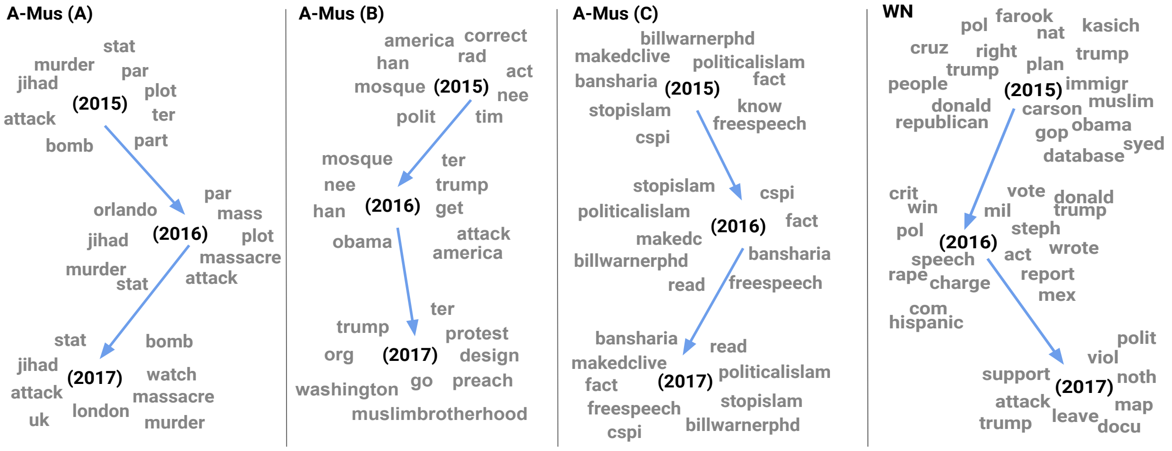


Figure 11.9: Sample dynamic topics associated with A-MUS and WN ideologies. A–MUS (A-C) denote three different dynamic topics discovered in A-MUS discourse while WN denotes one dynamic topic discovered in the WN discourse.

RQ2b: What are the prominent topics that evolved over the course of 2015-2017?

In order to anwer this research question, we train a dynamic topic model based on two layers of Non-negative Matrix Factorization [313]. Dynamic topic models developed to analyze the time evolution of topics in large document collections. We leverage the model to study how hate ideologies have evolved their topics of discourse over three windows: 2015, 2016, and 2017. By training a dynamic topic model on the three windows of 2015, 2016, and 2017, we were able to extract latent thematic topics across the three years among all hate ideologies. Since the output of training the model is a distribution of words changing temporally across the three years, we qualitatively investigate all the output topics to remove noisy topics. By leveraging a qualitative approach, we were able to discern four dynamic topics as depicted in Figure 11.9. Three of these topics belong to the A-MUS ideology and one belongs to the WN ideology. The first

persons by Anti-Semites online [312].

interesting observation lies in the recurrence of certain hashtags of the A-MUS ideology depicted in A-MUS (C) of Figure 11.9: #makedclisten, #stopislam, #billwarnerphd and #politicalislam. The third hashtag refers to the major conservative figure Bill Warner, who is known for his critiques of Islam. The last hashtag refers to a term coined by Bill Warner. These hashtags frequently accompany each other in this hate ideology's tweets, reaffirming that hate groups attempt to use logic (demonstrated by Warner's arguments) to convince people of their message. Another one of A-MUS ideology's topics – jihad, attack, murder, massacre, bomb, plot, mass – has evolved to reference major events shown in A-MUS (A) of Figure 11.9. In 2016, the keyword "orlando" appears in this topic, referring to the tragic Pulse nightclub shooting. Members of this hate ideology appropriated the shooting to validate their own politics, as demonstrated by the tweet *"People remember Islamists and Islamism kill. This is a war. You must act on it or it may kill you. #orlando #islam"*. Another A-MUS topic demonstrates an increase in political tendencies in the ideology, possibly triggered by the 2016 presidential election. In 2016, the topic began to include "obama" and "trump," and then in 2017, the topic included "trump," "preach," "protest," and "washington." This is illustrated in A-MUS (B) of Figure 11.9.

BN also likes to make references to real-world events. A topic that included "freddiegray" in 2015 evolved to include "altonsterling" and "philandrocastile" in 2016. All of these keywords are references to black men–Freddie Gray, Alton Steling, and Philandro Castile–who were murdered by the police, and are a testament of the black American community to the police brutality that they endure.

The KKK experiences a stagnant topic throughout the three windows: all, hood, ku, klux, klan, well, unite, brother, and white. This could be a one-to-one correspondence with the tweet *"KKK alive and well white brother hood unite Ku Klux Klan www.kkk.com,"*, which is retweeted frequently in each year. This attempt at recruitment has a very inviting tone, and connects back to the observation made that the KKK is driven by affiliation rather than power, as in the rest of the groups. The topics discovered by dynamic topic modeling for WN ranged from immigration and politics to riots and recruitment. One of its topics discussed the politics of the 2016 presidential election simultaneously with immigration issues and is depicted in WN in Figure 11.9. In 2015, several GOP presidential candidates were referred to by topic keywords such as "donald," "trump," "kasich," "carson," and "cruz." These political figures shared topic space with the discussion of Muslim-related immigration. "Database" appeared as a keyword in this topic in 2015, in reference to Trump's proposed Muslim registry. This keyword was accompanied by the words "syed" and "farook," referencing one of the shooters in the 2015 San Bernardino attack. One WN directly links the Muslim registry to this tragic event: *"Syed Farook would have been in the Trump database"*. Discussion of immigration propagated to 2016 in this topic, however it shifted from Muslim-oriented immigration to Hispanic immigration, as the key words "hispanic" and "mexican" are introducted to the topic model. The WN tweet *"If you think we're worrying now about Hispanic percentages just wait a few years"* demonstrates the WN's perceived threat of Hispanic immigration. Another interesting WN topic discovered by dynamic topic modeling concerns recruitment. The keywords "#projectsiege" "university," "texas," "washington," and "ca" all appear in this topic. The hashtag is a WN recruitment effort that targets college students. Violence and confrontation are also discussed by WN. The topic started in 2015 by consisting of "protest," "rally," and "battle" and evolved to specifically discuss the riots started by WN in Berkeley in 2017 using the terms "antifa" and "berkeley." The ideology viewed the event as a heroic act on the part of WN: *"CHEERS Today was an awesome day seeing Heroic Men of the #AltRight n beat back #Evil antifa blm SCUM#berkeley"*.

11.5.3 Other Forms of Expression

The rise of social media has inspired a new wave of research on different forms of expression other than text, such as understanding the ecosystems of memes in social networks [314], the utilization of emojis as a means to express emotions online [315], and the application of hashtags to increase visibility and popularity [316]. These online utitilies provide users with an ability to communicate their thoughts and feelings in a brief and sometimes cryptic manner that is understood within their social network. Thus, these alternate forms of expression allow us to add another dimension to our understanding of how hate groups operate online. In this section, we analyze cryptic emojis, news sites, and hashtags that refer to offline violent and collective action events in order to provide insight to how hate ideologies leverage these tools as vehicles to communicate their beliefs.

RQ3a: How are hate groups leveraging emojis and new sites to spread hateful content?

Encoding Emojis as Hate Speech. Hate speech can be expressed in a variety of ways, including symbolic emojis. Symbolic hate symbols have ranged from closed fists to animals, such as frogs, as found in the Anti-Defamation League's hate symbol database [317]. Because social media platforms have become increasingly aggressive in censoring hate speech [318, 319], emojis may play a more vital role in online hate speech. For example, emojis used as hateful symbols may be used in lieu of hateful text to bypass hate detection systems. Thus, in this section, we attempt to analyze the usage of potential cryptic messages behind what appears as benign emojis. We conduct a mixed-method investigation for all emojis in our dataset and outline the emojis that we found to be the most cryptic. [14]. Table 11.7 depicts the top five emojis and their corresponding percentages per hate ideology. We additionally use Word2Vec [320], a neural net that outputs vector representations of text, to infer the common contexts for which these emojis are highly used.

Across each hate ideology, we find that the emoji that occurs the most is the 😂 emoji, with the exception of BN and A-Immgr. The 😂 appears in our dataset in the following percentages: 16.1% for WN, 34.9% for KKK, 22.0% for A-LGBT, 11.5% for A-Mus, 18.5% for Neo-Nazi, and 13.4% for RacSkin. Tweets that occur alongside this emoji are frequently used to ridicule or mock others such as political figures, other users, or news articles. Some examples in our dataset include *"I am drunk on your white tears of pain @usr ! 😂"* and *"@usr 😂 😂 😂 Yall are f*cking losers. Why dont you go shave your head and get back to your macaroni picture frame"*. We also note that this emoji is most commonly used in contexts of the following words: 'lol', 'lmao', 'dude', 'haha', and '😉'. Another interpretation of the use of this emoji as a hate symbol could be to "gloat about human suffering" [321].

Other notable emojis in our dataset include the 🐸 emoji and the 🥛, which comprises 8.0% and 4.2% of all emojis used in WN tweets, respectively. The 🐸 emoji seems to associate with the alt-right, commonly used with "#AltRight" and the phrase "Unite the Right". The frog may also allude to "Pepe the Frog", a meme that was characterized as a hate symbol after its offensive use in the 2016 election season [322]. The 🥛 emoji gained virality as a hate symbol in 2017. This emoji, when used in context of hateful speech, "reinforces notions of white superiority and idealized visions of masculinity" [323]. Within our WN dataset, we find that of all users who provide a name, 2% of users embed the 🐸 emoji in their names. Of tweets that include the 🐸, we note that the 🥛

[14] Because our current dataset is stemmed and lemmatized, we obtain these emojis from the original text before processing.

WN	BN	KKK	A-LGBT	A-MUS	Neo-Nazi	A-Immgr	RacSkin
😂 (16.1%)	🔥 (32.2)	😂 (34.9%)	😂 (22.0%)	😂 (11.5%)	😂 (18.5%)	👍 (4.4%)	😂 (13.4%)
🤔 (8.4%)	👨 (3.9%)	🖕 (6.2%)	😄 (20.9%)	✔ (8.0%)	✔ (7.1%)	👇 (3.7%)	📷 (11.5%)
🐸 (8.0%)	🏆 (2.7%)	😭 (5.3%)	😁 (12.8%)	👉 (7.7%)	👍 (3.8%)	💥 (3.6%)	💜 (9.6%)
🥛 (4.2%)	👥 (2.7%)	💀 (3.3%)	😣 (12.2%)	🎯 (3.7%)	💨 (2.7%)	😡 (3.6%)	😎 (7.7%)
™ (2.7%)	👴 (2.6%)	✊ (2.8%)	😒 (10.0%)	🚨 (2.8%)	😭 (2.5%)	😂 (3.6%)	🔴 (5.8%)
24,188	14,388	3,390	409,106	51,257	2,869	2,863	52

Table 11.7: Top five emojis and their percentages per hate ideology.

emoji appears in 37% of user's names with the 🐸 emoji. Independently, the 🥛 emoji appears in 0.71% of WN user's names. The usage of the 🥛 emoji and the 🐸 emoji in these user's names likely shows their support for white supremacy as discussed in [323] and [322] . Additionally, the 👌 emoji has been noted as a hate symbol and is used to support white supremacy [324]. Used as a hate symbol, this emoji appears to form the letters "WP", standing for "white power" [324]. Within our WN dataset, the 👌 emoji appears in 0.82% of WN user's names.

The ✊ emoji, which appears in 2.8% of KKK tweets, is frequently used in contexts that show support and brotherhood for the KKK. An example tweet is *"@usr @usr @usr I too will kill one of those porch monkeys 🐵 🔫 they are so ungrateful. Please let me join. White power!* ✊*"*. We find that the ✊ emoji is used most commonly in contexts of the words: " 😊", "unite", "white", "kkk", and "brother". As such, this emoji additionally supports our previous findings that the KKK is driven by affiliation and brotherhood.

Fake News and Hate Speech. Media richness in tweets may provide further insights on how hate groups use urls to spread and attract their content to group members or new users. Similarly to our study on emojis, we conduct a small mixed-method investigation on news sites for all hate ideologies. Some of the domains mentioned in hate group tweets include the following news sites: *fairus*, *breitbart*, *lifenews*, and *amren*. We use Media Bias Fact Check [325], the most comprehensive, online news fact checking resource, to

Ideology	VCA Hashtag	Hashtag Type	Persistence (Days)	Prevalence	Non-VCA Avg. Persistence (Days)	Non-VCA Avg. Prevalence
A-IMMGR	#kateslaw	V	4.33	0.09	1.13	0.66
	#boycottstarbucks	CA	1.0	0.5		
A-LGBT	#vvs17	CA	2.68	0.29	1.13	0.64
	#prolifecon	CA	2.78	0.07		
	#marchforlife	CA	2.36	0.07		
	#whywemarch	CA	2.24	0.09		
	#alexandriashooting	V	2.0	0.06		
A-MUS	#parisattacks	V	5.64	0.77	1.07	0.63
	#womensmarch	CA	2.15	0.24		
	#garlandshooting	V	4.40	0.71		
	#orlando	V	2.06	0.12		
	#manchester	V	3.4	0.09		
	#marchagainstsharia	CA	5.11	0.39		
BN	#altonsterling	V	2.57	0.28	1.17	0.57
	#freddiegray	V	2.5	0.03		
	#ferguson	V	1.76	0.07		
	#philandocastile	V	3.17	0.37		
RacSkin	#charlottesville	V	2.0	0.44	1.0	0.80
	#berkeley	V	1.0	0.02		
KKK	#justiceforjessica	V	2.0	0.01	1.01	0.86
WN	#charlottesville	V	2.43	0.63	1.04	0.79
	#katesteinle's	V	2.43	0.63		
	#berkrally	V	2.0	0.44		
	#antishariamarch	CA	2.08	0.33		
	#berkeley	V	2.54	0.07		

Table 11.8: Persistence and prevalence values for hashtags that reference offline violence and collective action events and the average values for hashtags that do not reference such events. Hashtags that refer to violent events are denoted as V and hashtags that refer to offline collective action events are denoted as CA.

identify the reputation and/or partisanship of the sites in our dataset. Among the total returned top news sites for the hate ideologies, we note that *more than half of the sites were questionable sources.* According to Media Bias Fact Check [325], a source is deemed questionable when it "exhibits one or more of the following: extreme bias, consistent promotion of propaganda/conspiracies, poor or no sourcing to credible information, a complete lack of transparency and/or is fake news" [325]. This indicates a potential relationship between online hate and fake news; however this needs further investigation from the research community.

RQ3b: How are hate groups utilizing hashtags, and how do these hashtags persist?

Social media is a common online space where hate groups can congregate. However, their discourse on Twitter is in no way limited to the online sphere. Hate ideologies have leveraged hashtags as a tool to reference offline events that promote their messages.

During our studies, we looked at hashtags that correspond to offline violence and collective action (VCA hashtags), and how their lifespan activity compares to other hashtags (non-VCA hashtags).

We manually identify 25 VCA hate group hashtags across all hate ideologies excluding Neo-Nazi. In order for a hashtag to be classified as VCA, it must reference an offline violent event or offline collective action. For example, all of the VCA-identified BN hashtags (#altonsterling, #freddiegray, #ferguson, and #philandocastile) reference controversial police shootings of black men. Other offline violent acts reference include the Pulse night club mass shooting in Orlando, the Charlottesville attack, and the 2017 bombing of a concert in Manchester. In terms of collective action, several marches are referenced by these hashtags, such as A-MUS's #womensmarch and WN's #antishariamarch. Hashtags that refer to conferences that discuss hate group-specific topics are also qualified as VCA hashtags. For example, A-LGBT's #vvs17 hashtag references the 2017 Values Voter Summit that acts as a platform to discuss the A-LGBT issues. It is hosted by the Family Research Council, which the SPLC has identified as a hate group.

For non-VCA hashtags, 15% of each hate ideology's hashtags were randomly sampled from hashtags that lie in the middle 80% of hashtags in terms of hashtag use count. For example, some of A-Immgr's non-VCA hashtags include #stopsanctarycities, #nopityforillegals, and #cawildfires. The reason that different methodologies were used in choosing VCA and non-VCA hashtags lies in the difference between their counts. Only 25 VCA hashtags were identified across all hate ideologies studied. Therefore to minimize the discrepancy in counts, we chose a random sample for the non-VCA hashtags.

Persistence and prevalence are interesting values to look at when comparing VCA and non-VCA hashtags because they can indicate the longevity and activity-levels of a hashtag. These are metrics defined by [326] in the context of Internet routing that are used to measure churn, or the instability of a network structure. We used persistence to measure how long an average a hashtag is active in continuous time slots (days). Prevalence measures the fraction of days that the hashtag was used during its lifetime. Table 11.8 lists the persistence and prevalence values measured for VCA and non-VCA hashtags across the seven hategroups that employed VCA hashtags. Generally, VCA hashtags persist longer but prevail less than non-VCA hashtags. This is confirmed by the t-test, accounting for differing sample sizes and variances, which produces $p < 10^{-8}$ for persistence and $p < 10^{-7}$ for prevalence.

11.6 Discussion and Critique of Methodology

We presented a novel empirical study that aims at studying the language of the most prevalent hate communities in the U.S. Next, we reflect on the implications of this work to hate speech detection and online communities more broadly. We also discuss the limitations of our work.

11.6.1 Implications for Hate Speech Detection Systems

Our results show a key trend related to the usage of formal analytic language in hate groups as evidenced by the high analytical scores. Current hate speech detection systems rely on datasets and lexicons that are profanity-heavy. In this chapter, we showed that language that dehumanizes communities based on their protected characteristics does not necessarily invoke explicit hate speech keywords but can be persuasive and formal. We have also shown that these hate groups have utilized cryptic emojis that can signal hate without explicit hate keywords. One promising direction is the use of Sequence to Sequence models [327] to infer the meaning of hate symbols [15]. It would be interesting to investigate whether the same concepts could help human beings understand cryptic hate emojis used by hate communities. The aforementioned observations suggest that hate speech does not have to be emotional or explicit. This should reflect in a shift in how we design the next generation hate speech detection systems which are currently trained on content filtered by specific keywords.

11.6.2 Critique of Methodology

Because we use the Twitter Firehose, which guarantees delivery of 100% of Twitter content that matches a selected criterion, we are not exposed to sampling limitations in terms of tweet quality or count when collecting hate group tweets [93, 94]. However, investigating a sample of three hate groups per ideology has the limitation of not capturing the whole realm of online hate communities. Hate groups do not only use social networking websites, they also place their content on video platforms, online funding websites, regular websites, or the dark web. Despite that, our choice to analyze the hate groups with the largest number of followers has enabled us to amass a unique dataset with the highest outreach to online audiences.

11.7 Conclusion

In this chapter, we analyzed the temporal dynamics of 24 hate groups that comprise eight hate ideologies prevalent in the US over the course of 3 years. Our analyses revealed that hate groups show diverse trends over time likely reflective of their modes of operation and their ideology, which is likely influenced by the current socio-economic environment. We showed that some hate groups focus more on appealing via emotions while some tend to appeal more analytically. In terms of emotion, we found that anger was predominately present over other negative emotions in every hate ideology's online presence. We also found that power is a primary driver of most hate ideologies, except for the KKK which is most driven by affiliation due to its unique social structure. Furthermore, we demonstrated that our studied hate ideologies contain semantic overlap with each other, and these shared semantics are indicative of shared topic discourse. Alternate forms of online expression were also investigated, revealing that several benign emojis are used by hate groups to denote hate group affiliation and hateful, cryptic messages and that hashtags that refer to offline events persist longer but prevail less than other hashtags used by hate groups. We hope that our results can be used to improve our sociological understanding of how hate groups operate online, what they are motivated by, and how their philosophies differ.

Chapter 12

Conclusion and Future Directions

12.1 Conclusion

Mitigating biases is critical to both communities' well-being and ethical and fair technology. Online social platforms have provided an outlet to targets of prejudice to voice injustice [7, 8, 274]. Unfortunately, online prejudices have caused a concern between society members that link these online prejudices to offline violence. According to a survey conducted by the Anti-Defamation League (ADL) in 2018, 59% of the respondents believe online hate and harassment make hate crimes more common [328]. In this dissertation, we argue that there is a need to bridge between research communities of hate speech detection and hate speech characterization. We show in this thesis that online hate speech language is much more nuanced and could contain implicit language.

Given that the machines' understanding of the world is built on a statistical foundation, it is natural to see why machines learn through a biased lens mirroring societal constructs. This propagation of bias in downstream applications of NLP models carries real-life consequences. In this thesis, we also show that gender bias does exist in widely used NLP systems such as Neural Relation Extraction systems.

This dissertation has made impact with respect to its intellectual contributions as well as its societal contributions. Work from this dissertation was published in premier computer science venues focusing on web and social media as well as Natural Language Processing. Our study from that pertains to hate target language (Chapter 10) that was published in the AAAI Conference on Web and Social Media in 2018 has been cited 36 times in less than one year. Additionally, the aforementioned work alomg with the work presented in Chapter 9 were featured in a tutorial on Characterization, Detection, and Mitigation of Cyberbullying in the AAAI Conference on Web and Social Media in 2018 [329]. Our work is the first work to challenge the prior literature on nuances of hate language including hate speech actors, target language, and language of communities. We have collected and made publicly available a dataset of 28K hate speech tweets and collected 4.7M tweets for the most prevalent hate ideologies in the U.S. Additionally, we provided a comprehensive literature survey and critique pertaining to gender bias in NLP to the research community in Chapter 6 to encourage more researchers to tackle the problem of gender bias in NLP. This survey was accepted in the top NLP conference of the Association of Computational Linguistics (ACL), in 2019. Additionally, we contributed WikiGender, a distantly supervised dataset of 40K sentences with a human annotated test set that has an even split of male and female sentences specifically curated to analyze gender bias in relation extraction systems.

We conclude with a brief discussion on future research directions made possible by this dissertation. We identify two main research categories: bias detection and repair, and harmful speech and counter speech.

12.2 Future Directions

12.2.1 Bias Detection and Bias Repair

As the influence of AI on daily life increases, lack of awareness of biased predictions made by automatic systems may cause the systems to perpetuate unfairness potentially, and broad adoption of these models can work to magnify stereotypes or implicit biases. We outline three promising research directions.

Other forms of bias. NLP models absorb and learn from human generated corpora that are prone to exhibit social biases such as gender-stereotypes [118, 119]. These social stereotypes propagate through word embeddings, a widely-used vector representation of words in the semantic space and a crucial building block of many NLP systems. One neglected research direction is the exploration of a large variety of other forms of biases such as non-binary gender, disability and cultural biases exhibited by geographic location. Additionally, bias in corpora collected from social media, such as Twitter, Reddit, and Facebook has not been explored.

Implicit Bias Detection. A problem that faces bias detection systems is that data is noisy and bias is sometimes implicit. For example, the tweet "I'm not sexist but I cannot stand women commentators" is an instance of gender bias, even though the first half is misleading [13]. In the case of implicit bias, the discourse is obscured by sarcasm and the lack of hateful terms. One tweet example that was observed in hate group language is "He Told The World: The Immortal Words of Adolf Hitler (VIDEO)". This post constitutes implicit bias but does not use any hate terms in current hate lexicons. Insights from this dissertation point to a promising research direction which is the detection of implicit bias by developing NLP models that learn the author's generic stereotypic associations and personality traits, such as lower empathy, which has been shown to be associated with implicit bias.

Bias Repair. Feldman et al. distinguish classification algorithms that achieve fairness by debiasing data versus adjusting the classification method through a process called "repair" [154]. There is a need to study whether these methods propagate or amplify the bias in the data. While some prior work has attempted to study classic Machine Learning Models such as Naive Bayes, Logistic Regression and Support Vector Machines (SVM) [154], there is still a need to analyze the relationship between dataset features, algorithms, and repair performance. There is an imminent need to understand the geometry of societal bias in machine-induced representations, and shed a light on how black-box deep learning models capture demographic information in data (e.g., the gender of the person in an image). Additionally, there is a need to develop computational methods that correct these biases.

12.2.2 Harmful Speech and Counterspeech

Determination of the optimal response to online hate speech is complex. While platforms could enforce sanctions such as content removal or account suspension of users who do not follow hateful conduct policies, these users could persist in posting hate speech through their own account, or through the creation of new accounts. Counterspeech is a direct response to hate or harmful speech that can be practiced by almost anyone, requiring neither law nor institutional support. Further, it has been shown to have a favorable effect on individuals targeted by hate speech [330]. For deeper understanding, computational approaches are needed to study counterspeech at scale. There is a need to conduct largescale studies that investigate the efficacy of different counterspeech techniques and their scenario-specific application for maximum impact. The next step would be to design automated dialogue systems that are grounded by the lessons learned from the previous study. The objective of the dialogue system would be to either form an appropriate response to the harmful text, or to recommend a human intervention, for example through notification of someone in the instigator's or target's network.

Another line of research, that is inspired by the results of this dissertation, is the investigation of the correlation between online misinformation, such as fake news, and online misuse, as measured by hate speech engagement. Our prior work has shown that one of the reasons online users participate in hate speech is that they believe that hate speech is truth telling [209]. There is a need to investigate online news credibility in harmful speech conversations to discern the impact of online misinformation on online misuse.

Motivated by the increase in U.S. mass shootings witnessed, a novel research direction is early instigator identification via mining social media data to use this information to predict who or what entities are likely to be involved in offline violence, such as shootings or even violent protests, and where these events are likely to occur. For instance, the New York Times revealed that a social media account associated with the Pittsburgh shooter in 2018 was filled with anti-Jewish slurs and conspiracy theories [331]. While challenging, the prediction of violence based on instigator social media data is an important problem that has the potential to be life-saving.

Bibliography

[1] Reuters, "Amazon scraps secret AI recruiting tool that showed bias against women." `https://reut.rs/2Po4ZJi`, 2019. [Online; accessed September-2019].

[2] H. W. Rittel and M. M. Webber, *Wicked problems*, *Man-made Futures* **26** (1974), no. 1 272–280.

[3] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Debiaswe." `https://bit.ly/2ruopBZ`, 2018. Accessed on 12.10.2018.

[4] K. Crawford, "The Trouble With Bias." `https://bit.ly/2ruopBZ`, 2017. Keynote at Neural Information Processing Systems (NIPS'17).

[5] Lexico powered by Oxford dictionary, "Definition of bias in English." `https://www.lexico.com/en/definition/bias`, September, 2019.

[6] T. Gale, "Consequences of Prejudice." `https://www.encyclopedia.com/social-sciences/news-wires-white-papers-and-books/consequences-prejudice`, September, 2019.

[7] A. Olteanu, I. Weber, and D. Gatica-Perez, *Characterizing the Demographics Behind the #BlackLivesMatter Movement*, in *AAAI 2016 Spring Symposium Series*, 2016.

[8] M. De Choudhury, S. Jhaver, B. Sugar, and I. Weber, *Social Media Participation in an Activist Movement for Racial Equality*, in *ICWSM'16: Proceedings of the 10th International AAAI Conference on Web and Social Media*, 2016.

[9] UNICEF, "Gender equality – Accelerating progress and opportunities for everyone." `https://www.unicef.org/gender-equality`, 2019.

[10] UNFPA, "Gender equality." `https://www.unfpa.org/gender-equality`, 2019.

[11] Organization of American States, Council of Europe, Permanent Mission of France to the United Nations and Permanent Mission of Argentina to the United Nations, "The Convention of Belem do Para and the Istanbul Convention: a

response to violence against women worldwide." https://www.oas.org/es/mesecvi/docs/CSW-SideEvent2014-Flyer-EN.pdf, March, 2014.

[12] Council of Europe, Committee of Ministers, CM document, "Gender Equality Commission (GEC) - Gender Equality Strategy 2014-2017." https://search.coe.int/cm/Pages/result_details.aspx?ObjectID=09000016805c7246, June, 2015.

[13] J. Qian, M. ElSherief, E. Belding, and W. Y. Wang, *Leveraging intra-user and inter-user representation learning for automated hate speech detection*, in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, vol. 2, pp. 118–123, 2018.

[14] J. Qian, M. ElSherief, E. Belding, and W. Y. Wang, *Hierarchical cvae for fine-grained hate speech classification*, in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3550–3559, 2018.

[15] J. Qian, M. ElSherief, E. Belding, and W. Yang Wang, *Learning to decipher hate symbols*, in *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL'19)*, 2019.

[16] Hollaback, "Read and Share Stories. When it comes to street harassment, you are not alone." http://www.ihollaback.org/share/, 2015. [Online; accessed July-2015].

[17] S. Crabtree and F. Nsubuga, "Women feel less safe than men in many developed countries." http://www.gallup.com/poll/155402/women-feel-less-safe-men-developed-countries.aspx, 2012. [Online; accessed July-2015].

[18] Stop Street Harassment, "Unsafe and Harassed in Public Spaces: A National Street Harassment Report." http://www.stopstreetharassment.org/wp-content/uploads/2012/08/2014-National-SSH-Street-Harassment-Report.pdf, 2014. [Online; accessed July-2015].

[19] Stop Street Harassment, "Definitions." http://www.stopstreetharassment.org/resources/definitions/, 2015. [Online; accessed July-2015].

[20] Walk Score, "Walkable Neighborhoods." https://www.walkscore.com/walkable-neighborhoods.shtml, 2015. [Online; accessed July-2015].

[21] A. Williams, E. Robles, and P. Dourish, *Urbane-ing the city: Examining and refining the assumptions behind urban informatics*, *Handbook of research on urban informatics: The practice and promise of the real-time city* (2009) 1–20.

[22] C. B. Leinberger, "Now coveted: A walkable, convenient place." `http://ctmainstreet.org/wp-content/uploads/2012/07/1a-Opinion-Walkable-Place.pdf/`, 2012. [Online; posted May-2012]".

[23] R. E. Lee, K. McAlexander, and J. Banda, *Reversing the obesogenic environment.* Champaign, IL: Human Kinetics, 2011.

[24] D. Quercia, R. Schifanella, and L. M. Aiello, *The shortest path to happiness: Recommending beautiful, quiet, and happy routes in the city*, in *Proceedings of the 25th ACM Conference on Hypertext and Social Media*, pp. 116–125, September-2014.

[25] D. Quercia, L. M. Aiello, R. Schifanella, and A. Davies, *The digital life of walkable streets*, in *Proceedings of the 24th International Conference on World Wide Web*, pp. 875–884, May-2015.

[26] D. Van Dyck, G. Cardon, B. Deforche, and I. De Bourdeaudhuij, *Do adults like living in high-walkable neighborhoods? Associations of walkability parameters with neighborhood satisfaction and possible mediators*, *Health & Place* **17** (2011), no. 4 971–977.

[27] J. Speck, *Walkable City: How Downtown Can Save America, One Step at a Time.* Macmillan, 2013.

[28] L. J. Carr, S. I. Dunsiger, and B. H. Marcus, *Validation of walk score for estimating access to walkable amenities*, *British Journal of Sports Medicine* **45** (2010), no. 14 1144–1148.

[29] D. T. Duncan, J. Aldstadt, J. Whalen, S. J. Melly, and S. L. Gortmaker, *Validation of walk score for estimating neighborhood walkability: an analysis of four us metropolitan areas*, *International journal of environmental research and public health* **8** (2011), no. 11 4160–4179.

[30] D. Estrin, K. M. Chandy, R. M. Young, L. Smarr, A. Odlyzko, D. Clark, V. Reding, T. Ishida, *et. al.*, *Participatory Sensing: Applications and Architecture [Internet Predictions]*, *IEEE Internet Computing* **14** (2010), no. 1 12–42.

[31] U. Demiryurek, F. Banaei-Kashani, and C. Shahabi, *Transdec: A spatiotemporal query processing framework for transportation systems*, in *IEEE 26th International Conference on Data Engineering (ICDE)*, pp. 1197–1200, 2010.

[32] Scientific American, "8 Apps That Turn Citizens into Scientists." `https://www.scientificamerican.com/article/8-apps-that-turn-citizens-into-scientists/`, 2013.

[33] Safetipin, "Safetipin — Supporting Safer Cities." `http://www.safetipin.com/`, 2016.

[34] L. Kazemi and C. Shahabi, *GeoCrowd: Enabling Query Answering with Spatial Crowdsourcing*, in *Proceedings of the 20th International Conference on Advances in Geographic Information Systems*, (Redondo Beach, CA), pp. 189–198, November, 2012.

[35] H. To, C. Shahabi, and L. Kazemi, *A Server-assigned Spatial Crowdsourcing Framework*, *ACM Transactions on Spatial Algorithms and Systems* **1** (2015), no. 2 29–56.

[36] P. Cheng, X. Lian, Z. Chen, R. Fu, L. Chen, J. Han, and J. Zhao, *Reliable Diversity-based Spatial Crowdsourcing by Moving Workers*, *Proceedings of the VLDB Endowment* **8** (2015), no. 10 1022–1033.

[37] H. To, S. H. Kim, and C. Shahabi, *Effectively Crowdsourcing the Acquisition and Analysis of Visual Data for Disaster Response*, in *IEEE International Conference on Big Data*, pp. 697–706, 2015.

[38] B. S. Manoj and A. H. Baker, *Communication Challenges in Emergency Response*, *Communications of the ACM* **50** (2007), no. 3 51–53.

[39] D. S. Hochbaum, *Approximating Covering and Packing Problems: Set Cover, Vertex Cover, Independent Set, and Related Problems*, in *Approximation Algorithms for NP-hard Problems*, pp. 94–143, PWS Publishing Co., 1996.

[40] B. Hull, V. Bychkovsky, Y. Zhang, K. Chen, M. Goraczko, A. Miu, E. Shih, H. Balakrishnan, and S. Madden, *CarTel: A Distributed Mobile Sensor Computing System*, in *Proceedings of the 4th International Conference on Embedded Networked Sensor Systems*, pp. 125–138, 2006.

[41] D. Deng, C. Shahabi, and U. Demiryurek, *Maximizing the Number of Worker's Self-selected Tasks in Spatial Crowdsourcing*, in *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, (Orlando, FL, USA), pp. 324–333, November, 2013.

[42] H. Yu, C. Miao, Z. Shen, and C. Leung, *Quality and Budget Aware Task Allocation for Spatial Crowdsourcing*, in *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, May, 2015.

[43] J. Wang, Y. Wang, D. Zhang, F. Wang, Y. He, and L. Ma, *Psallocator: multi-task allocation for participatory sensing with sensing capability constraints*, in *Proceedings of the ACM CSCW*, pp. 1139–1151, 2017.

[44] J. Wang, Y. Wang, D. Zhang, L. Wang, H. Xiong, A. Helal, Y. He, and F. Wang, *Fine-grained multitask allocation for participatory sensing with a shared budget*, *IEEE Internet of Things Journal* **3** (2016), no. 6 1395–1405.

[45] H. To, L. Fan, L. Tran, and C. Shahabi, *Real-time Task Assignment in Hyperlocal Spatial Crowdsourcing Under Budget Constraints*, in *IEEE International Conference on Pervasive Computing and Communications (PerCom)*, (Kona, Big Island, HI, USA), March, 2016.

[46] L. Anselin and A. Getis, *Spatial Atatistical Analysis and Geographic Information Systems*, *The Annals of Regional Science* **26** (1992), no. 1 19–33.

[47] S. Lloyd, *Least Squares Quantization in PCM*, *IEEE Transactions on Information Theory* **28** (1982), no. 2 129–137.

[48] H. Robbins, *Some Aspects of the Sequential Design of Experiments*, in *Herbert Robbins Selected Papers*, pp. 169–177. 1985.

[49] M. Sherman, *Spatial Statistics and Spatio-temporal Data: Covariance Functions and Directional Properties*. John Wiley & Sons, 2011.

[50] W. R. Tobler, *A Computer Movie Simulating Urban Growth in the Detroit Region*, *Economic Geography* **46** (1970), no. 1 234–240.

[51] United Nations, "Declaration on the Elimination of Violence against Women." `http://www.un.org/documents/ga/res/48/a48r104.htm`, 1993.

[52] UNFPA, "Gbv topic overview." `http://www.unfpa.org/gender-based-violence`, 2016.

[53] United Nations, "UN Women Violence against Women (VAW) stats." `http://www.unwomen.org/en/what-we-do/ending-violence-against-women/facts-and-figures`, 2016.

[54] World Health Organization, "World Health Organization (WHO) violence stats." `http://www.who.int/gho/women_and_health/violence/en/`, 2010.

[55] D. Esty and R. Rushing, *The promise of data-driven policymaking*, *Issues in Science and Technology* **23** (2007), no. 4 67–72.

[56] GBVIMS, "Gender-based Violence Information Management System (GBVIMS) tools." `http://www.gbvims.com/gbvims-tools/`, 2016.

[57] E. Buchbinder and Z. Eisikovits, *Battered Women's Entrapment in Shame: A Phenomenological Study.*, *American Journal of Orthopsychiatry* **73** (2003), no. 4 355.

[58] CNN, "Sexual assault and the Trump tape: 1 million women say it's #notokay." `http://money.cnn.com/2016/10/08/technology/notokay-twitter-donald-trump/`, 2016.

[59] J. Bartlett, R. Norrie, S. Patel, R. Rumpel, and S. Wibberley, "Misogyny on Twitter." `https://www.demos.co.uk/files/MISOGYNY_ON_TWITTER.pdf`, 2016.

[60] R. Fulper, G. L. Ciampaglia, E. Ferrara, Y. Ahn, A. Flammini, F. Menczer, B. Lewis, and K. Rowe, *Misogynistic language on Twitter and sexual violence*, in *ChASM'14: Proceedings of the ACM Web Science Workshop on Computational Approaches to Social Modeling*, 2014.

[61] H. Purohit, T. Banerjee, A. Hampton, V. L. Shalin, N. Bhandutia, and A. P. Sheth, *Gender-based violence in 140 characters or fewer: A #BigData case study of Twitter*, *arXiv preprint arXiv:1503.02086* (2015).

[62] P. Karuna, H. Purohit, B. Stabile, and A. Hattery, *On the Dynamics of Local to Global Campaigns for Curbing Gender-based Violence*, *arXiv preprint arXiv:1608.01648* (2016).

[63] G. Lotan, E. Graeff, M. Ananny, D. Gaffney, I. Pearce, *et. al.*, *The Arab Spring | The Revolutions were tweeted: Information flows during the 2011 Tunisian and Egyptian revolutions*, *International Journal of Communication* **5** (2011) 31.

[64] S. González-Bailón, J. Borge-Holthoefer, A. Rivero, and Y. Moreno, *The dynamics of protest recruitment through an online network*, *Scientific reports* **1** (2011) 197.

[65] J. W. Creswell, *Research design: Qualitative, quantitative, and mixed methods approaches.* Sage publications, 2013.

[66] I. de Sola Pool and M. Kochen, *Contacts and influence*, *Social networks* **1** (1978), no. 1 5–51.

[67] P. Domingos and M. Richardson, *Mining the network value of customers*, in *SIGKDD'01: Proceedings of the 7th ACM International Conference on Knowledge Discovery and Data mining*, 2001.

[68] M. Cha, A. Mislove, and K. P. Gummadi, *A measurement-driven analysis of information propagation in the Flickr social network*, in *WWW'09: Proceedings of the 18th International ACM conference on World Wide Web*, 2009.

[69] S. Hill, F. Provost, and C. Volinsky, *Network-based marketing: Identifying likely adopters via consumer networks*, *Statistical Science* **21** (2006), no. 2 256–276.

[70] J. Bollen, B. Gonçalves, G. Ruan, and H. Mao, *Happiness is assortative in online social networks*, *Artificial life* **17** (2011), no. 3 237–251.

[71] P. S. Dodds, K. D. Harris, I. M. Kloumann, C. A. Bliss, and C. M. Danforth, *Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter*, *Public Library of Science (PloS) | one* **6** (2011), no. 12 e26752.

[72] L. Coviello, Y. Sohn, A. D. Kramer, C. Marlow, M. Franceschetti, N. A. Christakis, and J. H. Fowler, *Detecting emotional contagion in massive social networks*, *Public Library of Science (PloS) | one* **9** (2014), no. 3 e90315.

[73] R. Fan, J. Zhao, Y. Chen, and K. Xu, *Anger is more influential than joy: Sentiment correlation in Weibo*, *Public Library of Science (PloS) | one* **9** (2014), no. 10 e110184.

[74] A. Leavitt, E. Burchard, D. Fisher, and S. Gilbert, *The influentials: New approaches for analyzing influence on twitter*, *Web Ecology Project* **4** (2009), no. 2 1–18.

[75] D. Kempe, J. Kleinberg, and É. Tardos, *Maximizing the spread of influence through a social network*, in *SIGKDD'03: Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data mining*, 2003.

[76] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance, *Cost-effective outbreak detection in networks*, in *SIGKDD'07: Proceedings of the 13th ACM International Conference on Knowledge Discovery and Data mining*, 2007.

[77] W. Chen, Y. Wang, and S. Yang, *Efficient influence maximization in social networks*, in *SIGKDD'09: Proceedings of the 15th ACM international conference on Knowledge Discovery and Data mining*, 2009.

[78] M. Granovetter, *Threshold models of collective behavior*, *American Journal of Sociology* **83** (1978), no. 6 1420–1443.

[79] D. M. Romero, B. Meeder, and J. Kleinberg, *Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on Twitter*, in *WWW'11: Proceedings of the 20th International Conference on World Wide Web*, 2011.

[80] S. Goel, A. Anderson, J. Hofman, and D. J. Watts, *The structural virality of online diffusion*, *Management Science* **62** (2015), no. 1 180–196.

[81] B. State and L. Adamic, *The Diffusion of Support in an Online Social Movement: Evidence from the Adoption of Equal-Sign Profile Pictures*, in *CSCW'15: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 2015.

[82] C. Chung and J. W. Pennebaker, *The psychological functions of function words*, *Social communication* (2007) 343–359.

[83] S. Nilizadeh, A. Groggel, P. Lista, S. Das, Y.-Y. Ahn, A. Kapadia, and F. Rojas, *Twitter's glass ceiling: The effect of perceived gender on online visibility*, in *ICWSM'16*, pp. 289–298, 2016.

[84] J. Zhang, X. Hu, Y. Zhang, and H. Liu, *Your age is no secret: Inferring microbloggers' ages via content and interaction analysis*, in *ICWSM'16: Proceedings of the 10th International AAAI Conference on Web and Social Media*, 2016.

[85] T. Hu, H. Xiao, J. Luo, and T. T. Nguyen, *What the Language You Tweet Says About Your Occupation*, in *ICWSM'16: Proceedings of the 10th International AAAI Conference on Web and Social Media*, 2016.

[86] B. Perozzi and S. Skiena, *Exact Age Prediction in Social Networks*, in *WWW'15: Proceedings of the 24th International Conference on World Wide Web*, 2015.

[87] J. Marquardt, G. Farnadi, G. Vasudevan, M.-F. Moens, S. Davalos, A. Teredesai, and M. De Cock, *Age and Gender Identification in Social Media*, in *CLEF'14: the 5th Conference and Labs of Evaluation Forum*, 2014.

[88] H. Fan, Z. Cao, Y. Jiang, Q. Yin, and C. Doudou, *Learning Deep Face Representation*, *arXiv preprint arXiv:1403.2802* (2014).

[89] Twitter, "FAQs about Retweets." `https://support.twitter.com/articles/77606`, 2016.

[90] L. A. Bell, *Storytelling for social justice: Connecting narrative and the arts in antiracist teaching.* 2010.

[91] J. E. Davis, *Narrative and social movements*, *Stories of change: Narrative and social movements* (2002) 3–29.

[92] J. P. Dimond, M. Dye, D. LaRose, and A. S. Bruckman, *Hollaback!: the role of storytelling online in a social movement organization*, in *CSCW'13: Proceedings of the 16th ACM conference on Computer Supported Cooperative Work and Social Computing*, 2013.

[93] Z. Tufekci, *Big Questions for Social Media Big Data: Representativeness, Validity and other Methodological Pitfalls*, in *ICWSM'14*, 2014.

[94] F. Morstatter, J. Pfeffer, H. Liu, and K. M. Carley, *Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose*, in *ICWSM '13*, 2013.

[95] S. I. Ahmed, S. J. Jackson, N. Ahmed, H. S. Ferdous, M. R. Rifat, A. Rizvi, S. Ahmed, and R. S. Mansur, *Protibadi: A Platform for Fighting Sexual Harassment in Urban Bangladesh*, in *Proceedings of the 32nd annual ACM Conference on Human factors in Computing Systems*, (Toronto, ON, Canada), pp. 2695–2704, May, 2014.

[96] M. ElSherief and E. Belding, *The Urban Characteristics of Street Harassment: A First Look*, in *UrbanGIS'15: First International Workshop on Smart Cities and Urban Analytics associated with the 23rd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, (Seattle, WA, USA), 2015.

[97] M. ElSherief, E. Belding, and D. Nguyen, *#NotOkay: Understanding Gender-based Violence in Social Media*, in *Proceedings of the the 11th International AAAI Conference on Web and Social Media (ICWSM'17)*, (Montreal, Canada), May, 2017.

[98] N. Karusala and N. Kumar, *Women's Safety in Public Spaces: Examining the Efficacy of Panic Buttons in New Delhi*, in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, (New York, NY, USA), pp. 3340–3351, ACM, 2017.

[99] S. Krug, *Don't Make Me Think: A Common Sense Approach to Web Usability.* New Riders Publishing, 2009.

[100] A. Lee, *Gender, Everyday Mobility, and Mass Transit in Urban Asia*, *Mobility in History* **8** (2017), no. 1 85–94.

[101] D. Massey, *Space, Place and Gender.* John Wiley & Sons, 2013.

[102] P. A. Campos, K. L. Falb, S. Hernández, C. Díaz-Olavarrieta, and J. Gupta, *Experiences of Street Harassment and Associations with Perceptions of Social Cohesion Among Women in Mexico City*, *Salud Pública de México* **59** (2017), no. 1 102–105.

[103] J. Clark Blickenstaff, *Women and Science Careers: Leaky Pipeline or Gender Filter?*, *Gender and Education* **17** (2005), no. 4 369–386.

[104] E. Holbrook, *Lean In: Women, Work, and the Will to Lead*, *Risk Management* **60** (2013), no. 4 44–45.

[105] M. Hilbert, *Digital Gender Divide or Technologically Empowered Women in Developing Countries? A Typical Case of Lies, Damned lies, and Statistics*, in *Women's Studies International Forum*, vol. 34, pp. 479–489, 2011.

[106] B. Nardi, *Inequality and Limits*, in *Proceedings of the First Workshop on Computing within Limits*, (Irvine, CA, USA), 2015.

[107] S. Jayachandran, *The Roots of Gender Inequality in Developing Countries*, *Annual Review of Economics* **7** (2015), no. 1 63–88.

[108] S. I. Ahmed, N. Ahmed, F. Hussain, and N. Kumar, *Computing beyond gender-imposed limits*, in *LIMITS'16: Proceedings of the Second Workshop on Computing within Limits*, (Irvine, CA, USA), June, 2016.

[109] P. Sengers, K. Boehner, S. David, and J. Kaye, *Reflective design*, in *Proceedings of the 4th ACM Decennial Conference on Critical Computing: Between Sense and Sensibility*, pp. 49–58, 2005.

[110] P. Schmitt and E. Belding, *Navigating Connectivity in Reduced Infrastructure Environments*, in *Proceedings of the Second Workshop on Computing within Limits*, (Irvine, CA, USA), June, 2016.

[111] K. Toyama, *Preliminary Thoughts on a Taxonomy of Value for Sustainable Computing*, in *Proceedings of the Second Workshop on Computing within Limits*, (Irvine, CA, USA), June, 2015.

[112] D. J. Patterson, *Haitian Resiliency: A Case Study in Intermittent Infrastructure*, in *Proceedings of the Second Workshop on Computing within Limits*, (Irvine, CA, USA), June, 2015.

[113] J. Shi, J. Wan, H. Yan, and H. Suo, *A Survey of Cyber-Physical Systems*, in *2011 International Conference on Wireless Communications and Signal Processing (WCSP)*, (Nanjing, China), pp. 1–6, November, 2011.

[114] E. A. Lee, *Cyber Physical Systems: Design Challenges*, in *2008 11th IEEE International Symposium on Object and Component-Oriented Real-Time Distributed Computing (ISORC)*, (Orlando, FL, USA), pp. 363–369, May, 2008.

[115] K. Toyama, *Geek Heresy: Rescuing Social Change from the Cult of Technology.* PublicAffairs, 2015.

[116] C. A. Moss-Racusin, J. F. Dovidio, V. L. Brescoll, M. J. Graham, and J. Handelsman, *Science Faculty's Subtle Gender Biases Favor Male Students*, *Proceedings of the National Academy of Sciences* **109** (2012), no. 41 16474–16479.

[117] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang, *Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods*, in *North American Chapter of the Association for Computational Linguistics (NAACL'18)*, 2018.

[118] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, *Man Is to Computer Programmer As Woman Is to Homemaker? Debiasing Word Embeddings*, in *Neural Information Processing Systems (NIPS'16)*, 2016.

[119] A. Caliskan, J. J. Bryson, and A. Narayanan, *Semantics Derived Automatically from Language Corpora Contain Human-Like Biases*, *Science* **356** (2017), no. 6334 183–186.

[120] N. Garg, L. Schiebinger, D. Jurafsky, and J. Zou, *Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes*, *Proceedings of the National Academy of Sciences* **115** (2018), no. 16 E3635–E3644.

[121] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang, *Men Also Like Shopping: Reducing Gender Bias Amplification Using Corpus-Level Constraints*, in *Empirical Methods of Natural Language Processing (EMNLP'17)*, 2017.

[122] L. Douglas, "AI is not Just Learning our Biases; It Is Amplifying Them." `https://bit.ly/2zRvGhH`, 2017. Accessed on 11.15.2018.

[123] K. Burns, L. A. Hendricks, T. Darrell, A. Rohrbach, and K. Saenko, *Women Also Snowboard: Overcoming Bias in Captioning Models*, *European Conference on Computer Vision (EECV'18)* (2018).

[124] R. Tatman, *Gender and Dialect Bias in YouTube's Automatic Captions*, in *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing (ACL'17)*, pp. 53–59, 2017.

[125] J. H. Park, J. Shin, and P. Fung, *Reducing Gender Bias in Abusive Language Detection*, in *Empirical Methods of Natural Language Processing (EMNLP'18)*, 2018.

[126] K. Lu, P. Mardziel, F. Wu, P. Amancharla, and A. Datta, *Gender Bias in Neural Natural Language Processing*, 2018.

[127] A. G. Greenwald, D. E. McGhee, and J. L. Schwartz, *Measuring Individual Differences in Implicit Cognition: The Implicit Association Test*, *Journal of Personality and Social Psychology* **74** (1998), no. 6 1464.

[128] B. A. Nosek, F. L. Smyth, N. Sriram, N. M. Lindner, T. Devos, A. Ayala, Y. Bar-Anan, R. Bergh, H. Cai, K. Gonsalkorale, S. Kesebir, N. Maliszewski, F. Neto, E. Olli, J. Park, K. Schnabel, K. Shiomura, B. Tulbure, R. Wiers, M. Somogyi, N. Akrami, B. Ekehammar, M. Vianello, M. Banaji, and

A. Greenwald, *National Differences in Gender-Science Stereotypes Predict National Sex Differences in Science and Math Achievement*, *Proceedings of the National Academy of Sciences* **106** (2009), no. 26 10593–10597.

[129] C. May, A. Wang, S. Bordia, S. R. Bowman, and R. Rudinger, *On Measuring Social Biases in Sentence Encoders*, in *North American Chapter of the Association for Computational Linguistics (NAACL'19)*, 2019.

[130] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, *Deep Contextualized Word Representations*, in *North American Chapter of the Association for Computational Linguistics (NAACL'18)*, 2018.

[131] T. Manzini, Y. C. Lim, Y. Tsvetkov, and A. W. Black, *Black is to Criminal as Caucasian is to Police: Detecting and Removing Multiclass Bias in Word Embeddings*, in *North American Chapter of the Association for Computational Linguistics (NAACL'19)*, 2019.

[132] H. Gonen and Y. Goldberg, *Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But Do Not Remove Them*, in *North American Chapter of the Association for Computational Linguistics (NAACL'19)*, 2019.

[133] S. Kiritchenko and S. M. Mohammad, *Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems*, in *7th Joint Conference on Lexical and Computational Semantics (SEM'18)*, 2018.

[134] L. Dixon, J. Li, J. Sorensen, N. Thain, and L. Vasserman, *Measuring and Mitigating Unintended Bias in Text Classification*, in *Association for the Advancement of Artificial Intelligence (AAAI'17)*, 2017.

[135] R. Rudinger, J. Naradowsky, B. Leonard, and B. Van Durme, *Gender Bias in Coreference Resolution*, in *North American Chapter of the Association for Computational Linguistics (NAACL'18)*, 2018.

[136] K. Webster, M. Recasens, V. Axelrod, and J. Baldridge, *Mind the GAP: A Balanced Corpus of Gendered Ambiguous Pronouns*, in *Transactions of the ACL (TACL'18)*, 2018.

[137] T. B. Hashimoto, M. Srivastava, H. Namkoong, and P. Liang, *Fairness Without Demographics in Repeated Loss Minimization*, .

[138] J. Zhao, Y. Zhou, Z. Li, W. Wang, and K.-W. Chang, *Learning Gender-Neutral Word Embeddings*, in *Empirical Methods of Natural Language Processing (EMNLP'18)*, 2018.

[139] K. Lee, L. He, M. Lewis, and L. Zettlemoyer, *End-to-End Neural Coreference Resolution*, in *Empirical Methods of Natural Language Processing (EMNLP'17)*, 2017.

[140] K. Lee, L. He, and L. Zettlemoyer, *Higher-Order Coreference Resolution with Coarse-to-Fine Inference*, in *Empirical Methods of Natural Language Processing (EMNLP'18)*, 2018.

[141] N. Madaan, S. Mehta, T. Agrawaal, V. Malhotra, A. Aggarwal, Y. Gupta, and M. Saxena, *Analyze, Detect and Remove Gender Stereotyping from Bollywood Movies*, in *Conference on Fairness, Accountability and Transparency (FAT'18)*, pp. 92–105, 2018.

[142] M. O. R. Prates, P. H. Avelar, and L. C. Lamb, *Assessing gender bias in machine translation: a case study with Google Translate*, *Neural Computing and Applications* (Mar, 2018).

[143] E. Vanmassenhove, C. Hardmeier, and A. Way, *Getting Gender Right in Neural Machine Translation*, in *Empirical Methods of Natural Language Processing (EMNLP'18)*, 2018.

[144] A.-M. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, and N. Kourtellis, *Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior*, in *Association for the Advancement of Artifical Intelligence (AAAI'18)*, 2018.

[145] Z. Waseem and D. Hovy, *Hateful symbols or hateful people? predictive features for hate speech detection on twitter*, in *Proceedings of the NAACL student research workshop*, pp. 88–93, 2016.

[146] B. Schmidt, "Rejecting the Gender Binary: A Vector-Space Operation." `https://bit.ly/1OhXJM0`, 2015. Accessed on 11.15.2018.

[147] M. Nickel and D. Kiela, *Poincarè Embeddings for Learning Hierarchical Representations*, in *Advances in Neural Information Processing Systems (NIPS'17)*, pp. 6338–6347, 2017.

[148] D. Roth and W.-t. Yih, *A linear programming formulation for global inference in natural language tasks*, in *Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004*, 2004.

[149] M. Yatskar, L. Zettlemoyer, and A. Farhadi, *Situation Recognition: Visual Semantic Role Labeling for Image Understanding*, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (ICVPR'16)*, pp. 5534–5542, 2016.

[150] A. M. Rush and M. Collins, *A Tutorial on Dual Decomposition and Lagrangian Relaxation for Inference in Natural Language Processing*, *Journal of Artificial Intelligence Research* **45** (2012) 305–362.

[151] B. H. Zhang, B. Lemoine, and M. Mitchell, *Mitigating Unwanted Biases with Adversarial Learning*, in *AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES'18)*, 2018.

[152] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, *Generative Adversarial Nets*, in *Advances in Neural Information Processing Systems (NIPS'14)*, 2014.

[153] T. Calders and S. Verwer, *Three Naive Bayes Approaches for Discrimination-Free Classification*, *Data Mining and Knowledge Discovery* **21** (2010), no. 2 277–292.

[154] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, *Certifying and Removing Disparate Impact*, in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'15)*, 2015.

[155] M. Hardt, E. Price, and N. Srebro, *Equality of Opportunity in Supervised Learning*, in *Advances in Neural Information Processing Systems (NIPS'16)*, 2016.

[156] I. Misra, C. Lawrence Zitnick, M. Mitchell, and R. Girshick, *Seeing Through the Human Reporting Bias: Visual Classifiers from Noisy Human-Centric Labels*, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (ICVPR'16)*, pp. 2930–2939, 2016.

[157] J. Kleinberg, S. Mullainathan, and M. Raghavan, *Inherent Trade-offs in the Fair Determination of Risk Scores*, in *Computing Research Repository (CoRR'16)*, 2016.

[158] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger, *On Fairness and Calibration*, in *Advances in Neural Information Processing Systems (NIPS'17)*, pp. 5680–5689, 2017.

[159] A. Beutel, J. Chen, Z. Zhao, and E. H. hsin Chi, *Data Decisions and Theoretical Implications when Adversarially Learning Fair Representations*, in *2017 Workshop on Fairness, Accountability, and Transparency in Machine Learning*, 2017.

[160] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, *Model cards for model reporting*, in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 220–229, ACM, 2019.

[161] E. M. Bender and B. Friedman, *Data statements for natural language processing: Toward mitigating system bias and enabling better science*, *Transactions of the Association for Computational Linguistics* **6** (2018) 587–604.

[162] S. Reddy and K. Knight, *Obfuscating Gender in Social Media Writing*, in *Proceedings of the First Workshop on NLP and Computational Social Science*, pp. 17–26, 2016.

[163] Y. Elazar and Y. Goldberg, *Adversarial Removal of Demographic Attributes from Text Data*, in *Empirical Methods of Natural Language Processing (EMNLP'18)*, 2018.

[164] Y. Li, T. Baldwin, and T. Cohn, *Towards Robust and Privacy-Preserving Text Representations*, in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL'18)*, pp. 1650–1659, 2018.

[165] D. Hovy and S. L. Spruit, *The Social Impact of Natural Language Processing*, in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL'16)*, vol. 2, pp. 591–598, 2016.

[166] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning.* fairmlbook.org, 2018. `http://www.fairmlbook.org`.

[167] A. Chouldechova and A. Roth, *The Frontiers of Fairness in Machine Learning*, *arXiv preprint arXiv:1810.08810* (2018).

[168] C. Richards, W. P. Bouman, L. Seal, M. J. Barker, T. O. Nieder, and G. T'Sjoen, *Non-Binary or Genderqueer Genders*, *International Review of Psychiatry (IRP'16)* (2016).

[169] N. Kilbertus, M. Rojas-Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf, *Avoiding Discrimination Through Causal Reasoning*, in *Neural Information Processing Systems (NIPS'17)*, 2017.

[170] A. Herbelot, E. Von Redecker, and J. Müller, *Distributional Techniques for Philosophical Enquiry*, in *Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pp. 45–54, Association for Computational Linguistics, 2012.

[171] C. Avin, B. Keller, Z. Lotker, C. Mathieu, D. Peleg, and Y.-A. Pignolet, *Homophily and the Glass Ceiling Effect in Social Networks*, in *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science (ITCS'15)*, pp. 41–50, ACM, 2015.

[172] L. Fu, C. Danescu-Niculescu-Mizil, and L. Lee, *Tie-breaker: Using Language Models to Quantify Gender Bias in Sports Journalism*, in *Proceedings of the IJCAI workshop on NLP meets Journalism*, 2016.

[173] N. Schluter, *The Glass Ceiling in NLP*, in *Empirical Methods of Natural Language Processing (EMNLP'18)*, 2018.

[174] A. Olteanu, S. Vieweg, and C. Castillo, *What to Expect when the Unexpected Happens: Social Media Communications Across Crises*, in *CSCW'15*, 2015.

[175] J. Crump, *What are the Police doing on Twitter? Social Media, the Police and the Public*, *Policy & Internet* **3** (2011), no. 4 1–27.

[176] L. A. Silva, M. Mondal, D. Correa, F. Benevenuto, and I. Weber, *Analyzing the Targets of Hate in Online Social Media*, in *ICWSM'16*, 2016.

[177] M. S. Hamm, *Conceptualizing Hate Crime in a Global Context*, *Hate crime: International perspectives on causes and control* (1994) 173–194.

[178] J. Levin and J. MacDevitt, *Hate crimes: The rising tide of bigotry and bloodshed.* Springer, 2013.

[179] J.-M. Xu, K.-S. Jun, X. Zhu, and A. Bellmore, *Learning from Bullying Traces in Social Media*, in *NAACL'12*, 2012.

[180] Statista, "Twitter, Number of Monthly Active Twitter Users Worldwide from 1st Quarter 2010 to 2nd Quarter 2017 (in millions)." `https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/`, 2017.

[181] E. Spertus, *Smokey: Automatic Recognition of Hostile Messages*, in *AAAI'97*, 1997.

[182] K. Dinakar, B. Jones, C. Havasi, H. Lieberman, and R. Picard, *Common Sense Reasoning for Detection, Prevention, and Mitigation of Cyberbullying*, *ACM Transactions on Interactive Intelligent Systems (TiiS)* **2** (2012), no. 3 18.

[183] W. Warner and J. Hirschberg, *Detecting Hate Speech on the World Wide Web*, in *ACL'12: Proceedings of the 2nd Workshop on Language in Social Media*, 2012.

[184] P. Burnap and M. L. Williams, *Hate Speech, Machine Classification and Statistical modeling of Information Flows on Twitter: Interpretation and Communication for Policy Decision Making*, *Internet, Policy and Politics Conference* (2014).

[185] S. Sood, J. Antin, and E. Churchill, *Profanity Use in Online Communities*, in *CHI'12*, 2012.

[186] A. Razavi, D. Inkpen, S. Uritsky, and S. Matwin, *Offensive Language Detection using Multi-level Classification*, *Advances in Artificial Intelligence* (2010) 16–27.

[187] G. Xiang, B. Fan, L. Wang, J. Hong, and C. Rose, *Detecting Offensive Tweets via Topical Feature Discovery over a Large Scale Twitter Corpus*, in *CIKM'12*, 2012.

[188] P. Burnap and M. L. Williams, *Us and Them: Identifying Cyber Hate on Twitter across Multiple Protected Characteristics*, *EPJ Data Science* **5** (2016), no. 1 11.

[189] H.-C. Huang, J.-M. Xu, K.-S. Jun, A. Bellmore, and X. Zhu, *Using Social Media Data to Distinguish Bullying from Teasing*, *Biennial meeting of the Society for Research in Child Development* (2013).

[190] A. Schmidt and M. Wiegand, *A Survey on Hate Speech Detection using Natural Language Processing*, in *SocialNLP'17: Proceedings of the 5th International Workshop on Natural Language Processing for Social Media*, 2017.

[191] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, *Abusive Language Detection in Online User Content*, in *WWW'16*, 2016.

[192] H. Zhong, H. Li, A. C. Squicciarini, S. M. Rajtmajer, C. Griffin, D. J. Miller, and C. Caragea, *Content-Driven Detection of Cyberbullying on the Instagram Social Network.*, in *IJCAI'16: Proceedings of the 25th International Joint Conference on Artificial Intelligence*, 2016.

[193] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati, *Hate Speech Detection with Comment Embeddings*, in *WWW'15*, 2015.

[194] N. D. Gitari, Z. Zuping, H. Damien, and J. Long, *A Lexicon-based Approach for Hate Speech Detection*, *International Journal of Multimedia and Ubiquitous Engineering* **10** (2015), no. 4 215–230.

[195] C. Van Hee, E. Lefever, B. Verhoeven, J. Mennes, B. Desmet, G. De Pauw, W. Daelemans, and V. Hoste, *Detection and Fine-grained Classification of Cyberbullying Events*, in *RANLP'15: International Conference Recent Advances in Natural Language Processing*, 2015.

[196] P. Burnap and M. L. Williams, *Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making*, *Policy & Internet* **7** (2015), no. 2 223–242.

[197] H. Hosseinmardi, S. A. Mattson, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra, *Detection of Cyberbullying Incidents on the Instagram Social Network*, *arXiv preprint arXiv:1503.03909* (2015).

[198] M. L. Williams and P. Burnap, *Cyberhate on Social Media in the Aftermath of Woolwich: A Case Study in Computational Criminology and Big Data*, *British Journal of Criminology* **56** (2015), no. 2 211–238.

[199] Wired, "Inside Google's Internet Justice League and its AI-Powered War on Trolls." `https://goo.gl/Nvf6ZA`, 2016.

[200] G. E. Hine, J. Onaolapo, E. De Cristofaro, N. Kourtellis, I. Leontiadis, R. Samaras, G. Stringhini, and J. Blackburn, *Kek, Cucks, and God Emperor Trump: A Measurement Study of 4chan's Politically Incorrect Forum and Its Effects on the Web*, in *ICWSM'17*, 2017.

[201] T. Davidson, D. Warmsley, M. Macy, and I. Weber, *Automated hate speech detection and the problem of offensive language*, in *ICWSM'17*, 2017.

[202] Facebook, "Controversial, Harmful and Hateful Speech on Facebook." `https://goo.gl/TWAHdr`, 2016.

[203] Twitter, "Hateful Conduct Policy." `https://goo.gl/NxR4sR`, 2016.

[204] E. Chandrasekharan, U. Pavalanathan, A. Srinivasan, A. Glynn, J. Eisenstein, and E. Gilbert, *You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech*, in *CSCW'18*, 2018.

[205] S. Benesch, *Countering Dangerous Speech to Prevent Mass Violence during Kenya's 2013 Elections*, tech. rep., United States Institute of Peace, 2014.

[206] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali, *Mean Birds: Detecting Aggression and Bullying on Twitter*, in *WebSci'17*, 2017.

[207] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali, *Hate is not Binary: Studying Abusive Behavior of #GamerGate on Twitter*, in *HT'17*, 2017.

[208] J. Cheng, M. Bernstein, C. Danescu-Niculescu-Mizil, and J. Leskovec, *Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions*, in *CSCW'17*, 2017.

[209] M. ElSherief, V. Kulkarni, D. Nguyen, W. Y. Wang, and E. Belding, *Hate lingo: A target-based linguistic analysis of hate speech in social media*, in *Twelfth International AAAI Conference on Web and Social Media*, 2018.

[210] R. Faris, A. Ashar, U. Gasser, and D. Joo, *Understanding Harmful Speech Online*, tech. rep., Berkman Klein Center for Internet and Society at Harvard University, 2016.

[211] CNN Tech, "Twitter Launches New Tools to Fight Harassment." `https://goo.gl/AbYbMv`, 2016.

[212] RSDB, "The Racial Slur Database." `http://rsdb.org/`, 1999.

[213] S. W. List and C. Filter, "List of Swear Words and Curse Words." `https://www.noswearing.com/dictionary`, 2011.

[214] A. Sellars, *Defining Hate Speech*, tech. rep., Berkman Klein Center for Internet and Society at Harvard University, 2016.

[215] E. Wulczyn, N. Thain, and L. Dixon, *Ex Machina: Personal Attacks Seen at Scale*, in *WWW'17*, 2017.

[216] D. Hovy, T. Berg-Kirkpatrick, A. Vaswani, and E. Hovy, *Learning whom to trust with mace*, in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1120–1130, 2013.

[217] S. Nilizadeh, A. Groggel, P. Lista, S. Das, Y.-Y. Ahn, A. Kapadia, and F. Rojas, *Twitter's Glass Ceiling: The Effect of Perceived Gender on Online Visibility*, in *ICWSM'16*, 2016.

[218] A. Mislove, S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, and J. Rosenquist, *Understanding the Demographics of Twitter Users*, in *ICWSM'11*, 2011.

[219] I. B. Docs, "The Science behind the Service." `https://console.bluemix.net/docs/services/personality-insights/science.html#science`, 2015.

[220] K. Lee, J. Mahmud, J. Chen, M. Zhou, and J. Nichols, *Who Will Retweet This?: Automatically Identifying and Engaging Strangers on Twitter to Spread Information*, in *ACM IUI'14*, 2014.

[221] J. Chen, E. M. Haber, R. Kang, G. Hsieh, and J. Mahmud, *Making Use of Derived Personality: The Case of Social Media Ad Targeting*, in *ICWSM'15*, 2015.

[222] N. K. Sharma, S. Ghosh, F. Benevenuto, N. Ganguly, and K. Gummadi, *Inferring Who-is-Who in the Twitter Social Network*, in *WOSN'12: Proceedings of the 2012 ACM Workshop on Online Social Networks*, 2012.

[223] N. K. Sharma, S. Ghosh, F. Benevenuto, N. Ganguly, and K. Gummadi, *Inferring Who-is-Who in the Twitter Social Network*, *ACM SIGCOMM Computer Communication Review* **42** (2012), no. 4 533–538.

[224] C. L. Ridgeway, *Gender, Status, and Leadership*, *Journal of Social issues* **57** (2001), no. 4 637–655.

[225] K. Yu, Z. Lu, and J. Stander, *Quantile Regression: Applications and Current Research Areas*, *Journal of the Royal Statistical Society* **52** (2003), no. 3 331–350.

[226] S. Jung, J. An, H. Kwak, J. Salminen, and B. Jansen, *Assessing the accuracy of four popular face recognition tools for inferring gender, age, and race*, in *International AAAI Conference on Web and Social Media*, 2018.

[227] A. Chakraborty, R. Sarkar, A. Mrigen, and N. Ganguly, *Tabloids in the era of social media?: Understanding the production and consumption of clickbaits in twitter*, *Proc. ACM Hum.-Comput. Interact.* **1** (Dec., 2017) 30:1–30:21.

[228] A. Chakraborty, J. Messias, F. Benevenuto, S. Ghosh, N. Ganguly, and K. Gummadi, *Who makes trends? understanding demographic biases in crowdsourced recommendations*, in *International AAAI Conference on Web and Social Media (ICWSM'17)*, 2017.

[229] B. J. Frey and D. Dueck, *Clustering by passing messages between data points*, *Science* **315** (2007), no. 5814 972–976.

[230] L. R. Goldberg, *The Structure of Phenotypic Personality Traits*, *American psychologist* **48** (1993), no. 1 26.

[231] I. Bluemix, "The Service in Action." `https://console.bluemix.net/docs/services/personality-insights/applied.html#applied`, 2017.

[232] N. A. Turiano, L. Pitzer, C. Armour, A. Karlamangla, C. D. Ryff, and D. K. Mroczek, *Personality Trait Level and Change as Predictors of Health Outcomes: Findings from a National Study of Americans (MIDUS)*, *Journals of Gerontology Series B: Psychological Sciences and Social Sciences* **67** (2011), no. 1 4–12.

[233] A. E. van Vianen, U.-C. Klehe, J. Koen, and N. Dries, *Career Adapt-abilities Scale—Netherlands Form: Psychometric Properties and Relationships to Ability, Personality, and Regulatory Focus*, *Journal of Vocational Behavior* **80** (2012), no. 3 716–724.

[234] S. Tok, *The Big Five Personality Traits and Risky Sport Participation*, *Social Behavior and Personality: an international journal* **39** (2011), no. 8 1105–1111.

[235] J. B. Hirsh, S. K. Kang, and G. V. Bodenhausen, *Personalized Persuasion: Tailoring Persuasive Appeals to Recipients' Personality Traits*, *Psychological science* **23** (2012), no. 6 578–581.

[236] N. Myszkowski and M. Storme, *How Personality Traits Predict Design-driven Consumer Choices*, *Europe's Journal of Psychology* **8** (2012), no. 4 641–650.

[237] P.-H. Arnoux, A. Xu, N. Boyette, J. Mahmud, R. Akkiraju, and V. Sinha, *25 tweets to know you: A new model to predict personality with social media*, in *Eleventh International AAAI Conference on Web and Social Media*, 2017.

[238] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. Seligman, *et. al.*, *Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach*, *PloS one* **8** (2013), no. 9.

[239] B. Plank and D. Hovy, *Personality Traits on Twitter-or-How to Get 1, 500 Personality Tests in a Week*, in *8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (EMNLP)*, 2015.

[240] I. B. Docs, "Personality Models." `https://console.bluemix.net/docs/services/personality-insights/models.html#outputBigFive`, 2015.

[241] J. Tanton, *Encyclopedia of Mathematics*. Facts On File, 2005.

[242] G. Toegel and J.-L. Barsoux, *How to Become a Better Leader*, *MIT Sloan Management Review* **53** (2012), no. 3 51.

[243] J. Price, L. Sloman, R. Gardner, P. Gilbert, and P. Rohde, *The Social Competition Hypothesis of Depression*, *The British Journal of Psychiatry* **164** (1994), no. 3 309–315.

[244] D. Schwartz, K. A. Dodge, and J. D. Coie, *The Emergence of Chronic Peer Victimization in Boys' Play Groups*, *Child development* **64** (1993), no. 6 1755–1772.

[245] M. Camodeca, F. A. Goossens, C. Schuengel, and M. M. Terwogt, *Links between Social Information Processing in Middle Childhood and Involvement in Bullying*, *Aggressive behavior* **29** (2003), no. 2 116–127.

[246] H. R. Johnson, M. J. Thompson, S. Wilkinson, L. Walsh, J. Balding, and V. Wright, *Vulnerability to Bullying: Teacher-reported Conduct and Emotional Problems, Hyperactivity, Peer Relationship Difficulties, and Prosocial Behaviour in Primary School Children*, *Educational Psychology* **22** (2002), no. 5 553–556.

[247] D. K. Linton and J. L. Power, *The Personality Traits of Workplace Bullies are Often Shared by Their Victims: Is there a Dark Side to Victims?*, *Personality and Individual Differences* **54** (2013), no. 6 738–743.

[248] S. Benesch, D. Ruths, K. P Dillon, H. Mohammad Saleem, and L. Wright, *Considerations for Successful Counterspeech*, tech. rep., Evaluating Methods to Diminish Expressions of Hatred and Extremism Online as part of The Kanishka Project of Public Safety Canada, 2016.

[249] K. Munger, *Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment*, *Political Behavior* **39** (2017), no. 3 629–649.

[250] Forbes, "Fighting Social Media Hate Speech With AI-Powered Bots." `https://www.forbes.com/sites/kalevleetaru/2017/02/04/fighting-social-media-hate-speech-with-ai-powered-bots/#6d64da7a27b1`, 2017.

[251] J. Cheng, C. Danescu-Niculescu-Mizil, and J. Leskovec, *Antisocial Behavior in Online Discussion Communities*, in *ICWSM*, pp. 61–70, 2015.

[252] H. M. Saleem, K. P. Dillon, S. Benesch, and D. Ruths, *A Web of Hate: Tackling Hateful Speech in Online Social Spaces*, in *Proceedings of the 1st Workshop on Text Analytics for Cybersecurity and Online Safety*, 2016.

[253] N. Wolfson, *Hate Speech, Sex Speech, Free Speech*. 1997.

[254] Y. Mehdad and J. R. Tetreault, *Do Characters Abuse More Than Words?*, in *SIGDIAL'16*, 2016.

[255] S. O. Sood, E. F. Churchill, and J. Antin, *Automatic Identification of Personal Insults on Social News Sites*, *Journal of the Association for Information Science and Technology* **63** (2012), no. 2 270–285.

[256] P. Burnap, O. F. Rana, N. Avis, M. Williams, W. Housley, A. Edwards, J. Morgan, and L. Sloan, *Detecting Tension in Online Communities with Computational Twitter Analysis*, *Technological Forecasting and Social Change* **95** (2015) 96–108.

[257] Z. Waseem, T. Davidson, D. Warmsley, and I. Weber, *Understanding Abuse: A Typology of Abusive Language Detection Subtasks*, *arXiv preprint arXiv:1705.09899* (2017).

[258] J. Eisenstein, A. Ahmed, and E. P. Xing, *Sparse Additive Generative Models of Text*, in *ICML'11*, 2011.

[259] Y. Sim, N. A. Smith, and D. A. Smith, *Discovering Factions in the Computational Linguistics Community*, in *Proceedings of the ACL 2012 Special Workshop on Rediscovering 50 Years of Discoveries*, 2012.

[260] W. Y. Wang, E. Mayfield, S. Naidu, and J. Dittmar, *Historical Analysis of Legal Opinions with a Sparse Mixed-Effects Latent Variable Model*, in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, 2012.

[261] A. Ritter, S. Clark, Mausam, and O. Etzioni, *Named Entity Recognition in Tweets: An Experimental Study*, in *EMNLP'11*, 2011.

[262] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, "The Development and Psychometric Properties of LIWC2015." `https://goo.gl/1n7y5A`, 2015.

[263] D. Chen, N. Schneider, D. Das, and N. A. Smith, *Semafor: Frame Argument Resolution with Log-linear Models*, in *Proceedings of the 5th International Workshop on Semantic Evaluation*, 2010.

[264] C. F. Baker, C. J. Fillmore, and J. B. Lowe, *The Berkeley Framenet Project*, in *the 36th Annual Meeting of the ACL and 17th International Conference on Computational Linguistics*, 1998.

[265] J. Ruppenhofer, M. Ellsworth, M. R. Petruck, C. R. Johnson, and J. Scheffczyk, *FrameNet II: Extended theory and practice*, 2006.

[266] A. Søgaard, B. Plank, and H. Martinez Alonso, *Using Frame Semantics for Knowledge Extraction from Twitter*, in *ICWSM'15*, 2015.

[267] M. Cikara, M. M. Botvinick, and S. T. Fiske, *Us versus Them: Social Identity Shapes Neural Responses to Intergroup Competition and Harm*, *Psychological Science* **22** (2011), no. 3 306–313.

[268] D. K. Citron and H. Norton, *Intermediaries and hate speech: Fostering digital citizenship for our information age*, *Boston University Law Review* **91** (2011) 1435.

[269] R. Baumeister and M. Leary, *The Need to Belong: Desire for Interpersonal*, *The Use of Oral Narrative in North American Families* **159** (1995).

[270] J. Kagan, *Loneliness: Human Nature and the Need for Social Connection*, *American Journal of Psychiatry* **166** (2009), no. 3 375–376.

[271] J. T. Cacioppo and W. Patrick, *Loneliness: Human nature and the need for social connection.* WW Norton & Company, 2008.

[272] D. Ahn and D.-H. Shin, *Is the social use of media for seeking connectedness or for avoiding social isolation? mechanisms underlying media use and subjective well-being*, *Computers in Human Behavior* **29** (2013), no. 6 2453–2462.

[273] Y. Lu and J. K. Lee, *Stumbling Upon the Other Side: Incidental Learning of Counter-attitudinal Political Information on Facebook*, *New Media & Society* **21** (2019), no. 1 248–265.

[274] M. ElSherief, E. Belding, and D. Nguyen, *# NotOkay: Understanding Gender-Based Violence in Social Media*, in *Eleventh International AAAI Conference on Web and Social Media*, 2017.

[275] E. Bakshy, S. Messing, and L. A. Adamic, *Exposure to Ideologically Diverse News and Opinion on Facebook*, *Science* **348** (2015), no. 6239 1130–1132.

[276] C. R. Sunstein, *Republic.com.* Princeton university press, 2002.

[277] C. R. Sunstein, *Echo chambers: Bush v. Gore, impeachment, and beyond.* Princeton University Press Princeton, NJ, 2001.

[278] M. A. Wong, R. Frank, and R. Allsup, *The Supremacy of Online White Supremacists–An Analysis of Online Discussions by White Supremacists*, *Information & Communications Technology Law* **24** (2015), no. 1 41–73.

[279] J. A. Schafer, *Spinning the Web of Hate: Web-based Hate Propagation by Extremist Organizations*, *Journal of Criminal Justice and Popular Culture* **9** (2002), no. 2 69–88.

[280] J. Berger, *Nazis vs. ISIS on Twitter. A Comparative Study of White Nationalist and ISIS Online Social Media Networks*, *GW Program on Extremism* (2016).

[281] FBI, "Hate crimes." `https://www.fbi.gov/investigate/civil-rights/hate-crimes`, Jan, 2018.

[282] L. G. McNamee, B. L. Peterson, and J. Peña, *A call to educate, participate, invoke and indict: Understanding the communication of online hate groups*, *Communication Monographs* **77** (2010), no. 2 257–280.

[283] K. C. Thompson, *Watching the Stormfront: White Nationalists and the Building of Community in Cyberspace*, *Social Analysis: The International Journal of Social and Cultural Practice* **45** (2001), no. 1 32–52.

[284] M. ElSherief, S. Nilizadeh, D. Nguyen, G. Vigna, and E. Belding, *Peer to Peer Hate: Hate Speech Instigators and Their Targets*, in *Proceedings of the 12th International AAAI Conference on Web and Social Media*, ICWSM'18, 2018.

[285] E. Chandrasekharan, M. Samory, S. Jhaver, H. Charvat, A. Bruckman, C. Lampe, J. Eisenstein, and E. Gilbert, *The internet's hidden rules: An empirical study of reddit norm violations at micro, meso, and macro scales*, *Proceedings of the ACM on Human-Computer Interaction* **2** (2018), no. CSCW 32.

[286] J. Hunsinger, *Reflexivity in e-science: Virtual communities and research institutions*, *ACM SIGGROUP Bulletin* **25** (2005), no. 2 38–42.

[287] P. Fortuna and S. Nunes, *A survey on automatic detection of hate speech in text*, *ACM Computing Surveys (CSUR)* **51** (2018), no. 4 85.

[288] D. Houtman and S. Aupers, *'Stormfront is like a Second Home to Me': Social Exclusion of Right-Wing Extremists*, in *Paradoxes of Individualization*, pp. 85–102. Routledge, 2016.

[289] M. Caiani and L. Parenti, *European and American Extreme Right Groups and the Internet.* Routledge, 2016.

[290] R. Graham, *Inter-ideological Mingling: White Extremist Ideology Entering the Mainstream on Twitter*, *Sociological Spectrum* **36** (2016), no. 1 24–36.

[291] A. Klein, *Slipping Racism into the Mainstream: A Theory of Information laundering*, *Communication Theory* **22** (2012), no. 4 427–448.

[292] J. Morgan, "These Charts Show Exactly How Racist and Radical the Alt-right has Gotten this Year." `https://wapo.st/2FBKzdK`, September, 2016.

[293] R. Ottoni, E. Cunha, G. Magno, P. Bernadina, W. Meira Jr, and V. Almeida, *Analyzing Right-wing YouTube Channels: Hate, Violence and Discrimination*, in *WebSci'18*, 2018.

[294] Schafer and J. R., *Seven stage hate model: The psychopathology of hate groups*, Feb, 2003.

[295] R. McVeigh, *Power devaluation, the ku klux klan, and the democratic national convention of 1924*, *Sociological Forum* **16** (2001), no. 1 1–30.

[296] L. Quillian, *Prejudice as a response to perceived group threat: Population composition and anti-immigrant and racial prejudice in europe*, in *American Sociological Review*, vol. 60, pp. 586–611, 08, 1995.

[297] E. G. Bormann, J. F. Craan, and D. C. Shields, *In Defense of Symbolic Convergence Theory: A Look at the Theory and its Criticisms after Two Decades*, *Communication Theory* **4** (1994), no. 4 259–294.

[298] B. Jackson, *Linking the Immediate with the Mass-mediated Theatre in Organizations: The Case for Symbolic Convergence Theory*, in *Proceedings of the 15 th International Standing Conference on Organizational Symbolism*, 1997.

[299] S. W. Littlejohn and K. A. Foss, *Theories of Human Communication.* 2010.

[300] SPLC, "What We Do." `https://www.splcenter.org/what-we-do`, 2018.

[301] D. Andone, "Southern Poverty Law Center's List of Hate Groups." `https://www.cnn.com/2017/08/17/us/hate-groups-us-map-trnd/index.html`, Aug, 2017.

[302] M. Suttmoeller, S. Chermak, and J. D. Freilich, *The Influence of External and Internal Correlates on the Organizational Death of Domestic Far-Right Extremist Groups*, *Studies in Conflict & Terrorism* **38** (2015), no. 9 734–758.

[303] B. Barnett, *Untangling the Web of Hate: Are Online "hate Sites" Deserving of First Amendment Protection?* Cambria Press, 2007.

[304] Pew Research Center, "State of the News Media." `http://www.pewresearch.org/topics/state-of-the-news-media/`, June, 2017.

[305] Southern Poverty Law Center, “Intelligence report: The year in hate and extremism.” https://bit.ly/2DdTTmE, 2017.

[306] R. L. Nydick and H. J. Weiss, *Let’s put the seasonality and trend in decomposition*, *INFORMS Transactions on Education* **12** (2012), no. 3 147–152.

[307] M. E. Duffy, *Web of Hate: A Fantasy Theme Analysis of the Rhetorical Vision of Hate Groups Online*, *Journal of Communication Inquiry* **27** (2003), no. 3 291–312.

[308] M. G. Lacy, *Exposing the spectrum of whiteness: Rhetorical conceptions of white absolutism*, *Annals of the International Communication Association* **32** (2008), no. 1 277–311.

[309] R. Kiros, Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler, *Skip-thought vectors*, in *Advances in neural information processing systems*, pp. 3294–3302, 2015.

[310] D. Kingma and J. Ba, *Adam: A method for stochastic optimization*, *ICLR* **2015**.

[311] D. M. Blei, A. Y. Ng, and M. I. Jordan, *Latent dirichlet allocation*, *JMLR* **3** (2003) 993–1022.

[312] Anti-Defamation League, “Echo.” https://www.adl.org/education/references/hate-symbols/echo.

[313] D. Greene and J. P. Cross, *Exploring the political agenda of the european parliament using a dynamic topic modeling approach*, *Political Analysis* **25** (2017), no. 1 77–94.

[314] S. Zannettou, T. Caulfield, J. Blackburn, E. D. Cristofaro, M. Sirivianos, G. Stringhini, and G. Suarez-Tangil, *On the origins of memes by means of fringe web communities*, in *IMC*, 2018.

[315] G. Guibon, M. Ochs, and P. Bellot, *From emojis to sentiment analysis*, 2018.

[316] Z. Ma, A. Sun, and G. Cong, *On predicting the popularity of newly emerging hashtags in twitter*, *JASIST* **64** (2013) 1399–1410.

[317] C. Antonelli, *The Diffusion of Advanced Telecommunications in Developing Countries.* OECD Development Center Studies, Paris, France, 1991.

[318] New York Times, “Gatekeepers or censors? how tech manages online speech.” https://www.nytimes.com/2018/08/07/technology/tech-companies-online-speech.html, Aug, 2018.

[319] Twitter, “Creating new policies together.” https://blog.twitter.com/official/en_us/topics/company/2018/Creating-new-policies-together.html.

[320] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, *Distributed representations of words and phrases and their compositionality*, in *Advances in Neural Information Processing Systems 26* (C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, eds.), pp. 3111–3119. Curran Associates, Inc., 2013.

[321] A. Wilkinson, "The 'Tears of Joy' Emoji is the Worst of All – It's Used to Gloat about Human Suffering." `https://bit.ly/2gFguOH`, Nov, 2016.

[322] Anti-Defamation League, "Pepe the Frog." `https://bit.ly/2r3fqIv`.

[323] I. Gambert and T. Linné, "How the Alt-right Uses Milk to Promote White Supremacy." `https://bit.ly/2QV2qn2`, Oct, 2018.

[324] Southern Poverty Law Center, "Is that an OK Sign? A White Power Symbol? Or Just a Right-wing Troll?." `https://bit.ly/2NxXxth`.

[325] Media Bias/Fact Check, "Questionable sources." `https://mediabiasfactcheck.com/fake-news/`.

[326] V. Paxson, *End-to-end Routing Behavior in the Internet*, *IEEE/ACM Transactions on Networking* **5** (1997), no. 5 601–615.

[327] I. Sutskever, O. Vinyals, and Q. V. Le, *Sequence to sequence learning with neural networks*, in *Advances in neural information processing systems*, 2014.

[328] Anti-Defamation League, "Online Hate and Harassment: The American Experience." `https://www.adl.org/onlineharassment`, 2018.

[329] C. Chelmis and D.-S. Zois, "ICWSM 2018 Tutorial Characterization, Detection, and Mitigation of Cyberbullying." `http://www.cs.albany.edu/~cchelmis/icwsm2018tutorial/`, 2018.

[330] L. Wright, D. Ruths, K. P. Dillon, H. M. Saleem, and S. Benesch, *Vectors for Counterspeech on Twitter*, in *Proceedings of the First Workshop on Abusive Language Online*, pp. 57–62, 2017.

[331] New York Times, "11 Killed in Synagogue Massacre; Suspect Charged With 29 Counts." `https://www.nytimes.com/2018/10/27/us/active-shooter-pittsburgh-synagogueshooting.html?fbclid=IwAR0NBSD3egj2crE2A1q1l1bUgC78pmaypXLdoOCtcNrn-CnFOebIO592I8U`, September, 2018.