Automating Hate: Exploring Toxic Reddit Norms with Google Perspective

Nicholas Chevrier

Submitted to the University of Ottawa in partial fulfillment of the requirements for the Master of Arts degree in Communication

Department of Communication
Faculty of Arts
University of Ottawa

Abstract

The Canadian Online Harms Legislation (COHL) proposal identifies proactive Automated Moderation as a solution to classifying and removing online content which violates norms such as hate. Emerging automated moderation algorithms include Google Perspective, a machine learning model which scores hateful features in text content as "toxicity." This study identifies that hateful community content norms are currently emerging on volunteer user moderation platforms such as Reddit. To operationalize these concepts, a Theoretical Framework is constructed using Gorwa's (2019) Platform Governance models and Massanari's (2017) overview of Toxic Technoculture communities. While previous research exploring community toxicity is discussed, there is a gap in research which analyzes the Post, Comment, and Image Meme contributions of Reddit Moderator users to hateful community content norms. As such, an analysis of the Reddit community R/Metacanada is constructed which compares the toxicity of Moderator and user contributions using Google Perspective. The results of the applied Mann-Whitney U test analysis indicate that r/Metacanada Moderators and users contribute content at similar toxicity levels. Supplementing these tests, RQ1 then structures a qualitative analysis of false negative results which may emerge in the automated classification of multi-modal image content. Identifying that hate in online memes is structured through layered Signifier and Signified elements, a critical discussion is established which interprets potential marginalizing effects of the COHL's automated moderation applying Noble's (2018) theory of Technological Redlining. As such, this thesis immerses itself within the contemporary context of online content regulation, drawing upon existing conceptualizations and methodological approaches, offering a critical discussion of regulating hate content using automated algorithms.

Keywords: Automated Moderation, Governance, Reddit, Google Perspective, Toxicity

Acknowledgments

Gratefully, the people around me never underestimated me and my abilities. To Hope, who has spent countless nights listening to my crazy ideas, I love you! To my mom and dad, who are still wondering what I am talking about to this day, but always listened, I love you both! To my friends who made me laugh at every moment and those who helped me when I hit bugs, I love all of you! I would not have been able to accomplish putting together this work without the endless support of all, and I am deeply grateful for their encouragement. To my late Grandfather, who remained deeply interested in my research until his passing. He would always ask whether I had submitted my thesis. Well, it's done! Thank you for your support and encouragement to pursue education. The seeds you have planted will continue to grow and bring joy to this world. I love you and I miss you.

Table of Contents

Chapter 1: Introduction

The Internet has a hate problem. While social media platforms have developed moderation strategies through which user contributions are filtered, some hateful content evades these mechanisms. As platforms further expand their scale and affordances, so do the layers of these moderation solutions. Through wielding recent advancements in the field of Machine Learning, algorithmic models are being developed to classify hate content at scale. This includes Google's Perspective API, an open-source machine learning automated moderation algorithm which classifies text. Perspective aims to "make it easier to host better conversations online'' (Perspective, 2021a) by assigning "toxicity" scores to platform content. Developed by Google's Jigsaw Subsidiary in 2017, Perspective's toxicity scores have subsequently been used to operationalize hate features in online content research (Mall et al., 2020; Mittos et al., 2020; Zannettou et al., 2020) as well as for comment platforms such as The New York Times (NYT Team, 2020).

Concurrently, state regulatory frameworks have proposed to mandate online platforms to implement automated moderation solutions. In Canada, a recent federal proposal for an Online Harms Legislation (COHL) maintains that platforms "must take all reasonable measures, to identify harmful content that is communicated" (Heritage Canada, 2021, para. 10). This framework suggests that platforms will be required to monitor and proactively remove user content which contains identified harms, including hate speech. The proposal identifies "the use of automated systems" (para. 10) to classify and remove this content in real time. This work establishes itself within this context of both platforms and governments determining that hate speech as a threat while proposing regulatory frameworks which mandate the use of automated moderation.

To analyze the proposed integration of proactive automated moderation solutions, this research explores how hateful content norms currently emerge on a platform. Reddit

maintains a unique community moderation system amongst major social media platforms. First implemented in 2008 (Ohanian, 2008), the platform's volunteer user moderation system provides anonymous users unique privileges to remove content and ban other users from an assigned community called a 'Subreddit'. Moderation authority is delegated to users who create a Subreddit or are added by its existing Moderators. As they perform this task of classifying content, Moderators can also contribute content to these communities through the same account.

The platform has experienced significant growth since this implementation, now hosting over 100,000 user moderated Subreddit communities (Reddit Press, 2021) and is currently the 7th most visited website in Canada (Alexa, 2021). Reddit posts contain text, images or external links to which users respond via text comments, resulting in almost 3.5 billion pieces of user content created in 2020 (Reddit Press, 2021).

Recently, following the 2020 Black Lives Matter protests, hundreds of Reddit Moderators wrote an open letter to CEO Steve Huffman accusing Reddit of "providing a platform for racist users and hateful communities" (Reddit, 2020a, para. 4). The contents of the letter argued that the platform affordances had been manipulated by bad actor Moderators while a lack of administrative action had established hateful content norms. Reddit administration responded to these concerns through a Content Policy update, claiming that "communities and users that incite violence or that promote hate based on identity or vulnerability will be banned" (Reddit, 2020b, para. 1).

Despite Reddit's apparent recommitment to removing hate from its platform, absent from this list of banned Subreddits is r/Metacanada, a community hosting over 38,000 users which describes itself as "the only not-retarded Canadian Subreddit" (Gruzd et al., 2020, p. 5). Launched as a satirical alternative to the default r/Canada Subreddit, the community engages in discussions on political and cultural topics (Gruzd et al., 2020). While current

research analyzing the community is minimal, the subreddit does share many users with the controversial r/The_Donald community (Nithyanand et al., 2017, p. 14). r/Metacanada was thus selected a case study to operationalize Google Perspective's toxicity scores within the context of Canadian online content regulation.

A Theoretical Framework is constructed which first grounds the analysis within ongoing discussions of the process and effects of moderation expanding digital communities. Gorwa's (2019) Platform Governance approach constructs models to conceptualize the mechanisms which influence the arrangement of comments. The authors' Self-governance, Co-governance and External governance models are applied in the analysis of the automated moderation of the content of hateful Reddit communities.

Meanwhile, Massanari (2017) argues that Reddit's design supports "Toxic Technocultures" (p. 330), communities whose users exploit platform affordances to share hateful content. Massanari explores how Reddit's non-human actors, which includes governance and algorithmic mechanisms, allow human users to establish Toxic Technocultures on the platform.

Furthermore, automated algorithms like Google Perspective appear to have varying effects on different communities of users. Noble (2018) defines Technological Redlining as "the ways . . . digital decisions reinforce oppressive social relationships and enact new modes of racial profiling" (p. 7). These instances have been repeatedly framed as unique "glitches" (p. 153), resulting in little public awareness of their frequency, nor leading to any equity-focused modifications of the algorithms shaping these occurrences (Noble, p. 121). It was thus important to construct a critical analysis of the functionalities of Google Perspective.

Past work demonstrates that Moderators of Toxic Technocultures are engaged in the process of establishing community norms through both visible and invisible actions (Gilbert, 2020). Specifically, hateful norms emerge in small volunteer moderated communities through

some use of multiple platforms affordances (Chandrasekharan et al., 2018). Google Perspective has previously been applied to classify user (Mall et al., 2020) and community (Mittos et al., 2020) toxicity trends in content, however this past research does not analyze the visible contributions of Reddit Moderator users. Recognizing the proactive automated moderation of hate content mandated by the COHL, the primary goal of this research is analyzing the content of Moderators who maintain governance influence in volunteer user moderation platforms like Reddit. Massanari (2017) suggests that Moderators of Toxic Technocultures post a sizable portion of the hateful content to their Subreddits. As past research (Mittos et al., 2020; Zannettou et al., 2020) exploring hate speech on the Reddit platform displays its presence in posts, and comments, the following hypotheses are tested:

H1: Toxic Technoculture Moderators post higher toxicity posts than non-Moderator users

H2: Toxic Technoculture Moderators post higher toxicity comments than non-Moderator users

As well, novel classification methodologies were identified hateful image memes through an analysis of text contained in the image. For these reasons, the following hypothesis was developed:

H3: Toxic Technoculture Moderators post higher toxicity image texts than non-Moderator users

To test these hypotheses, the posts, comments, and images posted to the r/Metacanada between June 1-30th, 2020 were collected. Next, the text contained in the community's memes were extracted using Google Vision API's OCR functionality. Each post title, comment, and extracted image-with-text meme text were then scored with Google Perspective's toxicity model. Using Mann-Whitney U Tests, differences in the toxicity scores means of r/Metacanada users and Moderators across post, comment, and image text were analyzed.

However, each of the hypotheses were rejected, indicating that r/Metacanada

Moderators and users contribute similar levels of toxicity. Overall low toxicity is also observed in Positively Skewed distributions formed by the toxicity scores gathered from each content type. These findings establish a discussion of how the large-scale moderation of content may interpret the content of Reddit communities which appear to have hateful norms.

Next, this research explores how the application of an automated moderation algorithm in the analysis of multi-modal content can lead to misclassifications. First, two r/Metacanada memes which received low toxicity scores in the quantitative analysis were selected. Then a preliminary response to the following exploratory research question was constructed:

RQ1 How is hate signified in image memes containing non-toxic image text?

To do so, the two memes were analyzed using a multi-modal discourse analysis methodology developed by Yoon (2016). This approach applies social semiotics to visual and textual content to interpret how themes are structured within intertextual memes. This analysis demonstrates that hateful r/Metacanada memes are constructed through layered denotative visual and textual elements. The implications of these findings are further explored in the context of the COHL's mandate for automated moderation, offering a preliminary discussion of the potential of these initiatives to impose Technological Redlining.

This research begins by constructing a Theoretical Framework which discusses emerging Platform Governance models, Toxic Technocultures via ANT, and Technological Redlining. This establishes a Literature Review which discusses moderation's influence of community norms, then explores Reddit's volunteer user moderation system in application. This is performed through reviewing the platform's 2020 Transparency Report as well as previous work which analyzes the norm setting influence of Subreddit Moderators. An overview of the function and applications of Automated Moderation is then constructed, leading to a review of Google Perspective's applications in research and emerging limitations

of the tool. This section also discusses related work in the field of image classification through the extraction of image meme text. The Literature Review concludes with a research argument for the analysis, justifying the research questions and the identified hypotheses.

The fourth chapter provides a summary of the methodological approach used to construct this analysis. The ensuing Analysis section first presents a summary of the data and then the results of testing H1, H2 and H3. Then, responding to RQ1, a critical analysis of two r/Metacanada memes, revealing how semiotic layers structure hate in multi-modal images. A discussion chapter interprets these findings, applying the lenses of the identified theories. This work concludes with a summary of findings, identifies limitations of the analysis, suggests opportunities for future research and offers an interpretation of potential future developments in the application of automated moderation in the classification of online hate content.

Chapter 2: Theoretical Framework

This chapter constructs a Theoretical Framework which establishes a background of the mechanisms of online platforms, conceptualizes the emergence of hateful reddit communities and then discusses potential effects of the automated moderation of content.

2.1 Platform Governance

Gorwa (2019) defines platforms as "online, data-driven apps and services" (p. 4). Many of these services such as Facebook Messenger and Google Search are developed and maintained by large corporations who profit off "engineering what has become the global infrastructure of free expression" (p. 5). To do so, platforms have established numerous governance mechanisms including "systems of rules, norms, and civic labour governing an online community" (p. 9). While these corporations maintain extensive control of these mechanisms, they are increasingly influenced by "local, national, and supranational mechanisms of governance" (p. 5) as they aim to operate in various legal jurisdictions around the globe. Platform governance is thus an approach which requires "an understanding of technical systems" while considering the "inherently global arena within which these platform companies function" (p. 9).

Gorwa's (2019) conceptualization of platform governance proposes that 3 models of governance have emerged. First, the currently "dominant" model of "Self-governance" establishes a "relatively laissez-faire relationship between governing institutions and platform companies" (p. 13). Enabled by the liability provisions towards moderating user content by Section 230, platforms engage in minimal transparency and make decisions with "minimal external oversight" (p. 13). In addressing the various concerns relating to user content such as its use to spread hate speech, regulators "rely on goodwill" (p. 13) from platforms while offering "limited recourse" when this hate turns into violence outside these digital spaces. However, Gorwa argues that platforms' voluntary response to some of these issues suggest

that they "can be nudged in the right direction without complex regulatory interventions" (p. 13). Reddit's internal modification of their hate content policy is an example of one of these initiatives. This process was undertaken by the platform's internal governance mechanisms and implemented through "technical changes" and "transparency efforts" (p. 13) through which they retain oversight. Nevertheless, Gorwa argues that these initiatives "do little to provoke systemic change or modify platform business models" despite internal platform practices "which may be fundamentally problematic" (p. 14).

The next model identified by Gorwa is "External governance" which considers how government intervention influences platforms. This includes "specially crafted legislation around existing policy levers" such as Germany's NetzDG which according to Gorwa are "born out of apparent frustration with the Self-governance model" (p. 14). NetzDG requires platforms to remove "evidently unlawful" (p. 14) material including hate speech within 24 hours of a user report and provide regular transparency reports on these decisions. Heldt (2019) identifies that NetzDG "might have incentivized . . . platforms to remove hate speech faster" but a lack of transparency in the reports provided by platforms establishes "that there is no certainty about its effect" (p. 2) on limiting hateful content.

Current regulatory mandates like NetzDG appear to be an influence on External regulatory frameworks, specifically to the government of Canada's recently proposed Online Harm Legislation. While NetzDG requires platforms to moderate content following a user complaint, the technical paper by Canada proposes they "must take all reasonable measures, to identify harmful content that is communicated on its [platform]" (Heritage Canada, 2021, para. 10). This is a far more proactive approach than NetzDG mandates, and the proposed legislation suggests the use of automated systems to achieve these demands. Harmful content described in this proposal includes child exploitation, terrorism, inciting violence, and hate speech which "is likely to cause harms identified by the Supreme Court of Canada" (Heritage

Canada, 2021, para. 8). Platforms must also assure that the automated tools implemented in this proactive approach to moderation "do not result in differential treatment of any group based on a prohibited ground of discrimination within the meaning of the Canadian Human Rights Act" (Heritage Canada, 2021, para. 8).

Lastly, Co-governance models "seek to provide some values of democratic accountability without making extreme changes to the status quo" (Gorwa, 2019, p. 16) of platforms. This process involves platforms establishing "more granular forms of user participation in policy decisions" (p. 16). Gorwa suggests that Co-governance "leads away from major, corporatized platforms and towards various platform cooperatives, decentralized systems, and other forms of community self-management" (p. 16). It is suggested that "scale issues" and "network effects" are limitations of these initiatives. However, Reddit's moderation system follows this Co-governance model, providing volunteer users the affordances to construct communities on the platform.

Gorwa et al. (2020) model online hate content as a "governance puzzle" (p. 1). Exploring the intersection of these 3 models of platform governance seems essential to understanding how hate emerges on platforms as well as the attempts to regulate this content. As such, observing Reddit's Co-governance affordances constructs a critical discussion of the mechanisms which exert governance on the Reddit platform. The Literature Review thus explores the "systems of rules, norms, and civic labour" (Gorwa, 2019, p. 9) which exert governance on the Reddit platform, observing these elements within the scope of an External governance mandating the automated moderation of hate content. The ensuing analysis applies these concepts, constructing a case study which demonstrates how the 3 models of platform governance influence the emergence of hate and engage in the "puzzle" of removing this content.

2.2 Toxic Technocultures

By aiming to model the automated regulation of hate content through a governance lens, exploring how hate content becomes normalized on platforms remains essential. Massanari (2017) argues that a combination of Reddit's affordances has allowed users to construct "Toxic Technocultures" (p. 330). Toxic Technocultures are defined as "toxic cultures that are enabled by and propagated through sociotechnical networks such as Reddit, 4chan, Twitter, and online gaming" (p. 333). They are conceptualized by Massanari through the lens of Actor-Network Theory (ANT) which emerged from the works of Callon (1986), Latour (1996), and Law (1992) in the field of science and technology studies (Ritzer, 2005). Critical of an "implicit idealism" (Couldry, 2008, p. 95) afforded to human actors, ANT claims that "knowledge is a social product rather than something generated by through the operation of a privileged scientific method" (Law, 1992, p. 381). This approach explores "the creation and maintenance of coextensive networks of human and nonhuman elements" (Walsham, 1997, p. 466) otherwise known as "actors". Walsham defines actors as "both human beings and nonhuman actors such as technological artefacts" (p. 468).

ANT proposes these various actors possess agency (Latour, 1996) in shaping social processes via a "heterogeneous network of aligned interests" (Walsham, p. 468). Through an asymmetrical conceptualization of their influence, ANT challenges the "rigid separation of humans and nonhumans" (Walsham, p. 476), exploring how their multiple ties establish social reality. A frequent criticism of ANT encountered in the literature is its inability to make a "distinction between human action and the behaviour of things [which] is an abdication of human responsibility" (Walsham, p. 469). However, ANT rejects a consideration of the "priori substance" of these actors, suggesting that it is "via the networks in which they associate that actants derive their nature" (Ritzer, 2005, p. 1). ANT thus explores how "relatively stable networks of aligned interests are created and maintained, or

alternatively, to examine why such networks fail to establish themselves" (Walsham, p. 469). In application, ANT provides a framework to "describe the observable interactions in a network of human and non-human actants'' in order to "explain how claims become facts as they connect to sources within the network" (Pantumsinchai, 2018, p. 766).

Consequently, ANT provides Massanari (2017) a functional framework through which the social effects of the evolving actor-networks formed by Reddit's human and non-human actors are conceptualized as Toxic Technocultures. Users within these spaces, often technically proficient, utilize non-human platform affordances to create communities containing morally hateful content. However, Massanari identifies that Reddit is currently a far more accessible platform for users, arguing its "design, policies and norms" (p. 336) have supported the emergence of popular Toxic Technocultures.

In Massanari's (2017) ANT approach to Reddit's Toxic Technocultures, non-human actors include the Subreddit creation system, its governance structure, and policies around offensive content (p. 330). Specifically, the author performs an ethnographic study of two Toxic Technocultures: a leaked celebrity image scandal known as #TheFappening and Gamergate, an online harassment campaign against women in the video game industry. In both cases, users utilized Reddit to create Subreddits to share objectionable content which was further spread to other communities via the aggregation algorithm. The salience and private nature of the content shared in both instances resulted in significant traffic for Reddit, and the platform acted reluctantly to take action. r/TheFappening was permitted to host stolen nude images of celebrities until an image of a female minor was shared, resulting in its banning by Reddit administrators. Meanwhile, the Subreddit formed in the emergence of Gamergate, r/Kotakuinaction, continues to exist despite presenting many harassing behaviours (Jhaver et al., 2018). Massanari (2017) notes this is a reoccurring approach to governance taken by Reddit administration, and that issues are often ignored until public

pressure leads to changes (p. 342).

Toxic Technocultures use “memetic logics to radicalize and spread their message of hate” (Massanari, 2018, p. 167) while Moderators explicitly removed any resistance to this discourse. This is believed to have been “extremely effective in pushing far-right ideas and racist speech” (p. 171) on Reddit and beyond. Massanari adds that “it is often the most powerful Moderators/users who post objectionable content in the first place” (as cited in Zimdars & Mcleod, 2020, p. 148). This demonstrates that some human Moderators engage with Reddit’s non-human affordances to spread hate content within the network.

Massanari (2017) writes that “the ways in which Toxic Technocultures develop, are sustained, and exploit platform design is imperative” (p. 342) to understanding how toxic content, such as hate speech, emerges on platforms. The Literature Review thus explores further findings on the r/The_Donald Subreddit’s toxicity as well as how memes contribute to hate on the Reddit platform. This establishes an analysis of how Toxic Technoculture Moderators use content affordances to curate hateful Subreddit communities. This requires the operationalizing of the human and non-human actors which form the Reddit platform, thus contributing to ANT’s exploration of discourse as a “product or an effect of a network of heterogeneous materials” (Law, 1992, p. 381).

2.3 Technological Redlining

While the COHL proposes automated content classification as the solution to Toxic Technocultures, marginalizing effects of these initiatives have been identified by authors such as Noble (2018). Their theory of Technological Redlining explores the marginalizing effects of automated content classification algorithms. Technological Redlining relates the contemporary effects of automation to the historic systemic denial of services such as housing and banking to racialized communities, a concept known as “redlining” (2018, p. 19). Tasked with preventing mortgage foreclosures following the Great Depression, the

United States' Home Owners' Loan Corporation (HOLC) produced 'security maps' of urban neighbourhoods to assess the credit worthiness of homeowners (Nardone et al., 2021). Basing these assessments off factors including "prior home values, presence of industry, and racial demographics" (Nardone et al., 2021, p. 1), the HOLC assigned the following grades to neighbourhoods in over 200 cities:

A. Green: "Best"
B. Blue: "Still Desirable"
C. Yellow: "Definitely Declining"
D. Red: "Hazardous"

Allen (2019) indicates that "Redlined neighbourhoods were usually urban areas housing people of color" (p. 236). As these maps informed mortgage, insurance, and tax algorithms, they established a cycle of divestment from minority communities labelled as "hazardous" throughout the country. While initiatives such as the 1968 Fair Housing Act aim to ban the practice, Nardone et al. (2021) write that redlining has "reinforced pre-existing segregation in many places and (is) associated with present-day levels of racial segregation, poverty, and income inequality" (p. 1). Redlining presents a crucial example of the risks of the data supplied to algorithms, how this information is interpreted, and the policies which are established as a result.

According to Noble, Technological Redlining refers to "the ways . . . digital decisions reinforce oppressive social relationships and enact new modes of racial profiling" (2018, p. 19). While redlining mechanisms were constructed through pencil crayons, paper and calculus, the effects of social media and automated algorithms are more elusive as there is less "oversight or intervention" (as cited in Bulut, 2018, p. 295) in the development and implementation of code than with the policies of public institutions such as the HOLC.

Noble explores this concept with several examples, including discursive representations of race and gender within Google search results. Searching for innocuous terms such as "Black girls" (p. 24), this process revealed that Google's search algorithm often

returned sexually and violently explicit content. The author indicates these images evoke stereotypical representations of African Americans, specifically the “jezebel whore”, reinforcing a “White male gaze that functions as the dominant paradigm on the Internet” (p. 59). Noble argues this analysis reveals the “power of algorithms in controlling the image, concepts, and values assigned to people” (p. 184). The threat of Technological Redlining thus challenges the perception of the internet “as a democratic space” (p. 100), as platforms such as Google rapidly increase their algorithmic complexity while maintaining a significant monopoly within the online information access process. Noble dismisses platforms’ attempts to frame these issues as “glitches”, writing that “this process reflects a corporate logic of either willful neglect or a profit imperative that makes money from racism and sexism” (p. 26).

Noble identifies that Google has conveniently positioned their search results as operating under the “auspices of free speech” (p. 143) rather than being subjected to the regulations facing a public information institution such as a library. Legal frameworks including Section 230 have enabled Google to adjust their algorithm as required while facing essentially no legal recourse for the impacts of these decisions. A notable example of these actions are Google’s algorithmic filters to prevent the sale of Nazi memorabilia within Germany and France (p. 95). On the request of these governments, Google demonstrated the ability to selectively remove this content from search when faced with regulatory requests. Noble writes that this is an indication that algorithmic “search results are deeply contextual and easily manipulated, rather than objective, consistent, and transparent” (p. 95).

Noble (as cited in Bulut, 2018) argues that technological redlining often emerges “under the guise of a more perfect and objective adjudication process when it is anything but” (p. 295). This suggests that the COHL’s mandate for the automated moderation of online platforms may have influences which parallel the historical effects of redlining in housing.

While Noble's research focuses primarily on Google's search algorithm they call for "a full-on re-evaluation of the implications of our information resources being governed by corporate controlled advertising companies" (p. 5). The author argues that while platforms have potential to act as a "democratizing force", they "must be contextualized in the historical conditions that both create it and are created by it" (p. 231). This research thus constructs a critical exploration of the COHL's automated moderation mandate. To do so, the Literature Review explores why misclassifications occur in content automation. This serves to construct an analysis which uses an automated moderation solution to perform research while maintaining a critical lens of its potentially marginalizing effects. As such, this research explores how Google's emerging role in the automated moderation of content may contribute to Technological Redlining when applied at scale.

Chapter 3: Literature Review and Research Argument

The following Literature Review chapter begins by exploring the influence of content moderation, before discussing hateful norms which have emerged on Reddit. This leads to a discussion on the types and integrations of automated moderation tools. A discussion on both the applications and limitations of Google Perspective as well as image-with-text classification concludes the section. Methodological components of the identified research are discussed throughout to provide further context and justification for the approach in the resulting analysis.

3.1 Moderation

3.1.1 Community Norms

The effects of platform governance mechanisms have been explored through analyzing the content moderation of platforms. Grimmelmann (2015) writes that the goal of moderation (p. 61) is setting "strong shared norms among participants" (p. 4). Norms "target every form of strategic behaviour" and are an "emergent property of social interactions" (p. 61). Social platforms like Reddit are "acutely sensitive to group norms" (p. 62), with invasive actors such as "spammers" and "trolls" threatening the establishment of beneficial community norms.

The author recognizes that platforms maintain "limited power over group norms'', allowing them to "nudge norms in one direction or another, possibly unpredictably" (p. 61). This occurs both directly and indirectly. Direct norm setting occurs "by articulating them". Examples include the establishment of Codes of Conduct or Rules and responding to user behaviour (p. 62). Alternatively, indirect norm setting is performed through less visible moderation practices of "Excluding" (p. 56), "Pricing" (p. 57), and "Organizing" (p. 58). "Excluding" and "Pricing" practices provide means to limit user access through technical and economic means respectively.

Moderator "Organizing" influences how community content is arranged and maintained. "Organizing" is practiced by Moderators through the "Deletion", "Editing", "Annotation", and "Filtering" (p. 59) of user content. The availability of these "Organizing" tools to Moderators and the potential influence of their implementation varies widely across platforms. Grimmelmann (2015) identifies that "different communities strike the balance differently" (p. 63) in implementing moderation strategies, while noting that their influence on norm setting within communities "can spur users to individual effort at the cost of social cohesion" (p. 63).

3.1.2 Volunteer User Moderation

Volunteer user moderation systems, such as Reddit's, integrate unpaid platform users to perform these "Organizing" tasks. There is a diversity of influences from these systems. Seering et al. (2020) perform interviews of volunteer Moderators from Reddit, Twitch and Facebook Groups. A conceptual map of the "social roles" (p. 6) volunteer Moderators assumed within these communities is then constructed. The coded social roles include "Nurturing and Supporting, Overseeing and Facilitating, Governing and Regulating, Managing, and Fighting for Communities" (p. 10).

Twenty-two metaphors functioning as sub-categories to these social roles are identified by Seering et al. (2020). While several Reddit Moderators expressed the "Nurturing and Supporting" social role including metaphors such as "Teacher", and "Gardener" (p. 10), Reddit Moderators formed a plurality of volunteer users assuming the "Governing and Regulating" (p. 14) social role. This role includes metaphors such as "Judge", who have authority on "whether an action warrants punishment" (p. 14) as well as "Governor", a Moderator who "leads with a general sense of consent from the 'governed', though typically not through a democratic mandate" (p. 15). Seering et al. (2020) note that these differences in social roles may be explained by the "politically charged aspects of Reddit's culture and

history" (p. 16). As displayed, volunteer Reddit Moderators take on "fluid and changing" (p. 16) social roles.

As such, these users appear to have varying influences across platforms. Cook et al. (2021) perform a cross-platform analysis of differences in user perceptions of commercial and volunteer moderation approaches. A survey was administered to Reddit, Facebook, Twitch, YouTube, and Twitter users. Survey responses to questions on social media use and perceptions of implemented moderation solutions display that compared to volunteer moderation platforms, users felt that commercial content platforms should take more responsibility in moderation (p. 1). Cook et al. (2021) find that users of volunteer moderation-based platforms such as Reddit "seem generally satisfied with how much responsibility both Moderators and the company have in terms of content moderation" (p. 4). Meanwhile, these users expressed demand for platforms to assume more responsibility in moderation, to "reduce the responsibility put on the audience" (p. 4). Despite these demands, it is unclear in Cook et al.'s. (2021) findings what changes users believe should be implemented or why they developed these perceptions.

It is identified that users of volunteer moderation platforms feel more confident in their knowledge of platforms rather than users of commercial platforms (p. 5). Cook et al. (2021) also discover statistical significance in users who choose to engage in toxicity management practices such as blocking, reporting and positive encouragement, and their knowledge of a platform's moderation system (p. 6). Increasing the transparency of a platform's approach to moderation may establish less hateful communities.

3.1.3 Reddit Transparency Report

Reddit describes their moderation system as a "layered, community-driven approach" which gives users the "ability to vote" on content, "establish community-specific norms" and "share some responsibility" in the function of the platform (Reddit, 2020c, para. 3). The

platform maintains a Content Policy; a "set of principles-based rules that apply to all users and content on Reddit" (2020c, para. 4). Reddit administrators enforce this policy while Subreddit Moderators are asked to "apply the Content Policy to their communities in addition to their own specific rules" (2020c, para. 5).

Reddit's 2020 Transparency Report (2020c) contains an update to the Content Policy and offers a quantitative breakdown of administrator content removals across the platform. Arguing that this update demonstrates a "direct stance against hate and racism" (2020c, para. 6), Reddit provides their renewed stance on hate content in Rule 1 of the Content Policy:

> Remember the human. Reddit is a place for creating community and belonging, not for attacking marginalized or vulnerable groups of people. Everyone has a right to use Reddit free of harassment, bullying, and threats of violence. Communities and people that incite violence or that promote hate based on identity or vulnerability will be banned (Reddit, 2020c, para. 7)

Reddit discussed some effects of this policy, providing examples of content that should be removed under their guidelines. These include Subreddits "dedicated to mocking people with physical disabilities", posts describing "a racial minority as sub-human and inferior to the racial majority", comments which argue that the "rape of women should be acceptable" and memes that suggest "it is sickening that people of color have the right to vote" (Reddit, 2020b, para. 5). This approach to hate content has significantly changed from the one expressed by Reddit CEO Steve Huffman in 2018, who previously indicated that Reddit's "approach to governance is that communities can set appropriate standards around language for themselves" (as cited in Statt, 2018, para. 5).

Reddit suggests that the content removal breakdown provided in their 2020 Transparency Report (2020c) reflects the influence of their Content Policy update on the moderation actions taken by Reddit administration. A pie diagram (para. 10) visualizes that

Reddit administrators remove only 2% of all posted content while Moderators account for 4%, revealing that approximately 96% of all user posted content remains on the platform. While spam and copyright removals accounted for 99.76% of administrator removals, 55 942 pieces of "hateful content" were removed. It is unclear in the transparency report how Reddit administrators determined if content was hateful and how these items were identified at scale across its numerous communities.

Internal Reddit administrators, not Moderators, maintain the ability to ban Subreddits. Administrators banned 82,858 Subreddit communities for content policy violations, with the majority (77%) of action taken for simply being unmoderated. Excluding these unmoderated spaces, the most prominent reason Reddit administrators removed a community was for "hateful content", accounting for 6915 Subreddits, or 8% of all removals (para. 13). Meanwhile, Subreddit Moderators who evaded these bans by creating new communities comprised 7% of removals, followed by 5% of removals due to "Harassment" (para. 15).

Reddit administrators also engage in the removal of some posts, removing 65,359 (0.017% of total posts) for violating the updated content policy. "Minor sexualization" (52%) is the most prevalent reason for removal followed by "involuntary pornography" (16%) and next "hateful content" (11%, or 7048 posts). More specifically, the Transparency Report displays that image posts comprised most removals (52%) undertaken by Reddit administrators (para. 16). Despite the Report's goal to "provide transparency about content that was removed" (Reddit, 2020c, para. 2), the methodology employed by Reddit administrators in classifying images remains unknown. It is thus unclear how the platform operationalized hateful content and which tools were used to discover these occurrences at scale. Reddit administrators also removed 112,452, or 0.045% of all comments. Most comments were removed for the indicated reason of "hateful content" (43%), followed by "harassment" (22%), and "violent content" (21%) (para. 18). Again, the Transparency Report

does not expand on how these concepts are operationalized.

Moreover, “hateful content” was the primary reason for Reddit administrators to permanently ban a user, accounting for 67,620 account sanctions. This was followed by 21,946 account sanctions for “minor sexualization” and 17,268 for “harassment” (para. 14). This demonstrates that Reddit administrators remain engaged in banning accounts despite accounting for a significantly small number of total removals. The 2020 Transparency Report indicates that Reddit Moderators are far more engaged in this process, removing 131,132,344 pieces of content, compared to administrators removing a mere 202,749 (para. 20). While this information is contained in the Transparency Report, the reasons for Moderator removals are not indicated as Moderator actions are not required to be in response to violations of the platform’s content policy.

3.1.4 Reddit Norms

Lacking a similar report on Moderator content actions, the study of community norms provides insight of Moderator engagement in constructing these communities. Fiesler et al. (2018) argue that online communities are influenced both by “rules instituted by a platform” as well as informal “norms that form through behaviour” (p. 72). The influence of Reddit’s content policy is limited, with Moderators shaping community norms through the rules they construct and content they remove from their respective Subreddits. In an analysis of the rules of 100,000 Subreddits, only 164 communities made explicit reference to the Reddit Content Policy (Fiesler et al., 2018, p. 74). These findings suggest it is “difficult to infer directionality when it comes to the relationship between Subreddit rules and site-wide policies” (p. 80).

The formalization of norms through the establishment of community rules is argued to be “context dependent” (p. 80), emerging from the topic of the community rather than the guidance of Reddit administrators. For example, political Subreddits were more likely to have

rules concerning hate speech (67% of Subreddits), than Art based communities which focused on formalizing norms related to copyright materials. Given the variety of discovered norms, Fiesler et al. (2018) argue that Moderators of individual Subreddits have "a great deal of influence over the culture of the small communities they lead" (p. 80).

Chandrasekharan et al. (2018) argue that "norms are central to how online communities are governed" (p. 1) and discuss three types of norms guiding Reddit moderation. Gathering 2.8M comments removed by Moderators of Reddit's top 100 Subreddits, the authors constructed clusters of similar comments using a Latent Dirichlet Allocation topic modelling algorithm (p. 13). The authors identify "Macro", "Meso", and "Miso" norms across various communities emerging from the removal decisions made by Moderators (p. 3).

"Macro" norms are "universal to most parts of Reddit" and are enforced by "the Moderators on most Subreddits" (p. 18). Macro norm violations include "personal attacks, misogyny, and hate speech in the form of racism and homophobia" (p. 32:3). Meso norms "are shared across certain large groups of Subreddits" (p. 2). Notably, user content which mocks religion and expresses hostility towards immigrants (p. 3) are identified at the "Meso" scale only. This is an important discovery given that these occurrences are likely to violate Reddit's Content Policy towards hate speech, but removing this content is not an established site wide norm. Micro norms are enforced in "individual, relatively unique Subreddits" whose Moderators remove comments which "violate highly specific norms" (p. 3). Echoing the discoveries of Fiesler et al. (2018), Micro norms are "context-dependent" and are "distinctive to the particular Subreddits they emerge from" (p. 21). Violations of Subreddit micro norms include using "high school level science" to respond to questions on an academic focused community such as r/AskScience or promoting "diet advice" within health support communities (p. 20).

While the authors identify that Macro, Meso and Micro norms may be used to "derive normative guidelines" (p. 3) on which to base moderation practices, they acknowledge that some norms are "problematic" and are difficult to challenge within a Subreddit community. Specifically, criticizing Subreddit Moderators is identified as a Macro norm violation, despite hate-related norm violations such as mocking immigrants being enforced only at the Meso scale. Thus, Reddit Moderators maintain considerable influence in the setting of community specific norms while maintaining control of content which may criticize these decisions. Therefore, it is important to explore how community specific norms are enforced by Reddit Moderators.

Gilbert (2020) demonstrates how Reddit Moderators enforce community norms both visibly and invisibly. They analyze the tasks performed by the Moderators of r/AskHistorians, a Subreddit which provides "users with academic-level answers to questions about history" (p. 1). Gilbert collects data through both observation of the community's discussion as well as semi-structured interviews with both Moderator and non-Moderator users (p. 7), finding that the Subreddit functions as a "public history site" (p. 7) by establishing "rules and norms that differ from those of the wider platform" (p. 24).

Visible moderation is observed both in the removal of comments and the comments Subreddit Moderators may leave to explain why this decision was made. Gilbert notes that by enforcing the norms of the r/AskHistorians community, several of the threads are left with a small fraction of the total comments users contributed to a post. The author identifies that this often establishes a "feedback loop" (p. 4). Users who have a comment removed often seek an explanation from Moderators contributing additional comments, which likely violate the rules and norms established by Moderators.

Offering explanations for these removals is another form of visible work undertaken by r/AskHistorians Moderators, who often "sticky" (p. 9) comments that explain their

removal decisions to the top of a thread. Despite these attempts to justify Moderators' decisions using established community rules and norms, Gilbert identifies they are often interpreted as "censorship" (p. 1) by Reddit users. This is thought to create challenges for r/AskHistorians Moderators, who aim to encourage "participation by marginalized populations" (p. 3) while enforcing strict rules and norms which differ widely from other Subreddits.

Conversely, invisible moderation is observed by Gilbert in both the labour that some Moderators undertake to construct a response to historical questions posted on the Subreddit (p. 17) as well as general contributions to "maintaining community health" (p. 18). The author recounts an interview with a Moderator who spent considerable time constructing a response to a likely contentious question related to an image depicting sexual violence in wartime (p. 17). While the Moderator user provided an academically cited response to the question, they claimed to be "confronted with upright hostility because of the content of [their] answers or because of moderation" (p. 18) decisions they made. Thus, the r/AskHistorians Moderators engage in numerous invisible tasks to assure these user responses do not harm the norms of the community. Moderators respond to inquiries through direct messages to community users, offer personal feedback to users to improve future submissions, as well as engage in discussions with other Moderators regarding decisions on content and the direction of the community (p. 18).

As discussed, moderation work on Reddit occurs both visibly and invisibly to observing users and external researchers. While Gilbert's extensive ethnographic work showcases a moderation community committed to reinforcing community-specific guidelines, the influence of visible and invisible moderation affordances is an interesting aspect of the regulation of hateful communities. Despite this, there appears to be a gap in the literature which establishes a critical interpretation of the visible actions of Reddit

Moderators.

3.2 Reddit Hate

3.2.1 r/The_Donald

Specifically, previous work demonstrates how the hateful content norms of Toxic Technocultures emerge through primarily invisible Moderator actions. Gaudette et al. (2020) explore how Reddit's r/The_Donald users constructed a "collective identity" which "spread anti-Muslim content" (p. 15). Their research collected the top 1000 upvoted posts from r/The_Donald in 2017, comparing them to a random sample of 1000 user comments. Using descriptive coding techniques, the authors discover that 11.6% of the most upvoted comments are hateful towards Muslims, while these comments comprised merely 1.6% of the random sample (p. 9). This is a notable discovery given that r/The_Donald's Moderators created a Subreddit rule indicating that "racism and Anti-Semitism will not be tolerated" (p. 6) within the community. Gaudette et al. note that despite these rules, the Moderators specify that "Muslim and illegal immigrant are not races" (p. 6), arguing that r/The_Donald effectively "condoned anti-Muslim content" (p. 15) within the community. Despite comments such as "NOT ALL MUSLIMS yeah, but damn near most of them it seems" (p. 10) promote hate, these comments were not removed by r/The_Donald Moderators. Instead, they were upvoted by r/The_Donald users, achieving higher visibility on comment threads, constructing an "othering" discourse which frames Muslims as "perceived enemies" (p. 15). Therefore, Gaudette et al. demonstrate that Reddit "facilitated" (p. 18) the promotion of Islamophobic sentiment by providing r/The_Donald's Moderators a platform to host and promote these user comments.

Meanwhile, Shepherd (2020) demonstrates how Reddit's voting algorithm was wielded by r/The_Donald Moderators to "amplify" (p. 12) their content to the platform's r/All page, intended to be a "collection of content from all of the other Subreddits" (p. 4).

The_Donald Moderators and users used the voting algorithm to have their content "disproportionately" (p. 3) displayed on r/All numerous times in the lead up to the 2016 election. For example, on October 28th, 2016, 100% of posts displayed on the r/All page were posts from r/The_Donald (p. 6). Shepherd indicates it is "extremely rare'' (p. 6) for this to occur, and that r/All is usually composed of Subreddits from a variety of popular Subreddits much larger than r/The_Donald, such as r/Politics and r/Pics.

Although Shepherd is uncertain how this was achieved, the author indicates that it is likely r/The_Donald Moderators used a combination of "stickied" Moderator posts and votes from the community's users to amplify this content to r/All. It is noted that Reddit made several failed attempts to modify the r/All algorithm, finally deciding to explicitly exclude multiple posts from r/The_Donald from appearing on the page (p. 7). Shepherd suggests that Reddit users browsing the general r/All page "would be much more likely to see content from r/The_Donald" (p. 7). This work reveals that Reddit's sorting algorithms are vulnerable to the spread of hateful content through the actions of a small group of dedicated Moderator users. Considering that r/The_Donald "used the interface's affordances to get around the unstated norms of the community" (p. 6), Reddit's moderation system appears vulnerable to hateful communities.

3.2.2 Hate Memes

In addition to posts and comments, Singer et al. (2014) discover that image content make up a substantial portion of contributions to the Reddit platform (p. 5). The analysis of user contributed image memes displays some of the hateful content norms maintained in some Reddit communities. Zannettou et al. (2018) define memes as "variants of a particular image, video, cliche, etc. that share a common theme and are disseminated by a large number of users" (p. 2). Performing the "first attempt to provide a multi-platform measurement of the meme ecosystem" (p. 17), perceptual hashing is used to create clusters of visually similar

memes from a dataset of 40 million images gathered from Reddit between July 1, 2016, and July 1, 2017. The authors discover that r/The_Donald users shared the most memes overall and the most memes containing hate content relative to the other Reddit communities in their study, noting that Reddit has an "extremely lax moderation" (p. 5).

Specifically, image memes classified as "Racism-Related" comprised 9.3% of gathered images from r/The_Donald (p. 8). These include the anti-Semitic "Happy Merchant" meme as well as multiple variations of the "Pepe the Frog" meme, which was designated a hate symbol by the Anti-Defamation League in 2016 (p. 9). While these results may be anticipated from a community such as r/The_Donald, it is notable that r/AdviceAnimals, an image meme community to which all Redditor users were subscribed by default, also contained Racist memes. Zannettou et al. (2018) suggest this highlights the "infiltration in otherwise non hateful communities" (p. 11) of hateful image memes. While the authors' work provides crucial insight into the presence of hateful memes on the Reddit platform, they identify that future research can incorporate "OCR techniques to capture associated text-based features that memes usually contain" (p. 17).

Expanding on this research, Finkelstein et al. (2020) apply the same image processing pipeline developed in Zannettou et al. (2018), analyzing the spread of anti-Semitic memes such as "The Happy Merchant" across online platforms. While fringe web communities such as 4chan and Gab both created and shared more "Happy Merchant" memes overall, the authors discovered that the r/The Donald Subreddit "is the most efficient actor" (p. 10) in spreading this anti-Semitic meme to other web communities. These findings demonstrate how 4chan boards such as /pol, that have been studied for its anti-Semitic and racist content, act as a "primary reservoir to incubate and transmit antisemitism to downstream Web communities" (p. 10) such as Reddit.

Finkelstein et al. (2020) also discover that "Racial and ethnic slurs are increasing in

popularity on fringe Web communities" (p. 10), such as 4chan's /pol. As the authors argue there is a "unidirectional meme flow" (p. 10) from these fringe communities to Reddit, mainstream platforms are exposed to the memes produced by communities with different content policies. The threat of this content is worsened when considering that "ethnic and antisemitic terms on Web communities is substantially influenced by real-world events" (p. 10), such as the 2016 Charlottesville Rally organized by white supremacist groups. Evidently, platform moderation systems remain exposed to developments occurring both outside their web communities as well as the violent and hateful actions of actors outside the boundaries of cyberspace. Finkelstein et al. conclude that the "scale, speed, and network effects" of online platforms have been "co-opted by actors that have harnessed it in worrying ways" (p. 11). Analyzing the memes shared by these users is thus important in the development of moderation systems.

3.3 Automated Moderation

3.3.1 Background

As discussed, automated moderation solutions are emerging as mechanisms in a platform's approach to content and community governance. Automated Classification Algorithms have been developed to classify content which may infringe upon platform content policies. Gorwa et al. (2020) define automated moderation "as systems that classify user-generated content based on either matching or prediction, leading to a decision and governance outcome" (p. 3). Social media platforms use these solutions "when scale problems make manual curation or intervention unfeasible" (p. 3). Gillespie (2020) challenges this conceptualization of "Scale", arguing that promoting the necessity of automated solutions is a "discursive justification for putting certain specific articulations into place" (p. 2). These articulations include the hiring of additional human employees to train automated algorithms, a practice Gillespie identifies as representing a "growth at all costs

imperative" (p. 2), which drives the decisions of platforms. The author proposes "Size" as a more efficient conceptualization of the issue facing platforms, noting that the development of automated moderation solutions "allow platforms to grow further", rather than providing solutions for the immense "Size" of content being shared by users (p. 2). While Gillespie suggests that "size can be changed" (p. 4), it must "enhance the intelligence of the Moderators, and the . . . communities" rather than supplement perceptions that the scale of platforms is too small to address these issues. Despite Gorwa et al. (2020) and Gillespie (2018) offering differing arguments on their justifications for automated solutions, it is evident that automated solutions are developed to filter substantial amounts of platform content as platform companies continue to expand their affordances.

Automated solutions "identify, match, predict, or classify some piece of content" (Gorwa et al., 2020, p. 1) which may contain any, or several of these features. They have been developed to classify copyright infringing, terrorism, and toxic speech content (p.1). While Gillespie (2018) notes that automated solutions claim to classify "problematic content" such as "pornography, harassment, and hate speech" (p. 77). In this pursuit, automated solutions aim to operationalize these features through the development of various algorithms drawing from a "range of techniques from statistics and computer science" (Gorwa et al., 2020, p. 3). This process requires platform developers to decide how to quantify and classify these features.

Gorwa et al. (2020) discuss how the feature of hate speech has been conceptualized by researchers through the umbrella term of "toxic speech" (p. 9). Many related content features such as "profanity", "personal attacks", "sleights", "defamatory claims", and "bullying" (p. 9) are also conceptualized by platforms under this general term of "toxicity". Given that these terms represent a variety of features, the term toxicity is being employed by platforms to compile content whose removal seeks less pressure from large media

corporations and government regulators such as the case with copyright content and terroristic content respectively. Therefore, there is a need to develop an understanding of the function of automated solutions which operationalize hate speech as toxicity.

3.3.2 Matching

Gorwa et al. (2020) describe two approaches to automated moderation, Matching and Classification (p. 3). Firstly, Matching systems require a database of pre-existing content to which newly uploaded content can be referenced to classify a match. These solutions often use a "hashing" system (p. 4) which reduces an image or text to a string which is then compared to a list of hashes representing identified prohibited content hashing is used as it is both "easy to compute" as well as "typically smaller in size than the underlying content" (p. 4). Despite these advantages, it is argued that hashing techniques "are not very useful for content moderation" (p. 4) as they remain vulnerable to slight modifications to user content.

Popular hash-matching algorithms include YouTube's ContentID system (Gorwa et al., p. 4) which detects copyright infringement within video content as well as Microsoft's PhotoDNA (p. 8), which matches user images to a database of child pornography. In the case of ContentID, copyright holders maintain a hash database of the content they wish to protect, which YouTube's algorithm automatically cross references with the extracted hashes of user videos. Gorwa et al. suggest that ContentID succumbs to the issue of "systematic overblocking" (p. 8), which sees non copyright infringing content being identified. This issue is further discussed by Solomon (2015) who argue that ContentID "deprives [users] of their rights under copyright law" (p. 255) by automatically flagging videos for copyright holders despite fair-use policies existing across various legal jurisdictions. ContentID's "systematic overblocking" often leads to YouTube creators losing the monetization of their content to the identified copyright holders, while the appeal process against these copyright holders is difficult to dispute (p. 255).

Meanwhile Farid (2018) discusses their work in collaboration with Microsoft to develop the PhotoDNA algorithm. It is argued that this hashing matching solution is highly effective in the classification of Child Pornography (CP) images as identifying unique content would require an algorithm which can extract the age of an individual as well as determine if it is sexually explicit. Farid argues that such an algorithm was "not feasible" at the time PhotoDNA was developed and the requirements for such an advanced algorithm would exceed the computational and network affordances of platforms (p. 594). Instead, PhotoDNA's matching algorithm has been implemented since 2008 using a database maintained by the National Center for Missing and Exploited Children. The algorithm is used by major online platforms including Google, Microsoft, Facebook, Twitter (Farid, 2018, p. 596) and Reddit (Robertson, 2021, para. 9) with all submitted images being reviewed by the system. Farid argues that PhotoDNA is a successful initiative, having identified over 10, 000,000 CP images "without any disputed take-downs" (p. 596).

Both ContentID and PhotoDNA, demonstrate the application of matching algorithms to classify features of content. While ContentID demonstrates some concerns, successes of an initiative like PhotoDNA suggest benefits in applying automated moderation to moderate content at scale.

3.3.3 Classification

Alternatively, an automated classification system "assesses newly uploaded content that has no corresponding previous version in a database" (Gorwa, et al., 2020, p. 4). These algorithms are developed through machine learning, which often requires an annotated training dataset containing examples of the desired classification features. Defined as ''the ability to learn without being explicitly programmed'' (Samuel, 1959, in Parekh & Patel, 2017, p. 964), machine learning creates a predictive model which analyzes an input and provides an output response. Gillespie notes that machine learning algorithms are being used

for the “identification and categorization of offensive social media content” (2018, p. 98). Widely used machine learning classification algorithms include Facebook’s Toxic Speech Classifiers, which address both hate speech and bullying, as well as Twitter’s Quality Filter, which classifies content for Spam and Harassment (Gorwa et al., 2020, p. 7).

The process of generalizing the features of user content presents concerns. Arguing that machine learning algorithms are “famously poor at making . . . difficult context-dependent judgement”, Gorwa et al. (2020) suggest they may lead to “overblocking” and potentially “curb important forms of expression” (p. 9). Specifically, to the classification of language, the authors note that identifying features within text “is incredibly complicated, personal and context dependent” (p. 10). As such, it is suggested that classification solutions may make it so the content of radicalized individuals who may use identity terms innocuously “more likely to be removed” (p. 11). Similarly, Gillespie (2018) identifies that “the lack of context, the evasive tactics of users, and the fluid nature of offense” (p. 98) makes it difficult to classify specific features such as hate within social media content. The author argues that these difficulties lead to classification solutions producing too many “False Positives” (p. 98) and “False Negatives” (p. 99) to become reliable replacements for human moderation. The author questions the lack of consensus in the detection rate of classification solutions, asking, “Is 73 percent good? Ninety-four percent? Ninety- nine?” (p. 104). It is thus important to explore the functions of emerging automated moderation solutions and develop further understanding of their classifying outputs at scale.

3.4 Automated Hate Moderation

3.4.1 Toxicity

As discussed, automated moderation has been applied to classify the contentious content of communities resembling Toxic Technocultures. Gorwa et al. (2020) identify that these solutions previously relied on keyword Blacklists curated by community Moderators.

While this approach prevents specific slurs from appearing within user content, it seems to struggle “as norms develop” missing “the wide range of contextual clues” (p. 5) present in user content. As such, machine learning is being applied to perform the “operationalization of a concept” (p. 5) rather than block specific words. This process involves making “generalisations about features of many examples from a given category into which unknown examples may be classified” (p. 5). Approaches to machine learning text analysis include “bag of word” solutions, which ignore semantic structure of text, as well as word embedding algorithms “which represent the position of a word in relation to all the other words that usually appear around it” (p. 5). While the technologies and applications of machine learning are numerous, they primarily establish “generalisations about features of many examples from a given category into which unknown examples may be classified” (p. 5).

Launched in 2017, Google’s Perspective API is an open-source algorithm which uses Natural Language Processing (NLP) to assign feature scores to online comments determining “the perceived impact (it) might have on a conversation” (Perspective, 2021a). NLP is a machine learning based “computerized approach to analyzing text that is based on both a set of theories and a set of technologies” (Liddy, 2001, p. 1). More specifically, Google Perspective’s NLP algorithm is a Convolutional Neural Network (CNN) which uses GloVe word embeddings to analyze text (Ribeiro et al., 2020, p. 9). CNNs are formed by “multiple layers of artificial neural networks” which are used to “recognize social gestures in an end-to-end architecture” (Albawi et al., 2018, p. 2). Meanwhile, GloVe word embeddings, short for Global Vectors for Word Representation, generate “word vectors fully unsupervised” relying on “word co-occurrence statistics” (Tifrea et al., 2018, p. 1), constructing a predictive model. While the complexities of GloVe are outside the scope of this research, Pennington et al. (2014) discovered it improves over previous word embedding models such as word2vec, noting that GloVe “outperforms other models” (p. 1541) in recognizing text features.

Implementing many NLP, CNN and GloVe embedding models requires specific technical knowledge, while Perspective's open-source API provides simplified access to the solution through the Google Cloud console (Perspective, 2021a).

Google claims that Perspective is a solution which "Enables healthy conversations", "Reduces toxicity and abusive behavior" and is accessible as it is "Free", "Self-Serve" and "Customizable" (2021a). Upon submission of a comment to the API, a score between .00 and 1.0 is returned, indicating the "likelihood that someone will perceive the text" (Perspective, 2021a) as contributing to one of the following attributes: Toxicity, Severe Toxicity, Insult, Profanity, Identity Attack, Threat, and Sexually Explicit. Specifically, Google developed Perspective's toxicity model to interpret if a comment is "rude, disrespectful, or unreasonable comment that [is] likely to make people leave a discussion" (Perspective, 2021a). Google recommends using Perspective as a tool for commercial content platform moderation, allowing them to "quickly prioritize and review" (Perspective, 2021a) the most toxic comments posted to their platform. It is also recommended that these toxicity scores can be used to display "feedback" to both readers and commenters, offering them an automated assessment of the content they are reading or contributing.

As discussed, machine learning algorithms require training data on which to build their predictive algorithm. Perspective was trained using text data collected from Wikipedia Talk page submissions as well as The New York Times online comment section. Comments were provided to an unspecified number of human coders, who performed a crowdsourced labelling of toxic content (Fortuna et al., 2020, p. 6789). While the complete training data used to train Perspective is not available to the public, Google offers public training datasets on their website, encouraging the development of third-party tools which integrate the tool.

Google Perspective is a widely used comment moderation platform, processing over 500 million comment requests daily as of February 2021 (Jigsaw, 2021). The applications of

the tool by some of these platforms demonstrates its functionality and customizability. The New York Times integrated Google Perspective in a human controlled moderation dashboard named “Moderator” (NYT Team, 2020). As user comments are submitted, Perspective assigns toxicity scores which are displayed in the Moderator tool. Next, a group of 15 human Moderators perform a review of each comment, with the Moderator dashboard indicating associated toxicity scores. From this dashboard, human Moderators can both reject and approve comments below or above a set toxicity threshold. The Times claims that Perspective has allowed them to “foster high-quality conversations around [their] journalism” (NYT Team, 2020, para. 9) in a time where many media platforms such as the CBC are testing turning off comments to limit hate content (Felon, 2021).

Spanish newspaper El País (2019) faced this decision as they encountered increased toxic comments on their platform. with shutting down their comments section, El País were contacted by Google to develop a comment moderation solution using the Google Perspective API. While Perspective was only available in English at the time, El País worked with Google in constructing Spanish training datasets using the comments from their platform. Following the integration of a human controlled toxicity solution like the NYT Moderator tool, El País identify they have observed a 7% reduction in comment toxicity while growing the number of comments on their platform by 19% (2019). El País indicate that they plan to use Perspective to observe long term toxicity trends for specific users and making editorial decisions on which articles are published based on potential toxicity. In both the examples of the NYT and El País, Perspective offers a flexible dynamic solution for platforms to classify hate speech in comments at scale.

3.4.2 Google Perspective in Research

Google Perspective’s toxicity model has also emerged as a tool in research, offering authors a tool which operationalizes hate in the study online content. The findings of these

studies reveal both its utility in studying Toxic Technocultures relative to other hate speech classifiers, as well as establish relevant findings on toxic content on Reddit.

Toxic trends appear to be observable in Community Subreddits. Mittos et al. (2020) apply the toxicity model to score over 1.3 million comments gathered from both Reddit and 4chan in an analysis of genetic testing discourse. They selected Google Perspective's toxicity model due to its suitability to examine lengthy comments (p. 455). It is argued that alternative methods of identifying toxicity, such as "hate speech libraries" (p. 455), are not suitable for Reddit content as both the sample size and character length of their respective training data are not sufficient for a large-scale comment analysis.

Following a manual annotation of Subreddits gathered in their dataset, these communities were grouped into 19 distinct categories including "Health", "News", "Religion", "Politics" and "Hate" (p. 454). Following a toxicity analysis of the comments in each identified category, the authors discover the "Hate" category as the most toxic, closely followed by "Politics". Specifically, to their research of Genetic Testing discourse, the authors identify that "Hate" Subreddits such as r/Altright and r/Kotakuinaction discuss this concept in "a highly toxic manner, often suggesting its use to marginalize or even eliminate minorities" (p. 453). Additionally, through a clustering of visually similar images the authors discovered that toxic comments are "genetic testing topics [and] are often accompanied by images and memes with clear racial or hateful connotations" (p. 460). Mittos et al.'s research demonstrates the utility of Google Perspective in processing large datasets and identifying toxicity in communities in which they anticipate observing hateful content.

Perspective was used to trace cross platform user toxicity by Zannettou et al. (2020), scoring a dataset of 125 million news article comments collected from 412K news articles (p. 1). The toxicity model was selected by the authors following a comparison of the tool with Hatesonar, a hate speech detection tool which uses logistic regression to classify comments

"as hateful, offensive, or neither" (p. 3). Following a human annotation of a sample of gathered comments, they discover that Google Perspective's toxicity models performed "substantially better" (p. 4) in limiting both false positive and false negative results. The authors explore how linking news articles to 6 chosen Subreddits impacts their respective user comments, shedding "light on what elements attract hateful comments on news articles" (p. 1). The authors discovered a correlation between posting an external news article on one of 6 selected Subreddits (r/AskReddit, r/politics, r/conspiracy, r/The_Donald, r/News, and r/Worldnews) and an increase of toxic comments on these external sites (p. 2). Specifically, the authors observe higher toxicity in comments on right-wing news sites, the most toxic being Infowars.com and the least toxic being the New York Times (p. 5). It is unclear in Zannettou et al.'s work if The New York Times had implemented Google Perspective during their data gathering process, its comment toxicity scores suggest active content moderation.

Additionally, the authors identified "significant increases in hateful commenting activity" (p. 1) following real world events such as the "Unite the Right" rally in 2016, organized by members of the alt-right (Blout & Burkart, 2021). These findings remain opaque in the authors' work given that they do not specify how they determine a comment to be hateful based on toxicity scores. Nonetheless, Zannettou et al. demonstrate the utility of Google Perspective to examine large-scale comment datasets and identify cross-platform toxic user behaviour.

As Zannettou et al.'s (2020) research focuses on comments on external news articles, Mall et al. (2020) employ Google Perspective, identifying 4 Reddit user toxicity trends. They argue that previous work "is focused on the development of machine learning classifiers" (p. 1) rather than identifying long term patterns in user toxicity. To construct this longitudinal analysis, the authors collected 10 submissions containing the most comments from Reddit's 10 most popular Subreddits between 2009 and 2017 (p. 3). Gathering over 4 million

comments using Python's Reddit API Wrapper (PRAW), each comment is scored with Google Perspective's toxicity model. Following this, a toxicity score threshold of 0.5 was applied, determining which comments in their dataset were considered hateful (p. 3). A manual review of the toxic comments by a human coder provided an agreement score of 86.7% (p. 4). The authors indicate that this agreement score "would ideally be higher" (p. 4) but proceed in their operationalizing of Google Perspective's toxicity scores given persistent subjectivity in accurately classifying content for hate speech.

Mall et al. classify 31.2% of Reddit user as "Fickle-Minded", meaning that they "switch between toxic and non-toxic commenting" (p. 1). Mall et al. interpret that these findings suggest users' toxicity scores likely change across both communities and the subject matter of the post. Their work does not explore these findings; however, they offer the following potential research question: "are some users more drawn to controversial topics than other users?" (p. 8).

The other user types follow as such: "Pacified" (25.8%), "Radicalized" (25.4%) and "Steady" (17.6%) (p. 7). While "Pacified" and "Steady" users demonstrate diminished or consistently low toxicity scores respectively, the large segment of "Radicalized" found in the dataset reveals how many Reddit users increase their toxicity scores over time. The authors note that given large-scale user toxicity patterns are quantifiable through machine learning algorithms such as Google Perspective, they can be useful in "identifying users who are steadily toxic or are becoming radicalized" (p. 8) rather than focusing on individual comments or discussions. It is further argued that through these solutions, platforms can act in "preserving the user friendliness of online communities by identifying continuously toxic users" (p. 3). While Mall et al.'s research showcases Perspective's functionality in identifying user toxicity patterns, their work does not explore the toxicity of Moderator contributed content. There is thus an opportunity to explore user toxicity patterns while acknowledging

the importance of Moderators within digital platforms.

3.4.3 Google Perspective Limitations

Previously identified limitations of Google Perspective demonstrate how aspects of Technological Redlining may emerge in its use. While the above applications of Google Perspective display its functionalities to classify content at scale, it has been suggested the toxicity algorithm demonstrates bias. In their discussion of the effects of biased algorithms, Reichert et al. (2020) write that "mistakenly flagging non-toxic content, especially content produced by marginalized voices, is an important social issue" (p. 1). Their work demonstrates that Perspective's toxicity model assigns higher toxicity scores to comments containing identity terms, constructing a bias against marginalized groups who may use identity terms in non-toxic discourse. The authors score 1.8 million user comments from the defunct platform Civil Comments using Google Perspective, applying a threshold of 0.5 toxicity to determine a toxic comment (p. 2). They discover that while 14.9% of comments referencing identity are scored as toxic, only 6.8% of comments containing no identity terms are classified as toxic (p. 3). It is suggested that non-toxic comments containing identity terms create false positive results due to Perspective's bias towards these terms.

To test this, the authors construct their own toxicity classification model from these comments using logistic regression, Neural Network, and Long Short-Term Memory models. Following a process of rebalancing their dataset to account for identity terms, they discover their model likely returns less false positives than Google Perspective. This is demonstrated on a dataset of Tweets authored by United States' politicians of which a manual review by the authors classified as non-toxic. Their classification model identifies 6.9% of tweets as toxic within this dataset while Perspective scored 10.3% as toxic, suggesting that Perspective's accuracy is threatened by false positive outputs (p. 7).

Despite the authors' efforts to account for bias through rebalancing the weight of

identity terms within their model, they identify that the "majority of tweets" (p. 8) misclassified by their model contain identity terms, while not containing observable toxicity. A concerning example of this being a tweet containing the following text and identified as toxic by Google Perspective: "When bias drives discipline, Black girls miss out on the chance to learn" (p. 7). Although Perspective is outperformed by the authors' classification models, their novel methodology revealed the same biases demonstrated by Google's increasingly used moderation algorithm.

Jiang and Vosoughi's (2020) audit of Google Perspective provides further examples of bias in its toxicity algorithm. The authors first annotated a random sample of 100 Twitter users with demographic information on "Age", "Race" and "Gender" (p. 3). Aiming "to find differences between performance across demographic categories" (p. 6), the authors did not discover statistical significance in toxicity scores. Rather, they identified that with a toxicity threshold of 0.5, "the overwhelming majority" of tweets "were short pieces of text input that contained profanity" (p. 5). This is notable given that tweets containing "slurs" and "attacks on identity", which the authors suggest would be hate speech in a legal context, received toxicity scores ranging from 0.3 to 0.5 (p. 6). These results are demonstrated by examples of tweets provided by the authors. The likely innocuous tweet, "Stay the fuck at home! Hahah" (p. 6), was assigned a toxicity score of 0.95 while a longer text calling immigrants both "illegals" and "INVADERS" was scored as 0.45 by Perspective (p. 6).

Jiang and Vosoughi (2020) argue these results indicate Perspective demonstrates bias towards marginalized groups, suggesting that profanity acts as "proxies for identity" (p. 6) for both race and age demographics. It is indicated that the "arbitrary conflation" of users using profanity both in toxic and non-toxic content is "what drives systemic disparate impact" (p. 6). Despite these concerns, it is noted that "Perspective API does not perform badly as much as it is unfocused" (p. 6), suggesting that future research must focus on "prioritizing the

transparency and accountability of algorithms, models, and systems" (p. 7) classifying online content.

Furthermore, research has explored how Perspective's Toxicity model is vulnerable to user attacks. Jain et al.'s (2018) research construct "adversarial attacks", which show how "malicious users may try to bypass toxicity detection by adding minor alterations to their original comment" (p. 1). It is argued that adversarial attacks impose a significant threat to text-based algorithms given that "Text is discrete, not continuous", that "small changes to text can significantly impact meaning" and that "language is dynamic" (p. 3). These hypotheses are tested through generating adversarial attacks on Google Perspective, interpreting how toxicity scores can be manipulated with slight changes to the text.

This dataset contained a total of 1,025,403 comments and 19,850 posts from Facebook and was scored by Google Perspective (p. 3). Adversarial attacks were generated using a "Carlini-Wagner attack" (p. 2) model for comments which scored above a 0.75 toxicity score. While this form of adversarial attack substitutes selected words within text, it maintains the readability of the original comment. Comments were considered misclassified when they are scored below 0.5 toxicity (p. 4). Jain et al. identify that an adversarial attack modification of 3 words in a comment evaded a toxic classification 25% of the time (p. 1). Following the adversarial modification, Google Perspective remains vulnerable to users looking to manipulate its toxicity scores. Jain et al. discover that Perspective "can be deceived by replacing words in toxic sentences while still preserving the original meaning" (p. 5).

Brown et al. (2019) expand on these findings using the same dataset gathered by Jain et al. (2018), providing specific examples of the generated adversarial attacks. To do so, the authors generate "Semantic attacks", "Acoustic attacks" and "Visual Attacks". "Semantic Attacks" replace a word with a similar one, the provided example being modifying "Hate" to

the term “Loathe” (p. 3). The authors discovered that these modifications resulted in a 33% reduction in the accuracy of Perspective’s toxicity model (p. 1). Meanwhile, “Acoustic attacks” modify “an individual character within words” (p. 3), such as changing “hate” to “hayte” maintaining its phonetic legibility. Acoustic Attacks result in a 72.5% reduction in Perspective’s performance (p. 5). As well, “Visual Attacks” replace letters in words with “visually similar characters”; the provided example being modifying “Hate” to “H@te”. Visual adversarial attacks resulted in a reduction of 30.4% accuracy in Perspective (p. 4). Brown et al. demonstrate Google Perspective’s susceptibility to multiple forms of adversarial attacks.

Despite the risks of bias and adversarial attacks, this chapter has demonstrated Google Perspective’s emergence as a popular tool in both content moderation and academic research. It is important to develop a critical understanding of this tool, exploring its applications in research while maintaining perception of Redlining effects when applied at scale.

3.3.4 Hate Meme Detection

There are also emerging techniques in research exploring the classification of hateful image content through operationalizing visual and textual features. For instance, Suryawanshi et al. (2020) identify that memes which are a “combination of text and image are difficult to regulate by automatic filtering” (p. 32). They argue that this is due to the ability of both hateful and non-hateful text and image content to “obscure” (p. 33) the meaning of a meme. A meme classification pipeline using Logistic Regression and Naive Bayes classifiers for text, and a Convolutional Neural Network for image analysis is constructed. Suryawanshi et al. determine that identification of hateful memes is improved “when both text and image modality associated with the meme was considered” (p. 38). A pre-existing dataset of 745 images whose text was manually annotated is used. Suryawanshi et al. suggest the

exploration of Optical Character Recognition (OCR) techniques to extract embedded text from hateful memes (p. 4) to explore classification techniques.

Sabat et al. (2019) construct a similar image classification methodology, noting that the combination of hate and text features of memes "may result in hate speech messages" (p. 1). The OCR software Tesseract is used to extract embedded text from the meme dataset, a BERT Natural Language Processing model is then applied to vectorize this data. Image content is encoded using a Convolutional Neural Network, later merging this extracted visual and textual data to "feed the multimodal features into a classifier" (p. 2). "Hate memes" (p. 2) are collected through Google searches while "non hate memes" are gathered from a pre-existing dataset of Reddit memes. The constructed textual and visual classification method achieves an accuracy score of 0.83 in classifying the gathered hate memes and non-hate memes.

It is argued that while these findings suggest "it is possible to automatize the task" of identifying hate memes, the potential of false positive and false negative classifications "would still require a human Moderator for many of them" (p. 4). The importance of some human involvement is made more explicit by the authors' suggestion that there is likely to be a "misuse of the system" (p. 4) as some users can manipulate content based on the classification they are likely to receive. Sabat et al. conclude that the implementation of meme classification solutions "should evaluate whether the computation cost of running the OCR and encoding the extracted text is worthy based on the reported gains in accuracy" (p. 4).

Similarly, Du et al. (2020) identify "image-with-text" memes (IWT) as "mechanism[s] through which ideologically potent and hateful content spreads" (p. 153). Arguing that IWT memes are an "alternative way of sharing text", the authors construct a classification pipeline using Tesseract to extract text features and a Residual Neural Network

for visual features from a training dataset of 7 million Twitter images. They identify that only 1% of memes in their dataset do not contain text (p. 154), suggesting that IWT memes comprise a sizable portion of image content shared on social media. A manual annotation observes that "hateful content [was] not often contained in IWT memes" (p. 159). Despite this, it is noted that consulting an IWT meme's image to classify it as hate content throughout this manual annotation process was rare. Based on this experience, Du et al. argue that "text-based hate speech classifiers may be useful in identifying hateful content in IWT memes" (p. 160).

Du et al. (2020) also explore the users who share IWT memes through matching the observed Twitter users to voter registration demographic information. They determine that self-identified Republicans, African Americans, and women share more IWT memes than other observed demographics. It is suggested this may lead to "issues with representation" (p. 162) as the discourse of marginalized groups is operationalized and filtered in the process of IWT classification. Additionally, the authors note the lack of precision achieved by the employed Tesseract OCR engine. While Du et al. avoided its use due to its "prohibitively expensive" (p. 158) cost, they discovered that Google Vision API's text extraction functionality provided far higher accuracy compared to Tesseract in identifying IWT memes (p. 156). Considering, these demographic considerations, exploring misclassifications in image-with-text classification is important to the discussion of automated content moderation.

3.5 Research Argument

As discussed, hateful content norms emerge on online platforms as toxic content evades a variety of moderation mechanisms. Recalling Gorwa's governance models, platforms appear to maintain significant self-governing influence in the moderation of this content. In Reddit's case, the Moderators of Subreddit communities such as r/The_Donald

appear to influence these norms through visible and invisible actions. Thus, Reddit's integration of a Co-governance volunteer moderation system has provided any user of the platform the affordances to construct a community with hateful content norms.

Meanwhile, External governance regulation such as the Canadian Online Harms Legislation focus on regulating hateful content rather than toxic communities or the platform affordances used in their constructing. The proactive approach described in the COHL's technical proposal mandates the use of automated systems to classify this content. As identified in the Literature Review, these automated algorithms such as Google Perspective have been used in research to operationalize hate features, identifying user (Mall et al., 2020) and community specific (Mittos et al., 2020) toxicity trends. However, the COHL calls for large-scaled proactive automated filtering of content across platforms. The goal of this research is to analyze how hateful community norms on a platform are quantified by automated classification systems at scale. Applying Google Perspective's Toxicity model, an opportunity emerges to perform a quantitative analysis of visible elements of Co-governance on the Reddit platform within the context of the COHL's mandate.

Recalling Massanari's (2017) theory of Toxic Technocultures, hateful Subreddit communities are constructed by Moderator users who manipulate the Reddit platform and maintain hateful community content norms. However, there is a gap in quantitative research exploring the content contributions of these Moderator users. This is despite the research by Mall et al. (2020) on Reddit user toxicity trends. Massanari suggests that "it is often the most powerful Moderators/users who post objectionable content in the first place" (as cited in Zimdars & Mcleod, 2020, p. 148). As reviewed, both Reddit posts and comments were discovered to contribute to hateful community norms. As discussed by Gilbert (2020), these content types are also the most visible of actions Moderators contribute to a Subreddit. Therefore, applying Massanari's assumptions that the Moderators of a Toxic Technoculture

contribute to toxic content norms, the following hypotheses are constructed:

H1 Toxic Technoculture Moderators post higher toxicity posts than non-Moderator users

H2 Toxic Technoculture Moderators post higher toxicity comments than non-Moderator users

The contributions of image memes to hateful community content norms were discussed. Also, methods to classify this hateful meme content through the operationalizing of textual features of IWT memes were explored. In result, the following hypothesis is constructed to analyze Reddit image content:

H3 Toxic Technoculture Moderators post higher toxicity image texts than non-Moderator users

Testing these hypotheses aims to display how the actions of Co-governance actors are quantified by an automated moderation algorithm like Google Perspective. Should the observed Moderators display higher toxicity across content types, the COHL may be effective in classifying the content of users who construct hateful communities. If the hypotheses are rejected, an opportunity emerges to discuss the limitations of automated approaches to classifying the content of platforms where hateful communities are maintained at the specific community level. Performing this analysis thus observes various human and non-human actors which comprise a Subreddit's heterogeneous Actor-Network.

As is argued by Noble (2018), the marginalizing effects of Technological Redlining emerge within misclassifications of platform content. Recalling the complex classification pipelines constructed for image-with-text analysis, applying a critical qualitative lens to the outputs of image-with-text analysis appears important in interpreting potential Redlining effects of algorithms such as Google Perspective. As such, a critical a case study of the automated classification of multi-modal platform content appears important. Thereby, the following research question was identified:

RQ1 How is hate signified in hateful image memes containing non-toxic image text?

Responding to RQ1 establishes a discussion of how a false negative misclassification may occur in image-with-text meme analysis. As such, this analysis explores potential misclassifications which may emerge as automated algorithms interpret the meaning of online content through operationalizing various features within an Actor Network. This is relevant as the COHL considers mandating the use of automated moderation algorithms. Overall, these research questions and hypotheses construct a critical analysis which interprets how platform governance models establish Toxic Technocultures while exploring potential Redlining effects of emerging in misclassifications of this content.

Chapter 4: Methodology

4.1 Research Approach

This study involves a mixed-methods analysis of a Subreddit community, r/Metacanada, aiming to observe Moderator contributions to hateful Reddit communities through the lens of an automated moderation algorithm. The Reddit archiving APIs Pushshift and PRAW are used to gather the Subreddit's post, comment, and image content. Expanding off the discussed findings of image-with-text meme analysis, the OCR functionalities of Google Vision are then applied, gathering a quantitative text variable from the collected images.

To test H1, H2, and H3, the hate is operationalized in Reddit content by applying Google Perspective's toxicity model to Post, Comment and Image Meme text content. The toxicity score means of the content posted by r/Metacanada Moderators are then compared to user contributions using Mann-Whitney U Test for each content type.

To respond to RQ1, a qualitative analysis of image memes containing low toxicity score text is performed using a multi-modal approach constructed by Yoon (2016). Applying Yoon's analytical grid to two r/Metacanada memes develops a critical analysis of the elements which construct hateful meanings in image-with-text memes.

The rest of this chapter is presented as follows: First, the data collection process is described, discussing which APIs were used to gather a comprehensive Reddit dataset. This identifies the collected variables gathered from the r/Metacanada community's content. Then, the approach to collecting image meme text using Google Vision is discussed.

Next, the data analysis section describes the process of scoring the gathered dataset with Google Perspective's Toxicity model. Yoon's qualitative multi-Modal meme approach is then presented, identifying how the elements constructing the meaning r/Metacanada's memes were analyzed. Lastly, limitations to this analysis are discussed and ethical concerns

emerging from the research approach are considered.

4.2 Data Collection

4.2.1 Community

This section identifies the observed Subreddit community, presenting the findings of related past work leading to the performed analysis. Aiming to capture Moderator contributions to toxic Reddit communities, the r/Metacanada Subreddit (R/Metacanada, 2021) was selected. Milton (2018) suggests the actions of r/Metacanada Moderators in the community's discourse amplify toxic discourse on the Reddit platform. r/Metacanada also provides an opportunity to study a Toxic Technoculture whose content "is accessible to persons in Canada" (Heritage Canada, 2021, para. 2) as described in the COHL proposal.

Gruzd et al. (2020) focused their study on the r/Metacanada Subreddit's content, exploring the community's "anti social behaviour" (p. 8). This selection of r/Metacanada was made following the observation of a "high level of toxic and nationalistic content'' (p. 5) encountered when manually reviewing the content of various Subreddits for their analysis. Using Google Perspective to score 22,560 r/Metacanada posts and comments submitted in the 2 weeks leading up to the 2019 Canadian election, the authors indicate they "expected a larger portion of posts to be toxic, but only a small fraction of them really are" (p. 12).

Gruzd et al. (2020) also explore the establishment of "cut-off values" to Perspective scores to explore what portion of the observed content may in fact be toxic. A threshold score of 0.7 identifies 15% of r/Metacanada posts as toxic, while this decreases to 10% when it is increased to 0.8, and lastly to 5.3% at 0.9 (p. 12). Following this analysis, the authors select a cut-off threshold of 0.8 toxicity in their application of Google Perspective (p. 12). While it is indicated that adjusting these cut off values establishes a balance between potential false positives and false negative classifications from Perspective, the effects of this decision are not considered.

A subsequent social network analysis of the r/Metacanada content determines that "some users tend to be the primary spreaders of anti-social acts" (p. 15). While the Moderator status of these users are not analyzed, Gruzd et al. suggest future research should establish a "content-driven approach" to analyze "the anti-social acts of these key users within the network in more detail" (p. 16). This research expands off this recommendation, observing the toxicity of the contributions of the Moderators of the r/Metacanada community.

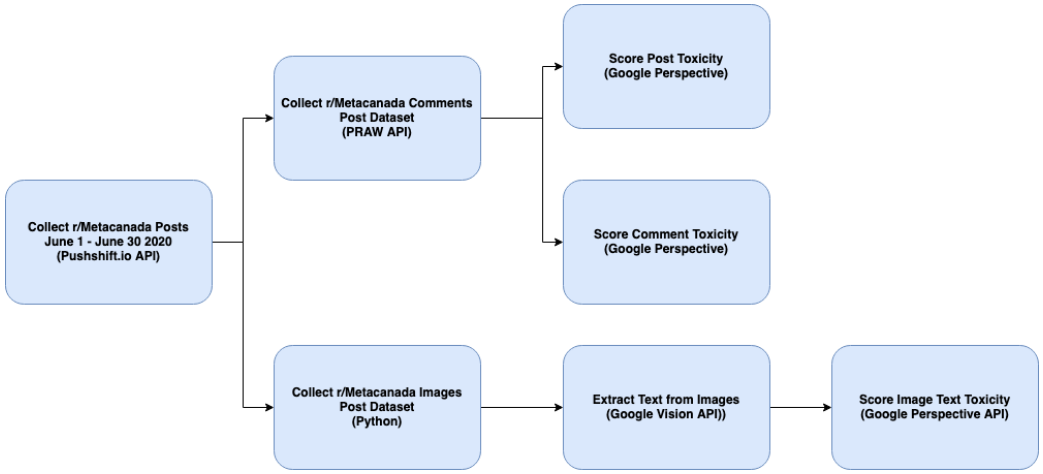
4.2.2 Data Gathering

Selecting a sample of content from the r/Metacanada community several time frames were considered. This includes the 2019 Canadian election and longer periods ranging from 3 to 6 months. Over time, the applied selective sampling process was shortened and influenced by Finkelstein et al.'s (2020) discovery that "the use of ethnic and antisemitic terms . . . is substantially influenced by real-world events" (p. 10). Bolsover (2020) discovered that online discourse surrounding the Black Lives Matter movement has demonstrated "worrying levels of polarisation, hate, incivility and conspiracy content" (p. 1). As such, the sampling period of June 1-30th, 2020, was established as it comprised the first full month of data concurring with this event (McCoy, 2020). It is believed that this 30-day period is sufficient in responding to the exploratory nature of the analysis and respects the scope of this research.

Collecting a comprehensive dataset of Reddit content over this period required accessing several APIs which were accessed through Python scripts. While Gruzd et al.'s Communalytic tool offered some relevant functionalities, developing custom scripts was more efficient and flexible as research progressed. As such, five Python scripts were developed to collect Reddit Posts, Comments, Images, extract text from these images and then score this content with Google Perspective's toxicity model. A link to a GitHub repository containing the developed Python scripts is provided in the Reference materials. Figure 1 provides a visualization of the data collection and analysis pipeline.

Figure 1

Data Collection Workflow



To begin, a Python script calling the Pushshift API was run on July 15, 2020, to gather all posts submissions to r/Metacanada from June 1st to June 30, 2020.

Pushshift archives Reddit content using Google's Big Query platform and Reddit's own API. A systematic overview of the methods used to study Reddit performed by Proferes et al. (2021) indicates that Pushshift is the 2nd most used data collection tool for the platform behind Reddit's own API (p. 3). Initial attempt to use Reddit's API proved difficult due to limited documentation being available and requiring more advanced Python scripts.

A total of 3210 posts from the r/Metacanada community submitted over the sample period were collected. Pushshift data includes the following metadata from each post, allowing for the identification of which user submitted a post, and links to comment replies as well as images.

- Post Text
- Post Author
- Comment ID URL
- Image Link URL

Next, the comment data was collected using a Python script calling the Python Reddit API Wrapper (PRAW). PRAW collects post comments and related metadata including

author, time posted and a unique comment id. Pushshift offered similar functionalities, however rate limits required applying a different tool. PRAW is used in a comparable manner by Zomick et al. (2019) and Buntain and Golbeck (2014) to gather large-scale Reddit comment datasets. This process was performed for all posts collected by Pushshift. This resulted in a dataset containing 14 401 comment texts posted by 2996 unique users.

4.2.3 Data Cleaning

A process of cleaning the data was undertaken to remove contributions with incomplete metadata. A total of 154 posts did not contain the author's username and were thus removed from the dataset. While it is unclear why these posts did not contain an author name, a [Deleted] tag indicates the users may have deleted their Reddit account.

Next, the comment dataset was reviewed for missing authors, user deleted comments and those removed by r/Metacanada Moderators as is indicated by a [removed] tag in place of text. In total, 658 comments deleted by users themselves, 1672 with missing authors and 207 which were removed by r/Metacanada Moderators. These 207 Moderator removed comments comprise 0.014% of all collected comments.

Separate post and comment datasets were then constructed for r/Metacanada Moderators and regular users. This establishes the categorical independent variable for the analysis. A total of 8 Moderator users were indicated in the Subreddit's sidebar during the collection period. Meanwhile, the post dataset identified a total of 692 unique users, while comments were posted by 2056 users.

4.2.4 Image Collection

Gathering the subsequent image meme dataset required an additional Python script to download images from links contained in the Post data. All posts in the cleaned dataset which contained a link URL ending in the image formats .PNG and .JPG were collected. The Python script allows images to be downloaded from some third-party hosting platforms

including imgur.com. In total, 31 Moderator and 628 user images were collected.

Following this, a process of identifying image-with-text memes within the image dataset was developed. While attempts to automate this process were taken, none met the requirements of this dataset. A manual classification of the images was thus performed guided by Du et al.'s (2020) research of image-with-Text memes. Images were kept if "the text displayed is vital to understanding the image" (p. 154). This approach filtered out images which were purely visual constructions despite potentially falling under the general definition of a meme. Images were kept if the "structure or content of the image could reasonably be imitated, altered, and re-shared" (p. 154). In total, 14 Moderator and 151 user images were identified as image-with-text-memes, removing 17 and 477 images from the respective datasets.

This produced an image-with-text meme dataset through which an OCR solution can extract text content from the images. Google Vision (2021) is a Computer Vision algorithm which provides these OCR functionalities through an automated text extraction model. Google Vision API provides an efficient solution, as it rapidly extracts embedded text from image memes while requiring limited technical comprehension of more complex Computer Vision techniques.

Gomez et al. (2020) took a similar approach, using Vision to extract text from multimodal memes. While their research approach constructed a complex classification pipeline, this analysis makes use of Google Vision's OCR functionalities to gather text which can then be scored by Google Perspective. This is a novel approach which has not been identified in past work despite it leveraging popular research tools.

A Python script called the Vision API, which extracted the text from the image-with-text memes. The text collected in this step was not further processed as this process aimed to mimic the automated classification of image-with-text memes at scale. This may influence

the findings of the analysis as Google Perspective's toxicity model is used to score the gathered text. However, the text gathered by Vision was reviewed for accuracy throughout the development of the Python script.

The described data collection process gathered a dataset which is reasonably representative of the r/Metacanada community's content across the multiple forms of platform content. Table 1 demonstrates the frequencies of collected posts and comments across r/Metacanada Moderator and regular users in the cleaned dataset.

Table 1

Frequencies of collected r/Metacanada Posts, Comments, and Images

Content Type	Moderators	Users
Posts	202	2488
Comments	481	11301
Images	14	151

4.3 Data Analysis

4.3.1 Google Perspective

A Python script was then used to call the Perspective API, assigning toxicity scores to post, comment, and image texts. Toxicity scores are assigned from 0.00 to 1.00, being the most toxic, offering a continuous dependent variable for the analysis. As Google Perspective offers several analysis models including Toxicity, Severe Toxicity, Insult, Profanity, Identity Attack, Sexually Explicit and Flirtation Threat, a model had to be chosen. Gruzd et al. (2020) discovered that the toxicity, severe toxicity, insult, and profanity models are highly correlated (p. 12). Thus, as observed in Perspective's research applications in the literature, the API's Toxicity model was selected. The Severe Toxicity model was also considered but its classification of "a very hateful, aggressive, disrespectful comment" (Perspective, 2021b) risks more false negative outputs. Thus, Perspective's Toxicity model offers predictive score

to determine if a comment is "rude, disrespectful, or unreasonable . . . that is likely to make people leave a discussion" (Perspective, 2021b).

Alternatives to Google Perspective such as Hatebase (2021) were also considered. Hatebase provides an average offensiveness score based on word embeddings matched to a growing list of slurs and other offensive terms. However, exploratory use of the API was difficult as documentation on its use was scarce.

4.3.2 Mann-Whitney U Test

A quantitative research method which compares differences in data between two groups was required to test the hypotheses. T-tests test for a "significant difference between two sample means" (Somekh & Lewin, 2005, p. 228). Applying a t-test determines "whether or not there is likely to be a real difference between the two groups" (Somekh & Lewin, 2005, p. 228). They are therefore an efficient and flexible method to analyzing differences in toxicity between r/Metacanada Moderators and users.

The Student's t-test is the most commonly used and requires data to be Parametric, or rather, form a normal distribution (Nachar, 2008, p. 13). To test the normality of the collected data, a Shapiro-Wilk Test of Normality (Mendes & Akin, 2003, p. 13) was applied to the Post text dataset. The output is presented in Appendix A. This returned p values of <0.001 for both Moderator and user posts is less than the 0.5 p value threshold indicated by recommended to reject the null hypothesis (Salkind & Rasmussen, 2007, p. 884). It is thus assumed that the collected r/Metacanada dataset likely does not follow a normal distribution and is non-Parametric.

As such, a non-Parametric t-test was required. The Mann-Whitney U Test was selected as it does not require a normalized distribution of data while offering outputs "concerning the difference between . . . groups" (Nachar, 2008, p. 13). The null hypothesis of Mann-Whitney U Test "stipulates that the two independent groups are homogeneous and

have the same distribution" (Nachar, 2008, p. 14). As such, the test is sufficient to respond to the hypotheses, while accounting for the non-parametric dataset.

JASP was again used to perform the Mann-Whitney U Tests for each hypothesis, thus the toxicity levels between Moderator and user post, comment and image texts were compared separately. The Mann-Whitney U Test requires an independent variable which is categorical; the status of a user being a Moderator or not was used. This was coded in the dataset through assigning a 0 (Moderator) or 1 (user) to each toxicity score for all content. The dependent variable is thus Google Perspective's Toxicity scores, a similar approach to what is applied by Guimarães et al. (2020) in their Mann-Whitney U Test analysis of the differences in comment replies across social media post types.

JASP offers a function to test a selected alternative hypothesis through the Mann-Whitney U Test. This was applied by selecting the Group 1> Group 2 option from within the software interface. The output of this analysis provides a Probability score p through which it can be accessed if the null hypothesis can be rejected. Guided by the work of Guimarães et al. (2020), a p value of >0.05 was selected as the measure of statistical significance for this analysis.

4.3.3 Critical Multi-Modal Meme Analysis

Following this, selecting which memes would best respond to RQ1 How is hate signified in image memes containing non-toxic image text? was required. To begin this process, all memes containing text assigned toxicity scores above 0.5 were removed from the dataset. This follows the guidance of Mall et al. (2020, p. 3) who identified that a manual coding process achieved an 86.7% agreement score when the 0.5 threshold was applied as a binary of whether text was toxic or not. Next, it was decided that selecting 1 image meme from both the Moderator and user images provided a sufficient response to the qualitative analysis. After all, RQ1 remains exploratory, aiming to extract general observations on how

the multi-modal features of image memes construct hateful messages. This focused approach is similar to Yoon's (2016) work which analyzes only 1 image meme.

A non-random purposive sampling approach was then applied to this low-toxicity text dataset to select image memes which appear to construct a hateful message. It is recognized that this purposive approach is influenced by the normative perceptions of hate of the researcher.

To reduce the impacts of personal biases, this process was guided was Reddit's restated approach to hate content focusing on contributions which promote "hate based on identity or vulnerability" (Reddit, 2020b, para. 1). Selected from the Moderator image dataset was a meme which constructs a comparison between Islamophobic perceptions of Muslims and police brutality was selected from the Moderator dataset. The score assigned by Perspective to the text of this image is 0.22, thus falling under Mall et al.'s 0.5 toxicity score threshold. While from the user dataset, a meme constructed using a screenshot from the 1975 film Jaws along with the mugshot image of a Black male known as "Wide Neck" which mocks the events of the murder of George Floyd was selected. The text from this meme was scored as 0.18 by Perspective's toxicity model, much lower than the identified threshold.

To analyze these selected images, Yoon's (2016) critical multi-modal approach to interpret image-with-text memes was applied. This method constructs a multi layered Visual Analysis grid through which the elements constructing image-with-text memes can be qualitatively analyzed. The first layer, denotative, considers "what, or who is being depicted" (Van Leeuwen, 2001, as cited in Yoon, 2016, p. 126) in a meme. This includes both the signifier text and visual modalities of an image-with-text meme as well as the signified meaning of these elements. While the second layer, connotation, considers "what ideas and values are expressed through what is represented, and through the way it is represented" (Van Leeuwen, 2001, as cited in Yoon, 2016, p. 126). Interpreting this second layer remains

influenced by the perceptions of hate norms of the researcher as mentioned above. However, the descriptive approach applied to both memes features provides readers flexibility in interpreting the hateful state of the analyzed images.

Using Yoon's visual analysis grid, the denotative signifier elements were identified through observing the visual elements contained in the meme. This includes the low-toxicity image text and all visual elements forming the meme. Next, the denotative signified elements were identified. This process aimed to construct a "mental representation" of the memes as identified by the author (Barthes 1967, as cited in Yoon, 2016, p. 107). Following this, connotative descriptions for both memes were interpreted to determine what is being represented in the meme. Yoon identifies this process requires a focus on "Intertextuality" by identifying how the meme's message was constructed "through the connection to other meanings carried by other images" (p. 107). As such, a qualitative analytical approach to describe how hate memes are constructed both through text and visual elements was established. Overall, this allowed this research to expand on Yoon's work through identify how both "hidden and explicit messages" (p. 107) construct hateful memes.

4.4 Limitations

This research used several tools to gather and analyze the r/Metacanada Subreddit dataset. Significant efforts were made to reduce the limitations of these methodological components, but remaining gaps in this approach influence the relevance of its findings. First, it analyzes the content of one group of Moderators on one online community hosted by one social media platform. Evidently, this limits the generalizability of its findings regarding the demonstrated complexity of identifying and addressing hateful communities and their content. Past work, including Mittos et al. (2020), make use of random Reddit datasets to establish toxicity baselines for communities. While this could be an avenue for future research, analyzing one Subreddit constructs a focused discussion of hateful communities

within emerging automated moderation mandates. This research thus aims to fill in some gaps in the understanding of toxic Moderator contributions to Reddit and expand on Gruzd et al.'s (2020) research of the same community using Google Perspective.

Next, while past works such as Mall et al. (2020) performed a manual review of Google Perspective's scores to observe its performance, this research has not undertaken this step of qualitatively reviewing its outputs. When considering the risks of false positives and false negatives of automated moderation identified in the Literature Review, these potential misclassifications are undoubtedly present in the gathered dataset. However, Mall et al. identified an agreement score of 86.7% (p. 4) between Google Perspective's toxicity classifications and the manual review performed by a human coder. The authors accept this score, indicating that toxicity is "subjective and thus hard to agree upon" (p. 4). Given this finding it is therefore assumed that Google Perspective is providing some level of accuracy in the classification of r/Metacanada content.

Further, it is acknowledged that this analysis establishes a novel approach to the classification of image meme text content. While this has provided minimal guidance on how to best use Google Perspective's toxicity model within text-based approaches to image meme analysis, it applies two research tools used in the classification of online content. As well, research by Sabat et al. (2019) and Du et al. (2020) performed the construction of extensive multi-layered classification pipelines whose development requires extensive technical knowledge. For these reasons, it is believed that leveraging the Google Perspective tool to score meme text provides an approach which is similar to past work and remains grounded in the evolving nature of platform content and affordances.

4.5 Ethics

Ethical considerations were made regarding the study of communities who share hate content, of online platforms and human involvement in research. Massanari (2018) identifies

risks in the study of far-right online communities, indicating that movements such as #Gamergate have targeted scholars with hate and harassment (p. 3). A possible solution for limiting this risk is to "anonymize all aspects of . . . research" (p. 5). As well, Massanari suggests that these risks are "contextually, culturally, and temporally dependent" (p. 6). Given the limited scope in which this analysis will be published and archived, as well as the relatively small scale of a community like r/Metacanada, this is not an ethical concern.

Proferes et al. (2021) provide further guidance on the ethics of performing research on Reddit. They indicate that "ethics bodies would be likely to require consent for surveys, interviews, or the use of data from closed communities" (p. 10). This analysis has not required any of these methodological approaches, does not interact with the r/Metacanada community or users and gathered online content which remains publicly available using popular collection tools.

This research thus respects the guidelines described by the Tri Council Policy Statement's Ethical Conduct for Research Involving Humans. As both Pushshift and PRAW collect publicly available Reddit content, the data collection process respects the Tri Council's "Public domain with no expectation of privacy'' (Canadian Institutes of Health Research, 2018, p. 15) policy. Analyzing this dataset for overarching community toxicity trends while not targeting specific users assured that no ethical approval was required to perform this research.

Chapter 5: Analysis

5.1 Data Summary

This initial summary section provides descriptive data of the r/Metacanada dataset, identifying the number of unique users contributing to each content type and their respective toxicity score means. Gruzd et al.'s (2020) 0.8 toxicity score threshold is then applied to assess the level of observed toxicity for post, comment and image texts.

For posts, the dataset gathered contributions from 3 of 8 r/Metacanada Moderators. As for users, there are 691 unique contributors. While it is possible that the other 5 r/Metacanada Moderators deleted their posts prior to the collection of the data, this may indicate that not all the Moderator users consistently contribute content to the community. As for users, the 691 unique contributors represent only 0.019% of the 35,000 users subscribed to r/Metacanada. This follows Gruzd et al.'s (2020) findings which identified that "an active group of users in the core of the network" (p. 13) contribute to the r/Metacanada community. Similar observations appear in the count of unique users who commented, identifying 4 of 8 r/Metacanada Moderators who started a thread. Meanwhile, 2056 non-Moderator users, or 5.87% of subscribers, contributed text comments to the collected posts

Next, Table 2 displays the Means and Standard Deviations of toxicity scores respective to each gathered content type. As can be observed, posts scored the least toxic (0.25), followed by comments (0.37) and most toxic were image meme texts (0.39).

Table 2

R/Metacanada Content Toxicity

	Posts	**Comments**	**Images**
Count	3050	11782	165
Mean	0.25	0.37	0.39
Std. Deviation	0.22	0.30	0.26

The distributions of these scores for each content type are presented in Histograms in

Figures 2, 3 and 4 below. As can be observed, the toxicity scores across content types appear to be positively skewed. There is a slight increase in comments and image texts scoring from 0.8-0.9 however most scores are clustered below 0.2 across content types. This indicates that toxicity scores are overall low. The skew of these distributions is further analyzed for content types alongside each hypothesis.

Figure 2

Toxicity Scores: Posts (N=3050)

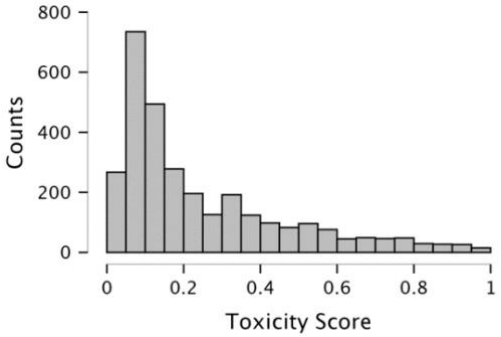


Figure 3

Toxicity Scores: Comments (N=11782)

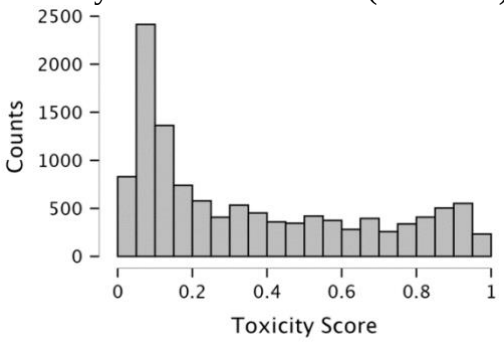
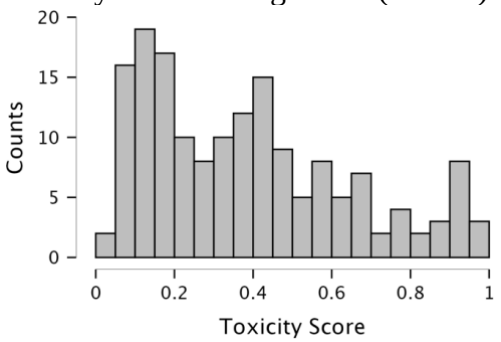


Figure 4

Toxicity Scores: Image-Text (N=165)



Following the observation of these distributions, Gruzd et al.'s toxicity threshold was

applied to analyze the long high-toxicity tails which were identified. Table 3 displays the percentage of posts in the dataset which score above the authors' 0.8 threshold.

Table 3

Percentage of Toxic Content (0.8 Threshold applied)

Content Type	Percent (>0.8)
Posts	3.7 %
Comments	14.8 %
Image Meme Text	9.7 %

As displayed, there are infrequent contributions to r/Metacanada which exceed Gruzd et al.'s 0.8 toxicity threshold. These frequencies do vary by content type, with comments containing the highest number of presumed "toxic" posts as scored by Perspective. Based on these differences and absent of a qualitative analysis of each text variable, it is difficult to determine at what threshold analyzed content is in fact hateful. To further analyze the content of the r/Metacanada community, the differences in observed toxicity scores between r/Metacanada Moderators and users are tested below.

5.2 Hypothesis 1: Post Toxicity

This section of the analysis demonstrates the result of the Mann-Whitney U Test for Moderator and user post toxicity, testing the following hypothesis:

H1: Toxic Technoculture Moderators post higher toxicity posts than non-Moderator users

The descriptive statistics for the Post dataset are displayed in Table 4, analyzing one month of r/Canada posts with Google Perspective's toxicity model. The Mean frequency is highest, then the Median followed by the Mode for both Moderators and users. This confirms the positive skew in the distribution of Post toxicity scores which was suggested by observing the histogram. The Median indicates that 50% of the data received less toxic scores than 0.15 for Moderators and 0.16 for users.

Table 4

Descriptive Statistics (Posts)

	Toxicity Score	
	Moderator	**User**
Count	202	2848
Mode	0.08	0.08
Median	0.15	0.16
Mean	0.24	0.25
Std. Deviation	0.22	0.22
Minimum	0.01	0.00
Maximum	0.99	0.99

Comparatively, the Mean of user post toxicity (0.25) is slightly higher than the Moderator post toxicity score Mean (0.24). The identical Standard Deviations scores of 0.22 for Moderators and users indicate that there is similar variance in the spread of toxicity scores for Post texts. As well, the Minimum and Maximum values demonstrate that both Moderators and user posts exhibit toxicity across similar ranges. These frequencies comprise nearly the entire range of Google Perspective's scores, with both groups sharing posts which range from nearly 0.0 toxicity to 0.99 toxicity.

Next, to test the statistical significance of the difference in toxicity, the Mann-Whitney U Test was applied. The results of this test for post toxicity are provided in Appendix B. As discussed, the identified alternative hypothesis is that Moderator posts are more toxic than non-Moderator users' posts. However, the p value of 0.786 is higher than the $p = 0.05$ confidence level. There was no statistically significant difference in means and the null hypothesis is not rejected.

Therefore, following the application of a Mann-Whitney U test to test if r/Metacanada Moderators post higher toxicity posts than non-Moderator users, H1 is rejected. There is no statistically significant difference in the average toxicity scores of Moderators versus non-Moderator users. Notably, this does not indicate that Moderator users do not contribute hate content, it simply means Moderators and non-Moderators do so fairly equally.

5.3 Hypothesis 2: Comment Toxicity

For the comment dataset the following hypothesis is tested, again using the Mann-Whitney U Test:

H2: Toxic Technoculture Moderators post higher toxicity comments than non-Moderator users

Table 5 contains the descriptive statistics for the comment dataset including Mean toxicity scores. Again, the positive skew of the distribution is confirmed, with a higher Median (0.24 for Moderators and 0.25 for users) than was observed for Posts.

Table 5

Descriptive Statistics (Comments)

	Toxicity Score	
	Moderator	**User**
Count	481	11301
Mode	0.15	0.06
Median	0.24	0.25
Mean	0.36	0.37
Std. Deviation	0.29	0.30
Minimum	0.01	0.000
Maximum	0.98	0.99

Similar to the Post scores, there is a small difference in the Mean toxicity scores of Moderator (0.36) and user (0.37) comments. As well, the Standard Deviations of r/Metacanada Moderator (0.29) and user (0.30) comment toxicity only differ by 0.1. Despite the larger sample size of the comment dataset relative to the post dataset, similar comparisons in the variance of Moderator and user toxicity can be observed. The similarities of these group's comment toxicity are further displayed in the Minimum and Maximum scores assigned to this content type. Again, Google Perspective assigned toxicity scores forming nearly the entire range of the model's outputs, from 0.01 to 0.98 for Moderators and 0.002 to 0.99 for users.

The results of the Comment Mann-Whitney U Test are displayed in Appendix C. A

comparative analysis of the toxicity scores assigned to 481 Moderator comments and 11, 301 user comments was performed. Again, the Mann-Whitney U Test was performed using the alternative hypothesis that r/Metacanada Moderator comments were higher in toxicity than the comments gathered from users. The displayed p value of 0.5514 is higher than the selected confidence level of p=0.05, therefore the null hypothesis, that there is no statistically significant difference between these groups, cannot be rejected. As such, H2 is rejected, indicating that r/Metacanada Moderators did not contribute comments which were statistically significantly higher in toxicity than those of non-Moderators users.

5.4 Hypothesis 3: Image-Text Toxicity

Likewise, the Mann-Whitney U Test is applied to test the following hypothesis regarding image texts:

H3: Toxic Technoculture Moderators post higher toxicity image texts than non-Moderator users

While the image-text dataset comprises a much smaller sample than the post and comment analysis, similar observations are observed in the descriptive statistics displayed in Table 6. As with post and comments, a positively skewed distribution is again observed, with the Median the largest of the 3, scoring 0.31 for Moderators and 0.36 for users.

Table 6

Descriptive Statistics (Images)

	Toxicity Score	
	Moderator	User
Count	14	151
Mode	0.15	0.06
Median	0.31	0.36
Mean	0.36	0.39
Std. Deviation	0.23	0.26
Minimum	0.09	0.04
Maximum	0.82	0.96

Meanwhile, Mean toxicity scores for the gathered image meme texts are close with 0.36 for Moderators and only slightly higher at 0.39 for users. Similar variances are observed in the distribution of image meme text, with the standard deviation of user image text only 0.02 higher on Google Perspective's toxicity scale.

When analyzing the toxicity of text of 14 Moderator and 151 user images, there is also a similar range to what was identified within Posts and Comments. User image text toxicity ranged from the Minimum of 0.03 with a Maximum toxicity score of 0.99. While the scores of Moderators produced a smaller range, from a minimum of 0.09 and a maximum of 0.82. This is likely due to the small sample size of Moderator posted images contained in the dataset and does not indicate a larger trend.

The output of the Mann-Whitney U Test comparing image text toxicity between r/Metacanada Moderators and users can be seen in Appendix D. As performed for the previous two tests, this Mann-Whitney U test integrates the alternative hypothesis that r/Metacanada Moderator users post memes containing higher toxicity texts than non-Moderator users. The output p value of 0.6272 demonstrates that this is not the case, as it higher than the confidence level of p=0.05. As such, the null hypothesis cannot be rejected as the alternative hypothesis does not achieve statistical significance following the Mann-Whitney U test. Therefore, r/Metacanada Moderators did not post memes containing statistically significantly higher toxicity image text than the community's users.

As such, following the testing of H1, H2 and H3, when observing the content of the r/Metacanada Subreddit over the collected period of June 1-30, 2020, the community's Moderators did not post higher toxicity content than its users. In fact, there was no statistical difference in the toxicity of their contributions compared to non-Moderator users. This is the case for the gathered post, comment and image text data and was tested for statistical significance using Mann-Whitney U tests. This section of the analysis applied Google

Perspective's toxicity model on the text content gathered from these content types. The next section expands on the observed r/Metacanada images, exploring challenges to automated moderation in the classification of multi-modal meme content.

5.5 RQ1: Image-With-Text Meme Analysis

This section of the analysis responds to the following exploratory research question through a critical multi-modal analysis of two image memes selected from the r/Metacanada dataset.

RQ1 How is hate signified in image memes containing non-toxic image text?

A preliminary response to this question is constructed using Yoon's (2016) analytical strategy of describing the Signifier and Signified denotative layers which construct the connotative meaning of an image meme. The two analyzed memes were selected due to the low toxicity scores assigned to the text overlaid within these images.

First, Table 7 demonstrates the results of the critical multi-modal analysis of an image-with-text meme shared by an r/Metacanada user. The image is provided in Figure 5. This image uses multiple signifier elements to construct a hateful message regarding the unjust murder of racialized individuals by law enforcement. The top image depicts a dark-skinned man smiling. No further context is available to determine the identity of this man solely using the elements present in the meme. Below this, there is an image of a white male smoking a cigarette. There is no indication present in the meme how the depictions of these two individuals are related. The text overlaid on the white male reads "We're gonna need a bigger knee". This text was assigned a score of 0.18 by Perspective's toxicity model, far below Gruzd et al.'s 0.8 threshold.

Table 7

r/Metacanada User Meme Analysis

	Denotative Signifier	**Denotative Signified**
The First Layer	An image of a dark-skinned man with a wide neck An image from a film of a light skinned man smoking a cigarette Text overlayed on the bottom image "We're gonna need a bigger knee" (0.18 Toxicity Score)	A mugshot style photo of a Black man A scene from the film Jaws of the Chief of Police character The white man needs a bigger knee because the black man's neck is wide
The Second Layer	Connotation	
	The combination of images indicates these images are related. The text located at the bottom of the meme suggests the white male is referring to the size of the dark-skinned man 's neck. The white male is suggesting that a larger knee is required to subdue the Black male	

Figure 5

r/Metacanada User Meme

Google Vision Text	**Perspective Toxicity Score**
WE'RE GONNA NEED A BIGGER KNEE	0.18

An analysis of the signified elements of this meme begins to reveal how a combination of images and text construct a hateful message using intertextuality. First, the top image is a mugshot photo of a Black male named Charles McDowell referred to as "Wide Neck" according to the meme database KnowYourMeme (KYM, 2021). The KnowYourMeme page demonstrates that following McDonnell's arrest, this photo has been

used to construct numerous memes referencing the size of the individual's neck. This includes popular posts on the r/BlackPeopleTwitter Subreddit as well as popularizing the Twitter hashtag #freewideneck calling for McDonnell's release. Next, the bottom image of the white male smoking a cigarette on a boat is of the Chief of Police character, Chief Brody, from the 1975 film Jaws. This specific scene from the film contains a quote from Chief Brody claiming that "we're gonna need a bigger boat" (Jaws, 1975) in reference to the size of the shark character in the film. While there is no KnowYourMeme page for this specific meme, it is a format available within the popular online meme generator imgflip (2021), which provides numerous examples of its use. The substation of the word "boat" from the film's quote to the word "knee" within the meme relates these images. Through the combination of visuals and text content, it is signified that Chief Brody is referring to the size of McDonnell's neck.

Deconstructing the Connotative meaning of this image-with-text meme thus required intertextual knowledge of two previous internet memes. Through their layering, along with the overlaid text, the meme argues that to subdue McDowell, a larger knee would be required due to the size of the individual's neck. As the meme was posted during the 2020 Black Lives Matter protests following the murder of George Floyd by a Minneapolis police officer kneeling on Floyd's neck, the meaning of this meme becomes clearer. This meme engages with this historical context by calling for the murder of McDowell in the same manner as George Floyd. This hateful message relies on the use of both the contained visual and textual elements, the latter of which received a low toxicity score of 0.182 from Google Perspective. Recalling that memes are constructed through signifier elements which establish a contextualized signified layer to the meme, the low toxicity text does not demonstrate the hateful message contained in this image. Recognizing that this image was posted by a r/Metacanada user and not removed by a community Moderator, nor a Reddit administrator,

demonstrates how multi-layered memes – which require various intertextual knowledge to understand its message of hate – are shared on the platform.

Next, the analysis of the image-with-text meme posted by a r/Metacanada Moderator is displayed in Table 8, while the image is provided as Figure 6. The signifier elements of this meme are a succession of text and images. First, there is the text “this doesn’t represent every Muslim…”. This text is assigned a Perspective Toxicity score of 0.348 when analyzed as an isolated piece of text. This phrase is followed by an image depicting fires in two large skyscrapers. There is no information contained in the meme which identifies what buildings these are or what is causing the fires. Next, there is the text “but somehow this represents every cop?”. This text receives an exceptionally low toxicity score of 0.057 from Perspective’s model when analyzed separately. Below this, there is an image of a police officer with their knee pressed on the neck of a dark skinned individual. The analysis of the full overlaid text of the meme “this doesn’t represent every Muslim…but somehow this represents every cop?” was assigned a toxicity score of 0.22.

Table 8

r/Metacanada Moderator Meme Analysis

	Denotative Signifier	**Denotative Signified**
The First Layer	Text at the top of the meme "this doesn't represent every Muslim..." An image of two buildings on fire Text "but somehow this represents every cop?" (0.22 Toxicity Score) An image of a police officer kneeling on a man's neck	Not all Muslims are represented by the events of September 11th An image of fire at the World Trade Center on September 11th 2001 Questioning if Police officers are represented by the events of the murder of George Floyd An image from the video depicting the 2020 murder of George Floyd
The Second Layer	**Connotation**	
	An argument is constructed through the combination of images and texts. The text above the image of the World Trade Center on fire relates the image to Muslims. The text above the image of the murder of George Floyd relates the depicted act to the representation of all police officers. These associations argue that if not all Muslims are related to the events of 9/11, then the murder of George Floyd is not representative of all Police officers	

Figure 6

r/Metacanada Moderator Meme

Google Vision Text	**Perspective Toxicity Score**
this doesn’t represent every Muslim. . . but somehow this represents every cop?	0.22

Analyzing the signified elements of this meme shared by an r/Metacanada Moderator further demonstrates how contextualized knowledge of visual and textual elements is required to deconstruct its hateful message. First, the above text referring to Muslims, likely

refers to discourse on social media platforms which emerges following some terrorist attacks. Popular hashtags such as #NotAllMuslims are used by activists to deter Islamophobic responses to these attacks as a message which "disavow[s] the terrorists and their sympathizers" (Bromwich, 2015). The image below this text depicts the World Trade Center on September 11, 2001, following two planes colliding with the buildings. This image is salient as it depicts events caused by terrorism which according to Watt (2012) "remain primary markers of the so-called Muslim world" (p. 12) within media. The text below the image expands on the previous text regarding Muslims, questioning if the second image is representative of all police officers. More context is gained from the meme when identifying that the image of a police officer kneeling on a Black male's neck is a screenshot from the video depicting the 2020 murder of George Floyd by a Minneapolis police officer.

Through the layering of these signifier elements, this meme constructs a connotative meaning which links stereotypes of Muslims as terrorists while questioning the public inquiries of systematic racism within police forces following the murder of George Floyd. This hateful message is thus twofold, drawing upon salient visualizations of unrelated tragedies to which a platform user is likely familiar due to their popularity in media. This meme constructs links between these events through the juxtaposition of images and text. It suggests that the inquiries into excessive force used by police officers towards racialized individuals are somehow related to discourse which argues against islamophobia and the perception of Muslims as terrorists. As such, this meme expresses hate towards Muslims by framing them as potential terrorists, while dismissing public inquiries into systemic racism emerging throughout the 2020 Black Lives Matter Protests. This constructs a hateful message regarding the treatment of Black Americans by police officers while furthering Islamophobic perceptions of Muslims.

To provide a preliminary response to RQ1, hate is constructed in r/Metacanada

memes through layering signifier elements of both textual and visual modalities. This analysis applied Yoon's (2016) Multi-Modal Semiotic Analysis methodology to discuss how hateful image content is constructed on Reddit. While this analysis was performed for only 2 image-with-text memes, the semiotic analysis of these images displays that hateful image content is constructed through layered intertextual elements. This analysis has thus demonstrated potential challenges in the automated classification of multi-modal content with automated solutions. The theoretical and applied implications of these findings are further discussed in the proceeding Discussion chapter.

Chapter 6: Discussion

6.1 Governing Hate

While statistically significant higher toxicity in R/Metacanada Moderator contributions was not discovered in H1, H2, and H3, the low levels of observed toxicity may indicate a challenge in the External governance moderation of hate content using automated moderation. This parallels the findings of Gruzd et al. (2020) in their analysis of the same Subreddit community. When applying the authors' 0.8 toxicity score threshold, only 7.7% of the analyzed r/Metacanada texts are classified as toxic. Similar to the authors' research, the analysis anticipated that the community's hateful content norms would be displayed when quantified through Google Perspective. Instead, the distributions for each content type skew positively, indicating low toxicity content. As this approach applied Google Perspective's toxicity model to explore Moderator contributions to hateful communities, Subreddits may contain hateful norms while offering low toxicity scores when observed at scale. Consequently, limitations to using automated moderation systems like Google Perspective with the goal of classifying hateful content are displayed.

The Subreddit communities established by Reddit's Self-governance model thus far appear to have provided flexibility for platforms to establish independent approaches to addressing hate content. This illuminates a potential conflict in the External governance of platforms such as Reddit who maintain Self-governance influence over their content moderation. In Reddit's case, this has resulted in the platform delegating Moderator responsibility to volunteer Moderators, relying on these users to filter most of the posted content. The platform made adjustments when faced with criticism for allowing hateful content, updating its hate content policy, and banning a list of communities including the r/The_Donald Subreddit. However, no changes were made to the affordances of Moderators to administer the platform's Subreddit communities. As hateful norms emerge at the

community specific “Miso” level, as discussed by Chandrasekharan et al. (2018), preventing users from wielding governance mechanisms to form hateful communities may require more substantive changes to these affordances.

Meanwhile, the Canadian Online Harm Legislation maintains a content focused approach, aiming to classify hate content through proactive automated moderation applied at scale. Considering that the r/Metacanada community is not notably “toxic’’ when observed at scale, applying an automated solution like Google Perspective may classify only a fraction of content as toxic despite hate emerging as an established norm in some Reddit communities.

Similarly, the analysis did not identify higher toxicity scores in r/Metacanada Moderator content relative to user content as tested in H1, H2 and H3. This was the case for post, comment and image text data gathered between June 1 and 30, 2020, and analyzed using Mann-Whitney U Tests. The hypotheses were exploratory, aiming to show a trend in the content of Moderator users of the Reddit platform. Moderator toxicity scores were analyzed as Reddit has maintained hateful communities through a Co-governance model which integrates contributing users as Moderators. Observing this content also demonstrated how the content of Reddit community’s may be classified by automated moderation at scale.

However, recalling Gilbert’s (2020) analysis of the r/AskHistorians community, Reddit moderation occurs both visibly and invisibly beyond content contributions. As such, the low toxicity contributions by r/Metacanada Moderators suggest limitations in analyzing Reddit communities through visible Moderator contributions. Rather, Moderator users also wield invisible affordances through which Subreddit norms appear to be established, such as content removal disposition and creating the community’s rules. Broad automated approaches to filtering hateful content at scale may not capture these invisible features that, for the Reddit platform, indicate to users that a Co-governed space accepts and encourages hate content which violates the site-wide Content Policy.

In r/Metacanada's case, minimal comment removals were observed in the analysis: 297 of 14,401 total comments were removed by Moderators. This may suggest that hateful norms emerge through a lack of active moderation rather than direct efforts from its Moderators. Regardless, Reddit's Co-governance approach to moderation seems to permit spaces which test the platform's Content Policy while sharing governance mechanisms with its users. External governance initiatives may have to adapt to these less visible factors of Co-governed spaces if they aim to effectively remove hate content.

Should the automated moderation approach described by the COHL proposal be implemented, there are some research opportunities related to platform governance which could expand off the analysis's findings. First, while overall toxicity scores in the analysis were low, a longitudinal analysis could be constructed to observe toxicity trends as proactive automated moderation is applied at scale. This analysis could be performed across numerous online communities to discover trends in toxicity levels. These could be the first steps in analyzing the influence of this regulation on online discourse and consider its effect on addressing hate content. Additionally, future analyses of Reddit's Co-governance model could establish qualitative frameworks to observe non-content contributions of its Moderators. The work of Fiesler et al. (2018) explores Subreddit rules at scale. Future research could explore similar, less invisible norm setting affordances to further understand their role in the governance of Subreddits. This may clarify how Moderators of Co-governed spaces engage with platform affordances and reveal how hate becomes an established norm.

6.2 Toxic Technocultures

The infrequent contributions and post and comment removals by r/Metacanada Moderator users also offers insight in the moderation of Reddit's Toxic Technocultures. While Massanari proposes that "it is often the most powerful Moderators/users who post objectionable content in the first place" (as cited in Zimdars & Mcleod, 2020, p. 148), this

was not observed in this analysis of r/Metacanada. In fact, Moderator contributions accounted for only 7.5% of posts, 4% of comments, and 8% of image posts. Meanwhile, the minimally observed content removals suggest little active filtering was performed by these users.

This low engagement was not anticipated as the Literature Review identified Reddit Moderators as active contributors within their community. In r/Metacanada's case, Moderators contributed sparingly via content, while appearing to allow users to participate in a low-moderated community on a platform with an evolving hate content policy. However, the collaborative influence of a small group of users create spaces where hate may become a community content norm. Infrequent Moderator contributions to the community's ongoing discourse suggest that Reddit's affordances beyond content submissions may be crucial to analyzing Toxic Technocultures.

Massanari (2017) indicates that a site's "design . . . governance structure and algorithmic logic" (p. 330) allows the emergence of Toxic Technocultures. As discussed, the Moderators of Toxic Technocultures such as r/The_Donald and seemingly r/Metacanada appear to leverage a combination of affordances to establish hateful content norms. The analysis of r/Metacanada's content captured some of these elements in Moderator contributions at scale. Operationalizing hate in this content using Google's Perspective's toxicity model did not discover higher toxicity trends in Moderator contributions. However, beyond this analysis, further exploration of Moderator contributions may advance knowledge on how the hateful norms of Toxic Technocultures are maintained. The ability of Reddit archiving APIs such as PRAW and Pushshift to produce large datasets could be leveraged to perform longitudinal analyses of multiple Subreddit communities. Then, the toxicity of Moderator post, comment, and text contributions could be compared across Subreddits to observe content trends in toxic communities. This approach could identify trends in Toxic Technoculture Subreddits while also observing how non-toxic community norms are

maintained by Moderator users.

While a qualitative analysis of an image meme posted by an r/Metacanada Moderator was performed, a more extensive analysis is required to understand how Toxic Technoculture norms are established. The work of Dignam and Rohlinger (2019) explores how the posts of Moderator and "elite users" of Reddit's r/TheRedPill community influenced the "politicization" (p. 594) of the Subreddit's misogynistic discourse. This analysis considered the visibility of the contributions through a Moderator applied "sticky" and assessed how its rhetorical arguments were interpreted by the community within its comments. A similar ethnographic qualitative approach may reveal where hateful Toxic Technoculture community norms are established. This could closely trace Moderator content, focusing on how these users engage with specific hateful topics, rather than observing them at scale as performed.

Moreover, the analysis explored Reddit's Toxic Technocultures using an automated moderation algorithm which detects "toxicity". When Perspective's toxicity model was applied to r/Metacanada's content, varying means were observed. This ranks from Posts (0.25) to Comments (0.37), to Image texts (0.39) as most toxic. Why this is the case remains unclear. Perhaps content modalities play some role in where users choose to contribute hateful ideas. It could also be that the length of text submitted to Perspective influences the API's toxicity model's outputs. Either way, this may have implications in the classification of content scored with its Toxicity model. For example, should comments skew higher in toxicity than Reddit posts, as observed by r/Metacanada, automated moderation would have varying effects based on content type. This demonstrates why the COHL's broad approach to regulating "hateful content" may be challenged by the distinctions in content of just one platform.

On a larger scale, Actor Network Theory's appreciation of how "non-human technological agents can shape and are shaped by human activity" (Massanari, p. 330)

provides a framework to explore how variations in toxicity scores may influence the functions of automated moderation. It seems that the integration of text analysis algorithms within a network introduces new non-human actors within these interactions expressed by an automated model's outputs. This analysis has demonstrated that this process has varied results. If Perspective's toxicity scores are applied to platform content at scale, understanding its outputs should thus move beyond homogenous conceptions of "content" and consider how content affordances interact within a platform. This could construct analyses which move from observing which users contribute hateful content to a platform, to focusing on how hateful norms are maintained through distinctions in posts, comments, and image memes. Recognizing and engaging in the heterogeneity of these human and non-human actors constructs more comprehensive community focused frameworks than the COHL's content classification approach to regulating hate content.

Further analysis of non-human actors could explore different content modalities using Perspective's toxicity models and perhaps identify distinctions between posts and comments from different platforms to observe how Perspective's scores may influence these respective networks. Social network analysis techniques could be applied to observe how toxicity spreads across comment reply chains. Topic Modelling could also be applied to quantify the discussions of a Subreddit to explore the transmission of certain topics through content modalities. Combined with a Toxicity Model, this could offer further understanding on how hateful ideas may spread within a platform and lead to the creation of Toxic Technocultures.

6.3 Technological Redlining

The RQ1 analysis demonstrated that image-with-text memes can construct hateful messages while containing text which are low in toxicity when scored by Google Perspective. This methodological approach was led by ongoing research in image-with-text meme analysis which quantifies image features within a classification pipeline. As observed through

Yoon's (2016) qualitative multimodal discourse analysis, a connotative understanding of memes is shaped by layered signifiers and signified denotative elements which are both visual and textual. The analysis interpreted how combinations of visual and textual features in image memes can express hate towards marginalized and minority. groups. Google Perspective's toxicity scores of these image texts (0.18, 0.22) provided a classification of only one signifier element, which constructs these images.

This possibly reveals a limitation with this analysis' experiment in constructing an image-with-text classification methodology. After all, image meme texts were operationalized to construct an analysis of r/Metacanada Moderator and user content using Google Perspective. This approach failed to construct a variable to operationalize the visual features which construct these images. Deriving the Connotative meaning of these memes requires a manual qualitative approach, which analyzes Signified elements structured within these memes through techniques such as intertextuality. Yoon's (2016) multimodal discourse analysis approach displays the layered elements which could be interpreted by an automated system aiming to accurately quantify images.

Thus, automated scale image-with-text classification requires the operationalization of both image and text features through automated solution. Implementing a solution of this scale begins to engage significant technical, legal, and financial considerations which exceed the scope of this analysis. In fact, Du et al. (2020) avoided the use of the Google Vision API in their meme analysis due to the significant computational and financial demands required to apply systems like Perspective to the scale of user content. Future analyses on the classification of multi-modal content should continue to explore the functionalities of APIs such as Google Vision which offer research capabilities. The manual classification of a larger image dataset could be performed. This could establish an analysis of the frequency of memes which would potentially be misclassified by an automated approach that

operationalizes the text contained in the image. Perhaps this could lead to a better understanding of the limitations of image-with-text analysis and explore how to develop a layered semiotic approach to the classification of image-with-text memes.

Accordingly, this research has observed how operationalizing one content modality may influence how connected features are subsequently classified by an automated moderation approach. In the case of Google Perspective, the algorithm provided a unimodal analysis of image content which is constructed by multiple modalities. The implications of multiple content modalities are a crucial element to consider as automated moderation is further implemented. In Reddit's case, recent instances of some users selecting usernames to harass Transgender users (Alford, 2021) suggest toxic or hateful behaviours move beyond the captured post and comment modalities. When considering the variety of social media platforms beyond Reddit, multi-modal content thus introduces the risk of numerous misclassifications within the automated moderation of content.

The performed analysis has demonstrated some limitations of automated moderation solutions, modelling potential false negatives within image-with-text classification and Google Perspective's classification of multi-modal content. The reasons for potential misclassifications are numerous, influenced by elements such as coordinated adversarial attacks, biases in NLP models and the application of these algorithms within multi-modal content environments. At scale, it seems that these misclassifications may be observed as "glitches" as discussed by Noble. However, the author indicates that the ability of automated moderation to "target or exclude" (As cited in Bulut, 2018, p. 295) content risks discriminating against certain groups. As such, determining how automated algorithms contribute to technological redlining at scale requires an observation of these potential misclassification "glitches".

In the case of automated image analysis, Du. et al. (2020) observed that certain

demographics were heavier sharers of memes. Should at scale image-with-text analysis moderation be implemented on a platform like Reddit, these users' contributions may be misclassified at a higher frequency than that of other demographics. It is also important to further explore the intersectionality of these demographics, as the influence of automated content misclassifications do not appear evident at scale.

Implementing and administering a Natural Language Processing model like Google Perspective with the goal of regulating hate content requires some elements of manual moderation to monitor misclassifications. This was observed in The New York Times implementation of the solution as well as setting a threshold for Google Perspective's toxicity model discussed by Gruzd et al. (2020). Selecting this threshold yields profound influence over the frequency of misclassifications as elements, such as frequent identity terms, can lead to higher perspective scores (Reichert et al., 2020). Thus, the lower a toxicity threshold is set, a higher risk for false-positive classifications emerges.

The misclassifications of user content which may emerge in the COHL's mandate for proactive automated content moderation risks discriminating against marginalized individuals by excluding some content while approving others. While the COHL currently remains a proposal, the decisions required for the implementation of its content automation mandate will influence the effects of Technological Redlining. These include which automated solutions to implement, how they are integrated within platforms and what classification decisions are derived from the classification outputs of algorithmic models at scale. The first steps to addressing the potential of this regulation's redlining influence may be to demand transparency on the technical specifics of this implementation. Knowing how features such as toxicity thresholds are implemented on a platform could better demonstrate how the speech of marginalized users may be detected. Subsequent testing of this threshold using non-hateful texts containing identity terms may serve to further explore the effects of regulatory mandates

such as the COHL through the lens of Technological Redlining.

Chapter 7: Conclusion

This thesis presented a mixed methods analysis of a hate community on Reddit guided by External governance mandates for the use of automated moderation algorithms. Focusing on the r/Metacanada Subreddit, the analysis applied relevant research strategies to explore the participation of Moderators within Reddit's Toxic Technocultures. As displayed, using Google Perspective's toxicity model to operationalize hate provided mixed results in the analysis of a community's hateful norms. The hypotheses that post, comment, and image text content shared by r/Metacanada Moderators would be higher in toxicity than users were rejected. However, a wider discussion on Moderator influence of Reddit community norms emerged. Gilbert (2020) indicates that Reddit communities establish norms through both the visible and invisible actions of Moderators. Observing the visible contributions of Moderators using Google Perspective, much of community norm setting may be occurring invisibly. A larger scale longitudinal qualitative analysis of a Subreddit would likely be required to discover how hateful communities are shaped by Moderator influence. This could observe how Moderator users engage in other "Organizing" activities on the Reddit platforms. It is perhaps through these observations where platform specific recommendations for the implementation of automated moderation solutions could be developed.

As well, this research explored the classification of image-with-text memes through operationalizing visual and textual features. It was demonstrated that the connotative meaning of a meme is constructed through layered denotative elements which comprise both image and text modalities. Then the analysis displayed that operationalizing meme text using Google Vision's OCR and Google Perspective risk false negative outputs. Future studies could expand off this work through manually coding all collected images in a dataset to analyze what portion of hateful memes contain non-toxic image text but construct a hateful message. This could discover additional limitations of image-with-text classification and

develop further understanding of how its analysis can best be used to detect this content.

This research discussed how Reddit's experiments in Co-governance has maintained hateful community content norms. Moderator users maintain full control over their Subreddits within Reddit, yet the platform continues to allow users to administer hateful communities. Consequently, regulating hate content is unlikely to change unless the influence of these moderation structures is examined. In this sense, potential limitations of the COHL's ability to regulate hate content on platforms with volunteer user moderations were demonstrated. This External governance measure may fail to filter the influence of users leveraging various platform affordances to construct hateful communities.

It is also important to note that sometime following the data collection period that the r/Metacanada subreddit was locked by moderators, blocking all user contributions. However, as indicated in the subreddit sidebar, it appears that the community has migrated to an alternative platform named omegacanada.win which offers seemingly identical content affordances as Reddit. This mirrors the events identified in Ribeiro et al.'s (2021) case study of the r/The_Donald's community migration to the same .win domain. A future analysis of how moderator content contributes to these migration efforts could expand on this research's exploration of co-governance actors. This approach could develop insight on the risk of external governance regulation in pushing users to alternative platforms which do not comply with content regulation principles.

Moreover, this research contributes to Massanari's (2017) conceptualization of the Reddit platforms manipulation by users to establish Toxic Technocultures. The analysis was grounded in the author's application of Actor-Network theory, operationalizing both human and non-human actors, exploring how hateful norms are established in these networks. Despite the hypotheses being rejected, this research shows online hate's existence as "a product or an effect of a network of heterogeneous materials" (Law, 1992, p. 381), thus

contributing to ANT's goal to observe the actors constructing social reality. As for Toxic Technocultures, the analysis suggests that hateful norms are established beyond Moderator contributions, which are difficult to observe using automated solutions at scale. While Perspective may in fact classify the most extreme examples of content through operationalizing slurs and common phrases, hate remains a dynamic and complex feature of which a binary classification is difficult to determine.

This analysis thus applied Noble's (2018) Technological Redlining theory. While potential implications of automated content moderation were explored within the scope of the analysis, concerns emerge in these discussions. Noble's research demonstrates that Google's monopolized search algorithm has influenced the representation of marginalized individuals in digital spaces. Now that Google Perspective is emerging as another layer of the company's offerings which aims to "make it easier to host better conversations online" (2021a), there are inevitable consequences with the tool's implementation. Whether it be the potential of language bias, user manipulation to spread hate content or other misclassifications which may emerge from applying the tool at scale, Perspective is likely to influence online discourse. Perhaps Google and platforms should not seek to make it simply "easier'' (2021a) to host better conversations, but rather engage deeply in the complex multi-stakeholder discussions evidently required to address these concerns. This should not take place in the vacuum-like approach identified in COHL's proposal, but instead engage marginalized groups across social classes and technical abilities to contribute to the development of platforms with fairness and a sense of humanity at the core of its content policy.

Furthermore, the technical scope of the methodology required the integration of multiple tools in which the researcher was unfamiliar with prior to this research. Leveraging Python to access various Reddit APIs as well as both Google Perspective and Google Vision likely introduced undiscussed biases to the dataset. More specifically, it is unclear if these

tools have introduced gaps in data or have been updated following the identified data analysis period, such as the likely case with Google Perspective. Regardless, all reasonable steps were taken to ensure the validity of the research approach. The similar levels of observed toxicity in this work's r/Metacanada dataset when comparing to those of Gruzd et al. (2020) provides reassurance that the collected data is representative of the community.

There are several opportunities for future research which could expand on this analysis. First, Google Perspective outputs must continue to be examined through a critical lens. While this work has demonstrated how misclassifications may occur in the analysis of multi-modal content, the tool receives consistent updates from Google. As platforms explore the integration of Perspective or similar moderation algorithms, there is an opportunity to examine its influence on community discourse. This research could perform a longitudinal study of toxicity levels prior to and following Perspective's integration, performing audits of the same data as the tool is updated, observing changes to scores over time. This could include a study of the demographics of users creating the content to further explore how biases may influence automated moderation systems at scale.

As well, there is a need to develop additional understandings of Reddit's hateful communities. Massanari's (2017) conceptualization of Toxic Technocultures provides a starting point for these discussions, mapping out how users leverage platform affordances to create hateful communities. A lack of access and ethical concerns, which would emerge in an attempt to perform ethnographic research or interviews involving toxic Moderators, limits the understanding of these users who volunteer to run hateful communities. Despite these concerns and the potential risk of providing an additional platform for these users to participate, understanding their motivations and their use of platforms could establish more in-depth discussion on their involvement in a community. This could also help understand how Moderators of toxic communities choose to use automated tools as they are further

integrated within platforms or mandated by government regulators. Therefore, it is important to remove barriers between researchers and the users they study while maintaining transparency and respecting ethical considerations.

In conclusion, hateful community content norms are maintained by the technical affordances of social media platforms. External solutions are unlikely to be found solely through advancements in machine learning algorithms which classify content features. In fact, applying automated moderation solutions like Google Perspective may render these problems more invisible, shielding the classification of content behind tools which impose bias and offer minimal transparency in their function. It is currently unclear if the COHL will be implemented, or how platforms will follow its mandates for automated moderation. Referring to the extensive list of regulations provided in the framework's proposal, Geist argues that the proposal has "patched together some of the worst from around the world" (2021, para. 13). Regulating the hateful content of communities like r/Metacanada requires more than the COHL's pastiche of regulatory principles, treating automated moderation as a solution to the harms which threaten digital spaces.

In brief, it is crucial to address online hate with caution and recognition of the diversity of users who participate in digital communities. The moderation arrangements of platforms such as Reddit should be further explored, as they define what community norms are acceptable. Regulatory frameworks such as the COHL must consider these influences when implementing their policies as they are essential to the pursuit of equitable digital communities.

Dataset Reference

Chevrier, N. (2021). Reddit and Google Perspective Python Scripts. Github. https://github.com/NicChevrier/Reddit-Toxicity-Analysis

References

Alexa. (2021). Top Sites in Canada. https://www.alexa.com/topsites/countries/CA

Alford, E. (2021, July 9). Transgender Redditors Are Being Driven From the Site by Transphobic Trolls Exploiting Reddit's "Follow" Function. Jezebel. https://jezebel.com/transgender-redditors-are-being-driven-from-the-site-by-1847256024

Albawi, S., Bayat, O., Al-Azawi, S., & Ucan, O. N. (2018). Social Touch Gesture Recognition Using Convolutional Neural Network. Computational Intelligence and Neuroscience, 2018, 1–10. https://doi.org/10.1155/2018/6973103

Allen, J. (2019). The color of algorithms: An analysis and proposed research agenda for deterring algorithmic redlining. Fordham Urb. LJ, 46, 219. https://ir.lawnet.fordham.edu/ulj/vol46/iss2/1

Blout, E., & Burkart, P. (2021). White Supremacist Terrorism in Charlottesville: Reconstructing 'Unite the Right.' Studies in Conflict & Terrorism, 1–22. https://doi.org/10.1080/1057610X.2020.1862850

Bromwich, J. E. (2015, November 20). Muslims Defend Islam From Being Hijacked by ISIS. The New York Times. https://web.archive.org/web/20210808130937/https://www.nytimes.com/2015/11/20/world/europe/muslims-defend-islam-from-being-hijacked-by-isis.html

Brown, S., Milkov, P., Patel, S., Looi, Y. Z., Dong, Z., Gu, H., Artan, N. S., & Jain, E. (2019). Acoustic and Visual Approaches to Adversarial Text Generation for Google Perspective. 2019 International Conference on Computational Science and

Computational Intelligence (CSCI), 355–360. https://doi.org/10.1109/CSCI49370.2019.00069

Bulut, E. (2018). Interview with Safiya Noble: Algorithms of Oppression, Gender and Race. Moment Journal, 5(2), 294-301. https://dergipark.org.tr/en/download/article-file/653368

Buntain, C., & Golbeck, J. (2014). Identifying social roles in reddit using network structure. Proceedings of the 23rd International Conference on World Wide Web, 615–620. https://doi.org/10.1145/2567948.2579231

Bolsover, G. (2020). Black Lives Matter Discourse on US Social Media during COVID: Polarised Positions Enacted in a New Event. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3688909

Canadian Institutes of Health Research (2018). Natural Sciences and Engineering Research Council of Canada, and Social Sciences and Humanities Research Council, Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans

Callon, M. (1986). The sociology of an actor-network: The case of the electric vehicle. In Mapping the dynamics of science and technology (pp. 19-34). Palgrave Macmillan, London.

Chandrasekharan, E., Samory, M., Jhaver, S., Charvat, H., Bruckman, A., Lampe, C., Eisenstein, J., & Gilbert, E. (2018). 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(CSCW), 1–25. https://doi.org/10.1145/3274301

Cook, C. L., Patel, A., & Wohn, D. Y. (2021). Commercial Versus Volunteer: Comparing user Perceptions of Toxicity and Transparency in Content Moderation Across Social

Media Platforms. Frontiers in Human Dynamics, 3,. https://doi.org/10.3389/fhumd.2021.626409

Couldry, N. (2008). Actor network theory and media: do they connect and on what terms? In A. Hepp, F. Krotz, S. Moores, & C. Winters (Eds.) Connectivity, Networks and Flows: Conceptualizing Contemporary Communications (pp. 93-110). Hampton Press, Inc.

Dignam, P. A., & Rohlinger, D. A. (2019). Misogynistic Men Online: How the Red Pill Helped Elect Trump. Signs: Journal of Women in Culture and Society, 44(3), 589–612. https://doi.org/10.1086/701155

Du, Y., Masood, M. A., & Joseph, K. (2020). Understanding visual memes: An empirical analysis of text superimposed on memes shared on twitter. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 14, pp. 153-164).

El País. (2019, March 21). How El País used AI to make their comments section less toxic. Google. https://blog.google/outreach-initiatives/google-news-initiative/how-el-pais-used-ai-make-their-comments-section-less-toxic/

Farid, H. (2018). Reining in Online Abuses. Technology & Innovation, 19(3), 593–599. https://doi.org/10.21300/19.3.2018.593

Felon, B. (2021, June 15). Why CBC is turning off Facebook comments on news posts for a month. CBC. https://www.cbc.ca/news/editor-blog-facebook-comments-1.6064804

Fiesler, C., McCann, J., Frye, K., & Brubaker, J. R. (2018, June). Reddit rules! characterizing an ecosystem of governance. In Twelfth International AAAI Conference on Web and Social Media.

Finkelstein, J., Zannettou, S., Bradlyn, B., & Blackburn, J. (2020, May). A quantitative approach to understanding online antisemitism. In Proceedings of the

International AAAI Conference on Web and Social Media (Vol. 14, pp. 786-797).

Fortuna, P., Soler, J., & Wanner, L. (2020, May). Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In Proceedings of the 12th language resources and evaluation conference (pp. 6786-6794).

Gaudette, T., Scrivens, R., Davies, G., & Frank, R. (2020). Upvoting extremism: Collective identity formation and the extreme right on Reddit. New Media & Society, 1-26. https://doi.org/10.1177/1461444820958123.

Geist, M. (2021, July 30). Picking Up Where Bill C-10 Left Off: The Canadian Government's Non-Consultation on Online Harms Legislation. Micheal Geist. https://www.michaelgeist.ca/2021/07/onlineharmsnonconsult/

Gilbert, S. (2020). "I run the world's largest historical outreach project and it's on a cesspool of a website." Moderating a Public Scholarship Site on Reddit: A Case Study of r/AskHistorians. Proceedings of the ACM on Human-Computer Interaction, 4(CSCW1), 1–27. https://doi.org/10.1145/3392822

Gillespie, T. (2018). Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media. Yale University Press.

Gillespie, T. (2020). Content moderation, AI, and the question of scale. Big Data & Society, 7(2), 205395172094323. https://doi.org/10.1177/2053951720943234

Google Vision (2021). Vision AI: Derive Image Insights via ML Cloud Vision API. Google Cloud. https://cloud.google.com/vision

Gomez, R., Gibert, J., Gomez, L., & Karatzas, D. (2020). Exploring Hate Speech Detection in Multimodal Publications. 2020 IEEE Winter Conference on Applications of

Computer Vision (WACV), 1459–1467. https://doi.org/10.1109/WACV45572.2020.9093414

Gorwa, R. (2019). What is platform governance? Information, Communication & Society, 22(6), 854–871. https://doi.org/10.1080/1369118X.2019.1573914

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data & Society, 7(1), 1-15. https://doi.org/10.1177/2053951719897945

Grimmelmann, J. (2015). The virtues of moderation. Yale JL & Tech., 17, 42.

Gruzd A, Mai P, Vahedi Z. (2020). Studying anti-social behaviour on Reddit with Communalytic. Advance. epub. doi: 10.31124/advance.12453749

Guimarães, S. S., Reis, J. C., Ribeiro, F. N., & Benevenuto, F. (2020, November). Characterizing toxicity on facebook comments in brazil. In Proceedings of the Brazilian Symposium on Multimedia and the Web (pp. 253-260).

Hatebase. (2021). About. https://hatebase.org/about

Heldt, A. (2019). Reading between the lines and the numbers: An analysis of the first NetzDG reports. Internet Policy Review, 8(2), 1-19. https://doi.org/10.14763/2019.2.1398

Heritage Canada. (2021, July 29). OHL Technical Paper. https://www.canada.ca/en/canadian-heritage/campaigns/harmful-online-content/technical-paper.html

ImgFlip. (2021). jaws Meme Generator—Imgflip. https://imgflip.com/memegenerator/7730956/jaws

Jain, E., Brown, S., Chen, J., Neaton, E., Baidas, M., Dong, Z., Gu, H., & Artan, N. S. (2018). Adversarial Text Generation for Google's Perspective API. 2018 International Conference on Computational Science and Computational Intelligence (CSCI), 1136–1141. https://doi.org/10.1109/CSCI46756.2018.00220

Jaws. (1975). You're Gonna Need a Bigger Boat Scene. https://www.youtube.com/watch?v=2I91DJZKRxs

Jhaver, S., Chan, L., & Bruckman, A. (2018). The view from the other side: The border between controversial speech and harassment on Kotaku in Action. First Monday, 23(2). https://doi.org/10.5210/fm.v23i2.8232

Jiang, J., & Vosoughi, S. (2020). Not Judging a user by Their Cover: Understanding Harm in Multi-Modal Processing within Social Media Research. Proceedings of the 2nd International Workshop on Fairness, Accountability, Transparency and Ethics in Multimedia, 6–12. https://doi.org/10.1145/3422841.3423534

Jigsaw. (2021). Google's Jigsaw Announces Toxicity-Reducing API, Perspective, is Processing 500M Requests Daily. Google. https://www.prnewswire.com/news-releases/googles-jigsaw-announces-toxicity-reducing-api-perspective-is-processing-500m-requests-daily-301223600.html

KYM. (2021). Charles McDowell's Wide Neck Mugshot. Know Your Meme. https://knowyourmeme.com/memes/charles-mcdowells-wide-neck-mugshot

Latour, B. (1996). On actor-network theory—a few clarifications. Soziale Welt-Zeitschrift fur Sozialwissenschaftliche for schung und praxis, 47(4), 1-13. http://www.nettime.org/Lists-Archives/nettime-l-9801/msg00019.html

Law, J. (1992). Notes on the theory of the actor-network: Ordering, strategy, and heterogeneity. Systems Practice, 5(4), 379–393. https://doi.org/10.1007/BF01059830

Liddy, E (2001). Natural Language Processing. In Encyclopedia of Library and Information Science, 2nd Ed. Marcel Decker, Inc

Mittos, A., Zannettou, S., Blackburn, J., & De Cristofaro, E. (2020, May). "And We Will Fight for Our Race!" A Measurement Study of Genetic Testing Conversations on

Reddit and 4chan. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 14, pp. 452-463).

Mall, R., Nagpal, M., Salminen, J., Almerekhi, H., Jung, S.-G., & Jansen, B. J. (2020). Four Types of Toxic People: Characterizing Online users' Toxicity over Time. Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society, 1–11. https://doi.org/10.1145/3419249.3420142

Massanari, A. (2017). #Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. New Media & Society, 19(3), 329–346. https://doi.org/10.1177/1461444815608807

Massanari, A. (2018). Rethinking Research Ethics, Power, and the Risk of Visibility in the Era of the "Alt-Right" Gaze. Social Media + Society, 4(2), 1-9. https://doi.org/10.1177/2056305118768302

McCoy, H. (2020). Black lives matter, and Yes, you are racist: the parallelism of the twentieth and twenty-first centuries. Child and Adolescent Social Work Journal, 37(5), 463-475. https://doi.org/10.1007/s10560-020-00690-4

Mendes, M., & Akin, P. (2003). Type I Error Rate and Power of Three Normality Tests. Information Technology Journal, 2(2), 135–139. https://doi.org/10.3923/itj.2003.135.139

Milton, J. (2018, October 19). Canada's largest subreddit accused of harbouring white nationalists. Ricochet. https://ricochet.media/en/2385/canadas-largest-subreddit-accused-of-harbouring-white-nationalists

Nachar, N. (2008). The Mann-Whitney U: A Test for Assessing Whether Two Independent Samples Come from the Same Distribution. Tutorials in Quantitative Methods for Psychology, 4(1), 13–20. https://doi.org/10.20982/tqmp.04.1.p013

Nardone, A., Rudolph, K. E., Morello-Frosch, R., & Casey, J. A. (2021). Redlines and Greenspace: The Relationship between Historical Redlining and 2010 Greenspace across the United States. Environmental Health Perspectives, 129(1), 1-9. https://doi.org/10.1289/EHP7495

Nithyanand, R., Schaffner, B., & Gill, P. (2017). Online political discourse in the Trump era. arXiv preprint arXiv:1711.05303.

Noble, S. U. (2018). Algorithms of oppression: How search engines reinforce racism. New York University Press.

NYT Team, (2020, April 30). To Apply Machine Learning Responsibly, We Use It in Moderation. Medium. https://open.nytimes.com/to-apply-machine-learning-responsibly-we-use-it-in-moderation-d001f49e0644

Ohanian, A. (2008, March 12). What's new on reddit: Make your own reddit. https://web.archive.org/web/20150905082338/http://www.redditblog.com/2008/03/make-your-own-reddit.html

Pantumsinchai, P. (2018). Armchair detectives and the social construction of falsehoods: An actor–network approach. Information, Communication & Society, 21(5), 761–778. https://doi.org/10.1080/1369118X.2018.1428654

Parekh, P., & Patel, H. (2017). Toxic Comment Tools: A Case Study. International Journal of Advanced Research in Computer Science, 8(5). 964-967. https://doi.org/10.26483/ijarcs.v8i5.3506

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1532–1543. https://doi.org/10.3115/v1/D14-1162

Perspective. (2021a). About the API - Key Concepts. https://developers.perspectiveapi.com/s/about-the-api-key-concepts

Perspective. (2021b). About the API - Attributes and Languages. https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages

Proferes, N., Jones, N., Gilbert, S., Fiesler, C., & Zimmer, M. (2021). Studying Reddit: A Systematic Overview of Disciplines, Approaches, Methods, and Ethics. Social Media+ Society, 7(2), 1-14. https://doi.org/10.1177/20563051211019004

Reddit. (2020a). Open Letter to Steve Huffman and the Board of Directors of Reddit, Inc– If you believe in standing up to hate and supporting black lives, you need to act. Reddit. https://www.reddit.com/r/AgainstHateSubreddits/comments/gyyqem/open_letter_to_steve_huffman_and_the_board_of/

Reddit. (2020b). Promoting Hate Based on Identity or Vulnerability. Reddit Help. https://reddit.zendesk.com/hc/en-us/articles/360045715951-Promoting-Hate-Based-on-Identity-or-Vulnerability

Reddit. (2020c). Transparency Report 2020—Reddit. Reddit. https://www.redditinc.com/policies/transparency-report-2020

Reddit Press. (2021). Reddit Press. https://www.redditinc.com/press

Reichert, E., Qiu, H., & Bayrooti, J. (2020). Reading between the demographic lines: Resolving sources of bias in toxicity classifiers. arXiv preprint arXiv:2006.16402.

Ribeiro, M. H., Blackburn, J., Bradlyn, B., Cristofaro, E. D., Stringhini, G., Long, S., Greenberg, S., & Zannettou, S. (2020). The Evolution of the Manosphere Across the Web. 15th International AAAI Conference on Web and Social Media. (pp. 1-12). http://arxiv.org/abs/2001.07600

Ribeiro, M., Jhaver, S., Zannettou, S., Blackburn, J., Stringhini, G., De Cristofaro, E., & West, R. (2021). Do platform migrations compromise content moderation? evidence from r/the_donald and r/incels. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW2), 1-24.

Ritzer, G. (2005). Encyclopedia of Social Theory. SAGE Publications, Inc. https://doi.org/10.4135/9781412952552

Robertson, A. (2021, August 10). Apple's controversial new child protection features, explained. The Verge. https://www.theverge.com/2021/8/10/22613225/apple-csam-scanning-messages-child-safety-features-privacy-controversy-explained

R/Metacanada. (2021). R/Metacanada. https://imgflip.com/memegenerator/7730956/jaws

Sabat, B. O., Ferrer, C. C., & Giro-i-Nieto, X. (2019). Hate Speech in Pixels: Detection of Offensive Memes towards Automatic Moderation. ArXiv:1910.02334, 1-5. http://arxiv.org/abs/1910.02334

Salkind, N. J., & Rasmussen, K. (Eds.). (2007). Encyclopedia of measurement and statistics. SAGE Publications.

Seering, J., Kaufman, G., & Chancellor, S. (2020). Metaphors in moderation. New Media & Society, 1-21. https://doi.org/10.1177/1461444820964968

Shepherd, R. P. (2020). Gaming Reddit's Algorithm: R/the_donald, Amplification, and the Rhetoric of Sorting. Computers and Composition, 56, 1–10. https://doi.org/10.1016/j.compcom.2020.102572

Singer, P., Flöck, F., Meinhart, C., Zeitfogel, E., & Strohmaier, M. (2014). Evolution of Reddit: From the Front Page of the Internet to a Self-referential Community? Proceedings of the 23rd International Conference on World Wide Web, 517–522. https://doi.org/10.1145/2567948.2576943

Solomon, L. (2015). Fair users or content abusers: The automatic flagging of non-infringing videos by content id on youtube. Hofstra L. Rev., 44, 237-268.

Somekh, B., & Lewin, C. (Eds.). (2005). Research methods in the social sciences. SAGE Publications.

Suryawanshi, S., Chakravarthi, B. R., Arcan, M., & Buitelaar, P. (2020, May). Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. In Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying (pp. 32-41).

Statt, N. (2018, April 11). Reddit CEO says racism is permitted on the platform, and users are up in arms. The Verge. https://www.theverge.com/2018/4/11/17226416/reddit-ceo-steve-huffman-racism-racist-slurs-are-okay

Tifrea, A., Bécigneul, G., & Ganea, O. E. (2018). Poincaré GloVe: Hyperbolic Word Embeddings. In Proceedings of the International Conference on Learning Representations (ICLR 2019). OpenReview. 1-24.

Walsham, G. (1997). Actor-network theory and IS research: current status and future prospects. In Information systems and qualitative research (pp. 466-480). Springer.

Watt, D. P. (2012). The Urgency of Visual Media Literacy in Our Post-9/11 World: Reading Images of Muslim Women in the Print News Media. Journal of Media Literacy Education, 4(1), 32-43.

Yoon, I. (2016). Why is it not just a joke? Analysis of Internet memes associated with racism and hidden ideology of colorblindness. Journal of Cultural Research in Art Education, 33. 93-123.

Zannettou, S., Caulfield, T., Blackburn, J., De Cristofaro, E., Sirivianos, M., Stringhini, G., & Suarez-Tangil, G. (2018, October). On the origins of memes by means of fringe web communities. In Proceedings of the Internet Measurement Conference 2018 (pp. 188-202).

Zannettou, S., El Sherief, M., Belding, E., Nilizadeh, S., & Stringhini, G. (2020). Measuring and characterizing hate speech on news websites. In 12th ACM Conference on Web Science (pp. 125-134).

Zimdars, M., & Mcleod, K. (2020). Fake News: Understanding Media and Misinformation in the Digital Age. MIT Press.

Zomick, J., Levitan, S. I., & Serper, M. (2019, June). Linguistic analysis of schizophrenia in Reddit posts. In Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology (pp. 74-83).

Appendices

Appendix A
Shapiro-Wilk Test of Normality (Posts)

		W	p
Moderator	0	0.839	< .001
User	1	0.810	< .001

Note. Significant results suggest a deviation from normality.

Appendix B

Independent Samples T-Test (Posts)

	W	df	p	Hodges-Lehmann Estimate
Toxicity Score	277977.000		0.786	-0.006

Note. For all tests, the alternative hypothesis specifies that group 0 is greater than group 1.
Note. Mann-Whitney U test.

Appendix C
Independent Samples T-Test (Comments)

	W	df	p	Hodges-Lehmann Estimate
Toxicity Score	2.7085e +6		0.5514	-0.0006

Note. For all tests, the alternative hypothesis specifies that group 0 is greater than group 1.
Note. Mann-Whitney U test.

Appendix D

Independent Samples T-Test (Image texts)

	W	df	p	Hodges-Lehmann Estimate
Toxicity Score	1002.0000		0.6272	-0.0197

Note. For all tests, the alternative hypothesis specifies that group 0 is greater than group 1.
Note. Mann-Whitney U test