ABSTRACT

Title of Dissertation: The Rhetoric of Atrocity:
Modeling Genocidal Speech

Rewina S. Bedemariam
Doctor of Philosophy, 2026

Dissertation Directed by: Professor Jennifer L. Wessel
Department of Psychology

This dissertation examines genocidal hate speech in the Tigray conflict through a multi-method lens, combining the development of a supervised classification model with human affective validation. In the first study, Tweets were harvested using a domain-specific lexicon, capturing messages containing ethnic slurs and conflict-related keywords. Unsupervised topic modeling and lexical feature analysis revealed distinct narratives, and building on these insights, a HateBERT classifier was fine-tuned to assign tweets to a ten-category taxonomy rooted in Atrocity Speech Law theory (Gordon (2017)). In a second study, members of the Tigrayan diaspora rated a subset of tweets on anger, fear, disgust, perceived threat, perceived violence and behavioral intentions. Repeated-measures analyses showed that direct calls for violence and dehumanizing metaphors elicited the strongest negative reactions, whereas economic insults provoked relatively mild responses. The convergence between model outputs and human judgments supports the taxonomy's validity. The findings underscore the potential of theory-informed NLP models for early warning and moderation, while highlighting the need for richer context, multilingual capabilities and continuous human feedback.

The Rhetoric of Atrocity:
Modeling Genocidal Speech

by

Rewina Sahle Bedemariam

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2026

Advisory Committee:
Professor Jennifer L. Wessel
Professor Hal Daumé III
Professor Paul Hanges
Professor James Grand
Professor Linda Zou

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my parents for their unwavering support, sacrifices, and belief in me throughout my Ph. D. journey. Their encouragement and resilience shaped not only this work, but the person I have become.

I am also profoundly grateful to my adviser, Professor Jennifer L. Wessel, whose guidance and patience helped me refine my ideas, strengthen my skills, and grow into an I/O psychologist.

I am deeply thankful to my friends for standing by me throughout this process. Their encouragement, humor, and presence made the long and difficult moments more bearable and the successes more joyful.

I would especially like to acknowledge the Tigray community, whose strength, solidarity, and stories have deeply influenced both my personal life and my scholarly commitments. Their resilience throughout the war and beyond served as a constant reminder of why this work matters.

To my partner Yared, I owe more gratitude than words can capture. Thank you for standing by me through every high and low, for your patience and unwavering belief in me. Your support carried me through moments of doubt and exhaustion, and your presence made this journey possible in ways that extend far beyond this dissertation.

Finally, to the reader. I hope that you find this work as engaging and meaningful as I have found it to be. This dissertation reflects not only years of research, but also years of grappling to make sense of war, suffering, and the ways in which words can both harm and reveal truth. It reflects my attempt to bring together theory and reality, and to approach a deeply painful subject with both intellectual clarity and honesty. If this work invites the reader to pause, question, or see familiar narratives in a new light, then it has fulfilled its purpose.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AI	Artificial Intelligence
AM	Accusation in a Mirror
CNN	Convolutional Neural Network
DA	Direct Advocacy
DM	Demonization
DS	Disloyalty
EAT	Economic Anti-Tigrayanism
ENDF	Ethiopian National Defence Force
IRB	Institutional Review Board
J	Justification
KDE	Kernel Density Estimate
LDA	Latent Dirichlet Allocation
LFA	Linguistic Feature Analysis
LLM	Large Language Model
LR	Logistic Regression
ML	Machine Learning
NLP	Natural language Processing
O	Other
P	Pathologization
PPV	Praising Past Violence
Q/E	Quotation/Explanation (Not Hate Speech)
TPLF	Tigray People's Liberation Front
TSC	Target-Sympathizer Conflation
V	Verminization

Chapter 1: Introduction

1.1 Background

Inflammatory hate speech has been found to incite mass killings, including genocide (Staub (2011); Benesch (2007)). By instilling in individuals the perception that certain groups of human beings are inferior and pose a mortal threat, influential figures can create an environment in which acts of atrocity are deemed acceptable and even necessary as a means of collective self-defense (Benesch (2007)). This form of speech has infamously preceded some of the most significant intergroup mass killings in history, such as the Holocaust and the 1994 genocide in Rwanda (Benesch (2007); Herf (2006); Des Forges (1999)). Regrettably, it continues to be prevalent in numerous countries that are at risk of collective violence, including Ethiopia, Myanmar, India and others.

Inflammatory speech has been shown to have particular rhetorical characteristics, even across different historical eras, languages, and cultures (Benesch (2007)). With the advancement of natural language processing (NLP) technology in recent years, substantial research has been conducted on automatic textual hate speech identification (Jahan and Oussalah (2023)). Several well-known contests (for example, May et al. (2019) and Schlechtweg et al. (2020), Ruppenhofer (2018)) have been organized in order to identify a better solution for automated hate speech identification (Zampieri et al. (2019), Zampieri et al. (2020), Wiegand et al. (2018)). Researchers are experimenting with various NLP and deep learning algorithms to detect and address hate speech, which has resulted in a major expansion of the field of hate speech analysis in recent years (Jahan and Oussalah (2023); Fortuna et al. (2022); MacAvaney et al. (2019); Schmidt and Wiegand (2017)). This has also prompted the investigation and comparison of various processing pipelines, such as feature set (eg., TF-IDF, Word2Vec), Machine Learning (ML) methods (e.g., supervised, unsupervised, and semi-supervised), classifica-

tion algorithms (e.g., Naives Bayes, Logistic Regression (LR), Convolution Neural Network (CNN), LSTM, BERT deep learning architectures, and others.

By drawing upon the research of other scholars on hate speech using NLP, I categorized the distinct dimensions of hate speech in the context of the Tigray war. I mostly focused on what Gordon termed genocidal speech (Gordon (2017)). I generated a taxonomy based on his theories of the characteristics of incitement to genocide. Using this generated taxonomy, I experimented with NLP models and generated a multiclass classification model.

1.2 Aims and Objectives

The primary aim of this research is to operationalize Gregory S. Gordon's atrocity-speech law typology in a computational framework. I aim to build a machine-learning classifier that can detect and categorize speech according to the legally significant incitement techniques he defined (Gordon (2017)). The specific objectives are to first develop a detailed annotation schema grounded in Gordon's taxonomy (e.g., direct calls for destruction, verminization, accusation in a mirror, justification during violence, victim-sympathizer conflation, etc.) Then, I aimed to collect and annotate a corpus of speeches (particularly tweets during the war in Tigray), propaganda, social media posts, and political discourse relevant to atrocity risk. Next, I aimed to train and evaluate a multi-label machine-learning classification model that can identify these different techniques in text. And finally, I sought to test the psychological relevance of the model by conducting human experiments, where participants react to text of tweets labeled with different incitement types.

1.3 Research Questions

To guide this research, I proposed the following research questions (RQs):

- RQ1: Can a multi-class ML classifier (built via theoretical lens) accurately distinguish between incitement techniques in text, and how will it perform (precision, recall, F1, etc)?
- RQ2: How do human participants psychologically respond to text labeled by the model as different types of atrocity speech? Do different hate speech types evoke different reactions?

1.4 Significance of the Proposed Study

A genocide determination is still pending in Ethiopia, where there has been a civil conflict centered in the Tigray region (B. Harris (2021)). Tigray's 6-7 million residents were effectively silenced from speaking out due to a two-year Internet outage. This has made it easier for the Ethiopian government and its allies to control narratives, foment hateful discourses and shape perceptions (B. Harris (2021); Zelalem (2022); Tackett (2023)). The Tigray war has had terrible repercussions such as famine, large scale sexual violence etc (Keaten (2023); Center (2023)) and the role of hate speech in exacerbating the conflict remains unstudied (Anna (2022); Muhumuza (2022); United Nations (2022)).

As of 2024 - 2025, the humanitarian consequences of the Tigray conflict remain catastrophic. The war (2020-2022) is estimated to have killed up to 600,000 people and displaced nearly 3 million, and its aftermath has been marked by famine-like conditions, mass displacement, and widespread sexual violence (Refugees International (2024)). Human Rights Watch's 2024 World Report found that, even after the Pretoria peace agreement, serious rights violations against civilians continued in Tigray throughout 2023, including ongoing killings, sexual violence, and the ethnic cleansing of Tigrayans in

western zones under occupying forces (Human Rights Watch (2024)).

Using the Tigray war as a case study, I sought to develop an empirical framework and methodology for developing psychosocial features set to aid in the machine classification of hate speech laden on Twitter and other social media. The number of studies on hate speech has increased in recent years, particularly those that use text mining and NLP to automatically detect hate speech (Waseem and Hovy (2016); Z. Zhang and Luo (2019); Silva et al. (2016)). However, there are not many publications that discuss how to spot hate speech in code-switched conversations and under-resourced languages (Jahan and Oussalah (2023)). This study contributes to bridging this gap by carrying out this study on code-switched language.

To do so, I took a holistic look at hate speech during conflicts and used theories such as Dangerous Speech Theory by Susan Benesch (Leader Maynard and Benesch (2016)), Moral Disengagment Theory by Bandura (Bandura (2016)), and Incitement to Genocide Framework by Gordon (Gordon (2017)). These frameworks helped in identifying linguistic analogs for these components and can be used with several computational techniques to measure these linguistic correlates that contribute to dehumanization.

The findings of the study can also enable large-scale studies of dehumanizing language against marginalized social groups. Therefore, it has implications for improving machines' abilities to automatically detect hate speech and abusive language online, which are typically underpinned by dehumanizing language. In sum, this paper bridges the gaps between computational modeling, psycholinguistics, and dehumanization research with implications for several disciplines.

1.5 Expected Contributions

I aimed to convert complex theories into empirical, testable categories by operationalizing Gordon's atrocity-speech law taxonomy and other related theories. This aids in bridging the knowledge gap between data and theory.

Additionally, by offering structured stimuli for experiments on the effects of various forms of inciting speech on human cognition, emotion, and moral judgment, the project can help us better understand how people perceive fear and threat.

Moreover, the classifier can make it possible to conduct extensive, quantitative analyses of rhetorical devices (such as dehumanization, justification, and metaphor) over time, across media, and in various cultural contexts.

It can also aid in the detection of hate speech and violent speech beyond binary models to fine-grained, legally-informed classification using contemporary NLP techniques.

The research can also direct and be incorporated into early warning systems to detect potentially harmful content.

Chapter 2: Literature Review

The purpose of this chapter is to provide a comprehensive review of theoretical, psychological, and computational literature relevant to the study of hate speech, genocidal rhetoric, dehumanization, and NLP approaches to automated detection. The chapter integrates scholarship from genocide studies, social psychology, linguistics, and ML to establish a multidisciplinary foundation for the study.

2.1 Definitions of Dehumanization and Related Concepts

Hate speech generally refers to expressions that denigrate, threaten, or incite hostility toward individuals or groups based on characteristics such as ethnicity, religion, nationality, or other identity markers (A. Brown (2017)). However, it is misleading to treat hate speech as a single, homogeneous phenomenon. Instead, it exists on a continuum, with relatively mild forms of discriminatory expression at one end and more extreme rhetoric at the other which is capable of driving collective atrocity (Gagliardone et al. (2015)).

At the most severe end of this lies genocidal speech. Genocidal speech encompasses forms of expression that explicitly incite, justify, or otherwise enable mass atrocities (Gordon (2017); Benesch (2012)). Although most instances of hate speech do not escalate into genocide, historical work makes it difficult to ignore how rarely genocidal violence occurs without a preceding period of sustained, hate-driven rhetoric (Chirot and McCauley (2006); Fein (1993); Stanton (1996)). This recurring pattern points to a close relationship among hate speech, dehumanization, and genocidal incitement, which, while conceptually distinct, operate in overlapping and mutually reinforcing ways (Haslam (2006); Bandura (1999); United States Holocaust Memorial Museum (2023)).

In research on hate speech and genocidal hate speech, dehumanization has repeatedly emerged as a particularly corrosive psychological process, especially in work that traces how hostile rhetoric trans-

lates into intergroup bias, discrimination, and, in extreme cases, violence (Mendelsohn et al. (2020)). Dehumanization refers to the perception that individuals or groups lack essential qualities of humanness, including both human nature and uniquely human attributes (Haslam (2006)), a conceptualization that has been especially influential in explaining how moral boundaries erode in contexts of sustained hostility. In practice, this denial of humanness can take several recognizable forms. Targeted groups may be portrayed as objects through what Haslam (2006) describes as mechanistic dehumanization, or as animals through animalistic dehumanization. In other instances, dehumanization operates more subtly, for example through the denial of secondary emotions, a process often referred to as infrahumanization.

Dehumanization is not always communicated overtly. It may function implicitly, such as when the emotional experiences of an outgroup are systematically disregarded, or explicitly through metaphorical language that frames targets in nonhuman terms. Xenophobic discourse provides a clear illustration of this process, particularly when immigrants are characterized as objects, invaders, or parasites, metaphors that strip targeted groups of moral standing while appearing rhetorically ordinary. Importantly, dehumanization plays a central role in making hate speech more actionable. By framing targeted groups as animals, diseases, or inanimate objects, dehumanizing rhetoric lowers psychological barriers to violence and facilitates moral disengagement (Haslam (2006)). As Gordon (2017) and Benesch (2021) observe, metaphors invoking vermin, cancer, or demonic figures recur in pre-genocidal discourse not merely as rhetorical excesses, but as cognitive tools that help normalize and justify exterminatory violence. For this reason, contemporary atrocity-prevention frameworks increasingly treat dehumanization as one of the most salient linguistic warning signs of impending mass violence (Straus (2001); United States Holocaust Memorial Museum (2023)).

In this study, I draw on Gordon (2017) theorization of genocidal speech, which refers to forms

of communication that provoke, instigate, abet, or explicitly order genocidal violence. These forms of speech vary in terms of speaker intent, rhetorical strategy, and their causal role in facilitating mass atrocity. Unlike broader forms of hate speech, which may remain at the level of insult or discrimination, genocidal speech is specifically concerned with language that helps bring about or legitimize the destruction of a targeted group. Within Gordon's framework, rhetorical strategies such as accusation in a mirror, pathologization, and explicit calls for destruction are not simply expressions of hostility. Instead, they help shape the discursive conditions under which genocide becomes imaginable, defensible, and ultimately actionable (Gordon (2017)).

This perspective aligns closely with work on dangerous speech, particularly Benesch's argument that certain forms of rhetoric become especially consequential when they portray a target group as an existential threat and frame violence as defensive or necessary (Benesch (2012), Benesch (2021)). Similar arguments appear across genocide studies and international law, where scholars have long emphasized that language plays an active role in preparing populations for mass violence, both psychologically and politically (Straus (2003); Tsesis (2002); Schabas (2000)). Dehumanizing metaphors, vilification, and claims of group disloyalty do more than exclude targeted populations from moral concern. They lower moral restraints and help legitimize extreme violence by reshaping who is seen as worthy of protection (Bandura (1999); Kelman (1973); Oberschall (2000)). From this standpoint, genocidal speech is defined less by the expression of hatred alone and more by its capacity to mobilize. It actively constructs the targeted group as an object whose elimination can come to appear justified or even necessary (Alvarez (2001); Leader Maynard (2014)).

My study focuses specifically on genocidal hate speech, understood as a form of rhetoric that combines the core features of hate with the escalatory and legitimizing dynamics associated with genocide (Gordon (2017)). Much of the existing research on hate speech detection has relied on binary

distinctions between hate and non-hate content. In contrast, recent work, including the present study, moves beyond this dichotomy by adopting multiclass classification approaches that differentiate among distinct forms of genocidal rhetoric, such as dehumanization, betrayal narratives, and economic scapegoating. Disaggregating these forms allows for the capture of meaningful variations in how genocidal discourse operates, while also improving analytical precision and detection performance (Aluru et al. (2021); Fortuna et al. (2022)). By treating genocidal hate speech as a set of interacting discursive strategies rather than a single category, this study aims to provide a more nuanced account of how hate escalates into violence.

My research has direct implications for improving automated systems designed to detect hate speech and abusive language online, where dehumanizing rhetoric is frequently present. Beyond detection, it also contributes to a broader understanding of how language reflects and reinforces social and political ideologies. Examining how marginalized groups are dehumanized offers insight into how moral boundaries are constructed and maintained within contentious sociopolitical environments.

In this study, I treat the detection of dehumanizing hate speech as a particularly challenging task in online environments, especially in the aftermath of triggering events such as war or political violence. During these periods, language data are generated rapidly, at large scale, and across a wide range of platforms. As a result the data tend to be noisy, repetitive, incomplete, and shaped by the broader constraints associated with big data. Addressing these challenges requires an interdisciplinary approach. In this study, I integrate methods from computer science, linguistics, and psychology to examine how dehumanization unfolds in real-world contexts, including active conflict settings (Pennebaker et al. (2003)). Psychological theory plays a particularly important role in this integration, as it provides a foundation for developing more interpretable and explainable ML systems, an ongoing challenge in many AI-based approaches to hate speech detection (Taylor and Taylor (2021)).

I also rely on insights from linguistics to understand how language contributes to the formation and transmission of stereotypes. Stereotypical beliefs are often reproduced through everyday linguistic choices, including labeling practices and patterns of description (Beukeboom and Burgers (2019), Burgers and Beukeboom (2020); Sutton and Douglas (2008)). Language functions as a primary mechanism through which social categories are learned and through which certain groups become marked as targets of bias (Caliskan et al. (2017); Pennebaker et al. (2003)). These processes frequently operate in subtle ways, shaping perceptions and evaluations even when they are not explicitly recognized. For example, the Social Categories and Stereotype Communication framework developed by Beukeboom and Burgers (2019) shows how systematic biases in language use shape perceptions of category entitativity, stereotype content, and essentialism. Examining frameworks of this kind allows me to better understand how dehumanization emerges and is sustained through everyday discourse.

At present, no single theory captures all dimensions of hate speech. As a result, this study draws on insights from psychology, linguistics, and genocide studies to develop a more comprehensive account of dehumanization in real-world contexts, particularly within social media discourse. Taken together, these perspectives offer complementary theoretical and methodological tools for studying how dehumanization functions at scale.

2.1.1 Theoretical Models and Frameworks

Although genocidal hate speech and more conventional forms of hate speech share surface similarities, particularly in their use of derogatory language and discriminatory attitudes toward targeted groups, the differences between them are substantial. These differences concern intent, severity, and, most importantly, consequence. Under international law, direct and public incitement to genocide is treated as one

of the most serious forms of hate speech and is explicitly criminalized (United Nations (1948).; United States Holocaust Memorial Museum (2019)). This legal distinction reflects a broader recognition that genocidal hate speech does more than communicate hostility. It works to dehumanize a targeted group and to cultivate a permissive environment in which genocide becomes imaginable and defensible, setting it apart from hate propaganda that may remain rhetorically extreme without explicitly ordering violence (Fyfe (2017); Timmermann (2005)).

At the core of this distinction is intent. Genocidal speech carries the specific intent, or mens rea, to destroy a protected group in whole or in part. This element of intent is central to the definition of genocide under the Genocide Convention and international criminal jurisprudence, and it marks a clear boundary between genocidal hate speech and other forms of hateful expression (Gordon (2017); Fyfe (2017)).

Conventional hate speech can, and often does, contribute to discrimination, harassment, and occasional acts of violence (Allport (1954); Perry (2001)). While these harms are real and should not be minimized, conventional hate speech usually stops short of calling for the systematic destruction of an entire group. That distinction matters. Genocidal rhetoric is defined precisely by its orientation toward elimination rather than exclusion (Gordon (2017)).

The legal and moral consequences attached to genocidal hate speech follow directly from this difference. As mentioned, advocating for genocide, or inciting violence against protected groups, constitute crimes under international law and are subject to international prosecution under the Genocide Convention (United Nations (1948); United States Holocaust Memorial Museum (2019)). While conventional hate speech may also be punishable under certain national legal frameworks, genocidal hate speech occupies a separate category because of its direct relationship to mass atrocity and its capacity to mobilize large-scale violence (Fyfe (2017); Timmermann (2005)).

Genocidal hate speech is also rarely produced in a vacuum. It tends to emerge in specific historical and sociopolitical contexts shaped by ethnic, religious, or political tension (Straus (2003)). Speakers often draw on long-standing grievances, familiar stereotypes, and narratives of victimhood to make violence appear justified or even necessary (Bandura (1999); Kelman (1973)). Conventional hate speech is likewise shaped by social and historical forces, but it is not inherently tied to genocidal ideologies or to aspirations of collective destruction (Benesch (2012)).

Taken together, these distinctions help explain why genocidal hate speech must be treated as analytically and practically distinct from more general forms of hate speech. Its defining feature is not only the intensity of hostility it expresses, but its ability to legitimize destruction and to set the conditions for mass violence (Alvarez (2001); Leader Maynard, 2014).

To conceptualize, categorize and eventually operationalize genocidal hate speech and the many types subsumed with in it, I turned to exploring theories and frameworks. However, as stated above there is no one theory that can capture the hate speech and dehumanization that occur in times of genocidal war single-handedly. As such my framework draws from numerous psychological, linguistic and genocide studies theories.

2.1.2 Atrocity Speech Law Framework

The major theoretical framework I drew from is Gregory Gordon's incitement to genocide speech (Gordon (2017)). However, other theories also share similar categories that are included with the theory and I have delineated those similarities and areas of convergence.

In Atrocity Speech Law: Foundation, Fragmentation, Fruition, Gregory S. Gordon offers a systematic account of how international law conceptualizes speech that contributes to mass violence.

Rather than relying on the vague and often inconsistent notion of "international hate speech law," Gordon proposes the broader framework of atrocity speech law. He uses this term to capture the set of legal rules governing speech connected to genocide, crimes against humanity, and war crimes (Gordon (2017)). I rely on this framework because it treats speech not simply as expression, but as a mechanism that can actively shape the conditions under which mass violence becomes possible.

A central component of Gordon's argument is that atrocity speech law must account for multiple forms of illicit speech, each associated with different mental states and causal relationships to violence. He identifies four distinct modalities. Incitement refers to speech that seeks to provoke genocide or mass violence, even if it does not ultimately succeed. Speech abetting, a category Gordon develops more fully than prior legal scholarship, captures rhetoric that encourages or reinforces atrocities while they are already underway, even when the speaker may not directly intend to cause them. Instigation occupies a more direct causal role. In this case, the speech not only aims at atrocity but also contributes to its occurrence, creating a clearer link between words and violence (Gordon (2017)). Ordering, the final modality, involves speech issued within a superior–subordinate relationship, such as a commander directing subordinates to commit atrocities, and may incorporate elements of both incitement and instigation (Gordon (2017)).

Gordon's work is particularly valuable for understanding how genocidal rhetoric operates in the lead-up to, and during, episodes of mass violence. By carefully distinguishing among different forms of incitement and related speech acts, he shows how perpetrators use language to normalize hatred, dehumanize targeted groups, and justify extreme violence. While most computational research on hate speech does not explicitly draw on Gordon's theory, recent work in NLP has begun to engage more directly with legal definitions of harmful speech. For example, Zufall et al. (2022) operationalize elements of the European Union's legal framework on criminal hate speech as an NLP task. They

demonstrated that aspects of legal reasoning can be translated into ML pipelines. At the same time, they emphasize the difficulty of mapping complex legal standards onto text in a way that remains both accurate and transparent (Zufall et al. (2022).

Similarly, my paper examines the different types of incitement to genocide speech identified by Benesch, Gordon and others, thus offering a comprehensive understanding of the phenomenon and its implications for genocide prevention efforts. Gordon outlines several types of incitement to genocide, which are discussed in more detail below.

2.1.2.1 Typology of Incitement Techniques

One type of incitement to genocide is accusation in a mirror. This occurs when a group makes accusations against another group, claiming that they are responsible for certain atrocities, and then uses these accusations as justification for their own acts of violence. An example of this can be seen in the Rwandan genocide, where Hutu extremists accused Tutsis of planning to exterminate the Hutu population. These accusations were then used to justify the killing of Tutsis by Hutu extremists. Benesch also identifies this technique as a hallmark of dangerous speech: she defines it as when speakers claim the target group poses an existential threat, thereby morally licensing violence as defensive action (Benesch (2021)). This is also related to a classifications in Albert Bandura's theory of moral disengagement of 'victimization' which involves viewing oneself as the victim in a situation, even when one is clearly the aggressor (Bandura (1999).

Another form of instigation to genocide is dehumanization. When a group is displayed as inferior, subhuman, or unworthy of fundamental human rights, dehumanization is taking place. Dehumanizing a group makes it simpler for others to defend violent crimes against them. Nazi Germany is a prime

example of this, as Jews were depicted as vermin or rats that had to be eradicated. Gordon categorizes dehumanization, including verminization, demonization, and pathologization, as forms of incitement to genocide. Benesch's framework similarly identifies dehumanization as one of the key "hallmarks" of dangerous speech: she notes that referring to a target as "animals, pests, or insects" makes violence psychologically easier to accept (Benesch (2021)). Albert Bandura's theory of moral disengagement also has this same classification. Dehumanization is delineated as the first mechanism of moral disengagement. Dehumanizing someone means that we deny the person a number of basic human qualities and we treat the person as if he or she is less than human. In this way, we are able to use dehumanization to rationalize doing some type of harm to the person. For instance, during times of war, soldiers often dehumanize their enemy in order to make killing or harming the enemy easier. Soldiers do so when they refer to their enemy using derogatory terms such as "rat", "cockroach", and so on.

Direct advocacy is a third type of incitement to genocide. This occurs when individuals or groups explicitly call for the extermination of another group. An example of this can be seen in the speeches made by leaders of the Khmer Rouge in Cambodia, who called for the elimination of all individuals associated with the former government. A related classification is what he termed "Predictions", which includes forecasting the elimination of the population. These predictions, exemplified by Ananie Nkurunziza's statement on RTLM, contributed to the dehumanization and targeted violence against Tutsis for example (Gordon (2017)).

Another type of incitement to genocide is glorification of past violence. This occurs when individuals or groups celebrate acts of violence against another group, often portraying the perpetrators as heroes or martyrs. It normalizes violence, erases moral ambiguity, and signals that similar acts are worthy of admiration rather than condemnation. Examples from the Rwandan genocide underscore the role of propaganda in lionizing perpetrators and normalizing genocidal acts. This can lead to an

increase in violence against the targeted group, as others are encouraged to engage in similar acts. An example of this can be seen in the glorification of Serbian paramilitary groups during the Bosnian War (Gordon (2017)).

Another form of incitement to genocide is the use of coded language. This occurs when individuals or groups use language that appears harmless on the surface, but which actually contains hidden messages that encourage violence against another group. An example of this can be seen in the Hutu Power movement in Rwanda, which used phrases like "cleansing" and "final solution" to refer to the extermination of the Tutsi population. Albert Bandura has a similar classification he termed as Euphemistic Language. It is the seventh mechanism of moral disengagement and involves using language to sanitize immoral actions and strip them of their moral implications. For example, a person may say they "terminated" an employee instead of saying they fired them. This allows them to avoid acknowledging the immorality of their actions and justify them as necessary for business or personal reasons. Other examples include words such as "cleanse", "clean the house", "remove parasites" etc.

The denial of past atrocities can also be a form of incitement to genocide. This occurs when individuals or groups deny that past atrocities have taken place, often in an attempt to justify or downplay ongoing violence against the targeted group. An example of this can be seen in the denial of the Armenian Genocide by the Turkish government (Gordon (2017)).

The justification of ongoing atrocities as necessary or humane represents another form of incitement to genocide. Nazi propagandists' attempts to portray mass murder as a defensive measure exemplify how perpetrators distort morality to justify their crimes against humanity. This classification is also present in Albert Bandura's theory which is termed social and moral justification. It involves justifying harmful actions as necessary for the greater good. For example, an army may justify exterminating a group because it is necessary for the "well being of the nation. This allows them to ignore

the harm they are causing to justify their actions as necessary for the greater good.

Another classification is Target Sympathizer Conflation where members of the majority group who help or sympathize with the targeted group are also persecuted. For example, during the Holocaust, non-Jews who hid Jews or simply opposed the genocide were murdered. In Rwanda, Hutus who opposed the genocide were labeled "traitors" and murdered.

2.1.2.2 Additional Categories and their Rationale

Existing "incitement to genocide" frameworks, particularly those outlined above (Gordon (2017)), provide a robust foundation by identifying rhetorical strategies such as dehumanization (for example, verminization, pathologization, and demonization), metaphors or euphemisms, conditional calls for destruction, and "accusation in a mirror." However, given the sociopolitical and economic dynamics specific to the Tigray war, I judged that these categories might not fully capture all relevant hate speech strategies. Therefore, I have introduced two additional labels: Disloyalty and Economic Anti-Tigrayanism to extend the taxonomy's conceptual reach.

The aforementioned literature on genocide propaganda covers many of the hallmark rhetorical elements common to mass violence, but it also suggests how new categories can emerge in specific sociopolitical contexts. Among these are economic scapegoating and framing a group as disloyal which, while overlapping with established dangerous speech frameworks, merit distinct analytical attention and categorization due to their historical resonance and frequency of use during the exploration of the dataset I worked with.

Economic resentment is a classic tool of scapegoating whereby a dominant group attributes national or economic decline to a marginalized minority. This strategy often occurs during periods of

political instability or economic collapse as explanations of hardships (Allport (1954); Staub (1989)). In the Ethiopian context, this pattern has been visible in anti-Tigrayan discourse, particularly in online and political rhetoric. Tigrayans have frequently been described as corrupt elites, illicit profiteers, or as draining national resources. These portrayals do more than express resentment: they frame the group as economically parasitic and morally suspect, positioning them as a threat to the broader public good.

Similar narratives have appeared repeatedly in other historical settings. In early twentieth-century Europe, antisemitic propaganda cast Jews as controlling financial systems, exploiting national economies, and engineering crises for personal gain, despite a lack of empirical evidence to support such claims (Herf (2006); Fine (2009)). In both cases, economic frustration was transformed into ethnic hostility through the repetition of long-standing stereotypes (Gordon (2017); Fisseha (2023)).

This mode of scapegoating aligns closely with Gordon's notion of "abetted" genocidal speech (Gordon (2017)), whereby speakers may not directly call for extermination but construct a climate in which violence becomes acceptable or even desirable. By framing the targeted group as the cause of collective suffering, economic scapegoating not only directs public frustration away from systemic issues but also legitimizes punitive responses such as mass displacement, property seizure, or physical extermination, which are all justified as necessary economic "corrections".

Similar acts took place during the war in Tigray. Many Tigrayans were forcibly removed from their homes throughout the conflict, with particularly severe patterns noted in Western Tigray. Human rights organizations and media have referred to the mass expulsion efforts carried out by local authorities and affiliated militias in that area as ethnic cleansing (Human Rights Watch (2022); Paravicini et al. (2021)).

Economic damage was also caused due to a siege enacted for long periods of time, severely limiting access to necessities like food, medicine and fuel (Ibrahim and Weis (2025)). Combined with

widespread infrastructure destruction and the looting of agricultural and commercial assets, the blockade contributed to the collapse of local economies, widespread hunger, and a humanitarian crisis affecting millions (Igoe (2021); UN World Food Programme (2022); World Health Organization (2023)).

Closely related to Economic Anti-Tigrayanism, is the theme of disloyalty which is depicting the targeted group as a "fifth column" that undermines the nation from within. This narrative is rooted in the belief that some ethnic or religious groups harbor secret allegiances that conflict with national interests (Benesch (2012)). In genocide propaganda, accusations of betrayal often precede or accompany violence. In Ethiopia, Tigrayans were frequently described in political discourse as traitors, spies, or collaborators with foreign forces. Large media outlets circulated narratives portraying Tigrayans as agents of destabilization, accusing them of disloyalty and various forms of subversion long before widespread violence broke out (van Reisen and Mawere (2024); Teklu (2022); OHCHR and EHRC (2021)). These portrayals not only foster paranoia but serve to preemptively justify repression. This type of speech positions the violence not as aggression but as necessary self-defense and it aligns directly with Bandura (1999)'s theory of moral disengagement, which highlights how in-group members come to justify violence by construing it as morally righteous and necessary for survival when the target is cast as a dangerous "other."

Moreover, Benesch and Maynard emphasize that dangerous speech often promotes fear. For example, asserting that an out-group plots an imminent attack (Benesch (2012); Leader Maynard and Benesch (2016)), making violence seem defensive. Meanwhile, the pattern of "questioning in-group loyalty" is well documented such that, in genocidal campaigns, victims are deemed dangerous and untrustworthy for the nation. These additional labels, economic scapegoating and loyalty betrayal, were chosen because they recur prominently in the discourse and map onto theorized mechanisms of group aggression. They also reflect what Staub (1989) identifies as key precursors to genocide:

difficult life conditions, culturally available ideologies that permit blame, and political leadership that exploits these grievances. Including "Disloyalty" and "Economic Anti-Tigrayanism" thus increases the ecological validity and sensitivity of my taxonomy, allowing detection of pre-violence, escalatory discourse that may function as early warning signals for mass atrocity risk.

Importantly, these categories are not novel inventions. The consistency with which economic resentment and loyalty framing appear across genocidal contexts, whether in Rwanda, Nazi Germany, or contemporary Ethiopia underscores their relevance as indicators of dangerous speech. Recognizing these patterns allows for more precise classification of genocidal rhetoric.

At the same time, I treat these labels as risk indicators rather than definitive calls for violence. Because they rely on more ambiguous or coded rhetoric rather than explicit "kill" commands, they reflect threat framing, delegitimization, and scapegoating. The complete list on the types of incitement to genocide classification and examples can be found in Appendix C.

2.1.3 Albert Bandura's Moral Disengagement Theory

Albert Bandura, a social cognitive psychologist, proposed the Moral Disengagement Theory as a means of explaining how people are able to justify morally reprehensible actions to themselves (Bandura (1999)). According to Bandura, people who engage in immoral actions often have to overcome psychological barriers in order to justify their behavior. The Moral Disengagement Theory describes the cognitive processes that individuals go through in order to disengage from moral standards and justify their actions.

The theory posits that there are eight mechanisms through which people can morally disengage. These mechanisms include dehumanization, social and moral justification, victimization, attributing

blame, displacing responsibility, minimizing or disregarding injurious effects, euphemistic language, and advantageous comparison. Each of these mechanisms serves as a psychological barrier to moral action and helps individuals to justify their immoral behavior.

Dehumanization is the first mechanism of moral disengagement. When we dehumanize someone, we strip them of their inherent human qualities and treat them as less than human. This allows us to justify harmful actions against them. For example, during war, soldiers may dehumanize their enemies by calling them “rats” or “cockroaches”. This makes it easier for them to kill or harm their enemies because they no longer see them as fellow human beings deserving of respect and dignity.

Social and moral justification is the second mechanism of moral disengagement. This involves justifying harmful actions as necessary for the greater good. For example, a company may justify polluting the environment because it is necessary for economic growth. This allows them to ignore the harm they are causing to the environment and justify their actions as necessary for the greater good (Bandura (1999)).

Victimization is the third mechanism of moral disengagement. This involves viewing oneself as the victim in a situation, even when one is clearly the aggressor. For example, a bully may view themselves as the victim of bullying and justify their actions as necessary for self-defense. This allows them to avoid taking responsibility for their actions and justify their behavior as necessary for their own survival (Bandura (1999)).

Attributing blame is the fourth mechanism of moral disengagement. This involves blaming others for one’s own immoral behavior. For example, a person may blame their abusive behavior on their partner’s provocation. This allows them to avoid taking responsibility for their actions and justify their behavior as a response to someone else’s actions.

Displacing responsibility is the fifth mechanism of moral disengagement. This involves shifting

responsibility for one's immoral behavior onto someone or something else. For example, a person may blame their addiction for their immoral behavior. This allows them to avoid taking responsibility for their actions and justify their behavior as a consequence of something beyond their control.

Minimizing or disregarding injurious effects is the sixth mechanism of moral disengagement. This occurs when individuals acknowledge that harm exists but downplay its seriousness. Damage is framed as small, unavoidable, or exaggerated. In practice, this allows people to avoid taking responsibility for the outcomes of their behavior. For example, organizations may admit that their actions cause harm, but characterize those effects as minimal or necessary in order to justify continuing as before.

Another mechanism operates through euphemistic language. In these cases, morally troubling actions are described using neutral or sanitized terms that remove their emotional and ethical weight. This linguistic distancing makes it easier to discuss harmful actions without fully confronting their moral implications, thereby reducing feelings of guilt or accountability (Bandura (1999)).

Advantageous comparison represents a further way moral standards are weakened. Rather than denying wrongdoing, individuals compare their behavior to actions they perceive as worse ("whataboutism"). A person may admit to cheating, for instance, but justify it by arguing that others cheated more often or more seriously. Through this comparison, the original behavior appears less troubling, and self-condemnation is reduced. The action remains unethical, but it no longer feels worthy of moral concern.

In conclusion, the moral disengagement theory provides a valuable framework for understanding how individuals can justify immoral behavior and disengage from their moral principles. By recognizing these mechanisms, individuals can become more aware of their own thought patterns and make more ethical decisions. The theory also highlights the importance of promoting a moral community that supports ethical behavior and encourages individuals to engage with their moral principles. Through education and awareness-raising we can create a more ethical society and prevent harmful behavior.

Utilizing Bandura's theory, a study titled "Toward Transformer-Based NLP for Extracting Psychosocial Indicators of Moral Disengagement" (Friedman et al. (2021)) presents a new approach to identify and extract indicators of moral disengagement in text using transformer-based NLP techniques. The article proposes a transformer-based NLP architecture that extracts psychosocial indicators of moral disengagement, such as dehumanization, euphemistic labeling, and displacement of responsibility, from text data. The proposed architecture proved capable of jointly identifying and extracting variables or factors described in language, qualitative causal relationships over these variables, and qualifiers and magnitudes that constrain these causal relationships. The approach involves the use of a novel knowledge graph schema and dataset that capture indicators of moral disengagement in text. The transformer-based NLP model is trained on this dataset to extract relevant psychosocial indicators of moral disengagement.

This study is highly relevant to my work as it provides some theoretical reinforcement and also methodological support for identifying hate-related psychological mechanisms in language. Friedman et al. (2021) operationalize Albert Bandura's theory of moral disengagement. These same mechanisms underlie many of the rhetorical strategies captured in my hate speech taxonomy for the Tigray conflict, such as dehumanization (e.g., verminization, pathologization), justification of violence, and coded euphemisms which are categories from Gordon (2017)'s theory that I aimed to operationalize. The significance of their work lies in demonstrating that psychological constructs of harm and incitement which are often discussed abstractly can be systematically extracted and modeled from naturalistic text data using transformers.

2.1.4 Dangerous Speech Theory

Benesch's paradigm of Dangerous speech framework has five constructs that influence the dangerousness of communication (Benesch (2021)). These factors include the speaker's power over the audience, the audience's receptiveness, the meaning of the speech such as a call to violence, the enabling social and historical backdrop, and the persuasive medium of dissemination.

Furthermore, the framework employs three indicators to identify dangerous speech: dehumanization of the target group through the use of an animal name or dog-whistling, implying that the in-group faces a grave threat from the out-group, and implying that elements of the out-group are tainting the in-group's purity or integrity (Benesch (2021)). These classifications have similarity with the theories of Bandura and Gordon. Table 2.1 illustrates this.

2.2 NLP Studies on Dehumanization

In a recent study, Saffari et al. (2025) take on a problem that has received surprisingly little focused attention in natural language processing research: the detection of dehumanizing language. Much of the existing work on hate speech detection has concentrated on more overt expressions of hostility, such as insults, slurs, or explicit threats. Dehumanization, while more subtle, is dangerous because it involves denying humanity to individuals or groups, a form of rhetoric that has been deeply implicated in historical and contemporary violence (Saffari et al. (2025)).

The authors conducted a systematic evaluation of four state-of-the-art large language models (LLMs): Claude, GPT, Mistral, and Qwen. They assessed how well each model can detect dehumanization in text. The experiments reveal a stark disparity in performance: only one model (Claude), under an optimized configuration, attains an F1 score exceeding 80%. The other models do worse,

Gordon's Classification	Similarity with other theories	Definition	Examples
Verminization	Dehumanization (Albert Bandura)	Classifies the target as something "whose extermination would be considered normal and desirable" (cockroaches, lice, rat, pest)	"TPLF is shell. Inside the shell live worms; Agame are lice-infested snakes."
Pathologization	Dehumanization Impurity (Susan Bench)	Expropriates pseudo-medical terminology/disease to justify massacre	"Filthy Agame. Clean our house of Agame."
Demonization	Dehumanization (Albert Bandura)	Centering on devils, malefactors, and other incarnations of "evil"	"Winning this war is winning against satan."
Direct Advocacy	-	Direct calls for destruction. Use words like loot, discriminate, put in concentration camp, beat, evict, kill, destroy	"Press forward until Tigrayans are wiped out."
Glorification of Past Violence	-	Glorifying past violence against the target	"The Derg was correct to remove Tigrayans."
Accusation in a mirror	Victimization (Albert Bandura)	Accuses the target of something that the perpetrator is doing or intends to do. Legitimizes retaliation in kind. Aggression as self defense. "Do unto others, as they would do unto you"	"If we don't win this war, they will kill us. Are we waiting until they kill us?"
Disloyalty	Threat Construction	Expression of mistrust, treachery and disloyalty of the group to the state. Fifth columnist, impure, foreign, infiltrator, barbarian.	"She couldn't hide her treacherous Agame ways.", "Agames are puppets of the West."
Justification	Justification	Justifying atrocities, blaming a group for their lot. Just deserts.	"Tigrayans support TPLF, so they should all be punished."
Economic Anti-Tigrayanism	-	Poor and thus useless or greedy and have stolen money and resources from the state.	"Tigrayans have stolen billions from our country."
Target-sympathizer conflation	-	Those who sympathize with target groups are also persecuted. In-group loyalty, members of dominant group who sympathize	"At such a time, the country is hurt by 'middle settlers.'"

Table 2.1: Classifications Used and Similarities to Other Theories.

particularly when challenged to distinguish dehumanizing language from more general hateful but non-dehumanizing content (Saffari et al. (2025))

A key contribution of their work is demonstrating that distinguishing dehumanization from related hate speech is particularly hard. Many hateful expressions contain insulting or derogatory language, but not all dehumanizing content is overt in that, some is metaphorical or context-dependent. The LLMs struggle especially when asked to make this fine-grained distinction (Saffari et al. (2025)).

The work of Saffari et al. (2025) builds directly on the foundation laid by Mendelsohn et al. (2020) who developed one of the first computational linguistic frameworks for dehumanization. Mendelsohn et al. (2020) identify key psychological components of dehumanization such as negative evaluation, denial of agency, moral disgust, and metaphors likening people to vermin and propose linguistic correlates for those components. They then apply their framework to analyze how LGBTQ people are represented in the New York Times over time, finding significant changes in how dehumanizing language is deployed (Mendelsohn et al. (2020); Haslam (2006); Goff et al. (2008); Kteily et al. (2015), Sherman and Haidt (2011)).

Their categorization for dehumanization of the NLP detection were as follows.

2.2.1 Negative Evaluations of Others

Prior work describes "extremely negative evaluations of others" as a major component of dehumanization (Haslam (2006)). Attribution of negative characteristics to members of a target group in order to exclude that group from "the realm of acceptable norms and values" is specifically the key component of delegitimization, a process of moral exclusion closely related to dehumanization (Bar-Tal (1990)).

Mendelsohn et al. (2020) calculated the valence of a text to analyze this variable. Valence cor-

responds to an individual's evaluation of an event or concept, ranging from negative/unpleasant to positive/pleasant (Russell (1980)). It involves calculating the average valence of words occurring in discussions of the target group. Words with the highest valence include love and happy, while words with the lowest valence include nightmare. They used paragraphs as the unit of analysis citing that a paragraph represents a single coherent idea or theme (Hinds (1977)).

2.2.2 Denial of Agency

Another component of dehumanization is the denial of agency to members of the target group (Haslam (2006)). According to Tipler and Ruscher (2014) there are three types of agency: the ability to (1) experience emotion and feel pain (affective mental states), (2) act and produce an effect on their environment (behavioral potential), and (3) think and hold beliefs (cognitive mental states). Dehumanization typically involves the denial of one or more of these types of agency (Tipler and Ruscher (2014)).

Mendelsohn and colleagues used Sap's extension of Connotation Frames for agency (Sap et al. (2017); Mendelsohn et al. (2020)). This lexicon considers the agency attributed to subjects of nearly 2,000 transitive and intransitive verbs. They also used the NRC VAD lexicon's dominance dimension which appeared to capture signals of denial of agency. They also used the NRC VAD lexicon's dominance dimension, which is a psychological affective resource that assigns valence (pleasantness), arousal (emotional intensity), and dominance (control) scores to over 20,000 English words. It is widely used in emotion analysis, including hate speech detection, to quantify the emotional tone of text (Mohammad (2018)). The highest dominance words are powerful, leadership, success, and govern, while the lowest dominance words are weak, frail, empty, and penniless (Mendelsohn et al. (2020)).

2.2.3 Moral Disgust

Disgust underlies the dehumanizing nature of metaphors and is itself another important element of dehumanization (Mendelsohn et al. (2020)). Disgust contributes to members of target groups being perceived as less-than-human and of negative social value (Sherman and Haidt (2011)). Buckels and Trapnell (2013) find that priming participants to feel disgust facilitates "moral exclusion of out-groups". Experiments by Sherman and Haidt (2011) and Hodson and Costello (2007) similarly find that disgust is a predictor of dehumanizing perceptions of a target group. Both moral disgust toward a particular social group and the invocation of non-human metaphors are facilitated by essentialist beliefs about groups, which Haslam (2006) presents as a necessary component of dehumanization.

Moral Foundations theory postulates that there are five dimensions of moral intuitions: care, fairness/proportionality, loyalty/ingroup, authority/respect, and sanctity/purity (Haidt and Graham (2007)). The negative end of the sanctity/purity dimension corresponds to moral disgust.

The dictionary for moral disgust includes over thirty words and stems, including disgust, sin, perversion, and obscene.

Mendelsohn et al. (2020), using word embeddings, constructed a vector to represent the concept of moral disgust by averaging the vectors for all words in the "Moral Disgust" dictionary, weighted by frequency. They identified implicit associations between a social group and moral disgust by calculating cosine similarity between the group label's vector and the Moral Disgust concept vector, where a higher similarity suggests closer associations.

2.2.4 Vermin as a Dehumanizing Metaphor

Metaphors and imagery relating target groups to vermin are particularly insidious and played a prominent role in the genocide of Jews in Nazi Germany and Tutsis in Rwanda (L. T. Harris and Fiske (2015)). According to Tipler and Ruscher (2014) the vermin metaphor is particularly powerful because it conceptualizes the target group as "engaged in threatening behavior, but devoid of thought or emotional desire." As with moral disgust, they created a Vermin concept vector by averaging the following vermin words' vectors, weighted by frequency: vermin, rodent(s), rat(s) mice, cockroaches, termite(s), bedbug(s), fleas (Mendelsohn et al. (2020)).

The newer studies by Saffari et al. (2025) can be seen as extending this work: where Mendelsohn et al. (2020) provided a theoretical and empirical toolkit for identifying dehumanization in media, Saffari et al. (2025) tested whether modern LLMs can operationalize that toolkit in automated detection. In doing so, they expose the gaps between psychological theory, media analysis, and the practical capacity of NLP systems. Their evaluation highlights where current LLMs fall short of the rich, multidimensional conceptualization of dehumanization that Mendelsohn et al. (2020) laid out (Saffari et al. (2025)).

2.3 The War in Tigray and the Proliferation of Hate Speech

Total war justifies the physical elimination of entire peoples since the destruction of the opponent entails not only the destruction of its armed forces but also of its society Förster (2007); Bicheno (2001) defines total war as a conflict where the entire population and resources of the combatants are committed to complete victory, which could potentially involve the destruction of the opponent's society. Straus (2012) also states that the study of genocide should be embedded in a broader study

of political violence, highlighting the need to examine the causes and logic of genocide in relation to group destruction. The violence of genocide is obscured and covered over by the violence of war (Straus (2012); Wyschogrod (2005)). While a country is at war, it is common for military activities to be justified by the idea that they are defending the country from external threats. Nevertheless, this justification may easily be extended to include the need to defend the country from internal threats (Bicheno (2001); Straus (2012)).

On November 4, 2020, the Tigray People's Liberation Front (TPLF) was accused of attacking the Ethiopian National Defence Forces (ENDF) in Tigray, the northernmost region composed mostly of ethnic Tigrayans which make up 5.7% (6 million) of the 121 million Ethiopian population (Unknown (n.d.)). TPLF leadership called the attack preemptive, stating they acted because the ENDF was preparing an attack (Wilmot et al. (2021)). As a response, the Ethiopian government launched what it called a "law enforcement operation" and declared a state of emergency in Tigray (Wilmot et al. (2021)). The same day, the federal government shut down internet and telecommunications across the region, creating an information vacuum (Wilmot et al. (2021)). Journalists, aid groups, and researchers have struggled to access the region since (Wilmot et al. (2021); Byaruhanga (2022)). Meanwhile, humanitarian agencies report that post-conflict aid efforts have been fraught with challenges. In 2023, investigations uncovered large-scale diversion of food aid, leading the UN World Food Programme and USAID to suspend aid deliveries nationwide; this suspension of aid (from mid-2023 until partial resumption in 2024) further exacerbated hunger in war-torn regions (Human Rights Watch (2024), Refugees International (2024)).

Amid the information and access constraints, competing information campaigns emerged to frame the conflict, including a deluge of hate speech and dehumanization (Associated Press (2021); United Nations (2022)). For example, the United States Agency for International Development warned

of dehumanizing rhetoric used by Ethiopian leadership, who have described Tigrayans as "cancers" and "weeds." (Anna (2021), Genocide Watch "Ethiopia" section). An investigation on the ill use of Facebook in developing countries revealed that groups associated with the Ethiopian government and state media incited violence against Tigrayans by using dehumanizing terms like "hyenas" and "cancer" to describe the ethnic minority group. Furthermore, the European Union special envoy Pekka Havvisto stated that Ethiopian leaders vowed to "wipe out the Tigrayans for 100 years" (Anna (2021)). As stated above, dehumanizing metaphors, such as referring to a group as a disease or invasive species, can lead others to develop hostile feelings towards that group of people and see them as "less human." (Benesch (2012)).

On November 3, 2021, Facebook removed a post from Ethiopia's Prime Minister Abiy Ahmed (where he called on citizens "to take up arms"), stating it had violated the platform's policies against inciting violence (BBC News (2021)). In disclosures made to the US Securities and Exchange Commission (SEC) and provided to Congress, it was revealed that armed groups in Ethiopia were using the platform to incite violence against ethnic minorities like Tigrayans (Mackintosh (2021)). The report identified a cluster of accounts affiliated with the militia group "Fano", an ethnic Amhara militia group accused of committing ethnic cleansing against Tigrayans in Western Tigray, including some based in Sudan (Human Rights Watch (2022)). They used their platform to "seed calls for violence", promote armed conflict, recruit, and fundraise.

Frances Haugen, a former data scientist at Facebook, told members of a Senate subcommittee that Facebook is "literally fanning ethnic violence" in Ethiopia (Akinwotu (2021)). As of late 2021, "Facebook still lacked misinformation and hate speech classifiers in Ethiopia, critical resources deployed in other at risk countries" (CFR Editors (2022)). Posts that often go unchecked include rhetoric that call for mob violence, ethnic clashes, and attacks against journalists and government representatives. In

one recent instance, an inflammatory Facebook post quickly got hundreds of shares and likes, and a day later the village cited in the post was ransacked, burnt to the ground, and its inhabitants murdered (Hale (2021)). The post remains on Facebook despite being reported by users.

Additionally, Melaku Alebel Addis, the Minister of Industry for Ethiopia, glorified past violence against Tigrayans to hundreds of thousands of his followers on Twitter. Viral Facebook posts accusing Tigrayans of past crimes and their associated guilt, as well as demonization by the federal government which depicted Tigray as the devil by putting an image of the Tigray regional flag have also been rampant. These disturbing instances raise concern, as the mass arrest, disappearance, and ethnic profiling of Tigrayans, including children under three years of age, have been widespread across Ethiopia since the beginning of the war (UN News (2021); Associated Press (2021)).

Even prior to the war, a study using different ML algorithms regarding hate speech in Ethiopia found that Tigrayans were disproportionately targeted by hate speech and were most vulnerable among the populations on the receiving end of the attacks (Mossie and Wang (2020)). Moreover, recent findings by the Distributed AI Research Institute (DAIR), which annotated 340 X (formerly Twitter) posts drawn from a dataset of 5.5 million. The researchers discovered that many tweets employed verminizing metaphors. For instance, one translated post read "Clean the cockroaches"□ a phrase deeply tied to historical dehumanizing propaganda against Tigrayans (The Distributed AI Research Institue (2025)). Identifying that post as genocidal content required knowledge of how Tigrayans had been referred to as "cockroaches" in other, state-aligned political rhetoric (The Distributed AI Research Institue (2025)). Such metaphors align closely with Gordon's verminization technique, where a target group is likened to pests that should be eradicated (Gordon (2017)).

The same DAIR study revealed numerous tweets denying or minimizing violent events. For example, a post using the hashtag #FakeAxumMassacre denied a massacre that was well documented by

Amnesty International, Human Rights Watch, and other sources (The Distributed AI Research Institue (2025)). This form of violent event denial is consistent with Gordon's notion of denial as an incitement modality: by denying atrocities, perpetrators reduce moral constraints and delegitimize the suffering of the victim group (Gordon (2017)).

Online rhetoric during the war also revealed attempts to morally justify violence. According to Amnesty International, Facebook's algorithm amplified posts portraying Tigrayans as a destabilizing threat to Ethiopia's security, framing them as dangerous or disloyal (Amnesty International (2023)). This justification mirrors Gordon's "justification during contemporaneous violence": violence is cast as defensive or necessary, rather than barbaric or criminal (Gordon (2017)).

Beyond organic user posts, there were coordinated tweet campaigns. Shortly after Internet and telecommunication access in Tigray was shutdown, diaspora groups launched "click-to-tweet" sites with pre-written messages designed to influence the narrative (DFRLab (2021)). These campaigns sought to mobilize international attention and framed Tigrayans in particular ways, shaping both global and domestic perceptions. Such orchestrated messaging has parallels with instigation in atrocity-speech law: using speech not just to degrade, but to mobilize and justify violence (Gordon (2017)).

Thus, the Tigray conflict (2020–2022) provides a powerful case study of how atrocity speech including dehumanizing rhetoric, denial, and calls for violence can proliferate on social media and contribute to real-world harm. Examining tweets and online content during this war can help illustrate the theoretical concepts in Gregory S. Gordon's atrocity-speech typology can map onto real, contemporary conflict dynamics, and underscores the need for sophisticated content-detection systems. In sum, I have sought to test if a nuanced and theory-based psychosocial feature set improves machine classification of genocidal speech from social media. Among the specific inquiries are the following:

Research Question 1. Can a multi-class ML classifier (built via theoretical lens) accurately

distinguish between incitement techniques in text, and how will it perform (precision, recall, F1, etc)?

To answer this, I developed a taxonomy of online genocidal speech comments using both qualitative open coding and the social and legal theories summarized above. The open coding technique is a coding method where the classes emerge from the material (Glaser and Strauss (1967)). After generating these taxonomies/classifications, I annotated data using this taxonomy and trained ML models. This way, I used a more nuanced categorization for extreme types of hate speech that could lead to mass atrocities and genocide.

Research Question 2. How do human participants psychologically respond to text labeled by the model as different types of atrocity speech? Do different hate speech types evoke different reactions?

To answer this, I sent out surveys to 30 Tigrayan Americans with different types of hate speech based on the taxonomy developed and they were asked to rate those tweets (which were classified by the ML model) on different dimensions such as perceived threat, perceived violence, emotional reactions etc. This way, we can test if there are actual perceived differences among different types of hate speech versus if there is overlap that can better be captured with more broader classifications/labels.

Chapter 3: Introduction

3.1 Methodology

This chapter explains the process of developing the dataset and outlines the steps of preprocessing the data and training and model performance.

3.1.1 Research Question

The first aim of this study was to determine how we can use legal, social psychological theories and linguistic markers to classify hate speech using NLP models. I aimed to train and assessed its performance on multi-class classification data.

3.1.2 Data Source and Collection Approach

To build a corpus of tweets relevant to the Tigray war, I used a keyword-based streaming strategy via the Twitter API, specifying a domain-specific set of slurs, insults, and hate terms that pertain to ethnic conflict. Keyword-based filtering (also called "keyword streaming") involves defining a list of terms (words or hashtags), and the API returns all tweets containing those terms in real time. This method is common in social media research examining crisis discourse and hate speech (Khazaie et al. (2021)).

I curated a set of 55 keywords tailored to the Tigray conflict. These included derogatory expressions, context-specific slurs, and politically charged insults that have been used in online discourse about Tigrayans. To identify the most relevant terms, I cross-referenced slang, local-language insults, and conflict-specific hate vocabulary with prior research. In particular, twenty of the terms (see Appendix A and Table C.1) came from a PeaceTech Lab study that enumerated inflammatory words and phrases used in Ethiopia during 2020, based on survey responses about language perceived to be par-

ticularly provocative (PeaceTech Lab (2020)).

Using this keyword list, I collected tweets from November 2021 to November 2022, a period which aligns with an intensification of the conflict and ends shortly after a permanent cessation of hostilities agreement was signed. On 2 November 2022, Ethiopian government forces (ENDF) and the TPLF agreed to a peace deal mediated by the African Union in Pretoria. This accord formally established a "permanent cessation of hostilities" (African Union (2022); Government of FDRE and TPLF (2022)).

Over this 12-month period, I collected approximately 1.75 million tweets, including both original tweets and retweets. For each tweet, I stored not only the text but also metadata: timestamp, tweet ID, whether the tweet was a reply or not, reply-to-user data, user ID, follower/following counts, user profile creation date, and geolocation data (when available). This approach is useful since by focusing on a carefully curated keyword list, I have ensured that the data collected is relevant to hate and conflict discourse, rather than generic or irrelevant tweets. The use of conflict-specific slurs helps to surface the most dangerous and dehumanizing rhetoric. Moreover, incorporating terms identified in the PeaceTech Lab study anchors the data collection in the local linguistic context, rather than relying solely on global or English-centric hate speech lexicons. This increases the likelihood of capturing contextually meaningful, high-risk speech. The chosen time frame (Nov 2021 – Nov 2022) captures a critical phase in the conflict, including the buildup to and signing of the Pretoria peace deal (Government of FDRE and TPLF (2022)). This enables analysis of how hate speech evolved around key political events. Additionally, collecting metadata along with text allows for multi-dimensional analysis: not just what was said, but by whom, when, and in what social context (e.g., influential users, network dynamics, retweet behavior). This data, while not used in this analysis will be made available for later research and researchers to expand on.

But there are some limitations to note.

For example, even with 55 terms, I may not have capture all hateful or dehumanizing speech, especially coded or emergent slurs not included in my list. And, given that my data mostly comes from Twitter, Twitter's API has rate limits and sampling biases, which may have excluded some content. Lastly, some hate speech used local languages, transliteration, or code switching that is difficult to enumerate exhaustively. Attempts to use Tigrinya and Amharic (languages native to Ethiopia) for ML have proven a bit more difficult.

3.1.3 Annotation Protocol

To create a high-quality, labeled dataset for detecting genocidal and dehumanizing speech, I developed a detailed annotation protocol grounded in the theoretical framework laid out in the literature review (Chapter 2).

Drawing on scholarship about genocidal hate speech, dangerous speech, and dehumanization, I designed guidelines that clearly define each category of interest, provide concrete examples, and instruct annotators on how to apply them. Initially, a qualitative analysis was conducted to examine theories on dehumanization, dangerous speech and genocidal speech.

Content analysis was used to discover essential terms from multiple definitions of hate speech, incitement to violence and existing theories of dehumanization, which guided the development of the study's conceptual framework. The annotation technique for guiding the team of human raters to accurately categorize the text for training emerged from this section.

Based on this annotation guideline, human annotators annotated by identifying the high-level features. The annotation guide helped annotators catch latent traits, such as "othering" language, which

has previously been found to be useful in collecting subtle kinds of hate speech (Alorainy et al. (2019)).

3.1.4 Annotator Recruitment and Training

I recruited five annotators fluent in English, Amharic, and Tigrinya. Before the annotation began, each was informed that the task would involve reading offensive and hateful content, which could be disturbing. The annotators received a comprehensive annotation guide (see Appendix D), which defined each category (e.g., "hate speech", "non-hate," specific incitement types), listed key tokens to watch for, and included sample annotated tweets to model the process.

To ensure consistency and reliability, I conducted a structured training phase. Annotators participated in interactive calibration sessions where they labeled a shared set of practice tweets, discussed their rationale, and resolved disagreements. I iteratively refined the guidelines based on these sessions. This training and calibration process are critical for ensuring the labeled dataset's validity and robustness.

3.1.5 Annotation Process

The data to be annotated was split into two spreadsheets: one containing English tweets, and the other containing tweets in Fidel script (Amharic and Tigrinya). Each annotator labeled approximately 2,000 tweets in each spreadsheet, for a total of around 4,000 tweets per annotator. During annotation, annotators first determined whether a tweet contained hate speech at all. If a tweet was classified as non-hateful ("No Hate"), it was not assigned a more specific incitement category. For hateful speech, annotators assigned one of ten incitement category codes, based on the typology drawn from atrocity-speech law and dangerous-speech frameworks.

3.1.6 Data Cleaning and Final Corpus

The final dataset consisted of English-language social media posts collected from conflict-monitoring initiatives and online platforms where hate speech proliferated during a mass atrocity risk period. Posts were aggregated from Twitter hashtags, and forums related to an ethnic conflict (e.g., mereja.com), yielding roughly 9,000 posts after cleaning (13k before cleaning). Each post was annotated by trained coders for hate speech presence and categorized into a detailed taxonomy of incitement types. The annotation schema was derived from atrocity speech research and dangerous speech frameworks, capturing subtle rhetorical strategies used to instigate violence. If a post was deemed non-hateful, it was labeled "No Hate" (and excluded from category classification). Hateful posts were assigned one of 10 category codes corresponding to specific incitement tactics. The categories included:

- Quotation/Explanation (Q/E): Not hate speech. These tweets usually cite or explain another source's inflammatory words (often to frame or justify an argument).
- Direct Advocacy (DA): Overt calls to harm or discriminate against the target group (e.g., explicit exhortations of violence or exclusion).
- Verminization (V): Describing the target as vermin or pests, a dehumanization subtype (e.g., calling a group "rats" or "cockroaches").
- Demonization (DM): Portraying the target as evil, demonic, or satanic to sanction extreme measures.
- Pathologization (P): Depicting the target as a disease or pathology afflicting society.
- Disloyalty (DS): Accusing the target group of treason, betrayal or lack of allegiance, to incite

distrust or punishment.

- Economic Anti-Tigrayanism (EAT): Scapegoating the target for economic woes (e.g., claiming the group controls resources or is responsible for unemployment).

- Accusation in a Mirror (AM): Falsely accusing the target of plotting the same violence that the speaker intends to carry out. This genocidal technique turns reality on its head, painting the victims as aggressors to justify "preemptive" violence by the in-group.

- Target-Sympathizer Conflation (TSC): Blending target groups with their sympathizers or allies, so that anyone showing empathy toward the victim group is labeled as part of the enemy. This extends hate to those who might defend or humanize the target.

- Other: Hateful content that fit none of the above uniquely (a catch-all for rare or uncategorizable cases). This category was eventually used to bundle categories that were too infrequent to make the classification and training process a bit better.

Each post received a binary hate speech label and, if hateful, one of the above categories. For quality control, all 5 annotators reviewed some posts, resolving disagreements via discussion with the lead researcher. Posts labeled "No Hate" or irrelevant were set aside from the categorization task to focus the model on distinguishing specific hate subtypes.

The final labeled dataset was class-imbalanced: the Q/E category was most frequent ($\sim$61% as many users quoted hateful propaganda), while minor categories like Target Sympathizer Conflation and Other constituted <5%. To mitigate extreme imbalance, I merged extremely rare categories and used the "Other" label for any combined minor classes. For example, two infrequent categories were merged into "Other" (Category 8 in the schema) prior to modeling. This yielded 10 final categories (including

Category (Code)	Description (Summary)	Count	Percentage of Dataset
Q/E (0)	Quotation/Explanation (no direct hate)	5,689	61.4%
Verminization (2)	Dehumanizing as vermin/insects	955	10.3%
Disloyalty (5)	Accusation of treason/treachery	754	8.1%
Direct Advocacy (1)	Explicit calls for violence	619	6.7%
Demonization (3)	Portraying as evil/demonic	400	4.3%
Accusation Mirror (7)	False accusation to justify attack	308	3.3%
Target-Sympathizer (9)	Attacking sympathizers ("traitors")	188	2.0%
Economic Anti-Tigr. (6)	Economic insults/scapegoating	162	1.7%
Pathologization (4)	Comparing to disease/virus	112	1.2%
Other (8)	Justification or praise of past violence	79	0.9%

(N = 9,266 tweets; categories are primary labels.)

Table 3.1: Category Distribution in Annotated Dataset

Other) numbered 0-9 for modeling convenience. This taxonomy was developed via open coding and informed by existing frameworks on incitement to genocide (e.g. Gordon (2017)), dangerous speech (Benesch (2012)) and others mentioned above.

As shown in Table 3.1, the majority (61.4%) of tweets fell under Quotation/Explanation (Q/E) non-hate tweets, indicating that a large portion of the dataset consists of tweets quoting hateful content or providing context without directly using hate speech. The remaining 38.6% of tweets were annotated with one of the explicit hate categories. Among the hateful categories, Verminization (10.3%) and Disloyalty (8.1%) were the most common, reflecting frequent use of dehumanizing slurs (e.g., "cockroaches," "worms") and accusations of treason against the target group. Direct Advocacy (6.7%), which captures direct calls for harm such as "kill" or "wipe out," was relatively rare but present, con-

sistent with prior observations that overt incitement is less common than implicit hate. Demonization (4.3%) and Accusation in a Mirror (3.3%) were moderately represented, while Target-Sympathizer Conflation (2.0%), Economic Anti-Tigrayanism (1.7%), and Pathologization (1.2%) appeared infrequently. The Other category (combining Justification and Praising Past Violence) was the smallest (<1%), indicating that explicit justification of ongoing atrocities or glorification of past genocide, though observed in the data, were very rare. These figures highlight a highly imbalanced dataset, with an extreme skew toward the neutral Q/E class and very few examples of some dangerous speech types (e.g., Other, Pathologization). This class imbalance posed a significant challenge for model training and evaluation, as discussed below.

Because the data I obtained came from social media platforms, the raw texts usually contain a lot of noise, such as punctuation and URLs that will affect my experiment results. I pre-processed my text data at the very first step. Given that Tigrinya and Amharic languages have their own idiosyncrasies, they were pre-processed a bit differently than my English corpora.

However, since the analysis used HateBERT, a variant of BERT- a modern transformer based model, it requires less aggressive text preprocessing than traditional ML models (Qiao et al. (2019)). I aimed for minimal cleaning such as lowercasing the text, and remove only truly irrelevant noise, keep punctuation, stopwords, and other context-providing tokens so HateBERT can fully utilize them. The goal was to clean the noise while preserving the context.

The preprocessing involved several steps.

Special tokens were handled by replacing the user's username with a [USER] token, and URLs with [URL]. The dataset was loaded using chardet, to ensure proper encoding. Missing *HATESPEECH* values were replaced with "NO", and missing *EXPLANATION* values were filled with "No Explanation Provided". Rows with missing TEXT values were discarded (meaning there was no tweet or it was

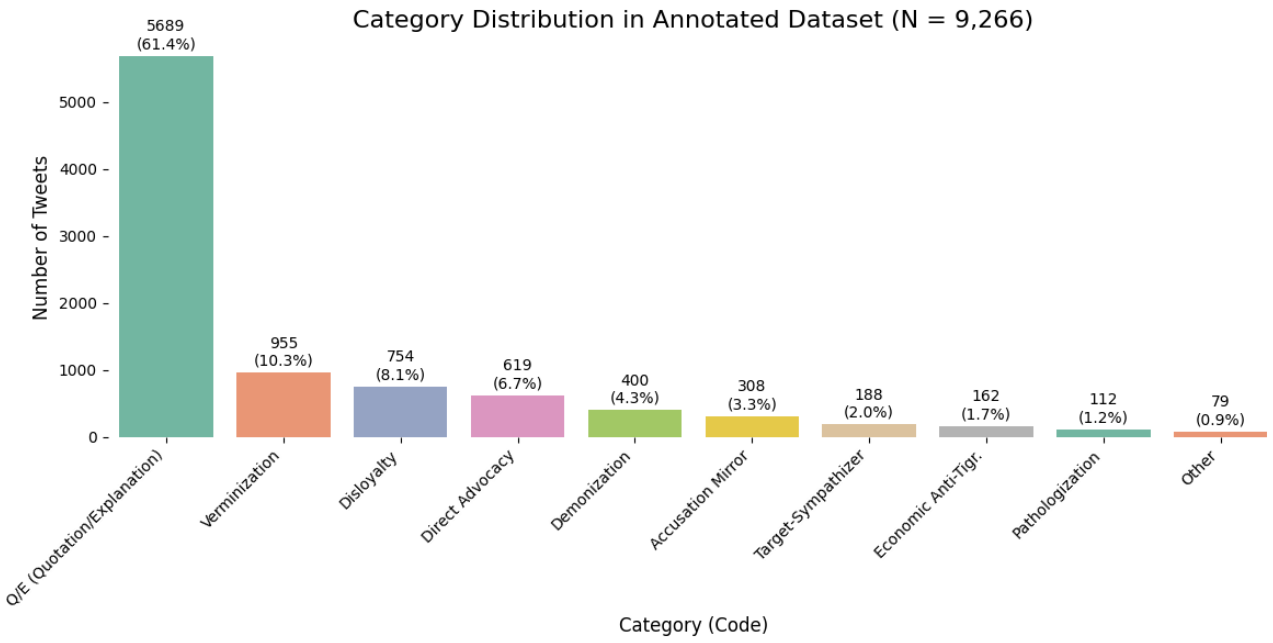

Figure 3.1: Distribution of Hate Speech Categories in the Dataset

just a character). Finally, underrepresented categories (TSC and other categories 8 and 9) were merged into a single "Other" category.

Figure 3.1 shows the distribution in the dataset and that the Q/E category is a majority class.

3.1.7 Single-label vs. Multi-label Categorization

In designing the classifier, I chose a single-label (multi-class) scheme where each tweet was assigned exactly one hate category rather than allowing multiple labels per tweet. Initially during annotation, I tried multi-label coding and observed specific issues with it. Annotators often hesitated to assign multiple categories or disagreed on how many categories applied. This aligns with prior findings that requiring multiple categories can overwhelm annotators and inflate "agreement by chance" (Pavlopoulos et al. (2022); Davani et al. (2023)). For example, one tweet could arguably fit both "dehumanizing" and "calls-for-violence," but asking coders to check both boxes led to inconsistent labeling. Consequently, I instructed annotators to select the single best-fitting category. This introduced some ambiguity, but

it avoided annotation deadlock. The practical issues I encountered and the associated disagreement, ultimately informed my decision to stick with single-label annotation.

While hate speech can overlap (e.g., one post could be both direct advocacy and verminization), recent surveys note that most classifiers still use single-label schemes even though practice may require overlapping subtypes (Tsoumakas and Katakis (2007); M.-L. Zhang and Zhou (2013)). Multi-label classification, which assigns all applicable categories to each instance, can better capture these overlaps but substantially increases modeling complexity (Kharde and Sonawane (2016)). Multi-label tasks often require transforming the problem (e.g., one-vs-rest binary classifiers or power-set methods) and typically need larger datasets (compared to what my study had) to adequately model label combinations (Tsoumakas and Katakis (2007); M.-L. Zhang and Zhou (2013)); Kharde and Sonawane (2016)). Therefore, I opted for a simpler single-label approach, following common practice that takes the most salient label per item, such as using majority vote when multiple annotations exist (Davani et al. (2023)). While this ignores some nuance, it greatly simplifies training and evaluation since standard multi-class classifiers and metrics can be used directly. I recognize this as a limitation, but at the time it was the most practical choice given my data size and annotation complexity.

Choosing single-label classification means potentially discarding overlapping signals. However, multi-label schemes have their own pitfalls. Allowing multiple labels during annotation can increase apparent agreement but complicates the interpretation of ground truth (Pavlopoulos et al. (2022)). For instance, if two annotators each see a different valid category and both labels are allowed, agreement on any single label may paradoxically drop, even if both partially agree on the content (Pavlopoulos et al. (2022); Kenyon-Dean et al. (2020)). Multi-label annotation also complicates inter-annotator agreement metrics since traditional measures like Cohen's Kappa are not straightforwardly applicable (Kenyon-Dean et al. (2020)). Given my limited annotator pool and the inherently subjective nature

of the categories, I found that forcing a single label per tweet produced more consistent coding. In effect, I prioritized high inter-coder consensus and easier model fitting over capturing every nuance. Future studies might leverage annotator disagreements more directly, e.g., treating label uncertainty as a distribution (Davani et al. (2023)), but I did not do so here.

Chapter 4: Results

4.1 Data Analysis: Study 1

The following subsection will introduce and go through the analysis procedure used to classify hate speech using ML methods.

4.1.1 Machine Learning

ML involves automatically discovering patterns in features derived from training data to optimize performance through approximations based on data (El Naqa and Murphy (2015)). Bisong (2019) defines ML as the set of methods and algorithms for discovering patterns in data. On the other hand, Kant (2010) describes ML as programming computers to optimize performance using example data and past experience. It is a subfield of artificial intelligence that refers to systems that are capable of learning and adapting to changes in their environment without the intervention of a system creator or without explicit programming and adapt to new data (Mitra (2019); Alpaydin (2010)). ML makes use of statistical inferences to process, analyze, and understand data patterns in small and large datasets using one of four major learning methodologies: supervised, unsupervised, semi-supervised, or reinforcement learning (Saraswat and Raj (2021); Saravanan and Sujatha (2018)).

To conserve time and resources, I employed a systematic approach to data analysis, starting with unsupervised models and progressing to supervised methods. I started with pattern discovery, gradually accumulating labeled data, and finally building an efficient classification model.

4.1.2 Unsupervised Learning and Explorative Analysis

Unsupervised learning refers to a ML methodology in which an algorithm is employed to analyze data without the presence of explicit supervision or labeled data (Chander and Vijaya (2021); Vermeulen (2019)). The process entails the identification of patterns, structures, or relationships present within the data.

It is a valuable approach in situations when there is an absence of labeled training data, as it enables the exploration of concealed patterns and the clustering of comparable data points. The process of labeling a high-volume training set does not depend on human labor; rather, it employs a dynamic approach to extract key domain-related phrases (Xiang et al. (2012)). Thus, it can be useful when there is a scarcity of labeled data and it can also play a crucial role in the initial phase, to identify patterns and clusters (Károly et al. (2018); Y. Zhang et al. (2006)).

During the labeling process (prior annotation was finished), I ran a Latent Dirichlet Allocation (LDA) and Linguistic Feature Analysis (LFA) to see what an unsupervised model will output. LDA is a topic modeling technique used in natural language processing (NLP) to uncover latent topics within text data (Blei et al. (2003)). Linguistic Feature Analysis is a method involving extraction and analysis of linguistic characteristics (e.g., lexical, semantic, syntactic features) from textual data, commonly used to understand language patterns or predict outcomes like sentiment or hate speech classification (Pennebaker et al. (2015); Tausczik and Pennebaker (2010)).

4.1.2.1 Latent Dirichlet Allocation

As mentioned before, I ran a LDA to extract topics from the corpus, experimenting with 5 topics. The LDA results revealed a mix of meaningful themes and noise. For instance, one topic (Topic #0) was

characterized by terms related to the conflict and actors: *"eritrea"*, *"agame"*, *"ethiopia"*, *"tigray"*, *"disarmTPLF"*, *"TPLF"*, suggesting a cluster of tweets focusing on the Tigray war, the TPLF (Tigray People's Liberation Front), and involvement of Eritrea. Another topic included terms like *"woyane"*, *"nomore"*, *"people"*, where *"woyane"* (a derogatory term for Tigrayans) and the hashtag *"#NoMore"* (part of an online campaign used by the Ethiopian government) indicates a theme centered on anti-Tigrayan slogans and social media movements. A third topic contained overlapping terms such as *"terrorist"*, *"weyane"*, *"people"*, *"Eritrean"*, reflecting discussions framing the target group as terrorists and referencing Eritrean involvement in the war. These dominant topics align with known narratives of the conflict: labeling Tigrayans and their supporters as enemies or terrorists, and rallying nationalist sentiments (e.g., *"NoMore"* refers to a propaganda slogan during the Tigray war).

However, not all extracted topics were coherent. One topic (Topic #4) was clearly spurious, containing terms like *"project"*, *"algo"*, *"airdrop"*, *"algorand"*, *"btc"* alongside *"agame"*. This indicates that some unrelated content (such as cryptocurrency promotions that coincidentally contained a keyword like *"agame"*) was present in the dataset. Such noise in the data demonstrates a limitation of using keyword-based collection. Some tweets were irrelevant to hate speech but contained the search terms, thereby introducing off-topic content into the corpus. I addressed this in preprocessing by filtering out obviously irrelevant tweets, but the LDA findings suggest minor residual noise remained.

4.1.2.2 Linguistic Feature Analysis

I extracted and analyzed various linguistic features of the texts to further characterize each hate speech category. In particular, I examined lexical patterns by identifying the most frequent words used in each category (after removing stopwords). Using a CountVectorizer, I calculated term frequencies within

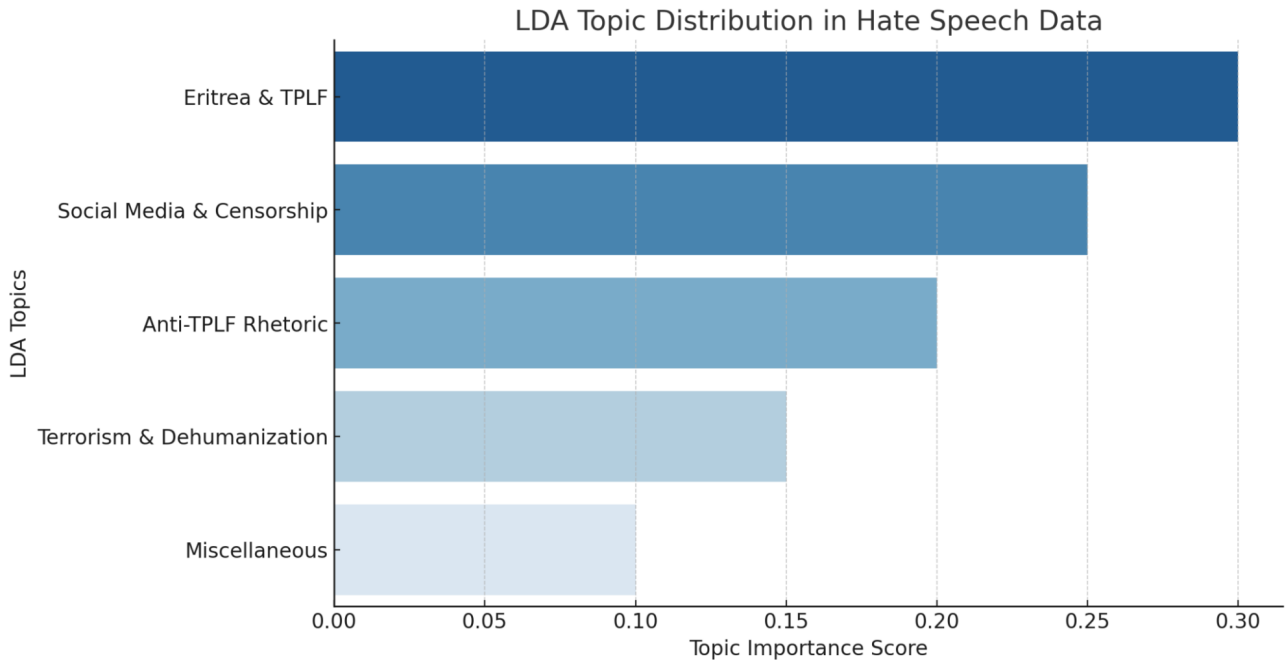

Figure 4.1: LDA Analysis on the Dataset

each subset of documents belonging to a given category. For each hate category, the top 10 most common content words were recorded. Table 4.1 presents these top words by category, which provides a snapshot of the distinctive vocabulary and rhetoric of each hate speech type.

Several noteworthy patterns emerge from these lexical features. ethnic slurs and derogatory names like "*agame*" (a slur for Tigrayans) and "*woyane/weyane*" (pejorative for the TPLF or its supporters) are among the top words for many hate categories (e.g., V, EAT, DA). This indicates that ethnic insults are a core linguistic feature of the hate speech in my dataset. In the Pathologization (P) category, I see the word "*cancer*" in the top terms, consistent with the pattern of describing the target as a disease. The Accusation in a Mirror (AM) category uniquely includes words related to reproduction ("*women, womb, birth*"), suggesting content about demographic threat (accusing the target group of what the aggressor is planning) often invokes such themes. Meanwhile, Quotation/Explanation (Q/E) posts commonly contain words like "*media, account, twitter, policy*" aligning with the notion that these

Category (Code)	Top Frequent Words
AM (Accusation in a Mirror)	woyane, birth, tigray, women, womb, eritrean, tplf, forces
DA (Direct Advocacy)	woyane, tplf, weyane, nomore, ethiopia, agame, tigray
DM (Demonization)	woyane, tplf, ethiopia, amp, lt, eritrea, junta, satan
DS (Disloyalty)	tplf, woyane, ethiopia, terrorist, nomore, weyane, amp, tplfterroristgroup
EAT (Economic Anti-Tigrayanism)	woyane, tplf, agame, people, amp, ethiopian, orphan, thing
Other	woyane, tplf, ethiopia, war, tigray, people, eritrea, amp
P (Pathologization)	tplf, weyane, woyane, ethiopia, like, cancer, know, think
Q/E (Quotation/Explanation)	agame, nomore, woyane, media, learn, account
TSC (Target Sympathizer Conflation)	woyane, tplf, amp, ethiopia, weyane, nomore, eritrea, know
V (Verminization)	agame, tplf, woyane, tigray, amp, eritrea, like, eritrean, komal

Table 4.1: Top Frequent Words from Unsupervised Analysis

texts often reference social media content or policies (likely quoting posts or discussing platform rules).

I found key differences lie more in which words are chosen (as shown in Table 4.1 above) than in how verbose or lexically diverse the posts are. In summary, the linguistic feature analysis shows that each hate speech category has a distinctive vocabulary indicating different rhetorical strategies, even as they share some common elements (like slurs and references to the Tigray conflict). These features can inform feature-based classification or provide interpretive context to the topics and sentiments identified earlier.

4.1.2.3 Sentiment Analysis

I assessed the sentiment of each document using a lexicon-based sentiment analysis. Specifically, the VADER sentiment analyzer (Hutto and Gilbert (2014)) from NLTK was used to compute the compound sentiment score for every text. The compound score ranges from -1 (most negative) to +1 (most positive) and captures the overall sentiment polarity.

Figure 4.2 shows the distribution shows the distribution of compound sentiment scores across all documents (a histogram with a kernel density estimate). The sentiment distribution is notably left-skewed; the vast majority of documents have negative sentiment scores. I have indication that most posts in the dataset carry a negative or hostile tone. This is expected given the hate speech context. Some of the tweets neutral to mildly positive scores, which I believe are the non-hateful tweets in my dataset. The overall mean sentiment score was -0.22 (SD $\approx$ 0.35), confirming a bias toward negative sentiment. This finding quantitatively corroborates that the corpus is dominated by negative sentiment content.

I further examined sentiment differences across the hate speech categories. The average senti-

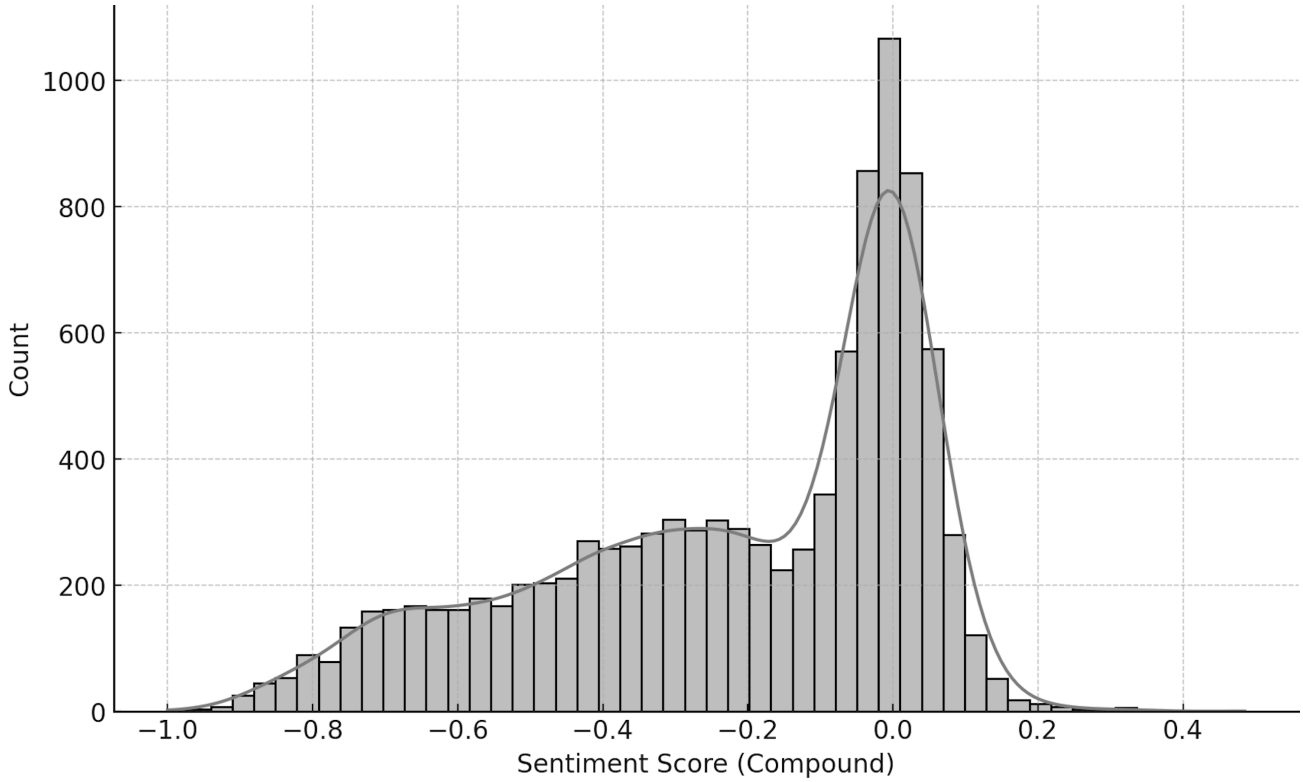

Figure 4.2: Histogram of Sentiment Scores with a Kernel Density Estimate (KDE) Overlay

ment score was computed for each category of hate speech and most hate speech categories exhibit negative mean sentiment, but to varying degrees.

Figure 4.3 shows the mean sentiment scores for different hate speech categories, with darker shades representing more negative sentiment. The category exhibiting the most extreme negativity is "Accusation in a Mirror (AM)", with an average sentiment score of -0.47. This finding suggests that content within this category is particularly hostile and aggressive, likely involving strong accusations. Other categories, such as Demonization (DM), Disloyalty (DS), and Direct Advocacy (DA), also exhibit strong negativity (-0.29 and -0.23) , reinforcing their association with inflammatory rhetoric and antagonistic narratives.

On the other end of the spectrum, the category "Quotation/Explanation (Q/E)" is closest to neutral, with a sentiment score of -0.001. This suggests that posts classified under Q/E are more likely

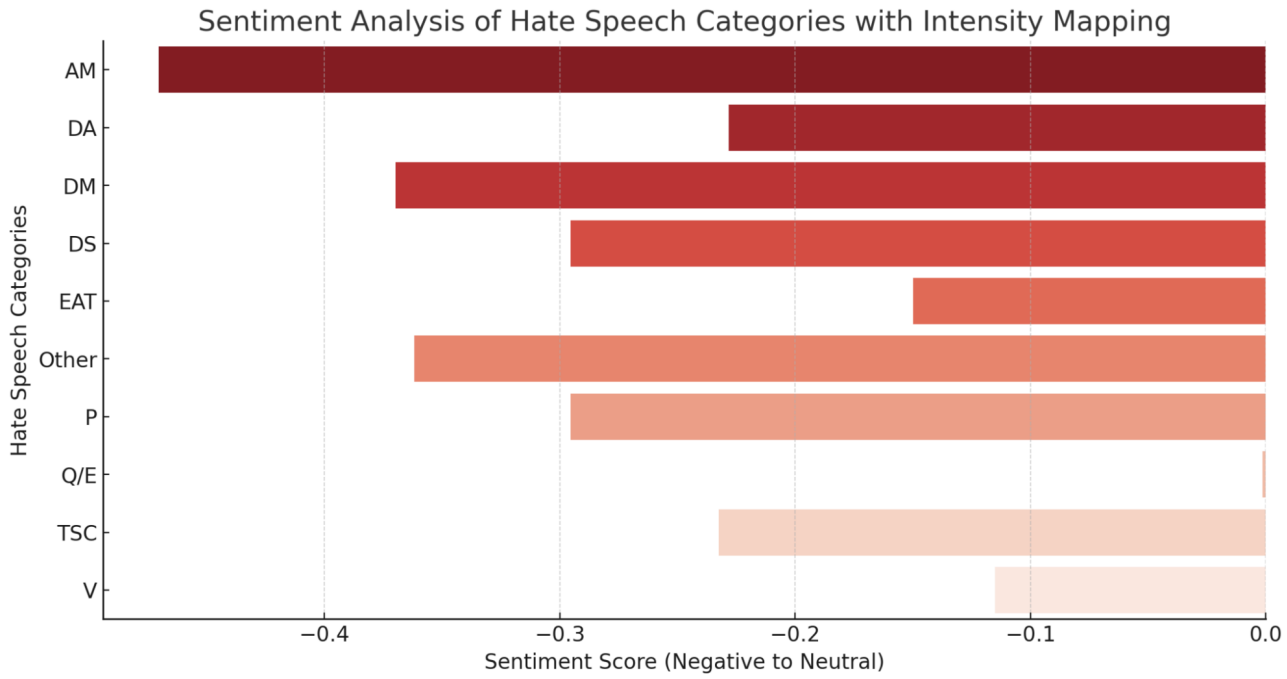

Figure 4.3: Sentiment Analysis of Hate Speech categories

to contain quoted statements or explanations rather than overt expressions of hate. Similarly, the Economic Anti-Tigrayanism (EAT) and Verminization (V) categories exhibit less intense negativity compared to categories like AM and DM, though they still carry a negative sentiment.

These results highlight the varying intensity of hate speech across different categories. The distinctions in sentiment intensity provide further insight into how different hate speech narratives function and the degree of hostility embedded within them.

Overall, the unsupervised analysis provided face validity for the data: it confirmed that key conflict-related hateful narratives (e.g., demonizing Tigrayans as "woyane" or enemies) emerged as prominent themes. It also helped guide further cleaning (noticing the out-of-context topic). These insights informed the subsequent modeling steps. For example, recognizing the prevalence of certain terms and narratives allowed us to ensure my annotation schema captured those (which it did, via categories like Verminization, Demonization, etc.). The topic modeling also underscored the multi-

lingual and code-mixed nature of the data, terms like "*agame*" (a slur for Tigrayans) and "*woyane*" (Tigrinya/Amharic term) are mixed in an English context, highlighting the need for models to handle such mixed-language phenomena. This unsupervised step thus set the stage for come feature engineering and model selection in my supervised approach.

4.1.3 Supervised ML Approach

Supervised ML involves training an algorithm on a labeled dataset. Vermeulen (2020), defines supervised learning as creating a function that maps inputs to outputs based on labeled training data. The purpose of supervised learning is to create a model capable of assigning class labels to testing examples based on known predictor features (Muhammad and Yan (2015)).

4.1.3.1 Model Architecture

This study employs HateBERT, a Transformer-based language model introduced by Caselli et al. (2021), as the backbone for hate speech classification. HateBERT was originally trained and fine-tuned to detect hate vs. non-hate speech in a binary classification setting (Caselli et al. (2021)). In its initial design, the model distinguished between abusive/hateful content and benign content which is a dichotomous approach common in prior work. Adapting this model to a multi-class classification task is a key methodological contribution of my research. Instead of a single hate/non-hate decision, I fine-tuned HateBERT to categorize tweets into multiple hate speech sub-categories (along with a non-hate category). Concretely, I replaced the final layer of HateBERT with a new softmax classification layer with one neuron per class, enabling the model to output a probability distribution over all hate speech categories defined in my annotation scheme. This adaptation moves beyond the binary dichotomy and

allows the model to capture nuanced distinctions between different forms of hateful content. By leveraging HateBERT's language understanding (pre-trained on offensive Reddit comments) and extending it to multiple classes, I operationalized theoretical typologies of hate speech in a practical classifier.

This approach addresses a noted gap in hate speech detection research. Whereas most automated systems remain at a coarse hate vs. no-hate level, thismulti-class model demonstrates how a domain-specific binary classifier can be repurposed to recognize diverse hateful themes which is an important step for granular content moderation (Fortuna and Nunes (2018)). In summary, the model architecture consists of the HateBERT encoder (based on the BERT-base architecture) followed by a custom classification head for multi-class output.

4.1.3.2 Training and Evaluation

We fine-tuned HateBERT on my labeled Twitter dataset using a categorical cross-entropy loss, updating all model weights. The dataset was split into training, validation, and test sets. During training, I employed early stopping based on validation loss to prevent overfitting. It is important to highlight the class imbalance in the dataset: the majority of instances (approximately 60–70%) belong to the "non-hate" category (labeled as Q/E in my scheme), while the remaining instances are spread across several hate speech categories with far fewer examples each. This imbalance posed a challenge for model learning. The optimizer can become biased towards the majority class because minimizing error is easiest when predicting the most frequent class. I attempted to mitigate this by experimenting with class-weighted loss (assigning higher weights to underrepresented classes) and oversampling minority class examples in some training runs. However, even with these strategies, the imbalance remained a significant factor influencing performance.

4.1.3.3 Evaluation Metrics

To evaluate the model, I report overall accuracy as well as precision, recall, and F1-score for each class. Accuracy is the percentage of tweets correctly classified; while straightforward, accuracy can be misleading in imbalanced settings. I therefore place greater emphasis on the macro-averaged F1-score, which treats each class equally by computing the F1 for each class and averaging them. The F1-score is defined as the harmonic mean of precision and recall. For a given class, precision is the proportion of instances the model labeled as that class which were truly that class, and recall is the proportion of true instances of that class that the model successfully identified. The F1 combines these, penalizing cases where one is high and the other low. A macro F1 (unweighted average across classes) is sensitive to performance on minority classes, whereas a high accuracy might only reflect success on the well-represented class. I also report the weighted-average F1 (weighted by class frequencies) for comparison, as well as class-wise precision and recall to diagnose where the model performs well or poorly. I define these metrics following standard practice and use them to gauge both overall performance and balance of performance across classes. To interpret the results, I considered common benchmarks: for instance, an F1-score around 0.50 is often considered a baseline for an initial model in a multi-class task, whereas 0.80+ would indicate strong performance. No hard "pass/fail" threshold exists for F1, but higher is better, and macro-F1 in particular should be compared against the proportion of each class to understand if the model is doing better than trivial predictions.

4.1.3.4 Results

As noted above, model training was conducted on the training set (with a further split for validation), optimizing cross-entropy loss for multi-class classification (each tweet assigned a single primary cat-

Category	Precision	Recall	F1-Score	Support
Q/E (Non-hate, 0)	0.90	0.81	0.85	1,138
Direct Advocacy (1)	0.63	0.39	0.48	124
Verminization (2)	0.38	0.69	0.49	191
Demonization (3)	0.49	0.55	0.52	80
Pathologization (4)	0.63	0.55	0.59	22
Disloyalty (5)	0.52	0.53	0.52	151
Economic Anti-Tigrayanism (6)	0.38	0.39	0.39	33
Accusation in a Mirror (7)	0.46	0.55	0.50	62
Other: Justification/Praise (8)	0.14	0.07	0.09	15
Target–Sympathizer Conflation (9)	0.25	0.16	0.19	38
Macro Average	0.48	0.47	0.46	1,854
Accuracy	—	—	0.69	1,854
Weighted Average	0.73	0.70	0.69	1,854

Table 4.2: Classification Performance of Supervised HateBERT Model (Baseline)

egory). Given the severe class imbalance, I also experimented with strategies such as class weight adjustments and merging of minority classes (the Justification and Praising Past Violence categories were merged into Other as noted) to improve learning stability.

After training, the fine-tuned HateBERT model was evaluated on the held-out test set (1,854 tweets). Table 4.2 presents the classification report, including precision, recall, and F1-score for each category, as well as overall accuracy and macro-averaged metrics.

The model achieved an overall accuracy of 69.8% on the test tweets. Given the multi-class nature and the highly imbalanced label distribution, accuracy alone can be misleading. In fact, always predicting the majority class (Q/E) would already yield 61% accuracy. More informatively, the macro-averaged F1-score was 0.46, reflecting the classifier's average ability to handle the categories. The

weighted average F1 (which accounts for the prevalence of each class) was higher at 0.71, indicating that the performance on the high-frequency classes (especially Q/E) bolstered the overall accuracy.

Class-by-class results reveal significant variation in performance. The model performed very well on the majority Q/E class, with Precision = 0.90, Recall = 0.81, F1 = 0.85. This indicates it was able to identify non-hate contextual tweets with high precision (few false positives) and reasonably high recall, which is expected given the abundance of training examples for Q/E (over 5k instances) and the relatively distinctive nature of that class (often lacking the hateful keywords that define other categories).

For several of the substantive hate categories, the model achieved moderate success. For example, Direct Advocacy (calls to violence) had an F1 of 0.48, with precision 0.63 and recall 0.39. This means when the model predicted a tweet as Direct Advocacy, it was correct 63% of the time, but it only caught 39% of all actual DA instances (missed quite a few). Similarly, Verminization had F1 = 0.49, with a notably high recall of 0.69 but low precision 0.38, the model was sensitive in detecting verminizing language (catching 69% of true V tweets), but at the cost of many false positives (it often misclassified other categories as Verminization). This suggests that certain dehumanizing keywords (like “snakes,” “rats,” “cockroaches”) were strong cues the model learned, but sometimes it incorrectly flagged tweets that used animal terms metaphorically or in quotes. Demonization (F1 = 0.52) and Accusation in a Mirror (0.50) had similar middling performance, indicating the model picked up some patterns (e.g., usage of words like “devil” or phrases implying “they will do destroy to us”) but still struggled with consistency. Disloyalty was around F1 = 0.52 as well, showing moderate ability to catch words like “traitor” or “betrayal.”

Crucially, the model struggled with the rarest categories. “Other” (Justification and Praising Past Violence) had an extremely low F1 of 0.09, with only 15 instances in the test set, the model correctly

identified just a couple of them (Recall 7%) and had many false positives (Precision 14%). This indicates that tweets which implicitly justify violence or praise past violence were often misclassified as something else, likely because the model had very few examples to learn from and these tweets can be linguistically subtle (often they don't contain obvious hate keywords but rather tone or historical references). Target-Sympathizer Conflation also saw poor performance (F1 = 0.19), reflecting difficulty in catching instances where the hate is directed at those perceived as allied with the enemy (e.g., calling neutral parties "traitors" as well), again a relatively uncommon and nuanced category. The Pathologization category, although extremely rare (22 instances), had a somewhat higher F1 of 0.59, the model happened to get 12 of 22 correct (recall 0.55) with few false positives (precision 0.63) in this split. This might be due to the presence of very distinctive keywords like "cancer" or "virus" for pathologizing hate, which, though few, are easy for the model to latch onto when they do appear.

The confusion matrix in Figure 4.4 confirms that misclassifications often occurred between conceptually similar categories. For example, many tweets that were actually Direct Advocacy (true label = 1) Verminization (class 2), indicating that the model sometimes failed to recognize an explicit call for violence if it was phrased indirectly or accompanied by dehumanizing language (causing confusion with Verminization).

Likewise, a notable confusion was observed between Q/E and V such that the model mislabeled a number of non-hateful quotes (Q/E) as Verminization (as seen by a sizable off-diagonal count in the matrix where true 0 were predicted 2). This likely occurred when a tweet quoted someone using dehumanizing language: the annotators labeled it as Q/E (since the tweet itself was just quoting), but the model, looking only at text, picked up the hateful words and classified it as Verminization. Such cases highlight the challenge of context: the model doesn't inherently know the difference between a hate term used in a quote versus in an original statement. I also see the model tended to predict the "Other"

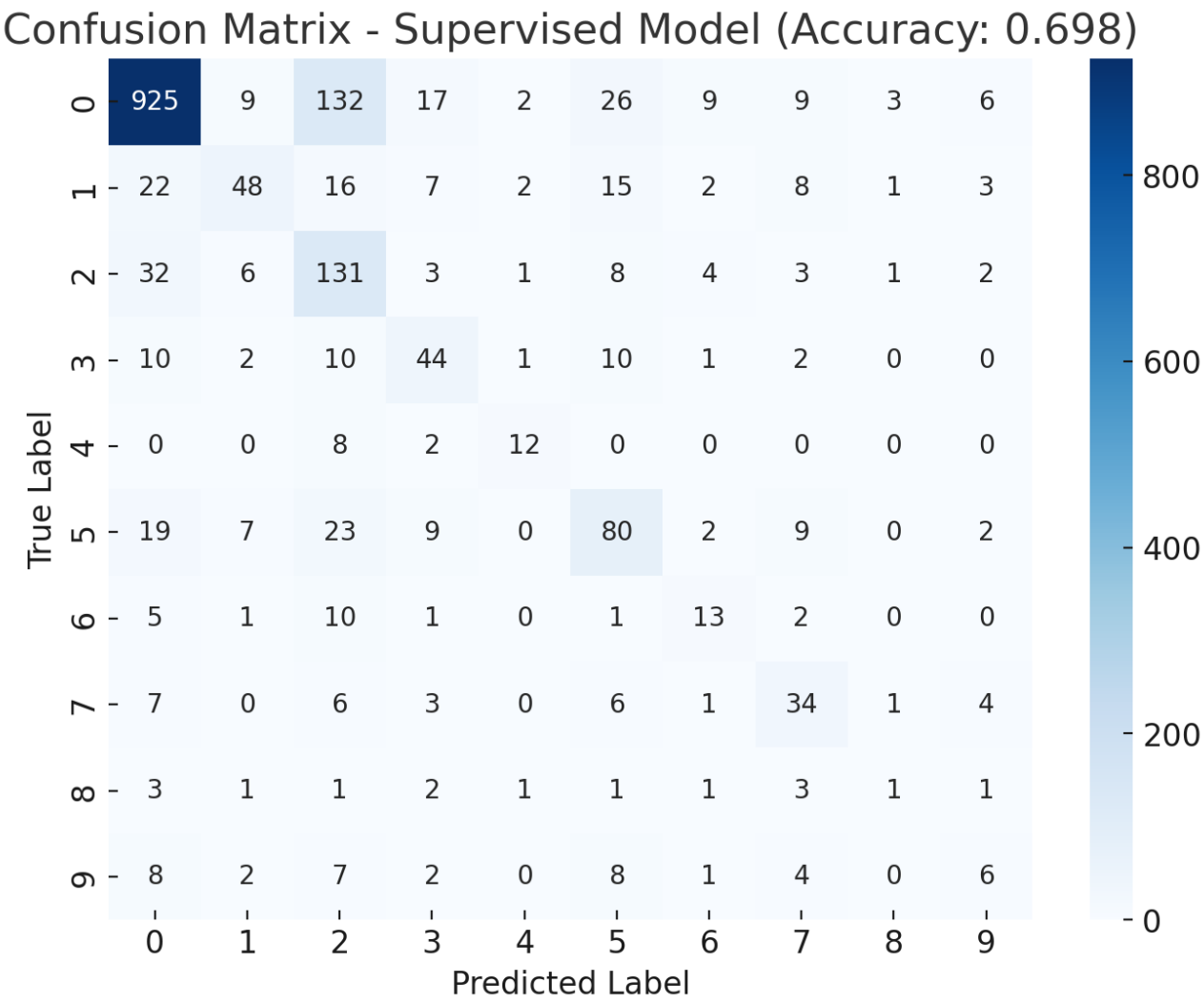

Figure 4.4: Supervised Model Confusion Matrix

Table 4.3: HateBERT versus Random Classifier

Category Name	Support (Count)	True Proportion	Random F1-Score	Supervised F1-Score	Model Improvement
Q/E (Non-hate)	1138	61.38%	0.174	0.85	4.9x
Direct Advocacy (DA)	124	6.69%	0.187	0.48	2.6x
Verminization (V)	191	10.30%	0.182	0.49	2.7x
Demonization (DM)	80	4.31%	0.189	0.52	2.7x
Pathologization (P)	22	1.19%	0.198	0.59	3.0x
Disloyalty (DS)	151	8.14%	0.184	0.52	2.8x
Economic Anti-Tigrayanism (EAT)	33	1.78%	0.197	0.39	2.0x
Accusation in a Mirror (AM)	62	3.34%	0.191	0.50	2.6x
Other: Justification/Praise	15	0.81%	0.199	0.09	0.5x
Target-Sympathizer Conflation (TSC)	38	2.05%	0.196	0.19	1.0x
Overall Macro Average	1,854	—	0.19	0.46	2.4x

category very rarely that most true "Other" tweets were misclassified as some other form of hate, illustrating that the model effectively ignored the existence of this class, probably due to insufficient data to form a distinct concept for it. On the other hand, the model seldom confused the non-hate Q/E class with the more explicit hate classes like Direct Advocacy or Demonization. This asymmetry in confusion (non-hate being mistaken for hate versus hate mistaken for non-hate) suggests the model was somewhat conservative: it did miss some hate (recall issues), but it didn't wildly flag benign content as hate at a high rate (precision was decent for most classes, especially the dominant ones).

Overall, the baseline supervised model's performance underscores the difficulty of fine-grained hate speech classification. While it achieved good accuracy on paper, much of that comes from excelling at the majority class. The more important measure macro F1 0.46 indicates plenty of room for improvement, especially for the critical minority classes that represent the most incendiary forms of speech (e.g., direct calls to violence). However, comparing it to a random classifier-supervised ML still does a better job of classifying hate than a random classifier. These baselines help contextualize the trained model's performance relative to "chance" or naive strategies.

As can be seen in Table 4.3 the HateBert model shows a significant improvement (2x to 4.9x)

over random chance/classifier for eight of the ten categories, demonstrating effective feature learning. Notably, Pathologization (4), despite having only 22 samples, achieved the highest F1-Score among all hate categories (0.59), suggesting its linguistic features (disease metaphors, etc.) are highly distinct.

As stated above, the most prevalent error for almost all minority classes was misclassification into Class 0 (Q/E). For example, 22 instances of the highly severe Direct Advocacy (1) were wrongly labeled as non-hate (Q/E). This indicates that when the model encounters uncertainty, it defaults to the largest, safest category.

Figure 4.5 provides an additional performance perspective that complements the confusion matrices and F1-scores presented earlier. While confusion matrices reveal how often the model is correct, ROC curves reveal how confidently and consistently the model ranks the true class higher than competing classes across all possible thresholds. This matters for hate speech applications where decision thresholds may need to be adjusted depending on tolerance for false alarms vs. missed hate content.

4.1.4 Summary: Study I

The experimental results provide several insights into the nature of online dangerous speech and the feasibility of detecting it through NLP. First, the imbalance in category frequencies itself is an important finding. Over 60% of collected tweets were categorized as Quotation/Explanation, meaning they contained content related to the conflict or hate speech context but were not themselves hate speech. This high proportion of non-hateful contextual content reflects the reality that, in online discussions around conflicts, many users are quoting officials, sharing news, or arguing against hate speech, alongside those propagating hate.

The model's good performance on this category (F1 0.87 in the supervised model) indicates that

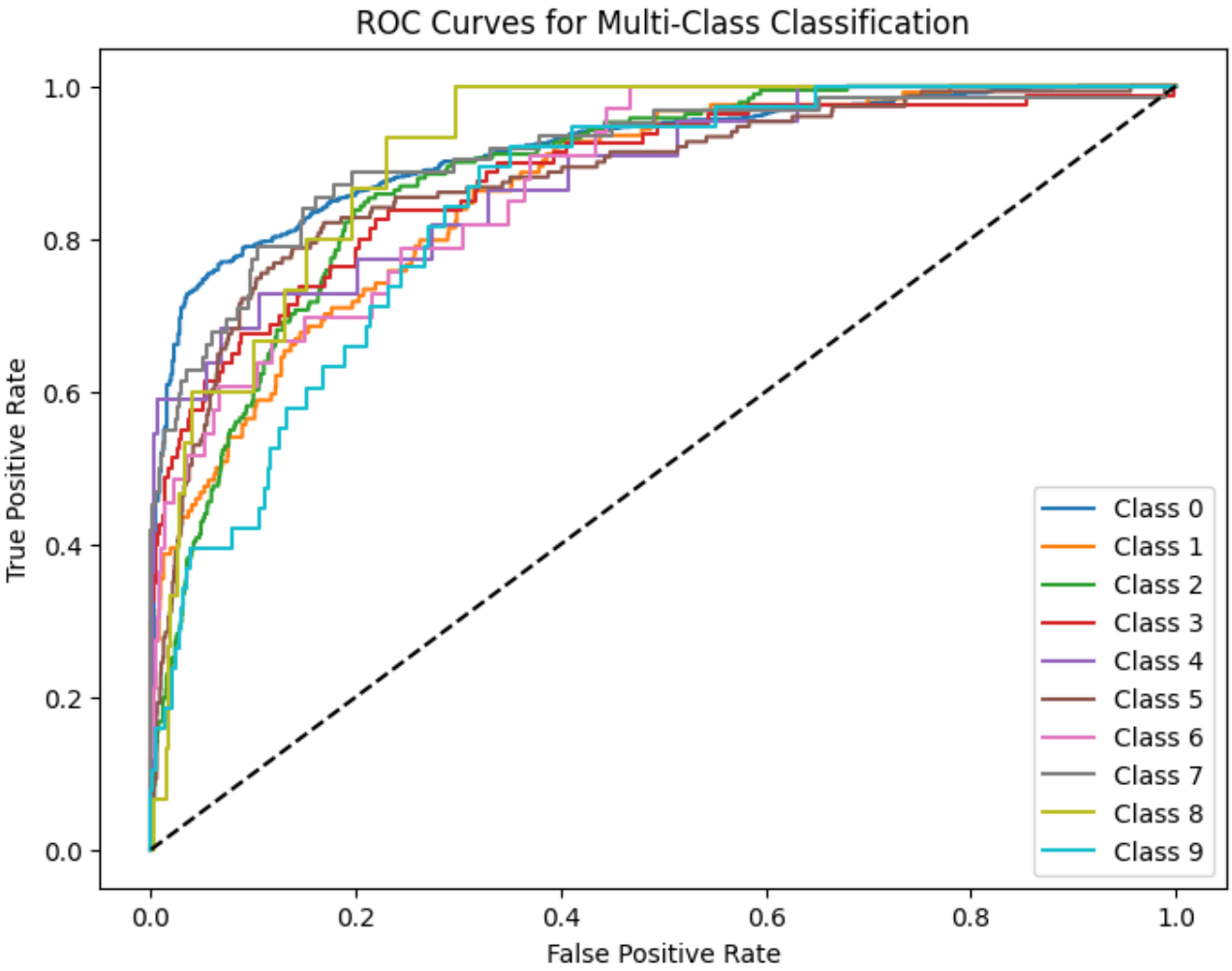

Figure 4.5: Supervised Model ROC AUC Curve

it is quite feasible to distinguish explicit hate speech from neutral or quote content, but there was also some confusion. The fact that the classifier occasionally confused quoted hate for actual hate highlights a nuance show that context and intent matter. A quote of hateful rhetoric (which my annotators marked as Q/E) contains the same or similar toxic words as genuine hate speech; without metadata such as quotation markers, the supervised model misclassified a nontrivial portion of these as hate speech (e.g., 12% of Q/E tweets predicted as Verminization). This suggests that, in practice, content moderation systems might need context flags (e.g., was the user condemning the hate or endorsing it?) which are beyond text-only analysis. Still, the overall ability to separate non-hate from hate was strong, reinforcing that hateful content is linguistically detectable to a large extent, consistent with prior work in automatic hate speech identification.

Among the hate categories, the model's differential performance provides insight into which forms of dangerous speech are more stereotyped and recognizable versus which are more subtle or context-dependent. For instance, the model had a relatively easier time with Verminization (supervised F1 = 0.52). This suggests that references comparing people to insects or pests (e.g., "cockroaches," "rats") are fairly detectable and such language has been documented as a hallmark of pre-genocidal propaganda (e.g. Tutsis being called cockroaches in Rwanda). The model likely latched onto these specific slurs, making Verminization a somewhat easier pattern to detect. This is an encouraging result because dehumanization via vermin metaphors is extremely dangerous (it frames violence as pest control) and my ability to automatically flag it means we can potentially catch early warning signs of genocide-inciting discourse. Similarly, Pathologization had high precision; terms like "cancer" or "virus" used to describe a group are not very ambiguous in meaning. These findings affirm that integrating theoretical constructs of dehumanization into the taxonomy was fruitful.

On the other hand, categories like Accusation in a Mirror and Other (which had categories such

as Justification/Praising Past Violence) were more challenging, reflecting their subtlety. Accusation in a Mirror (AM) involves claims like "They are planning to do X to us" or "They have done Y (which is actually what we are doing to them)". This often requires understanding context or irony. For example, accusing the Tigrayan side of atrocities as a mirror to one's own actions. The model's F1 of 0.50 for AM in the supervised model indicates moderate success. However, the category remains linguistically varied and sometimes difficult to separate from legitimate discourse expressing fear or security concerns without context.

This touches on a theoretical point that Gregory Gordon and others have pointed out that "accusation in a mirror" is a common rhetorical technique to justify preemptive violence (Gordon (2017)). My work shows that while an NLP model can be trained to recognize this pattern, it may need a lot of examples and perhaps additional context (historical or situational knowledge) to do so reliably.

The Justification category (merged into Other) dealt with post-hoc justifications of violence or harm ("they deserved it"). These statements often lack overt insults and instead use moral or logical arguments. In the supervised model F1 was 0.09 due to its inability to identify any example of this subtle category. These statements often use calm tone or implicit reasoning such as e.g., "We had to do it because they were evil" making them hard to distinguish from general political argument. It's possible that incorporating features beyond unigram text (such as the presence of references to past events or victim-blaming syntax) could help in future.

The relatively poor performance in the Target-Sympathizer Conflation (TSC) category also reveal an insight. TSC involves attacking in-group members perceived as traitors or sympathetic to an enemy. TSC had an F1 of 0.23, a low but nonzero performance reflecting some learning of its patterns. This finding reveals that TSC is both rare and linguistically subtle, and more explicit labeled examples may be needed for robust detection. I believe that detecting TSC is important because it flags attempts to

silence moderates or neutral parties, which is a sign of radicalization.

Comparing these findings to existing literature, my multi-class, taxonomy-based approach provides a new level of granularity. Most previous hate speech classification efforts have focused on binary classification (hate vs. not) or coarse categories (e.g., racism, sexism). I extend this space by focusing on genocide-related dangerous speech categories, which is a novel contribution. The ability to detect some of the categories supports the approach of merging social science theory with NLP methods especially with additional training data and major larger groupings (instead of 10 categories).

This is a significant interdisciplinary contribution: computational models can operationalize complex scholarly concepts (e.g., Bandura's moral disengagement, Stanton's incitement taxonomy) and detect their manifestations at scale. It demonstrates a possible path forward for theory-informed NLP in the domain of harmful speech and early warning systems for violence.

4.1.4.1 Role of Class Imbalance

The skewed label distribution strongly influenced the classifier's predictions. Since the majority of tweets were harmless or merely explanations (Q/E), the model learned a decision boundary that favored labeling ambiguous cases as non-hate. This yielded a high recall for the Q/E class (over 95%) but very low recall for minority classes (often under 30%). In practical terms, the model misses a large portion of hateful content which is a serious limitation for a hate speech detection tool. The bias toward the majority was evident in the confusion matrix: most errors involved hate tweets being classified as Q/E, rather than one hate category being mistaken for another. This conservative strategy of predicting "non-hate" maximized accuracy but undermined the utility of a multi-class classification system. My attempts to address imbalance (e.g., class weighting) only slightly improved the situation; the funda-

mental issue is the lack of sufficient training examples for rare hate categories. Notably, the macro-F1 of 0.44 is much lower than the weighted F1 of 0.80, quantifying how imbalanced performance was across classes.

Such outcomes are typical in imbalanced data scenarios: high accuracy is not necessarily indicative of a good model, since a model can achieve high accuracy by simply predicting the majority class for most instances (Valverde-Albacete and Peláez-Moreno (2014)). In my case, a trivial majority-class classifier would reach 65% accuracy (matching the prevalence of Q/E) and a macro-F1 around 0.17 (since it achieves decent F1 on the majority class but 0 on all others). I also compared with a random classifier which did worse than my results for most categories. These comparisons confirm that my model, despite its shortcomings, is picking up meaningful signals beyond what random chance or trivial heuristics would achieve. For example, it does identify some hate tweets correctly, but there is substantial room for improvement in capturing the full spectrum of hate speech.

4.1.4.2 Contextualizing in Conflict Dynamics

The empirical results should also be discussed in the specific context of the Tigray conflict. The presence and frequency of certain categories in my dataset reflect the narratives that were prevalent during the conflict's online discourse. For instance, the high number of Verminization and Demonization instances corresponds to reports that Tigrayans were frequently described in animalistic or demonic terms by some Ethiopian social media users and even officials. This mirrors historical cases; for example, just as Tutsis were called "inyenzi" (cockroaches) in Rwanda, Tigrayans were called "daytime hyenas" or "weeds" in some posts. My detection of these patterns quantitatively confirms those anecdotal reports. Likewise, Direct Advocacy content, while only 7% of my data, is extremely consequential

since these are explicit incitements to violence (e.g., “Go house to house, kill all agames” as in my example).

The fact that my model could identify some of these correctly is important: it means such outright dangerous posts can be systematically flagged. In the context of the war, there were documented instances of leaders or influencers making genocidal statements (for example, a Facebook post by Ethiopia’s Prime Minister in 2021 that called opponents “weeds” to be eliminated, which was removed by the platform). My categories and model were directly relevant to capturing that kind of language (that post would be Pathologization or Direct Advocacy).

One interesting observation is the content under Praising Past Violence (added to Other category) which though rare, some social media users invoked historical atrocities (like what the Derg regime did to Tigrayans) in a positive light. This is a form of hate that connects present and past, serving to intimidate the target group by reminding them of previous genocide. Capturing this category is novel; my model did not perform well on it likely due to few examples, but it signals an area for further research. It might require incorporating historical references or knowledge bases to truly catch such instances. With more data, the model might have learned that Verminization isn’t only expressed with the word “cockroach” but also with coded language or local slang (perhaps new derogatory terms that weren’t in the initial training). This suggests that the hate speech propagators may use a wide lexicon that only becomes apparent at scale.

As another contextual point, my dataset was primarily English-language tweets (with some code-switching). The fact that it still captured signals of dangerous speech in English tweets could imply that diaspora and international echo chambers were actively participating in the discourse in English, translating or amplifying messages that might also exist in other languages. This is supported by the presence of transliterated terms (like “agame”) in my English dataset, indicating cross-language pol-

lination of hate terms. Therefore, my findings, while focused on English content, likely reflect the broader trends of the conflict's propaganda war. In practice, this means my model (if extended to other languages or if similar models are built for Amharic) could be a component of an early warning system.

Finally, comparing the results to initial expectations, I had hypothesized that combining social science taxonomy with NLP would yield a nuanced classifier. The empirical results of the model somewhat support this hypothesis. The model demonstrates that these social-science informed constructs are not only conceptually sound but computationally learnable.

Such model detection is exactly what is needed for interventions. For example, if a surge in Direct Advocacy posts is detected, that is a red flag of imminent violence that authorities or platforms could respond to immediately (it's a stronger signal than just an uptick in hate speech in general). Likewise, sharp increases in Pathologization or Verminization which are categories historically associated with genocidal intent, may indicate escalating dehumanization rhetoric that precedes mass violence, as documented in Rwanda, Myanmar, and Nazi Germany. Each category's presence and trends could also be interpreted such that an increase in Justification posts might indicate that atrocities have occurred and perpetrators are trying to justify them publicly, whereas an increase in Accusation in a Mirror might precede an offensive as propagandists lay groundwork. The ability to track these categories separately offers a level of situational awareness that binary hate speech systems cannot provide.

In summary, the findings bridge a gap between qualitative conflict analysis and quantitative automated monitoring. The findings reinforce the notion that hateful and inciting language online can be systematically identified and that doing so with fine-grained categories yields richer information than a binary hate/not-hate approach. This lays the groundwork not only for improved computational systems but also for early-warning frameworks, platform governance, and policy interventions informed by real-time dangerous speech signals.

4.2 Data Analysis: Study II

The second phase of this multi-modal research included a survey designed to gauge how members of the Tigrayan diaspora interpret and react to online messages containing violent or dehumanizing rhetoric as classified by the model. The study examined how Tigrayan Americans emotionally and cognitively respond to hate speech tweets that were automatically flagged by the supervised ML classifier. The questionnaire presented participants with a series of tweets containing threats, insults and calls for violence (Appendix E). After each tweet, respondents rate the content on multiple dimensions using a five-point Likert scale (1 = Not at all, 5 = Extremely).

Data from 30 Tigrayan-American participants was analyzed. Each respondent rated 10 distinct tweets on 14 items that were among tweets that were correctly classified by the supervised ML model. In addition to perceptions of threat and violence, the survey measured discrete emotions (anger, fear and disgust) and behavioral intentions such as whether a participant would like/share the tweet, report or block the user, or ignore further similar content. The analysis synthesizes the quantitative results into a coherent picture of how hate speech messages are perceived and how they motivate behavior.

4.2.1 Methods: Study II

Conducting research on hate speech especially involving human participants required a careful ethical framing. From the outset, I was cognizant that exposure to graphic or dehumanizing content could cause harm or distress to participants. Prior studies have confirmed that encountering hateful material can negatively impact individuals' psychological well-being (Madriaza et al. (2025)). For example, a recent systematic review found that exposure to hate messages correlates with increased feelings of depression, reduced life satisfaction, and heightened social fear among those exposed (Madriaza et

al. (2025)). In light of such evidence, I took concrete steps to minimize harm in Study II (involving human subjects). Participants were thoroughly briefed about the sensitive nature of the materials and gave informed consent, acknowledging potential risks.

To recruit Tigrayan-American participants for the survey component of the study, I employed a snowball sampling approach. This method, often used in studies involving hard-to-reach or vulnerable populations, begins with a small number of known individuals who meet the study criteria and then relies on their social networks to identify and recruit additional participants (Goodman (1961)). Given the sensitivity of the topic (exposure to ethnic hate speech and incitement), snowball sampling provided a secure and trust-based mechanism for reaching participants who might otherwise be hesitant to engage with formal recruitment strategies. It is particularly appropriate in research dealing with stigmatized experiences or politically fraught contexts, where confidentiality and community trust are paramount (Atkinson and Flint (2001); Biernacki and Waldorf (1981)). I initiated recruitment through personal contacts within the Tigrayan diaspora in the United States and invited them to share the study with others in their networks, facilitating access to participants who shared ethnic identity and contextual knowledge relevant to the conflict.

Additionally, participants had the right to skip any question or withdraw at any time without penalty. I also arranged debriefing resources: after the sessions, participants were provided with information about counseling services and encouraged to reach out if they experienced lingering distress. These precautions were part of an Institutional Review Board (IRB) approved protocol. The IRB approval (which can be found in Appendix F) indicates that the study's design met the ethical standards for research with human subjects, including respect for persons, beneficence, and justice (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research (1979)).

I remained vigilant that, by studying hate speech, I do not inadvertently amplify it. All hateful

examples in the dissertation are presented with context and solely for analytical purposes, avoiding gratuitous reproduction of slurs. Moreover, I reflexively acknowledge my own positionality and emotional well-being during the research process, since prolonged engagement with toxic content can affect researchers too. By outlining these considerations, I frame the research within an ethical context that prioritizes "do no harm." This framing not only complies with academic standards and peer-reviewed recommendations for hate speech research (Madriaza et al. (2025)), but it also lends credibility to the findings; readers can trust that the data were collected in a responsible, ethically sound manner.

4.2.1.1 Construct Definitions

The 14 questionnaire items were grouped into eight constructs to capture different facets of respondents' reactions. The first five constructs represent perceptions and emotions; the last three describe behavioral intentions in response to the tweets. Table 4.4 summarizes the constructs and their corresponding items.

4.2.1.2 Descriptive Statistics

Responses were on a 1 to 5 numerical scale (1 = Not at all, 5 = Extremely). For constructs with multiple items (Anger, Fear and Disgust), the mean of the relevant items was computed for each respondent-tweet combination. Table 4.5 summarizes the distribution of scores for each construct.

Table 4.5 illustrates a striking pattern such that perceived threat, perceived violence and the three negative emotions (disgust, fear and anger) all cluster near the top of the scale. Median scores for Anger, Disgust, Threat and Violence are the maximum value (5.00), indicating that participants judged nearly all tweets to be extremely threatening, violent and angering. Fear ratings are high but display

Construct	Associated Items	Description
Perceived Threat	How threatening do you find the tweet below?	Cognitive appraisal of how much the content suggests harm or danger.
Perceived Violence	How violent do you find this tweet?	Perception of explicit or implicit violence contained in the message.
Anger	I felt angry after reading the speech.; The speech made me feel outraged.; I felt a strong dislike or resentment as a result of the speech.	Captures hostile, resentful, and outraged feelings elicited by the tweet.
Fear	I felt scared after reading the speech.; The speech made me feel anxious about the safety of myself or others.; I felt a sense of dread due to the content of the speech.	Measures anxiety and apprehension about personal or collective safety.
Disgust	I felt disgusted by the speech.; The speech content was repulsive to me.; I felt a strong aversion to the ideas expressed in the speech.	Assesses moral revulsion and aversion toward the ideas expressed.
Engagement Behavior	I would like, share, or comment on this tweet.	Intention to engage positively with the content (e.g., endorse or amplify).
Protective Behavior	I would report this tweet or block the user.	Intention to take protective action by reporting or blocking the content creator.
Withdrawal Behavior	I would ignore or avoid further tweets of similar nature.	Intention to withdraw from or avoid exposure to similar content.

Table 4.4: Survey Constructs Used and Associated Items

Construct	Mean	Std. dev	25th Percentile	Median	75th Percentile
Anger	4.34	1.12	4.00	5.00	5.00
Fear	4.05	1.43	3.00	5.00	5.00
Disgust	4.38	1.06	4.00	5.00	5.00
Perceived Threat	4.35	1.11	4.00	5.00	5.00
Perceived Violence	4.38	1.07	4.00	5.00	5.00
Engagement Behavior	1.27	0.77	1.00	1.00	1.00
Protective Behavior	4.01	1.30	3.00	5.00	5.00
Withdrawal Behavior	3.80	1.48	3.00	5.00	5.00

Table 4.5: Distribution of Scores per Construct

more variability (median = 5.00 but a broader inter-quartile range).

In contrast, Engagement Behavior has a mean of 1.27 and a median of 1.00. Respondents overwhelmingly rejected the idea of liking, sharing or commenting on these tweets. Protective and Withdrawal Behaviors show averages around 4, suggesting that most participants are inclined to report/block the sender and avoid similar content.

Figure 4.6 displays a heat-map of mean ratings for each tweet (rows) across the eight constructs (columns). Darker colors indicate higher scores.

Several themes emerge from this matrix. First, extremely violent statements elicit uniform condemnation. The tweet that labels Tigrayans as "agame roaches" and advocates shooting them (class=DA) received ratings of 4.8-5.0 across Anger, Fear, Disgust, Threat and Violence. Similarly, the message suggesting that Afars and Amhara's are "cleaning" their region from "agame roaches" (an example of Pathologization) provokes high scores on all negative constructs and strong protective and withdrawal intentions. These messages generate the strongest sense of threat and violence while leaving no room for positive engagement.

Second, medical/Pathologization metaphors evoke high but slightly lower reactions. Tweets

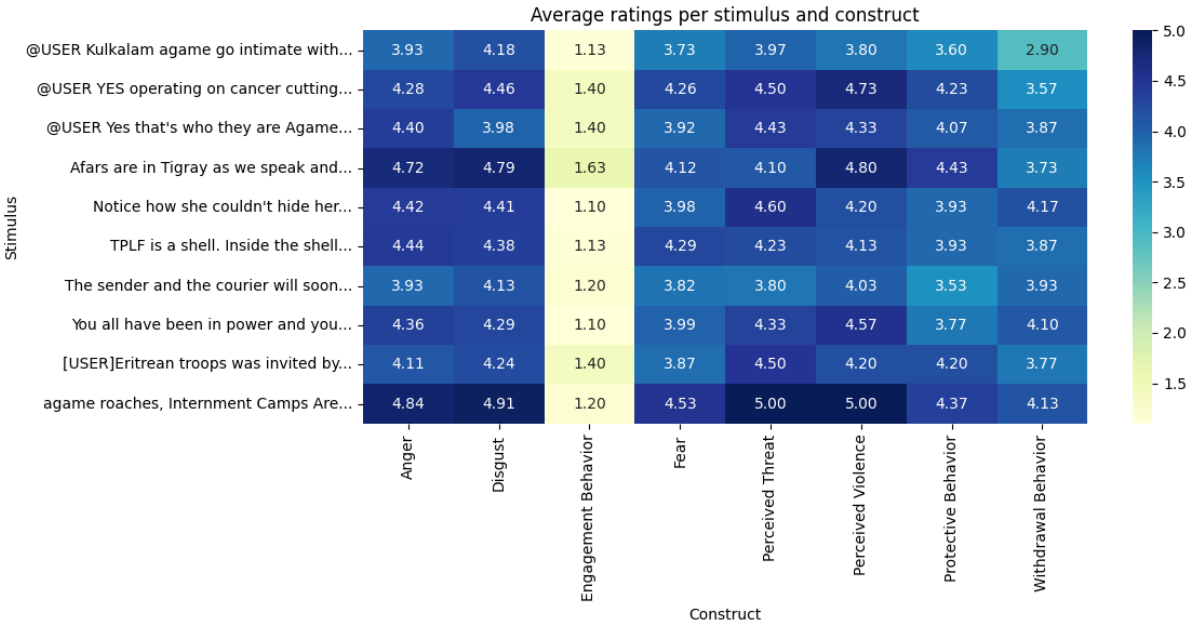

Figure 4.6: Patterns Across Tweets and Constructs

comparing Tigrayan fighters to a cancer to be cut out still elicit anger and disgust means above 4.3 but receive marginally lower fear and protective behavior scores. Participants view these statements as hateful yet not as imminently violent as explicit calls to shoot (direct advocacy). Third, insults with ambiguous threat yield lower fear. Messages such as "Kulkalam agame go intimate with ur arid land and cactus tree" (EAT) and "The sender and the courier will soon be sent to hell" (TSC) produce anger means around 3.9 - 4.0 and fear means closer to 3.8. Respondents still perceive these as derogatory but less physically dangerous, resulting in slightly lower protective and withdrawal intentions. These are tweets groups under Economic Anti-Tigrayanism and Target-Sympathizer conflation respectively. And lastly, behavioral intentions mirror emotional intensity. The patterns in Protective and Withdrawal Behaviors largely follow those of Anger and Disgust. Tweets rated as most threatening induce the strongest intention to report or block the user and to avoid similar content. Conversely, Engagement Behavior remains uniformly low across all stimuli, with means hovering around 1.1-1.6. Even comparatively "mild" insults do not motivate participants to engage positively with the tweet or

share it.

For anger, the most extreme tweets "Afars are in Tigray as I speak and Amhara militias are cleaning their region from agame roaches. Soon I will be coming back. And 3 forces will be looting Mekelle to the ground." and "agame roaches, Internment Camps Are Waste of Resources: Just Shoot Them" generate mean anger ratings above 4.7, whereas the least provocative message ("@USER Kulkalam agame go intimate with ur arid land and cactus tree.") averages around 3.9. This shows that Pathologization and Direct Advocacy tweets were classified as more angering than Economic Anti-Tigrayanism (EAT).

While fear ratings show greater dispersion with several tweets clustering near 4.0, the most threatening message of "shoot them" which is direct advocacy exceed 4.8. Disgust ratings are uniformly high, with even milder insults averaging around 4.0 but the direct advocacy tweet and Pathologization tweets got higher disgust reactions.

Perceived threat closely tracks anger and disgust, reaching its maximum for genocidal rhetoric and declining only slightly for insults without explicit violence. Perceived violence mirrors the threat pattern, but shows that statements involving physical harm (e.g. shooting or looting) are considered slightly more violent than metaphorical insults. All tweets score near the lowest possible value on intentions to like/share/comment (engagement behavior), underscoring participants' blanket refusal to amplify hateful content. Intentions to report or block the sender are highest for the most violent messages but remain relatively high ($\approx$ 3.5-4.5) even for less extreme insults. Mean withdrawal intentions vary between 2.9 and 4.3, indicating that participants generally choose to avoid further exposure, with the strongest avoidance corresponding to the most threatening tweets.

To check if the constructs are correlated to each other, I conducted correlations between constructs using respondent–tweet averages, as can be seen in Figure 4.7.

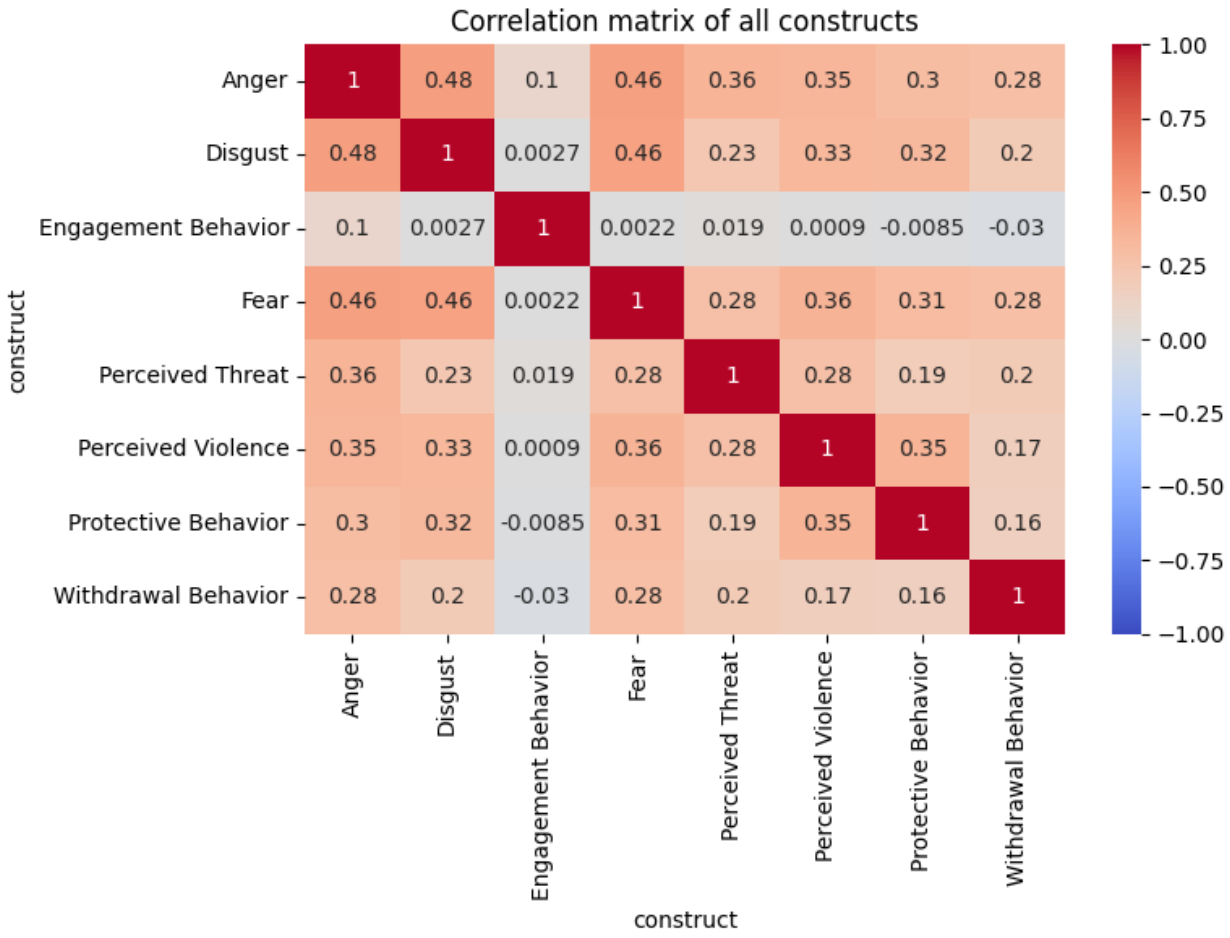

Figure 4.7: Correlations Among Constructs Used in the Survey

Anger, Fear and Disgust correlate moderately to strongly ($r \approx 0.46$-0.48), indicating that respondents often experience these emotions together when confronted with hate speech. Anger and Disgust exhibit the strongest link ($r \approx 0.48$).

Protective and withdrawal behaviors relate to negative emotions. Protective Behavior correlates positively with Anger, Fear and Disgust ($r \approx 0.30$-0.35), while Withdrawal Behavior shows slightly lower but still positive associations. These relationships could imply that stronger emotional reactions motivate both types of defensive action.

Engagement behavior is inversely related to negative reactions. Engagement Behavior displays

Category	Anger	Fear	Disgust	Perceived Threat	Perceived Violence	Engagement Behaviors	Protective Behaviors	Withdrawal Behaviors
DA (Direct Advocacy)	4.84	4.53	4.91	5.00	5.00	1.20	4.37	4.13
Verminization	4.44	4.29	4.38	4.23	4.13	1.13	3.93	3.87
Pathologization	4.50	4.12	4.79	4.10	4.80	1.63	4.43	3.65
Disloyalty	4.39	3.99	4.29	4.33	4.57	1.10	3.77	4.04
Justification	4.11	3.87	4.24	4.50	4.20	1.40	4.20	3.77
TSC (Target–Sympathizer Conflation)	3.93	3.82	4.13	3.80	4.03	1.20	3.53	3.93
EAT (Economic Anti-Tigrayanism)	3.93	3.73	4.18	3.97	3.80	1.13	3.60	2.90

Table 4.6: Emotional and Behavioral Responses by Class

near-zero or slightly negative correlations with all other constructs. This demonstrates that respondents who feel angrier, more fearful or more disgusted are less likely to engage positively with the tweet. Positive engagement with genocidal speech is virtually nonexistent in this sample.

To assess whether the model's categories align with how people actually respond to hate speech, average ratings from the survey were computed for each category and construct. Table 4.6 summarizes the mean scores (1-5 scale) across the eight constructs defined earlier.

Table 4.6 indicates that Direct Advocacy (DA) evokes the strongest reactions. The tweets calling for shooting Tigrayans scores at or near the maximum on Anger, Disgust, Perceived Threat and Violence, and elicits the highest protective and withdrawal intentions. The near-zero Engagement score underscores that participants unequivocally reject such exhortations. The model's classification of this tweet as Direct Advocacy is thus validated by participants' intense cognitive, emotional and behavioral responses.

Pathologization and Verminization provoke similar negativity. Comparing TPLF to a cancer or to lice triggers high anger, disgust and threat perceptions (means $\approx$ 4.5-4.8) and strong protective actions. Although the semi-supervised model treats these as separate classes, respondents react to both with comparable severity. This suggests that models might consider merging Pathologization

and Verminization into a broader "dehumanization" category when the goal is to predict psychological impact.

Disloyalty and Justification elicit moderate but still elevated responses. Tweets accusing Tigrayans of betrayal or attempting to rationalize violence produce slightly lower fear ($\approx$ 3.9-4.0) yet still high anger and disgust ($\approx$ 4.1-4.4). Protective and withdrawal behaviors remain elevated. The model correctly distinguishes these subtler narratives from direct calls to violence, but their real-world impact remains harmful.

Economic Anti-Tigrayanism is relatively less threatening. The sarcastic insult about "arid land and cactus tree" receives the lowest ratings across most constructs (Anger $\approx$ 3.9, Fear $\approx$ 3.7) and the lowest withdrawal intention (2.90). Participants perceive it as offensive but less dangerous. If the model treats such insults as a distinct class, it may be appropriate to assign them a lower priority for moderation.

Target Sympathizer Conflation sits between categories. The tweet threatening to send "the sender and courier to hell" evokes moderate anger (3.93) and moderate threat perceptions (3.80). This suggests that conflating sympathizers with traitors is perceived as harmful but not as virulent as direct advocacy. Models should recognize these hybrid forms but may need separate training data to distinguish them.

4.2.1.3 Repeated Measures ANOVA

Using a repeated-measures approach, each participant's ratings for a given construct were averaged within category and then analyzed across categories using AnovaRM from statsmodels. This accounts for the fact that the same respondents provided ratings across all hate speech categories.

Direct Advocacy and dehumanization categories provoke the strongest emotional and cognitive

reactions, while Economic Anti-Tigrayanism and Target Sympathizer Conflation elicit the weakest. These findings reinforce the robustness of the category effects and suggest that the observed differences are not driven by a few particularly reactive individuals but are consistent across the sample.

To see how anger, disgust, fear, etc. vary between specific hate speech types such as Verminization versus direct advocacy or Pathologization versus disloyalty, I examined the post-hoc pairwise comparisons and the tables of mean differences. That is, to identify specifically which categories differed after the repeated-measures ANOVA, I performed a series of post-hoc analyses tailored to the within-subject nature of the data. Below is an overview of the methodology, findings and their implications.

Because every respondent rated every hate speech category, the appropriate post-hoc comparison is a paired (within-subject) t-test rather than an independent sample test. For each construct, I averaged each participant's ratings for each category, then computed all pairwise differences (21 pairs) between categories. To control the family-wise error rate, i.e. the chance of flagging a difference as significant merely by chance, I applied the Holm correction, a step-wise method that is less conservative than Bonferroni yet maintains control over Type-I error. This approach ensures that the reported significant differences are not spurious while retaining power to detect real effects.

Table 4.7 lists category pairs with adjusted p-values < 0.05 after the Holm correction, along with the mean difference (first category minus second). Positive differences indicate that the first category elicited a higher rating.

Anger

Direct Advocacy tweets produced the strongest anger exceeding all other categories. These tweets did not simply look more "hateful" in abstract terms, they provoked intense self-reported anger across participants, significantly higher than many other categories. By contrast, Economic Anti-

Tigrayanism and more abstract justification tweets evoked markedly lower anger. This pattern aligns with the theory that metaphors that portray a group as vermin, disease or an existential threat are especially powerful triggers of moral outrage.

Disgust

The repeated-measures ANOVA revealed highly significant differences in disgust ratings across categories ($F \approx 8.54$, $p \approx 3.9\times10^{-8}$). Direct Advocacy tweets produced the strongest disgust (mean 4.91), significantly exceeding all other categories. Pathologization and Verminization followed closely; participants reacted almost as disgustedly to metaphors likening Tigrayans to disease or vermin as they did to calls for extermination. Disgust was moderately high for Disloyalty and Justification but significantly lower for Economic Anti-Tigrayanism and Target Sympathizer Conflation, which elicited the least revulsion. The pattern underscores that dehumanizing metaphors are perceived as morally repulsive, even when they stop short of explicit violence. Moreover, collapsing Pathologization and Verminization into a single "dehumanisation" class may reflect participants' indistinguishable disgust responses.

Fear

Although fear ratings varied significantly across categories ($F \approx 5.21$, $p \approx 5.9\times10^{-5}$), the differences were smaller than for anger or disgust. Direct Advocacy still elicited the highest fear (mean 4.53), but the gap between this and other categories was narrower. Economic Anti-Tigrayanism was rated lowest (mean 3.73), and fear for Verminization, Pathologization, Disloyalty and Justification fell between 3.9 and 4.3. Only a handful of pairwise contrasts remained significant after correction, such as direct advocacy vs. economic insults and vs. target sympathizer conflation. This suggests that respondents see genocidal rhetoric as morally outrageous and threatening, but not necessarily as an immediate personal danger, differentiating fear from cognitive appraisals of threat. Automated systems should not

equate fear with threat perceptions and might consider modeling personal fear separately from other harm indicators.

Perceived Threat

Perceived threat showed robust category effects (F $\approx$ 5.04, $p \approx 8.6\times10^{-5}$). Direct Advocacy scored at the ceiling (mean 5.00), indicating that exhortations to "shoot" or "exterminate" are seen as maximally threatening. Pathologization and Verminization provoked similarly high threat ($\approx$4.3 - 4.5), while Disloyalty and Justification were moderate ($\approx$4.3 - 4.5). Economic insults and Target Sympathizer Conflation were perceived as least threatening ($\approx$3.8 - 4.0). Post-hoc tests showed that direct advocacy was significantly more threatening than all other categories and that economic insults were significantly less threatening than several mid and high-level classes. These findings affirm the model's severity ordering and suggest that perceived threat is driven more by the explicitness of violence or dehumanization than by sarcasm or betrayal. I argue that for moderators, content that maximizes perceived threat merits the highest priority for removal

Perceived Violence

Perceived violence differed strongly across categories (F $\approx$ 6.13, $p \approx 7.5\times10^{-6}$). Direct Advocacy again topped the scale (mean 5.00), while Pathologization was only slightly lower ($\approx$4.77). Verminization and Disloyalty occupied a middle tier ($\approx$4.1- 4.6), Economic Anti-Tigrayanism was rated least violent ($\approx$3.80). Significance patterns resembled those for threat: direct advocacy vs. economic insults and vs. target sympathizer conflation differences were large and highly significant. The moderate correlation between perceived threat and violence (r$\approx$0.28) and the small mean difference could suggest that participants conceptually distinguish violence and threat but often rate them similarly. Models could therefore treat these constructs as distinct but closely linked dimensions of perceived harm.

Protective, Withdrawal and Engagement Behaviors

Intentions to report/block a tweet (protective behavior) and to avoid similar content (withdrawal behavior) varied modestly but significantly across categories ($F \approx 3.11$, $p \approx 6.4\times10^{-3}$; $F \approx 3.95$, $p \approx 9.9\times10^{-4}$). Participants were most willing to take protective or withdrawal actions against Direct Advocacy and Pathologization tweets and least likely to do so for Economic insults. Differences among the remaining categories were smaller and often non-significant, suggesting that once content crosses a certain harm threshold, respondents uniformly favor protective behaviors. Engagement behavior (liking/sharing) was uniformly low across all categories but showed a borderline significant effect in the repeated-measures ANOVA ($F \approx 2.19$, $p \approx 0.046$), indicating slightly higher (though still very low) engagement with Pathologization compared to Verminization. This reinforces the observation that participants reject hate speech content regardless of type and underscores the impracticality of models attempting to predict positive engagement with such messages.

4.2.2 Summary: Study 2

Direct Advocacy is uniformly the most extreme. It scores significantly higher than almost every other category across anger, disgust, fear, perceived threat and violence, often by large margins (≥ 0.70 on a 1-5 scale). This underscores the psychological severity of explicit calls to violence.

Dehumanization categories (Pathologization and Verminization) occupy an intermediate but still high tier. Pathologization is significantly more angering and disgusting than target sympathizer conflations and economic insults. Verminization also differs from economic insults on anger and fear. The similarity of their profiles suggests these two classes could be merged for practical purposes.

Economic insults are consistently the least harmful. These tweets evoked lower anger, disgust

and withdrawal, pointing to a distinct psychological footprint.

Justification and disloyalty occupy a middle ground. Disloyalty is more angering than economic insults and target sympathizer conflation, and justificatory rhetoric is less angering and disgusting than Pathologization. However, justification still yields relatively high perceived threat and violence.

Behavioral intentions respond to category differences. Protective behavior (reporting/blocking) is higher for Pathologization than for Verminization, while withdrawal (avoiding content) is notably higher for disloyalty compared to economic insults. Engagement (liking/sharing) shows little variation and is low across all constructs.

These post-hoc results reinforce and refine the conclusions drawn from the ANOVA: explicit advocacy of violence stands apart in its impact, dehumanizing metaphors (Verminization, Pathologization) are nearly as harmful, and economic insults lie at the bottom. By quantifying the magnitude of differences and adjusting for multiple comparisons, the analysis provides a robust empirical basis for prioritizing content moderation and model calibration.

These construct-specific findings illuminate how different facets of harm vary across hate speech categories. They support the supervised model's broad taxonomy while suggesting refinements. Dehumanization (including Pathologization and Verminization) and direct advocacy should be treated as high-priority classes, given their strong effects on disgust, threat and violence perceptions. Fear responses warrant separate consideration because they are less extreme and do not align perfectly with cognitive appraisals. Behavioral intentions reveal practical outcomes: categories that elicit high protective and withdrawal responses are those that most compel users to report or avoid content, guiding prioritization for moderation.

In sum, this survey demonstrates that hate speech tweets directed at Tigrayans elicit intense negative emotions, are seen as highly threatening and violent, and motivate strong defensive behaviors.

Participants almost never express a desire to engage with these tweets, opting instead to report, block or avoid the content. The strong relationships among perceived threat, perceived violence and the emotions of anger, fear and disgust underscore the importance of addressing genocidal rhetoric promptly. Incorporating human emotional feedback into AI-based moderation systems could enhance their ability to identify and mitigate harmful content, thereby better protecting vulnerable communities.

Construct	Pair (Category1 - Category2)	Mean difference	Significance (Holm)
Anger	Direct Advocacy – EAT	+0.91	$p \approx 8.6 \times 10^{-7}$
	Direct Advocacy – TSC	+0.91	$p \approx 1.1 \times 10^{-5}$
	Direct Advocacy – Justification	+0.73	$p \approx 3.9 \times 10^{-5}$
	Direct Advocacy – Pathologization	+0.34	$p \approx 4.1 \times 10^{-5}$
	Direct Advocacy – Disloyalty	+0.45	$p \approx 1.3 \times 10^{-4}$
	Pathologization – TSC	+0.57	$p \approx 2.99 \times 10^{-4}$
	EAT – Pathologization	–0.57	$p \approx 3.48 \times 10^{-4}$
	Disloyalty – EAT	+0.46	$p \approx 8.4 \times 10^{-3}$
	Justification – Pathologization	–0.39	$p \approx 9.0 \times 10^{-3}$
	EAT – Verminization	–0.51	$p \approx 1.2 \times 10^{-2}$
	Disloyalty – TSC	+0.46	$p \approx 1.4 \times 10^{-2}$
	TSC – Verminization	–0.51	$p \approx 2.7 \times 10^{-2}$
	Direct Advocacy – Verminization	+0.40	$p \approx 3.5 \times 10^{-2}$
Disgust	Direct Advocacy – Disloyalty	+0.69	$p \approx 2.1 \times 10^{-5}$
	Direct Advocacy – TSC	+0.78	$p \approx 4.9 \times 10^{-5}$
	Direct Advocacy – EAT	+0.73	$p \approx 1.98 \times 10^{-4}$
	Disloyalty – Pathologization	–0.40	$p \approx 4.9 \times 10^{-3}$
	Direct Advocacy – Justification	+0.67	$p \approx 5.3 \times 10^{-3}$
	Pathologization – TSC	+0.49	$p \approx 7.7 \times 10^{-3}$
	Direct Advocacy – Pathologization	+0.29	$p \approx 7.9 \times 10^{-3}$
	EAT – Pathologization	–0.44	$p \approx 1.0 \times 10^{-2}$
	Direct Advocacy – Verminization	+0.53	$p \approx 2.5 \times 10^{-2}$
Fear	Direct Advocacy – EAT	+0.80	$p \approx 6.7 \times 10^{-4}$
	EAT – Verminization	–0.56	$p \approx 5.8 \times 10^{-3}$
	Direct Advocacy – Justification	+0.67	$p \approx 1.8 \times 10^{-2}$
	Direct Advocacy – TSC	+0.71	$p \approx 4.6 \times 10^{-2}$
	Direct Advocacy – Disloyalty	+0.57	$p \approx 4.9 \times 10^{-2}$
Perceived Threat	Direct Advocacy – EAT	+1.03	$p \approx 1.4 \times 10^{-3}$
	Direct Advocacy – Disloyalty	+0.54	$p \approx 2.2 \times 10^{-3}$
	Direct Advocacy – Pathologization	+0.70	$p \approx 2.2 \times 10^{-3}$
	Direct Advocacy – TSC	+1.20	$p \approx 2.2 \times 10^{-3}$
	Direct Advocacy – Verminization	+0.77	$p \approx 2.5 \times 10^{-2}$
Perceived Violence	Direct Advocacy – EAT	+1.20	$p \approx 2.0 \times 10^{-5}$
	EAT – Pathologization	–0.97	$p \approx 1.3 \times 10^{-4}$
	Direct Advocacy – Disloyalty	+0.63	$p \approx 3.6 \times 10^{-4}$
	Direct Advocacy – TSC	+0.97	$p \approx 1.8 \times 10^{-2}$
	Pathologization – TSC	+0.73	$p \approx 1.9 \times 10^{-2}$
	Direct Advocacy – Verminization	+0.87	$p \approx 2.1 \times 10^{-2}$
	Direct Advocacy – Justification	+0.80	$p \approx 2.6 \times 10^{-2}$
Engagement Behavior	Pathologization – Verminization	+0.38	$p \approx 4.6 \times 10^{-2}$
Withdrawal Behavior	Disloyalty – Economic Anti-Tigrayanism	+1.14	$p \approx 1.2 \times 10^{-2}$

Table 4.7: Pairwise Comparisons of Constructs Across Hate Speech Categories. Mean Differences and Holm-Corrected p-values are Reported.

Chapter 5: Discussion

5.1 Advancing Theory

This study draws heavily on Gregory Gordon's "incitement to genocide" framework, and my empirical work demonstrates that key aspects of his theory can be operationalized and reliably identified. Gordon identifies several rhetorical forms such as verminization, pathologization, demonization, direct advocacy, accusation in a mirror, justification, praising past violence, and target sympathizer conflation that function as incitement techniques. In my annotated corpus, I implemented these categories as distinct labels and trained a classifier to detect them in Tigray war tweets. The supervised model's strong performance on categories like direct advocacy and dehumanization (verminization and pathologization) indicates that these rhetorical patterns are linguistically cohesive. For example, the classifier effectively identified tweets comparing Tigrayans to animals or diseases, aligning with Gordon's argument that such comparisons portray victims as beings "whose extermination would be considered normal and desirable" (Gordon (2017)). My survey results further support this finding: respondents consistently rated dehumanizing (pathologization and verminization) and direct advocacy tweets as highly threatening and morally outrageous, which reflects Gordon's claim that dehumanization lowers the moral threshold for genocide. In this sense, the study shows that core components of Gordon's framework can be successfully implemented and detected in both ML models and human judgments.

My findings also extend the theory by quantifying the relative impact of less studied categories. For example, Target sympathizer conflation in the survey findings showed that it can evoke some levels of anger and disgust and modest protective behavior. Gordon mentions that target-sympathizer conflation can lead to violence against moderates during genocide. But there has been little empirical work on how audiences perceive such messages. My survey shows that while these tweets do harm,

their emotional impact tends to be lower than that of dehumanization or direct advocacy, suggesting a continuum of danger rather than a binary. Similarly, justification rhetoric, where genocidal acts are defended as necessary or humane elicited only moderate outrage in my sample; this indicates that justification may be more effective at sustaining violence among already radicalized audiences than at inciting new perpetrators. By measuring these differences, I extend Gordon's taxonomy from a qualitative list to a quantitative hierarchy of harm.

There are also aspects where my results challenge or refine the theory. First, the classifier frequently confused pathologization and verminization, and the survey responses to both were indistinguishable. This suggests that audiences do not perceive diseases and vermin metaphors as qualitatively distinct; merging them into a single "dehumanization" class may better reflect psychological reality.

Second, participants reported high anger and disgust but relatively low fear in response to all categories, including direct advocacy. Because respondents belong to the targeted group (Tigrayans), their reactions reflect how out-group hate speech is perceived by victims. In this context, the strong anger and disgust scores paired with relatively low fear point to a response pattern more consistent with moral outrage and dehumanization than with acute fear or terror. Victims predominantly feel anger, disgust, and moral outrage, highlighting that the emotional impact is often about indignity and injustice rather than simple fear (Giner-Sorolla and Russell (2019); Saha et al. (2019)). Dehumanizing metaphors whether likening people to vermin or disease epitomize this injustice by casting victims outside the bounds of humanity, and victims consistently perceive such language as equally damaging in its denial of their dignity (Bandura (1999); Gordon (2017)). The repercussions of hate speech also permeate the social realm: it erodes trust between groups, undermines solidarity within the targeted community, and weakens the victimized group's cohesive identity (Staub (1989)). Over time, individuals subjected to hate speech may grapple with alienation, shame, and fragmentation of their social ties,

as well as serious psychological strain analogous to trauma (Pascoe and Smart Richman (2000); Saha et al. (2019)). Thus, the emotional response profile observed here suggests that hateful content directed at Tigrayans may primarily trigger moral condemnation, disgust-based revulsion, and anger-based indignation, rather than existential fear. From a theoretical standpoint, this implies that incitement in this context may operate less by instilling fear in the target group, and more by mobilizing moral emotions, stigmatization, and dehumanization potentially weakening social solidarity, increasing alienation, and justifying exclusion or violence. These insights complement and challenge incitement-focused theories, which often center on bystanders or perpetrators; by focusing on victims, we see that the harm in hate speech is as much inward and cumulative as it is outward and contagious (Benesch (2012); Leader Maynard and Benesch (2016)).

Third, the model struggled to identify euphemisms and coded language such as Other category (which had justification and praising past violence) and this is probably because of the limited number of tweets labeled and that these phrases require contextual/historical knowledge. This limitation suggests that Gordon's emphasis on euphemism is theoretically sound but computationally challenging; more contextual and pragmatic features are necessary to detect coded incitement. Finally, my attempt to extend the classifier to Amharic and Tigrinya tweets yielded poor performance due to lack of annotated data, and could signal that theories derived from English-language examples may not transfer seamlessly across languages or that it would need further refinment per nation/region it is applied to.

I also extend his framework by adding labels for Disloyalty and Economic Anti-Tigrayanism. While these labeles do not directly fall within the iincitement techniques he had listed, they were labels that emerged from looking at the tweets and discourse surrounding the Tigray war. By incorporating these labels, my taxonomy becomes more sensitive to the subtle, pre-violence conditions under which mass atrocities often germinate: delegitimization, scapegoating, internal-enemy framing, and

perceived economic threat. This extension retains Gordon's legal-scholarly framework, which emphasizes that incitement can take many forms beyond overt calls for killing but adapts it to the complex realities of modern ethnic conflict, resource competition, and identity-based tension.

By translating Gordon's incitement categories into machine-learning labels and testing them on real-world data, I was able to cooroborate the salience of dehumanization and direct advocacy, extend the framework with empirical measures of other categories (like disloyalty and Economic Anti-Tigrayanism) and identify areas where the theory may require refinement, such as collapsing dehumanization subtypes (pathologization, verminazation, demonization) in to one larger category. These findings not only bolster the theoretical understanding of genocidal rhetoric but also inform the development of more nuanced hate speech detection systems.

In sum, empirically, my research builds directly on Gregory Gordon's incitement-to-genocide framework and translates his legal categories into an empirical study of online hate. By collecting and annotating Tigray-war tweets into these categories, I operationalized his theory for ML. The supervised model learned to identify these patterns with good precision, and the confusion matrices showed that dehumanizing metaphors are linguistically cohesive. These results support Gordon's claim that certain metaphors are hallmarks of genocidal speech and demonstrate that they can be detected at scale.

My work also extends Gordon (2017) theory by quantifying the psychological impact of each incitement technique category through survey data. Participants rated direct advocacy and dehumanizing tweets as particularly angering, disgusting, and threatening, suggesting that these types of speech are perceived as especially severe and may warrant higher prioritization in monitoring frameworks. However, the study also highlights the importance of detecting less extreme rhetorical forms. For instance, target-sympathizer conflation which is a tactic Gordon identifies as part of the broader incitement landscape whereby moderates are labeled as traitors and subjected to violence, elicited lower

emotional responses than direct advocacy but higher than economic insults, placing it in a mid-range zone of harm. This suggests a continuum of rhetorical severity rather than a binary threshold, with implications for early detection. In Gordon's taxonomy, such rhetoric may not rise to the level of incitement per sé but can be situated within "abetting" or "instigation" categories, where speech facilitates or encourages atrocities indirectly (Gordon (2017)). Recognizing and flagging these subtler expressions is crucial because genocidal violence rarely begins with overt calls for killing; instead, it typically escalates through a sequence of legitimizing narratives and social conditioning (Benesch (2012); Leader Maynard and Benesch (2016)). Therefore, identifying early-stage categories such as target-sympathizer conflation or economic scapegoating supports preventative strategies aimed at disrupting the rhetorical lead-up to mass atrocity.

By bridging social-psychological theory, international criminal law and computational linguistics, my dissertation makes three contributions to hate speech scholarship. It demonstrates that Gordon's incitement categories have measurable linguistic signatures; it extends the framework by ranking categories by harm and proposing refinements (such as collapsing dehumanization subtypes and adding "economic insult" as a lower-severity class); and it also highlights the need for more context-sensitive studies and multilingual models to capture the full spectrum of dangerous speech.

5.2 Validation of Supervised Model

By eliciting ratings for representative tweets across seven categories of hate speech, I aimed not only to describe participants' responses but also to test and critique the supervised model trained to classify genocidal speech. Several key themes emerged.

First, not all hate speech elicits the same level of harm. The analyses consistently showed that

explicit advocacy of violence, tweets that unambiguously call for extermination, drew the most intense anger, disgust, fear and appraisals of threat and violence. Dehumanizing metaphors (pathologization and verminization) followed closely. Disloyalty accusations and justificatory rhetoric occupied a middle ground, while economic insults and target sympathizer conflations elicited the least severe reactions. These gradations matter, they suggest that Tigrayan respondents are sensitive not only to whether speech is hateful, but to how it is hateful, whether it pathologizes a group, accuses it of betrayal, justifies violence or calls for violence outright.

The supervised HateBERT model used the same taxonomic labels (verminization, pathologization, demonization, economic scapegoating, etc.). Dehumanizing metaphors like "roaches" and "cancer" feature prominently in the literature on dangerous speech because they move audiences to view mass violence as necessary. The survey therefore supports the classifier's emphasis on these categories - reducing false negatives in such classes is critical, whereas false negatives in economic scapegoating or disloyalty (rated around 4.0) have lower immediate risk.

Second, different psychological constructs reveal distinct nuances. Anger and disgust tracked closely with each other and with perceived threat and violence, indicating that moral indignation and appraisals of danger rise together in response to dehumanizing language. Fear was lower overall and varied less between categories, suggesting that respondents often view online hate speech as reprehensible but not necessarily as posing an immediate personal danger. Behavioral intentions further distinguish the categories: protective and withdrawal actions were more common for the most extreme and dehumanizing tweets, while engagement intentions (liking, sharing or commenting) remained uniformly near the floor across all categories. These patterns underscore the importance of measuring multiple dimensions of harm rather than assuming a single continuum.

Third, the survey findings both support and challenge the supervised model. On the one hand, the

model's taxonomy broadly aligns with participants' perceptions: tweets labelled as "Direct Advocacy" and "Pathologization" were indeed viewed as most harmful, while "Economic Anti-Tigrayanism" and "Target Sympathizer Conflation" were seen as relatively milder. This correspondence suggests that the lexical and thematic cues captured by the model have psychological validity. On the other hand, the model does not currently account for the severity of harm within categories. My results highlight that pathologization and verminization, though lacking explicit violence, elicited harm levels approaching those of direct advocacy. Conversely, justificatory rhetoric, although ideologically dangerous, provoked emotional responses closer to those for economic insults. The model also treats categories as mutually exclusive, whereas respondents' reactions suggest a continuum of harm influenced by overlapping rhetorical devices (e.g., a tweet may simultaneously pathologize and advocate violence). Thus, while the model's categories capture some semantic distinctions, they could benefit from incorporating human ratings to weight severity and handle category blending.

In sum, the survey demonstrates that the supervised approach is grounded in human judgements of incitement such that most of its labels correspond to how readers experience hate speech, and its improvements focus on the very categories (verminization and pathologization) that experts and survey participants alike regard as the most likely to lead to violence.

Additionally, the findings of this research map onto recognized stages in the escalation toward genocide by empirically showing how specific rhetorical forms function as early, mid, and late-stage incitement. Using Gordon (2017)'s framework of genocidal speech modalities, incitement, instigation, abetting, and ordering, the study demonstrates that different hate speech categories contribute to genocide differently depending on their severity, perceived emotional impact, and rhetorical purpose. For example, dehumanization and direct advocacy were shown to be linguistically cohesive and emotionally potent, eliciting high levels of disgust, anger, and threat perception in survey respondents.

These patterns align with mid-to-late stages of genocide escalation, where targets are morally excluded and violence becomes socially acceptable (Gordon (2017); Bandura (1999); Giner-Sorolla and Russell (2019)).

Conversely, categories like economic scapegoating and target-sympathizer conflation represent earlier phases of escalation. These rhetorical strategies sow division and justify exclusion without overtly calling for violence. Yet, as literature on dangerous speech shows, such messages lay the groundwork for atrocity by normalizing group blame and delegitimizing moderates (Benesch (2012); Leader Maynard and Benesch (2016)). The study's finding that these less extreme categories provoke moderate emotional responses supports their role in softening resistance to violence over time, functioning as rhetorical "priming" (Allport (1954); Staub (1989)). Thus, the research extends theoretical models by showing that incitement operates along a continuum and that early-warning systems should flag not only explicit incitement but also less inflammatory categories that enable escalation.

5.3 Moving Beyond Binary Classification

One major contribution of this research is the shift from simplistic binary hate speech detection toward a nuanced multiclass classification. Traditional approaches often dichotomize content as hate vs. non-hate, which obscures important variations in severity and type of hateful expression. For example, the HateBERT model (a BERT-based hate speech detector) was initially designed for binary classification of abusive content.

In contrast, my work recognizes that hate speech is not monolithic. Following recent scholarship calling to "move beyond traditional binary models", I implemented a multiclass framework that differentiates among categories of hateful discourse (Aldreabi et al. (2024)). This approach captures a

spectrum from subtle biases to explicit calls for violence, rather than lumping all instances into a single hateful class. Such granularity is valuable because mildly antagonistic or dehumanizing language can escalate into severe incitement if left unchecked (Aldreabi et al. (2024)). By identifying intermediate categories (for instance, content that contains dehumanizing metaphors versus outright genocidal threats), the classifier provides a more informative analysis of hate speech dynamics. This is especially important in contexts of ethnic or political conflict, where early-warning signs of atrocity may manifest as milder forms of hate speech before intensifying.

In sum, adopting a multiclass classification is a deliberate theoretical and practical advancement that addresses the complexity of hate speech. It aligns with recent studies that successfully categorized online hate into multiple tiers of severity (e.g. implicit bias vs. explicit incitement) to reveal variations that binary models would miss (Aldreabi et al. (2024)). I thus underscore that moving beyond a binary hate/not-hate paradigm is not only a methodological choice but a substantive contribution of this dissertation, yielding richer insights into the gradations of hate speech in the data.

5.4 Insights for Model Refinement

Despite this broad correspondence, the survey also exposes several limitations of the current model and suggests concrete avenues for enhancement delineated below.

The model treats category membership as a binary outcome, yet the data show substantial variability within categories. For example, some pathologization tweets were nearly as harmful as direct advocacy, whereas others evoked more moderate reactions. A supervised component could be trained on continuous harm scores (e.g. average anger or threat ratings) rather than solely on categorical labels. This would allow the model to assign a severity weight to each instance, providing more nuanced

outputs for downstream moderation.

Additionally, the significant overlap in responses between certain categories (e.g. pathologization vs. verminization) suggests that these might be merged into a broader dehumanization class for practical purposes. Conversely, the minimal difference between economic insults and target sympathizer conflation in most constructs implies that the latter class may not warrant separate treatment unless contextual factors (e.g. offline violence) are present.

Some tweets contained multiple hate strategies (e.g. demonization coupled with calls for violence). A strictly single-label model will force such messages into one category, potentially underestimating their harm. A multi-label classifier could capture the co-occurrence of rhetorical devices and better reflect the psychological impact revealed in the survey.

The current supervised model likely relies on lexical patterns and topic distributions. Incorporating features correlated with human affect such as sentiment scores, threat lexicons, or psycholinguistic markers (LIWC categories) may improve the classifier's ability to rank content by harm. For instance, weighting words associated with disgust or anger more heavily could help the model prioritize content that people find most repulsive.

Moreover, the alignment between certain model categories and human judgments suggests that user feedback could be leveraged to fine-tune decision thresholds. An active learning loop in which the model solicits ratings for borderline cases from domain experts could correct misclassifications and calibrate severity weights over time. For example, some justification tweets were perceived as less harmful than expected, reflecting the nuanced role of context and sarcasm. To address this, the model could incorporate discourse-level features (e.g. preceding tweets, reply structure) and pragmatic cues (e.g. quotation marks, rhetorical questions). Jointly modelling content and context would reduce false positives and better align the classifier's output with human harm judgments.

Lastly, many of the original tweets mix languages and rely on code-switching. The supervised component should therefore include multilingual embeddings and transliteration handling to ensure that dehumanizing metaphors or advocacy phrases are recognized regardless of script or orthography.

In sum, the survey findings provide an empirical grounding for the supervised model's categories while highlighting the need for severity weighting, multi-label classification and richer feature representations. Integrating human affective data and contextual analysis into model training would produce a more nuanced tool for identifying harmful speech.

5.5 Limitations

While the results are promising and lend credence to the theory used and empirically tests that a more nuanced classification of hate speech could be done, there were some limitations.

For one, the dataset was highly skewed, with the majority class (Q/E) comprising over 60% of examples and some critical categories like Justification and Praising Past Violence extremely rare (<1%) which were later subsumed under category 'Other'. This imbalance constrained the supervised learning such that the model had difficulty learning robust patterns for underrepresented categories, as evidenced by low recall and precision for those classes in the baseline. This limitation means that the classifier might miss some instances of the most dangerous rhetoric simply due to lack of training examples. In a real-world setting, one would want to continually update the training data with new examples of those rare but important categories to improve detection.

Second, annotating hate speech with fine-grained categories is challenging. Some tweets could reasonably fit multiple categories (e.g., a message could simultaneously demonize and call for violence). I enforced a single label per tweet for model training, which may oversimplify the content. The

confusion in the model between some categories and that human annotators also had disagreements or found overlaps shows that a multi-label approach may be a better way.

Additionally, my analysis focused on English-language tweets, even though much of the conflict-related hate speech occurred in Amharic, Tigrinya, and other Ethiopian languages. This is a significant limitation because it narrows the scope of findings. The English tweets often included local terms transliterated (e.g., "agame"), but the model was not trained on Amharic script or Tigrinya script content. Thus, my system would not catch hate speech written in native languages. Moreover, Ethiopia's online space is rife with code-switching, mixing English with Amharic phrases, etc. My preprocessing did not specifically tackle complex code-switched text (we treated transliterated words as English tokens). This might have led to loss of information or misclassification if part of a tweet was in another language. Tools for those languages (tokenizers, embeddings) were not integrated, which is a limitation in achieving a truly multilingual hate speech detector.

Moreover, the model was trained on data from a specific conflict (the Tigray war) during a specific time frame (2020-2021). The language used, slang terms, and targets of hate are domain-specific. This raises questions of generalizability. That is, if the model were applied to a different context (say, hate speech in a different country or a different conflict), its effectiveness is uncertain. This limitation suggests that continuous updates and possibly transfer learning would be needed to maintain performance over time or across conflicts.

This thesis focused solely on analyzing individual tweets. I did not include context such as who sent the tweet, whether it was a reply or part of a broader thread, or network information on how hate speech spreads (for example, through retweets or specific influential accounts). In actuality, hazardous speech frequently obtains power as a result of user networks' repetition and amplification. For example, a seemingly mild statement could become dangerous if it's echoing a known hate slogan or if the user

has a history of extremist posts. The model built judges only the text at face value, which is a limitation when deploying in a platform setting. It could be improved by context-aware architectures (like taking into account preceding tweets or the user's profile), but that was outside the current scope. Additionally, I did not address multi-modal content, i,e, hate messages might come with images or videos (memes, etc.), which my text-only approach cannot handle.

On the technical side the model does misclassify and make mistakes. False negatives (hateful content not identified) are a concern because they represent missed warnings; false positives (neutral content flagged as hate) are also problematic, especially if such a model were used in content moderation, it could lead to over-censorship or alarm fatigue. My results show that certain categories tended towards lower precision (like Verminization had many false positives in baseline). This limitation means the model's outputs should be interpreted with care, and a human-in-the-loop is needed for critical decisions.

We primarily used standard classification metrics and a held-out test set for evaluation. This is appropriate for measuring performance, but there are some limitations here. The test set is relatively small for some classes (e.g., only 15 instances of "Other"), making the evaluation of those classes' performance noisy. A few correct or incorrect predictions one way could change the F1 notably. Ideally, cross-validation or additional evaluation on an external dataset would strengthen confidence in results, but I lacked an external benchmark since my taxonomy is novel.

In summary, these limitations clarify that while my research made significant strides in classifying dangerous speech, it is not a complete solution. There are data-related constraints (imbalance, domain specificity), methodological issues (annotation subjectivity, lack of context), and deployment challenges (multilingual gaps, error implications) that must be addressed moving forward. Being aware of these constraints is essential for interpreting the results correctly and for guiding future enhancements

to this work.

A limitation regarding my second study is that while it captured the perceptions of the Tigrayan diaspora to tweets classified by my prior ML model- it had a small sample size. The survey comprises 30 respondents and 10 tweets too few to draw definitive conclusions about cross-category differences. More extensive human-rating data across diverse populations would improve confidence in the model's taxonomy.

The other fundamental limitation is that the instrument does not use a fully validated or psychometrically established scale. While I drew heavily from existing affective and threat-appraisal literature, the survey items were created specifically for this study and have not undergone formal validation such as factor analysis, reliability testing, or cross-sample replication. As a result, caution is required when interpreting the absolute values of the emotional ratings, because the internal structure of the scale is still exploratory.

Related to this, some constructs suffer from uneven item representation. Anger and disgust were measured with three items each, but perceived threat and perceived violence were measured with only one item per construct. Engagement, protective behavior, and withdrawal behavior were also measured with a single item. Single-item measures limit reliability because they cannot capture the full conceptual complexity of each construct.

Another limitation concerns construct overlap. Because items such as "The speech made me feel outraged" (anger) and "The speech content was repulsive to me" (disgust) are conceptually related, the scale may have inadvertently captured blended moral emotions. While the repeated-measures ANOVA detected clear differences across hate speech categories, the underlying constructs may not represent clean separations between anger, disgust, and threat. This could inflate correlations among the negative affect variables and reduce the ability to detect finer-grained emotional differences.

There are also structural limitations associated with the survey design and stimuli. Each participant evaluated only one tweet per category, drawn from examples classified by the ML model. Although these tweets were prototypical of each category, they still represent only a narrow slice of the rhetorical diversity present within the Tigray War discourse. Some categories had 2 or more examples while classes like Demonization did not have a tweet that was used in this study due to the co-existence of other classes (thus the need for multi-label approach). This restricts the generalizability of the results and may exaggerate or underestimate emotional reactions to entire categories based solely on one representative tweet.

Despite the above-mentioned limitations, the survey still provided meaningful evidence that the ML-derived categories correspond to differentiable emotional profiles. Nevertheless, future work should incorporate validated multi-item scales for each construct, expand the number of tweets per category, increase sample size, include contextual information, and implement a mixed-methods design that integrates both quantitative ratings and open-ended qualitative explanations.

Chapter 6: Conclusion and Future Work

Building on the findings and acknowledging the limitations, I propose several directions for future research and practical application. One immediate next step is to extend the dataset and models to local languages and multiple platforms. Given that much dangerous speech in conflicts occurs in languages like Amharic or Tigrinya (in the Ethiopian context), future work should collect and label data in those languages, or utilize translation techniques. Developing multilingual hate speech models (or fine-tuning multilingual transformers like mBERT or XLM-R) would allow detection across English and local language content. Additionally, cultural context must be considered because slang and metaphors differ by language. Collaborating with native speakers and annotators from the region can improve the taxonomy and catch nuances that an English-focused approach might miss. Cross-cultural validation of the taxonomy would also be valuable to test the generality of these dangerous speech indicators.

It has also been noted that my data and analysis suffered from data imbalance. To address data imbalance, future efforts should gather more examples/data of the rare categories. This could involve targeted data collection than what I had used which was general key word filtering using common slurs and terms. For instance, one may have to use search terms or ML strategies to specifically find more content that falls under Justification or Praising Past Violence. Active learning could be employed, where the model highlights uncertain tweets for human annotators to label, thus efficiently adding informative examples. Another approach is data augmentation: generating synthetic examples of underrepresented classes (carefully, to avoid introducing bias) or using transfer learning from related task. Ensuring a more balanced dataset will help the model improve on the most critical categories. Moreover, future datasets could benefit from multi-label annotation, allowing tweets to have more than one category when appropriate. This would provide richer supervision and likely improve model handling of overlaps (some recent studies show multi-label classifiers can learn nuanced distinctions better by

acknowledging overlap).

Our current model treats each tweet in isolation and relies on the transformer to capture context. Future research can experiment with context-aware models. For example, incorporating conversation context (previous tweets in a thread) or metadata (user info, time of post) might improve disambiguation. A hateful tweet that is a reply might quote someone; having that previous tweet could help determine if it's an attack or a rebuttal. Additionally, network features could be integrated. For example using graph neural networks where nodes are users and hate speech propagation is analyzed, to see if known hate actors can be identified and their content weighed differently. On the feature side, using lexical resources like the Hatebase dictionary or the PeaceTech Lab lexicon (which I partially used for keyword searches) could boost the model. These resources might provide synonyms or related hate terms that my data did not include, thus making the model more robust. Another promising direction is to incorporate emotion and sentiment analysis as features. Dangerous speech often carries extreme negative sentiment or contempt. If the model knew a tweet's emotion profile (e.g., high anger, low empathy), it might help classify borderline cases. Research by psychologists could inform features like moral rhetoric or use of justification language (there are computational linguistics tools to detect moral framing, which might catch justification of harm).

It is important for real-world adoption that models are interpretable, especially in sensitive contexts like identifying speech that might incite violence. Future work should incorporate explainable AI techniques, for instance, using SHAP or LIME to highlight which words or phrases in a tweet led the model to classify it as Demonization or Disloyalty. This not only builds trust with end-users (such as analysts or platform moderators) but also allows verification that the model is making decisions for the right reasons (e.g., it tagged a post as Verminization because it saw the word "cockroach" which is a valid reason and not because of an irrelevant word). I also recommend a deeper integration with

social science theories. For example, researchers could use the model on millions of posts to quantify how rhetorical patterns correlate with offline events (does a spike in Direct Advocacy online precede real-world violence). Such analysis can feed back into theory, refining our understanding of dangerous speech propagation. In essence, the model could become a theory-testing instrument, but to do so reliably, we must ensure it is transparent and well-calibrated.

We recommend continuous evaluation of the model in the field. This includes regularly measuring its performance on new data (to catch any degradation if hate actors change tactics) and conducting user studies (e.g., have human experts rate the model's classifications or use it in a simulated moderation scenario to see how well it assists them). Such feedback can highlight new limitations and guide iterative improvements. Another future avenue is to extend the taxonomy if needed: our categories covered many known dangerous speech forms, but future events might reveal new patterns (for instance, deepfake-based incitements or novel conspiracy theories that do not squarely fit existing labels). The framework of creating a taxonomy from theory and open coding can be repeated to update the category set. In sum, treating the model and taxonomy as living tools that evolve with the landscape will ensure the research continues to have real-world relevance.

Related to the above argument, Abdelkadir et al. (2025) give a set of concrete recommendations aimed at platforms, policymakers, and the research community. These are grounded in their findings and echo calls from prior literature on content moderation and hate speech (e.g. proposals by civil society after the Myanmar crisis (Fink (2018)).

One such recommendation is to hire culturally and linguistically diverse moderators. Social media companies should ensure dialectical diversity among moderators who speak the same language, and hire people with in-depth cultural and contextual knowledge of the regions they moderate. This is especially vital in linguistically diverse regions (Lee et al. (2023)). For example, if a platform is

moderating content in Ethiopia, it should employ not just any Amharic speaker, but moderators covering different dialects of Amharic and with knowledge of Ethiopian history and local issues. Research strongly supports this approach of having moderators from the relevant community can catch nuanced hate speech that outsiders miss (Braun and Clarke (2012)). It also helps address biases; as Davani et al. (2021) note, diverse annotator pools lead to more inclusive definitions of harm (Braun and Clarke (2012)).

Additionally, recent advances in large-scale language models (LLMs) including instruction-tuned and multimodal architectures such as GPT-4 (OpenAI (2023)), LLaMA-3 (Llama Team (2024)), and XLM-R-Large (Conneau et al. (2020)), offer fundamentally new capabilities for detecting, interpreting, and classifying hate speech. These models differ from earlier transformer encoders like BERT or mBERT (Devlin et al. (2018)) in their ability to perform complex reasoning, follow explicit natural-language instructions, incorporate contextual world knowledge, and generalize across tasks without supervised training. Instruction-tuned LLMs can be asked directly, without gradient-based training to classify a tweet into theoretically grounded categories such as dehumanization, pathologization, demonization, or direct advocacy of violence (T. Brown et al. (2020); OpenAI (2023)). This is particularly useful for genocide-related speech, which often requires multi-step reasoning and interpretation rather than simple pattern recognition.

A persistent barrier in this dissertation was the extreme imbalance across hate categories, some categories contained fewer than 50 examples. LLMs are uniquely effective as data augmentors capable of generating high-quality synthetic examples that preserve rhetorical intent and linguistic realism (Li et al. (2023)). LLMs can generate paraphrases of rare hate categories, culturally grounded variants of insults or dehumanizing metaphors, code-switched examples reflecting real patterns, adversarial examples that test model robustness etc.

By pursuing these recommendations, future research can build a more robust, generalizable, and impactful system for detecting and countering dangerous speech. The ultimate goal would be to integrate such NLP models into a broader strategy of violence prevention, where online hate is monitored and mitigated before it translates into offline harm and mass atrocities. Each improvement in accuracy, language scope, or explainability moves us closer to that goal, turning the lessons of past atrocities into proactive tools for peace.

Chapter A: Seedwords by PeaceTech Lab

The lexicon by PeaceTech Lab (2020), attempts to act as an initial guide to particular words and phrases identified as particularly inflammatory or having inflammatory potential.. The purpose of this vocabulary, according to PeaceTech Lab, is to provide information to people and groups who monitor and combat hate speech in Ethiopia. Some of the key words used that are also suggested by the PeaceTech Lab (2020) are the following:

ባንዳ (banda):

Definition of Term or Phrase: Banda (**ባንዳ**) means "traitor" in Amharic and Tigrinya. A traitor is "someone who betrays his country" or "someone who unites with the enemy of his country for his own benefit" even if "his action could negatively affect his country" (PeaceTech Lab (2020)).

The term was used during the Italian occupation of Ethiopia (1935/36-1941) in particular to refer to Ethiopians who fought for the occupier (then Italy) as mercenaries. *Bande* as an Italian word refers to irregular military units (Treccani (n.d.)).

This therefore conferred an additional historically negative dimension to the usage of the term.

In the current political context, it is used by most groups or communities to stigmatize political opponents or other ethnic groups as "sellouts," or the fifth column.

Respondents in the Peace lab report pointed out that it is "used by the government to refer to opposition groups" and that it is "used to refer to former Prime Minister Meles Zenawi and the Tigray people" or that "recently this word has been used by the government to indicate TPLF."

Why it is Offensive or Inflammatory: This word refers to political and ethnic allegiances in

Figure A.1: The Term **ባንዳ** (Banda) In Use

addition to power. Since the goal of calling someone a traitor is to isolate them from society and politics, it can be offensive when directed toward members of the same group. Because it intimidates people who want to collaborate with others by threatening them with humiliation and potential reprisal, its widespread use prevents people from working together across communities or political parties.

The phrase upholds the belief that other political parties are and always will be ”enemies,” which is widely taken to mean that there are incompatible objectives and that it is one’s responsibility to oppose or vanquish such adversaries. Related Hashtags: #traitor

Figure A.1 illustrates a sample post.

ወያኔ (weyane), woyane

Other Spellings and Associated Terms: Woyane, digital woyane

Definition of Term or Phrase: Woyane means “revolutionary” or “rebel” in Tigrinya, and this term normally has a positive connotation when used by Tigrayans. According to focus group participants, in that context it conveys “someone who fights for what he believes” and commits “heroic act(s) of the struggle for their belief.”

However, more recently the term has acquired a negative connotation, in particular when used in Amharic or by members of the Amhara and Oromo communities. For instance, survey respondents

Figure A.2: The Term ”Woyane” in Use

for the Peace Tech Lab indicated that “government officials use this term to refer to TPLF officials.” Used in this way, “the term comes closer to meaning ‘bandit,’” a participant pointed out, “like a robber and torturer,” because, according to an interviewee, it is “associated with [a] rebel group [from Tigray] that is doing a lot of atrocities.”

The term is also sometimes used more broadly by members of the other communities to denigrate “all Tigrayans and sometimes to mean TPLF or TPLF supporters” and is “loaded with the connotation of repressiveness.” By extension the term digital woyane is a reference to the promonitant role social media plays in Tigrayan organizing. It is used to both self-describe or designate nationalist Tigrayans active on social media.

Why it is Offensive or Inflammatory: The derogatory use of the term woyane targeting members of the TPLF is inflammatory when the intention is to portray a political group—which is currently in the opposition—as terrorists or a band of violent rebels. When applied generally to the Tigrayan community, this has a degrading effect because it indicates that they are ”without compassion or redeeming qualities as fellow members of Ethiopian society. Related Hashtags: #woyane , #digitalwoyane

Figure A.2 illustrates a sample post.

የቀን ጅብ (yeken jib), warrabessa guyyaa

Other Spellings and Associated Terms: yekeni jibi, yeken jeb, jeqen jib, jeken jiboch; **የቀን ጅብ ዘራፊዎች** (daytime robbers), yeqen jib zerafiwoch, yeken jib zerafiwoch

Definition of Term or Phrase: Yeken jib (**የቀን ጅብ**) and warrabessa guyyaa mean "daylight hyena" in Amhara and Afaan Oromo. As a workshop participant in the Peacetech Lab explained, "hyenas normally hunt at night, but yeken jib means a hyena that does hunting even on [sic] the day." According to survey respondents, this term is used to refer to "people who do the worst openly," like a "thief that has the audacity to steal in daylight" or a "daybreak robber."

While the term yeken jib was originally used to refer to "financially corrupt people" in general, survey data suggests that its use and meaning have evolved in recent years. Workshop participants pointed out that Prime Minister Abiy Ahmed mentioned it in a speech in which he used the term to reference members of the Tigrayan People's Liberation Front (TPLF), later repeating the term on national television. After this incident, participants explained that "it started to be used against previous regime officials and beneficiaries [...] that were assumed to [...] have amassed wealth under it." Many survey respondents pointed out that by extension, it began to be used as a derogatory term to refer to Tigrayans in general (PeaceTech Lab (2020)).

Why it is Offensive or Inflammatory: The term "daylight hyena" is insulting and demeaning, particularly when it is applied to the entire Tigrayan population. Participants in the program explained that hyenas prowl around at night and occasionally "takes a sheep or animal or human." Thus, at night, the hyenas steal or plunder. Most people assume that if a hyena is seen in the daylight, it must be a "starved hyena who will hurt humans." This gives it a more menacing undertone because it implies that one must defend oneself from hyena attacks.

According to those questioned, an individual who is accused of being a "daylight hyena" is someone who "has no shame, no conscience, no regard for morals, no regard for social values, he will go out and take [that] which isn't his." Even in its least provocative form, the term is problematic, according to survey respondents, especially when employed by certain party officials, who turn it into

Figure A.3: The Term "**የቀን ጅብ**" (Yeken Jib) in Use

a "politically charged" term that pits "the supporters of the new regime against the supporters of the old political regime." Related Hashtags: #yekenjib

Figure A.3 illustrates a sample post.

Translation: "Those yeken jib (daylight hyenas), we broke them, they don't belong to us, we prevented them entering to Addis Ababa without our permission, we blocked Tigrinya speakers (in all state media), the killers were speaking in Tigrinya, although there is no clash yet, the war has started already".

አጋሜ (agame)

Alternative Spelling and Associated Terms and Phrases: **አጋሜ**, anbetta belitta, anbeta belita

Definition and Context of Use: Agame (**አጋሜ**) is the name of a former province located in the regional state of Tigray, which was one of the country's poorest areas. According to focus group participants, people in the capital of Addis Ababa use the term to insult members of the Tigrayan community in general. It carries the negative connotation by referring to "illiterates, laborers, despised, low class." "It is used to humiliate the person by making them think they are low class, illiterate and greedy," survey respondents indicated.

አንበጣ በሊታ (anbeta girita) Alternative Spelling and Associated Terms and Phrases: anbetta belitta, anbeta belita

Definition and Context of Use: Anbeta girita means “locust-eater” in Amharic.

According to workshop participants “the word is used to degrade people from the Agame district in the Tigray region,” implying they are so poor that they have to eat grasshoppers. For some participants, it is a very degrading term, because it has “a connotation of undermining a person’s humanity, like they are sub-human, not human enough to make their own food.” They indicated that it is currently widely used to refer to Tigray. Historically this term was used to designate Tigray and Eritrea, and was used as a derogatory term particularly during the times of the student movement in the 1960s and 1970s.

ቆማል (qomale) Alternative Spelling and Associated Terms and Phrases: **ቅማላም** Definition and Context of Use: Qomale (**ቆማል**) means “body lice” in Tigrinya. It is inflammatory because it is dehumanizing and “it suggests the person is dirty and a parasite.”

ቁልቋል በሊታ (qulqual belita) Definition and Context of Use: Qulauql belita means “cactus eater” in Amharic. Some survey respondents explained that this term targets members of the Tigray community.

According to them, it is a derogatory reference to the climate in the northern part of the country where it is hotter and cactus fruit is consumed.

Validation workshop participants felt that it was mostly associated with “poverty and famine” and was more recently being used on social media in reference to the 1985- famine during which it is said those populations started consuming qulqual. “It is meant to belittle and to be a reference to social status.”

Chapter B: Theories Used and Their Similarities

Table B.1 maps the theories used and their similarities with each other. These theories were used for class/category generation.

Table B.1: Seed Keywords and Indicators from Stanton and Benesch Frameworks

Concept Name	**Description**	**Gregory Stanton's Incitement to Genocide**	**Benesch's Dangerous Speech Framework**	**Indicators**	**Example Words**
Distancing	Negation of Intimacy by the use of othering language	Target-Sympathizer conflation	Dehumanization	High pronoun usage in text, especially 3rd person plural nouns	They, them, their, she, he, us, we
Negative Passion	Use of negative sentiments and offensive language	Direct Advocacy	-	Emotions of anger, offensive, insulting, threatening, sexual, swear words	Damn, fuck, piss, kill, stop, hate, annoying, ugly, nasty, horny, uncircumcised
Devaluation Commitment	Commitment to hate the target by use of demeaning language	Dehumanization, Pathologization, Demonization, Verminization	Dehumanization	Use of subhuman, object, animal, or insect names to degrade person(s)	Cockroach, maggot, lice infested, locust eater
Subjectivity	Use of faulty arguments	Accusation in a Mirror, Justification	Threat Construction	Bias & propaganda using quantifiers and certainty	Always, never, all, many, much
Stereotyping	Hate directed on the target on the basis of a protected social group	Praising past violence, Disloyalty, Pathologization	Impurity	Presence of ethnic, racial, religious names	Agame, Komal, Kondaf

Chapter C: Seed Keywords Used for Tweet Extraction

Table C.1 shows the keywords used to collect Tweets, using Twitter/X's tweet streamer API.

Table C.1: Seed Keywords Organized into Six Columns

Col 1	Col 2	Col 3	Col 4	Col 5	Col 6
ነቀርሳ	አረም	እንቦጭ	የቀን ጅብ	ጡት ነካሽ	አረመኔ
መርዝ	ቆሻሻ	አምበጣ	ዐምበጣ	ኣምበጣ	አምበጣ በላ
ለማኝ	ባንዳ	ከሃዲ	ከሀዲ	ዓረም	አረም
ልክፍት	በሽታ	ዝንባም	ክፉ	ሴጣን	ደም ጠጪ
ሰው በላ	ዘይክንሹባት	ግብግብ	ወያኔ	ወያኔዎች	ትግሬ
ትግሬዎች	ተጋሩ	ጁንታ	ልቢ ትግራይ	ልቢ ትግራዋይ	ልቢ ትግራወይቲ
ኣጋመ	ዓጋመ	አጋመ	ኣጋሜ	ዓጋሜ	አጋሜ
ርዥራዥ	TPLF Orphan	agame	woyane	weyane	ማጥፋት
ማጽዳት	መንጥር	ቆንዳፍ	ቁልቋላም	ፀረለውጥ	

These taxonomy/categories were generated via open coding and through theories like incitement to genocide (Gordon (2017)), dangerous speech (Leader Maynard and Benesch (2016)) and dehumanization theories (Haslam (2006)).

Anti-Tigrayan Hate Speech Taxonomy

The following is a taxonomy of different categorize of hate speech.

- **Anti-Tigrayan Hate Speech**
  - **Incitement to Genocide**
    - discriminate
    - loot

- hot
- best
- evict/clear
- kill

- **Dehumanization**
- **Verminization**

  - agame
  - komal (lice)
  - kondaf
  - dirty/stinky
  - snakes

- **Demonization**

  - junta
  - evil
  - devil
  - daytime hyena

- **Pathologization**

  - blood sucker
  - cancer
  - weed
  - poison
  - dirt
  - parasite
  - disease
  - cursed womb

- **Economic Anti-Tigrayanism**

  - minorities have oppressed others
  - privileged
  - 6%
  - chronically poor
  - locust eater
  - beggar

- fly infected
- cactus eater
- greedy

- **Disloyal**
  - allegiance to a Tigrayan agenda
  - betrayer
  - lacking morals
  - chronic rebels
  - warmongers/intransigent
  - Tigrayan exceptionalism
  - foreign
  - backstabber
  - Banda
  - Heart of Tigrayans
  - deceptive
  - daytime hyena
  - liars
  - thieves

Chapter D: Annotation Guidelines

Hate Speech Categories and Examples

1. **Direct Advocacy** Direct calls for destruction are relatively rare and use words like *loot, discriminate, put in concentration camp, beat, evict, kill, destroy* ... There have been broadcasts to wipe out Tigrayas and destroy infrastructure, economy, morale, and so on.

   Examples:

   - Call for direct attack
   - "Press forward until Tigrayans are wiped out"
   - "Go House 2 House Kill All Agames"
   - Call to destroy Tigray infrastructures
   - "Cleanse your cities"
   - "After this, Jet will be sent [to Tigray] instead of budget"
   - "Eradicate TPLF from face of earth"

2. **Verminization** Classifies the target as something "whose extermination would be considered normal and desirable" (e.g., cockroaches to the Tutsis).

   Words to look for: *komal (lice), kondaf (barren), dirty, stinky, snakes, worms, agame, roach*

   Examples:

   - Comparing Tigrayan IDPs to snakes

- Lice
- Blen Mamo: “TPLF is shell. Inside the shell live worms”

3. **Demonization** Sinister figurative speech, often anthropocentric, centering on devils, malefactors, and other nefarious personages.

   Words to look for: *Daytime hyena, Junta, devil, satan, cursed*

   Examples:

   - Daniel Kibret used figurative speech such as “Satan” to refer to them
   - https://archive.ph/kCSzv
   - Snake

4. **Pathologization** Designates the target as a disease. Expropriates pseudo-medical terminology to justify massacre, dehumanizing victims as sources of filth and disease.

   Words to look for: *Cancer, disease, weed, straving (rehabtegna)*

   Examples:

   - Rehabtegna
   - Tweet example

5. **Disloyalty** Expression of mistrust, treachery, or disloyalty of the group to the state, nation, or continent. Includes remarks about Tigrayans working for the West and aid workers supplying arms.

   Words to look for: *backstab, foreign, banda, traitor, heart of Tigray (Libi Tigray), liars, terrorists*

Examples:

- "All of Woyane leaders' defeated minds and consciences can never think of anything better than evil and deceit. All the young people are being killed in a worthless war and they are still fighting."
- "You are one of the most criminal who always cry for only one ethnic group like Woyane and divide the nation."

6. **Economic Anti-Tigrayanism** Claims that Tigrayans are poor and useless or greedy and have stolen money/resources from the state.

   Words to look for: *Locust eater, beggar, fly infested, cactus eater, greedy, stole money, billions*

7. **Accusation in a Mirror** False claims that accuse the target of something the perpetrator is doing or intends to do. Justifies genocide by invoking collective self-defense.

   Example: Adebabay Media: "The traitors are one million people"

8. **Justification** Justifying ongoing atrocities; blaming a group for their lot; claiming the victims deserve what happens; scapegoating a group.

   Examples:

   - Amhara Media Corporation: "The people here [the Amharas] are special. There [in Tigray], even when you ask for water, they give you poison. If there were such people [as the Amharas] in Tigray, our country would be exceptional."
   - Menelik Television: "We are fighting to make the Tigray people Ethiopians. In the last 27 years, TPLF has instilled racism. To be honest, the Tigrayan support for TPLF is correct because they have been unfairly benefiting."

- "Fano, do not doubt for a moment that this aid is directly or indirectly going into the hands of the enemy who is trying to destroy you. So take action without hesitation and without being influenced by anyone."

9. **Praising Past Violence** Glorification of past atrocities, often referencing historical violence or comparing Tigray to Biafra, Tasmanians, etc.

   Example: https://archive.ph/uG3S1

10. **Target-Sympathizer Conflation** Those who sympathize with target groups are also persecuted. Non-Tigrayans who have supported them have been treated as junta or Woyane cadre.

    Example: Yonatan Tesfaye, Deputy Director of Ethiopian Media Corporation:

    > "A country is not harmed by enemies or bandits — we are already facing those! At such a time, the country is hurt by "middle settlers" ("fence sitters"), those that claim they are balanced, the thorns within the mud."

Key for Highlighting

- Red: Not English, completely irrelevant
- Yellow: Confusing because mixed English with another language (still categorize if possible)
- Green: Unsure which category tweet falls under (list 2 categories if possible)
- Blue: Quotes/explanation, no hate speech
- Orange: Unclear if hate speech due to historical/lack of context

Key for Dehumanization Categories

- DA = Direct Advocacy (calls to violence: kill, loot, destroy etc)
- V = Verminization (subhuman: worms, snakes, filth etc)
- DM = Demonization (threat/fear inducing: demons, devils, satan etc)
- P = Pathologization (cancer, malignant, infect us etc)
- DS = Disloyalty (lies, terrorists, traitors)
- EAT = Economic Anti-Tigrayanism (slaves, beggars, greedy, stole from us)
- AM = Accusation in a Mirror (they will/have done it to us, so we will do it to them, "fake genocide")
- J = Justification (they deserve the suffering/events)
- PPV = Praising Past Violence
- TSC = Target-Sympathizer Conflation
- Q/E = Quotation/Explanation

Note: If a slur is present alongside context, always take the context rather than the slur.

Chapter E: Genocidal Speech Survey

Survey: Perceived Threat, Violence, and Emotional Reactions

1. Perceived Threat

Question: How threatening do you find this tweet?

Response scale:

1. Not threatening at all
2. Slightly threatening
3. Moderately threatening
4. Very threatening
5. Extremely threatening

2. Perceived Violence

Question: How violent do you find this tweet?

Response scale:

1. Not violent at all
2. Slightly violent
3. Moderately violent

4. Very violent

5. Extremely violent

3. Emotional Reactions

Anger:

- I felt angry after reading the speech.
- The speech made me feel outraged.
- I felt a strong dislike or resentment as a result of the speech.

Fear:

- I felt scared after reading the speech.
- The speech made me feel anxious about the safety of myself or others.
- I felt a sense of dread due to the content of the speech.

Disgust:

- I felt disgusted by the speech.
- The speech content was repulsive to me.
- I felt a strong aversion to the ideas expressed in the speech.

4. Urgency to Act

Would you report a tweet, respond, etc.?

5. Behavioral Intentions

Engagement Behavior:

- I would like, share, or comment on this tweet.

Protective Behavior:

- I would report this tweet or block the user.

Withdrawal Behavior:

- I would ignore or avoid further tweets of similar nature.

Example of the Scale:

Pathologization Tweet: ”Afars are in Tigray as we speak and Amhara militias are cleaning their region from agame roaches. Soon we will be coming back. And 3 forces will be looting Mekelle to the ground.”

Response Scale: 1 = Not at all 2 = Slightly 3 = Moderately 4 = Very 5 = Extremely

Direct Advocacy Tweet: ”agame roaches, Internment Camps Are Waste of Resources: Just Shoot Them.”

Question	1	2	3	4	5
How threatening do you find this tweet?					
How violent do you find this tweet?					
I felt angry after reading the speech.					
The speech made me feel outraged.					
I felt a strong dislike or resentment as a result of the speech.					
I felt scared after reading the speech.					
The speech made me feel anxious about the safety of myself or others.					
I felt a sense of dread due to the content of the speech.					
I felt sad after reading the speech.					
The speech made me feel a sense of loss or sorrow.					
I felt hopeless after reading the speech.					
I felt disgusted by the speech.					
The speech content was repulsive to me.					
I felt a strong aversion to the ideas expressed in the speech.					
I would like, share, or comment on this tweet.					
I would report this tweet or block the user.					
I would ignore or avoid further tweets of similar nature.					

Table E.1: Survey Questions I

Question	1	2	3	4	5
Perceived Threat: How threatening do you find this tweet?					
Perceived Violence: How violent do you find this tweet?					
Anger: I felt angry after reading the speech.					
The speech made me feel outraged.					
I felt a strong dislike or resentment as a result of the speech.					
Fear: I felt scared after reading the speech.					
The speech made me feel anxious about the safety of myself or others.					
I felt a sense of dread due to the content of the speech.					
Sadness: I felt sad after reading the speech.					
The speech made me feel a sense of loss or sorrow.					
I felt hopeless after reading the speech.					
Disgust: I felt disgusted by the speech.					
The speech content was repulsive to me.					
I felt a strong aversion to the ideas expressed in the speech.					
Engagement Behavior: I would like, share, or comment on this tweet.					
Protective Behavior: I would report this tweet or block the user.					

Table E.2: Survey Questions II

Appendix F: IRB Exemption

The page following this one contains the letter confirming that the Institutional Review Board (IRB) of the University of Maryland (College Park) determined that this project was exempt from IRB review, in accordance with US federal regulations.

UNIVERSITY OF MARYLAND
INSTITUTIONAL REVIEW BOARD

1204 Marie Mount Hall
College Park, MD 20742-5125
TEL 301.405.4212
FAX 301.314.1475
irb@umd.edu
www.umresearch.umd.edu/IRB

DATE:	March 5, 2025
TO:	Jennifer Wessel , PhD
FROM:	University of Maryland College Park (UMCP) IRB
PROJECT TITLE:	[2194343-1] Genocidal speech detection using NLP
SUBMISSION TYPE:	New Project
ACTION:	DETERMINATION OF EXEMPT STATUS
DECISION DATE:	March 5, 2025
REVIEW CATEGORY:	Exemption category # 45CFR46.104(d)(2)(i)

Thank you for your submission of New Project materials for this project. The University of Maryland College Park (UMCP) IRB has determined this project is EXEMPT FROM IRB REVIEW according to federal regulations.

We will retain a copy of this correspondence within our records.

If you have any questions, please contact the IRB Office at 301-405-4212 or irb@umd.edu. Please include your project title and reference number in all correspondence with this committee.

This letter has been electronically signed in accordance with all applicable regulations, and a copy is retained within University of Maryland College Park (UMCP) IRB's records.

Bibliography

Abdelkadir, N. A., Yang, T., Kapania, S., Estefanos, M., Gebrekidan, F. C., Zelalem, Z., Ali, M., Berhe, R., Baker, D. K., Talat, Z., Miceli, M., Hanna, A., & Gebru, T. (2025). The role of expertise in effectively moderating harmful social media content. *International Conference on Human Factors in Computing Systems*.

African Union. (2022). Cessation of hostilities agreement between the government of the federal democratic republic of ethiopia and the tigray peoples' liberation front (tplf) [Press Release.]. https://www.peaceau.org/uploads/press-release-cessation-of-hostilities-pretoria-2-11-2022.pdf

Akinwotu, E. (2021). Facebook's role in myanmar and ethiopia under new scrutiny. https://www.theguardian.com/technology/2021/oct/07/facebooks-role-in-myanmar-and-ethiopia-under-new-scrutiny

Aldreabi, E., Harahsheh, K., Chhangani, M. D., Chen, C. H., & Blackburn, J. (2024). Beyond binary: Revealing variations in islamophobic content with hierarchical multi-class classification. *The International FLAIRS Conference Proceedings, 37*.

Allport, G. W. (1954). *The nature of prejudice*.

Alorainy, W., Burnap, P., Liu, H., & Williams, M. L. (2019). "the enemy among us": Detecting cyber hate speech with threats-based othering language embeddings. *ACM Transactions on the Web (TWEB)*, *13*(3), 1–26.

Alpaydin, E. (2010). *Introduction to machine learning* (2nd). MIT Press.

Aluru, S. S., Mathew, B., Saha, P., & Mukherjee, A. (2021). Deep dive into multilingual hate speech classification. *Dong, Y., Ifrim, G., Mladenić, D., Saunders, C., Van Hoecke, S. (eds) Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track. ECML PKDD 2020. Lecture Notes in Computer Science(), vol 12461. Springer, Cham.* https://doi.org/10.1007/978-3-030-67670-4_26

Alvarez, A. (2001). *Governments, citizens, and genocide: A comparative and interdisciplinary approach*. Indiana University Press.

Amnesty International. (2023). Ethiopia: Meta's failures contributed to abuses against tigrayan community during conflict in northern ethiopia. https://www.amnesty.org/en/latest/news/2023/10/meta-failure-contributed-to-abuses-against-tigray-ethiopia/

Anna, C. (2021). *Eu envoy: Ethiopian leadership vowed to 'wipe out' tigrayans*. https://apnews.com/article/europe-ethiopia-africa-ffd3dc3faf15d0501fd87cafe274e65a

Anna, C. (2022). *Anna, c (2022) ethiopians file lawsuit against meta over hate speech in tigray war*. https://www.pbs.org/newshour/world/ethiopians-file-lawsuit-against-meta-over-hate-speech-in-tigray-war

Associated Press. (2021). *Associated press (2021) us warns of 'dehumanizing rhetoric' on tigray (genocidewatch)*. https://www.genocidewatch.com/single-post/us-warns-of-dehumanizing-rhetoric-on-tigray

Atkinson, R., & Flint, J. (2001). Accessing hidden and hard-to-reach populations: Snowball research strategies. *The A-Z of Social Research*. https://doi.org/10.4135/9780857020024.n94

Bandura, A. (1999). Moral disengagement in the perpetration of inhumanities. *Personality and Social Psychology Review, 3(3), 193–209*. https://doi.org/10.1207/s15327957pspr0303_3

Bandura, A. (2016). *Moral disengagement: How people do harm and live with themselves*. Worth Publishers.

Bar-Tal, D. (1990). Causes and consequences of delegitimization: Models of conflict and ethnocentrism. *Journal of Social issues, 46(1), 65-81*.

BBC News. (2021). *Bbc news (2021) facebook deletes ethiopia pm's post that urged citizens to "bury" rebels*. https://www.bbc.co.uk/news/world-africa-59154984.

Benesch, S. (2007). Vile crime or inalienable right: Defining incitement to genocide. *Va. J. Int'l L, Vol 48, Pages 485*.

Benesch, S. (2012). *Dangerous speech: A proposal to prevent group violence*.

Benesch, S. (2021). Dangerous speech: A proposal to prevent group violence. *Noema Magazine, 2021*.

Beukeboom, C. J., & Burgers, C. (2019). How stereotypes are shared through language: A review and introduction of the social categories and stereotypes communication (scsc) framework. *Review of Communication Research*, *7*, 1–37. https://doi.org/10.12840/issn.2255-4165.017

Bicheno, H. (2001). *Oxford companion to military history*. Oxford University Press. http://www.cc.gatech.edu/~tpilsch/INTA4803TP/Articles/Total%20War-definition&discussion.pdf

Biernacki, P., & Waldorf, D. (1981). Snowball sampling: Problems and techniques of chain referral sampling. *Sociological Methods & Research, 10(2), 141-163*. https://doi.org/10.1177/004912418101000205

Bisong, E. (2019). *Building machine learning and deep learning models on google cloud platform (pp. 59-64)*.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3(Jan), 993–1022*. https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

Braun, V., & Clarke, V. (2012). *Thematic analysis*.

Brown, A. (2017). What is hate speech? part 1: The myth of hate. *Law and Philos 36, 419–468 (2017)*. https://doi.org/10.1007/s10982-017-9297-1

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ..., & Amodei, D. (2020). *Language models are few-shot learners*.

Buckels, E. E., & Trapnell, P. D. (2013). *Disgust facilitates outgroup dehumanization*.

Burgers, C., & Beukeboom, C. J. (2020). How language contributes to stereotype formation: Combined effects of label types and negation use in behavior descriptions. *Journal of Language and Social Psychology, 39(4), 438–456*.

Byaruhanga, C. (2022). *Ethiopia's tigray conflict: Nasa shows how a war zone faded from space*. https://www.bbc.com/news/world-africa-63315388

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). *Semantics derived automatically from language corpora contain human-like biases*.

Caselli, T., Basile, V., Mitrović, J., & Granitzer, M. (2021). Hatebert: Retraining bert for abusive language detection in english [arXiv preprint arXiv:2010.12472]. https://arxiv.org/abs/2010.12472

Center, W. (2023). *Ethiopia's tigray war and its devastating impact on tigrayan children's education*. https://www.wilsoncenter.org/blog-post/tigray-war-and-education

CFR Editors. (2022). *Facebook's content moderation failures in ethiopia*. https://www.cfr.org/blog/facebooks-content-moderation-failures-ethiopia

Chander, S., & Vijaya, P. (2021). Unsupervised learning methods for data clustering. In *Artificial intelligence in data mining* (pp. 41–64). Academic Press.

Chirot, D., & McCauley, C. (2006). *Why not kill them all?: The logic and prevention of mass political murder*. Princeton University Press.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. *Proceedings of ACL 2020, 5-10 July 2020, 8440-8451*. https://aclanthology.org/2020.acl-main.747

Davani, A. M., Atari, M., Kennedy, B., & Dehghani, M. (2023). Hate speech classifiers learn normative social stereotypes. *Transactions of the Association for Computational Linguistics*, *11*, 300–319. https://doi.org/10.1162/tacl_a_00550

Davani, A. M., Omrani, A., Kennedy, B., Atari, M., Ren, X., & Dehghani, M. (2021, August). Improving counterfactual generation for fair hate speech detection. In A. Mostafazadeh Davani, D. Kiela, M. Lambert, B. Vidgen, V. Prabhakaran, & Z. Waseem (Eds.), *Proceedings of the 5th workshop on online abuse and harms (woah 2021)* (pp. 92–101). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.woah-1.10

Des Forges, A. (1999). *Leave none to tell the story*. Human Rights Watch.

Devlin, J., M.-W., C., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding [arXiv preprint arXiv:1810.04805v2]. https://arxiv.org/abs/1810.04805

DFRLab. (2021). Eritrean officials promoted a report undermining evidence of violence in tigray, claiming an international disinfo campaign. https://dfrlab.org/2021/06/23/eritrean-report-uses-fact-checking-tropes-to-dismiss-evidence-as-disinformation/

El Naqa, I., & Murphy, M. J. (2015). What is machine learning? Springer International Publishing.

Fein, S. (1993). *Genocide: A sociological perspective*. Sage Publications.

Fine, R. D. (2009). Fighting with phantoms: A contribution to the debate on anti-semitism in europe. *Patterns Of Prejudice 43 (5), 459 - 479 (0031-322X)*.

Fink, C. (2018). Dangerous speech, anti-muslim violence, and facebook in myanmar. *Journal of International Affairs, Vol. 71, No. 1.5, pp. 43-52*.

Fisseha, A. (2023). Tigray: A nation in search of statehood? *International Journal on Minority and Group Rights (2023) 1–47*.

Förster, S. (2007). Total war and genocide: Reflections on the second world war. *Australian Journal of Politics & History, 53(1), 68–83*.

Fortuna, P., Domínguez, M., Wanner, L., & Talat, Z. (2022). Directions for nlp practices applied to online hate speech detection. *In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (pp 11794–11805)*.

Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, *51*(4), 1–30.

Friedman, S. E., Magnusson, I., Schmer-Galunder, S., Wheelock, R., Gottlieb, J., Patel, P., & Miller, C. (2021). Toward transformer-based nlp for extracting psychosocial indicators of moral disengagement. *Proceedings of the Annual Meeting of the Cognitive Science Society, 43*. https://escholarship.org/uc/item/9n71j1zh

Fyfe, S. (2017). Tracking hate speech acts as incitement to genocide in tracking hate speech acts as incitement to genocide in international criminal law. *Leiden Journal of International Law (2017), 30, pp. 523-548*.

Gagliardone, I., Gal, D., Alves, T., & Martinez, G. (2015). Countering online hate speech. *UNESCO Series on Internet Freedom*.

Giner-Sorolla, R., & Russell, P. (2019). Not just disgust: Fear and anger also relate to intergroup dehumanization. *Collabra: Psychology (2019) 5(1)*. https://doi.org/10.1525/collabra.211

Glaser, B., & Strauss, A. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Mill Valley, CA: Sociology Press.

Goff, P. A., Eberhardt, J. L., Williams, M. J., & Jackson, M. C. (2008). Not yet human: Implicit knowledge, historical dehumanization, and contemporary consequences. *Journal of Personality and Social Psychology, 94(2), 292*.

Goodman, L. A. (1961). Snowball sampling. *The Annals of Mathematical Statistics 32(1)*. https://doi.org/10.1214/aoms/1177705148

Gordon, G. S. (2017). *Atrocity speech law: Foundation, fragmentation, fruition*. Oxford University Press.

Government of FDRE & TPLF. (2022). Agreement for lasting peace through a permanent cessation of hostilities between the government of the federal democratic republic of ethiopia and the tigrayan people's liberation front (tplf) [Archived on the Wayback Machine.]. https://web.archive.org/web/20221111111115/https://addisstandard.com/wp-content/uploads/2022/11/AU-led-Ethiopia-Peace-Agreement.pdf

Haidt, J., & Graham, J. (2007). When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research*, *20*(1), 98–116.

Hale, L. (2021). *Social media misinformation stokes a worsening civil war in ethiopia*. https://www.npr.org/2021/10/15/1046106922/social-media-misinformation-stokes-a-worsening-civil-war-in-ethiopia

Harris, B. (2021). Biden administration considering tigray genocide determination (the national). *https://www.thenati news/2021/09/24/biden-administration-considering-tigray-genocide-determination/. Accessed Jan 21 2026.*

Harris, L. T., & Fiske, S. T. (2015). Dehumanized perception. *Zeitschrift für Psychologie*, *219*, 175–181. https://doi.org/10.1027/2151-2604/a000065

Haslam, N. (2006). Dehumanization: An integrative review. *Personality and Social Psychology Review, 10(3), 252–264*.

Herf, J. (2006). *The jewish enemy: Nazi propaganda during world war ii and the holocaust*. https://doi.org/10.1093/hgs/dcm024

Hinds, J. (1977). Paragraph structure and pronominalization. *Paper in Linguistics*, *10*(1-2), 77–99.

Hodson, G., & Costello, K. (2007). Interpersonal disgust, ideological orientations, and dehumanization as predictors of intergroup attitudes. *Psychological Science*, *18*(8), 691–698.

Human Rights Watch. (2022). *"we will erase you from this land": Crimes against humanity and ethnic cleansing in ethiopia's western tigray zone*. https://www.hrw.org/report/2022/04/06/we-will-erase-you-land/crimes-against-humanity-and-ethnic-cleansing-ethiopias

Human Rights Watch. (2024). *World report 2024 – ethiopia (events of 2023)*. https://www.hrw.org/world-report/2024/country-chapters/ethiopia

Hutto, C., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media, 8(1), 216-225*. https://doi.org/10.1609/icwsm.v8i1.14550

Ibrahim, A., & Weis, A. (2025). Tigray needs justice for peace to hold. https://foreignpolicy.com/2025/03/27/tigray-tribunals-war-crimes-rape-mass-murder-justice/

Igoe, M. (2021). 'new — and terrifying' findings from tigray. https://www.devex.com/news/devex-newswire-new-and-terrifying-findings-from-tigray-100243

Jahan, M. S., & Oussalah, M. (2023). A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing, 126232*.

Kant, S. (2010). Machine learning and pattern recognition. *Defence Science Journal, 60(4), 345–347*. https://doi.org/10.14429/dsj.60.502

Károly, A. I., Fullér, R., & Galambos, P. (2018). *Unsupervised clustering for deep learning: A tutorial survey*.

Keaten, J. (2023). *Un experts say ethiopia's conflict and tigray fighting left over 10,000 survivors of sexual violence*. https://apnews.com/article/ethiopia-tigray-war-crimes-sexual-violence-c2f47284aa80f164cb822f716cc2a0f6

Kelman, H. (1973). Violence without moral restraint: Reflections on the dehumanization of victims and victimizers. *Journal of Social Issues, 29, 25-61*. https://doi.org/10.1111/j.1540-4560.1973.tb00102.x

Kenyon-Dean, K., Newell, E., & Cheung, J. C. K. (2020, November). Deconstructing word embedding algorithms. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)* (pp. 8479–8484). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.681

Kharde, V., & Sonawane, S. (2016). Sentiment analysis of twitter data: A survey of techniques. *International Journal of Computer Applications 139(11):5-15*. https://doi.org/10.5120/ijca2016908625

Khazaie, A., Seghouani, N. B., & Bugiotti, F. (2021). Smart crawling: A new approach toward focus crawling from twitter. https://doi.org/10.48550/arXiv.2110.06022

Kteily, N., Bruneau, E., Waytz, A., & Cotterill, S. (2015). The ascent of man: Theoretical and empirical evidence for blatant dehumanization. *Journal of Personality and Social Psychology, 109, 901–931*. https://doi.org/10.1037/pspp0000048

Leader Maynard, J. (2014). Rethinking the role of ideology in mass atrocities. *Terrorism and Political Violence, 26(5), 821–841*. https://doi.org/10.1080/09546553.2013.796934

Leader Maynard, J., & Benesch, S. (2016). *Dangerous speech and dangerous ideology: An integrated model for monitoring and prevention*.

Lee, N., Jung, C., & Oh, A. (2023). Hate speech classifiers are culturally insensitive. *In Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP) (pp. 35–46)*. https://doi.org/10.18653/v1/2023.c3nlp-1.5

Li, Z., Zhu, H., Lu, Z., & Yin, M. (2023). *Synthetic data generation with large language models for text classification: Potential and limitations*.

Llama Team, A. (2024). The llama 3 herd of models [arXiv preprint arXiv:2407.21783]. https://arxiv.org/pdf/2407.21783

MacAvaney, S., Yao, H. R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). *Hate speech detection: Challenges and solutions*.

Mackintosh, E. (2021). *Facebook knew it was being used to incite violence in ethiopia. it did little to stop the spread, documents show (cnn)*. https://edition.cnn.com/2021/10/25/business/ethiopia-violence-facebook-papers-cmd-intl/index.html

Madriaza, P., Hassan, G., Brouillette-Alarie, S., Mounchingam, A. N., Durocher-Corfa, L., Borokhovski, E., Pickup, D., & Paillé, S. (2025). Exposure to hate in online and traditional media: A systematic review and meta-analysis of the impact of this exposure on individuals and communities. *Campbell systematic reviews, 21(1), e70018*. https://doi.org/10.1002/cl2.70018

May, J., Shutova, E., Herbelot, A., Zhu, X., Apidianaki, M., & Mohammad, S. M. (2019). Proceedings of the 13th international workshop on semantic evaluation. https://aclanthology.org/S19-2000/

Mendelsohn, J., Tsvetkov, Y., & Jurafsky, D. (2020). *A framework for the computational linguistic analysis of dehumanization*.

Mitra, M. (2019). Algorithms and machine learning. *American Research Journal of Electronics and Communication Engineering*.

Mohammad, S. (2018, July). Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In I. Gurevych & Y. Miyao (Eds.), *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 174–184). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1017

Mossie, Z., & Wang, J. H. (2020). *Vulnerable community identification using hate speech detection on social media*.

Muhammad, I., & Yan, Z. (2015). Supervised machine learning approaches: A survey. *ICTACT Journal on Soft Computing, 5(3)*.

Muhumuza, R. (2022). *Un genocide official: Hate speech is fueling ethiopia's war (ap news)*. https://apnews.com/article/technology-africa-uganda-ethiopia-government-and-politics-e7c4173e458dda8056c47

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979). The belmont report: Ethical principles and guidelines for the protection of human subjects of research. *U.S. Department of Health and Human Services*. https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/read-the-belmont-report/index.html

Oberschall, A. (2000). The manipulation of ethnicity: From ethnic cooperation to violence and war in yugoslavia. *Ethnic and Racial Studies*. https://doi.org/DOI:10.1080/014198700750018388

OHCHR & EHRC. (2021). Report of the ehrc/ohchr joint investigation into alleged violations of international human rights, humanitarian and refugee law committed by all parties to the conflict in the tigray region of the federal democratic republic of ethiopia. https://digitallibrary.un.org/record/3947207?v=pdf

OpenAI. (2023). Gpt-4 technical report [arXiv preprint arXiv:2303.08774]. https://arxiv.org/abs/2303.08774

Paravicini, G., Endeshaw, D., & Houreld, K. (2021). Ethiopia's crackdown on ethnic tigrayans snares thousands [Last accessed Jan 23, 2026.]. https://www.reuters.com/investigates/special-report/ethiopia-conflict-tigrayans/

Pascoe, E. A., & Smart Richman, L. (2000). Erceived discrimination and health: A meta-analytic review. *Psychological Bulletin, 135(4), 531–554*. https://doi.org/doi:10.1037/a0016059

Pavlopoulos, J., Laugier, L., Xenos, A., Sorensen, J., & Androutsopoulos, I. (2022, May). From the detection of toxic spans in online discussions to the analysis of toxic-to-civil transfer. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 3721–3734). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.259

PeaceTech Lab. (2020). Ethiopia hate speech lexicon [Retrieved November 19, 2022]. https://www.peacetechlab.org/ethiopia-lexicon

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of liwc2015*. https://repositories.lib.utexas.edu/handle/2152/31333

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology, 54(1), 547–577*.

Perry, B. (2001). *The name of hate: Understanding hate crimes*. Psychology Press.

Qiao, Y., Xiong, C., Liu, Z., & Liu, Z. (2019). *Understanding the behaviors of bert in ranking*.

Refugees International. (2024). *Scars of war and deprivation: An urgent call to reverse tigray's humanitarian crisis*.

Ruppenhofer, J. (2018). Overview of the germeval 2018 shared task on the identification of offensive language [Online available: https://epub.oeaw.ac.at/?arp=0x003a10d2 - Last access:21.1.2026]. In J. R. .-. M. S. .-. M. W. (Hrsg.) (Ed.). Verlag der Österreichischen Akademie der Wissenschaften. https://epub.oeaw.ac.at/?arp=0x003a10d2

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39(6), 1161*.

Saffari, H., Shafiei, M., Zhang, H., Harris, L. T., & Moosavi, N. S. (2025). Beyond hate speech: Nlp's challenges and opportunities in uncovering dehumanizing language. *In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (pp. 26953–26968). Association for Computational Linguistics*. https://doi.org/10.18653/v1/2025.emnlp-main.1370

Saha, K., Chandrasekharan, E., & De Choudhury, M. (2019). Prevalence and psychological effects of hateful speech in online communities. *Proceedings of the International AAAI Conference on Web and Social Media, 13(01), 1–12*.

Sap, M., Prasettio, M. C., Holtzman, A., Rashkin, H., & Choi, Y. (2017). Connotation frames of power and agency in modern films. *In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (pp. 2329–2334)*.

Saraswat, P. K., & Raj, S. (2021). A brief review on machine learning and its various techniques. *International Journal of Innovative Research in Computer Science & Technology*.

Saravanan, R., & Sujatha, P. (2018). A state of art techniques on machine learning algorithms: A perspective of supervised learning approaches in data classification. *In 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS) (pp. 945–949). IEEE*.

Schabas, W. A. (2000). *Genocide in international law*. Cambridge University Press.

Schlechtweg, D., McGillivray, B., Hengchen, S., Dubossarsky, H., & Tahmasebi, N. (2020, December). SemEval-2020 task 1: Unsupervised lexical semantic change detection. In A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May, & E. Shutova (Eds.), *Proceedings of the fourteenth workshop on semantic evaluation* (pp. 1–23). International Committee for Computational Linguistics. https://doi.org/10.18653/v1/2020.semeval-1.1

Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. *In Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media (pp. 1–10)*.

Sherman, G. D., & Haidt, J. (2011). Cuteness and disgust: The humanizing and dehumanizing effects of emotion. *Emotion Review, 3(3), 245–251*.

Silva, L., Mondal, M., Correa, D., Benevenuto, F., & Weber, I. (2016). Analyzing the targets of hate in online social media. *In Tenth International AAAI Conference on Web and Social Media*.

Stanton, G. H. (1996). The 8 stages of genocide. *Yale Program in Genocide Studies in 1998*. http://www.genocide-watch.com/images/8StagesBriefingpaper.pdf

Staub, E. (1989). *The roots of evil: The origins of genocide and other group violence*. https://doi.org/10.1207/s15327957pspr0303_2

Staub, E. (2011). *Overcoming evil: Genocide, violent conflict, and terrorism*. https://doi.org/10.1093/acprof:oso/9780195382044.001.0001

Straus, S. (2001). Contested meanings and conflicted imperatives: A conceptual analysis of genocide. *Journal of Genocide Research*. DOI:10.1080/14623520120097189

Straus, S. (2003). Darfur and the genocide debate. *Foreign Affairs, Vol. 84, No. 1 (Jan. - Feb., 2005), pp. 123-133 (11 pages)*. https://doi.org/10.2307/20034212

Straus, S. (2012). *"destroy them to save us": Theories of genocide and the logics of political violence*.

Sutton, R. M., & Douglas, K. M. (2008). Celebrating two decades of linguistic bias research: An introduction. *Journal of Language and Social Psychology, 27(2), 105–109*. https://doi.org/10.1177/0261927X07313642

Tackett, C. F. A. (2023). *After years under a shutdown, tigray is slowly coming back online*. https://www.accessnow.org/tigray-shutdown-slowly-coming-back-online/

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology, 29(1), 24–54*. https://doi.org/10.1177/0261927X09351676

Taylor, J. E. T., & Taylor, G. W. (2021). Artificial cognition: How experimental psychology can help generate explainable artificial intelligence. *Psychonomic Bulletin & Review, 28(2), 454–475*.

Teklu, T. A. (2022). Religious legitimization of retributive genocide: The case of tigray. https://www.tghat.com/2022/04/04/religious-legitimization-of-retributive-genocide-the-case-of-tigray/

The Distributed AI Research Institue. (2025). Ethiopia: Social media platforms "failed to adequately moderate genocidal content" during tigray war, study finds. https://www.business-humanrights.org/en/latest-news/ethiopia-social-media-platforms-failed-to-adequately-moderate-genocidal-content-during-tigray-war-study-finds-x-formerly-twitter-did-not-respond/

Timmermann, W. K. (2005). The relationship between hate propaganda and incitement to genocide: A new trend in international law towards criminalization of hate propaganda? *Leiden Journal of International Law , Volume 18 , Issue 2 , June 2005 , pp. 257 - 282*. https://doi.org/10.1017/S0922156505002633

Tipler, C., & Ruscher, J. B. (2014). *Agency's role in dehumanization: Non-human metaphors of outgroups*.

Treccani. (n.d.). Banda. in vocabolario treccani online [Retrieved Jan 22nd, 2026.]. https://www.treccani.it/vocabolario/banda2_%28Sinonimi-e-Contrari%29

Tsesis, A. (2002). *Destructive messages: How hate speech paves the way for harmful social movements*. NYU Press.

Tsoumakas, G., & Katakis, I. (2007). Multi-label classification: An overview. *International Journal of Data Warehousing and Mining, 3(3), 1–13*.

UN News. (2021). *Ethiopia: Mass arbitrary arrests target tigrayans, says un rights office*. https://news.un.org/en/story/2021/11/1105892

UN World Food Programme. (2022). Ethiopia - tigray emergency food security assessment: Tigray crisis response, august 2022 [Last accessed Jan 21st 2026.]. https://reliefweb.int/report/ethiopia/ethiopia-tigray-emergency-food-security-assessment-tigray-crisis-response-august-2022

United Nations. (1948). Convention on the prevention and punishment of the crime of genocide. https://www.ohchr.org/en/instruments-mechanisms/instruments/convention-prevention-and-punishment-crime-genocide

United Nations. (2022). *Ethiopia still in grip of spreading violence, hate speech and aid crisis (un news)*. https://news.un.org/en/story/2022/06/1121772

United States Holocaust Memorial Museum. (2019). Incitement to genocide in international law [Last accessed Jan 21st 2026.]. https://encyclopedia.ushmm.org/content/en/article/incitement-to-genocide-in-international-law

United States Holocaust Memorial Museum. (2023). A strategic framework for helping prevent mass atrocities [Last accessed Jan 21st 2026.]. https://www.ushmm.org/m/pdfs/A_Strategic_Framework_for_Helping_Prevent_Mass_Atrocities_.pdf

Unknown. (n.d.). *Ethiopia – the world factbook* [Retrieved Jan 23, 2026.]. https://www.cia.gov/the-world-factbook/countries/ethiopia/

Valverde-Albacete, F. J., & Peláez-Moreno, C. (2014). 100% classification accuracy considered harmful: The normalized information transfer factor explains the accuracy paradox. *PLoS ONE 9(1): e84217*. https://doi.org/10.1371/journal.pone.0084217

van Reisen, M., & Mawere, M. (2024). *Tigray. the hysteresis of war*. Langaa, Bamenda.

Vermeulen, A. F. (2019). *Unsupervised learning: Using unlabeled data*.

Vermeulen, A. F. (2020). *Supervised learning: Using labeled data for insights*.

Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. *In Proceedings of the NAACL Student Research Workshop (pp. 88–93)*.

Wiegand, M., Ruppenhofer, J., Schmidt, A., & Greenberg, C. (2018). Inducing a lexicon of abusive words–a feature-based approach. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers) (pp. 1046–1056)*.

Wilmot, C., Tveteraas, E., & Drew, A. (2021). *Dueling information campaigns: The war over the narrative in tigray*. https://mediamanipulation.org/case-studies/dueling-information-campaigns-war-over-narrative-tigray/

World Health Organization. (2023). Herams tigray baseline report 2023: Operational status of the health system [Last accessed Jan 21, 2026.]. https://www.who.int/publications/m/item/herams-tigray-baseline-report-2023-operational-status-of-the-health-system

Wyschogrod, E. (2005). *The warring logics of genocide*.

Xiang, G., Fan, B., Wang, L., Hong, J., & Rose, C. (2012). Detecting offensive tweets via topical feature discovery over a large scale twitter corpusg. *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (pp. 1980–1984)*.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). *Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval)*.

Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., ..., & Çöltekin, Ç. (2020). *Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020)*.

Zelalem, Z. (2022). *Feature-six million silenced: A two-year internet outage in ethiopia*. https://www.reuters.com/article/ethiopia-internet-shutdown-idAFL8N2ZM09X

Zhang, M.-L., & Zhou, Z.-H. (2013). A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering, 26(8), 1819–1837*. https://doi.org/10.1109/TKDE.2013.39

Zhang, Y., Orgun, M. A., & Lin, W. (2006). Unsupervised learning aided by hierarchical analysis in knowledge exploration. *In Sixth International Conference on Intelligent Systems Design and Applications (Vol. 1, pp. 661–665). IEEE*.

Zhang, Z., & Luo, L. (2019). *Hate speech detection: A solved problem? the challenging case of long tail on twitter*.

Zufall, F., Hamacher, M., Kloppenborg, K., & Zesch, T. (2022). A legal approach to hate speech – operationalizing the eu's legal framework against the expression of hatred as an nlp task. *Pro-*

ceedings of the Natural Legal Language Processing Workshop 2022 (pp. 53–64). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.nllp-1.5