The Role of Expertise in Effectively Moderating Harmful Social Media Content

Nuredin Ali Abdelkadir
Computer Science and Engineering
University of Minnesota
Minneapolis, Minnesota, USA
The Distributed AI Research Institute
Minneapolis, Minnesota, USA
ali00530@umn.edu

Tianling Yang
Weizenbaum Institute
Technische Universität Berlin
Berlin, Germany
tianling.yang@tu-berlin.de

Shivani Kapania
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
skapania@andrew.cmu.edu

Meron Estefanos
The Distributed AI Research Institute
Stockholm, Sweden
meron@dair-institute.org

Fasica Berhane Gebrekidan
Independent Researcher
Nairobi, Kenya
fasikaberhane@gmail.com

Zecharias Zelalem
Freelance Journalist
Montreal, Quebec, Canada
zechariaszelalem@gmail.com

Messai Ali
Independent Researcher
Stockholm, Sweden
messai6776@gmail.com

Rishan Berhe
Independent Researcher
New York, New York, USA
rishanaberhe@gmail.com

Dylan Baker
The Distributed AI Research Institute
Seattle, Washington, USA
dylan@dair-institute.org

Zeerak Talat
University of Edinburgh
Edinburgh, United Kingdom
z@zeerak.org

Milagros Miceli
The Distributed AI Research Institute
Berlin, Germany
Weizenbaum Institute
Technische Universität Berlin
Berlin, Germany
m.miceli@tu-berlin.de

Alex Hanna
The Distributed AI Research Institute
Oakland, California, USA
alex@dair-institute.org

Timnit Gebru
The Distributed AI Research Institute
Oakland, California, USA
timnit@dair-institute.org

Abstract

Social media platforms played a significant role in spreading genocidal content in the 2020-2022 Tigray war, where the deadliest genocide of the 21^{st} century was committed. While linguistic expertise is clearly needed to adequately moderate such content, we ask: What additional expertise is needed? Why and to what extent do experts disagree on what constitutes harmful content, and what is the best way to resolve these disagreements? What do social media platforms do instead? We examine these questions through a 4-month study with 7 experts labeling 340 X (formerly Twitter) posts, and by interviewing 15 commercial content moderators. We find in-depth cultural knowledge and dialects to be most important for accurate hate speech annotation – knowledge which social media platforms do not prioritize. Even amongst experts, disagreements are high (71%), dropping to 40% after deliberation meetings. Based on these results, we present 7 recommendations to improve hate speech annotation and moderation practices.

CCS Concepts

• **Human-centered computing** → **Social media**; **Empirical studies in collaborative and social computing**.

Keywords

Expertise, Content Moderation, Data Annotation, Expert Disagreement, Harmful Content, Social Media Platforms

ACM Reference Format:
Nuredin Ali Abdelkadir, Tianling Yang, Shivani Kapania, Meron Estefanos, Fasica Berhane Gebrekidan, Zecharias Zelalem, Messai Ali, Rishan Berhe, Dylan Baker, Zeerak Talat, Milagros Miceli, Alex Hanna, and Timnit Gebru. 2025. The Role of Expertise in Effectively Moderating Harmful Social Media Content. In *CHI Conference on Human Factors in Computing Systems (CHI*

CHI '25, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3714010

'25), April 26–May 01, 2025, Yokohama, Japan. ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/3706598.3714010

1 INTRODUCTION

A number of researchers, journalists and victims of genocide have documented the extent to which social media platforms have spread genocidal content [18, 79, 93], leading to lawsuits against companies like Meta by victims of genocide and civil society leaders [5, 48]. This failure to adequately moderate genocidal content, especially in languages and regions major social media platforms do not invest in, has been discussed in a number of contexts [18, 75, 78]. For example, the 2020-2022 war in the Northern Ethiopian region of Tigray has been described as the worst genocide of the 21^{st} century thus far [63], with hateful content rampant on social media platforms during the war [33, 52, 70]. However, while operating in Ethiopia, a country with a population of 128 million[1] people speaking an estimated 100 languages [17], Facebook only supported "two of those languages for integrity systems" according to documents released by whistleblower Frances Haugen [65].

Investing in enough content moderators with appropriate linguistic knowledge is a baseline requirement for moderating hateful content. But in this paper we ask:

(1) What additional expertise do content moderators need to appropriately moderate hateful content?
(2) To what extent do moderators with the necessary expertise agree on how social media posts should be categorized following the platforms' guidelines? For instance, do experts agree on which posts should be marked as violent, abusive, or neutral? If not, what are the sources of their disagreements and how do they resolve these disagreements?
(3) What types of expertise do social media platforms value, and how do those compare with the necessary expertise identified by our work? What processes do social media platforms use to resolve disagreements amongst content moderators, and how do those compare to processes we identify to be important in resolving disagreements?

We explore these questions in the context of the 2020-2022 Tigray war, and perform a 4-month study with 7 expert annotators and interview 15 commercial content moderators. In collaboration with our annotators, we first created a dataset of 5.5M X (formerly Twitter) posts that could pertain to the Tigray war. Our annotators, consisting of journalists, activists, data archivists, refugee advocates, and former content moderators with linguistic, cultural and dialectical knowledge, then jointly labeled 340 of these posts (see Section 3 for details). We found that dialectical knowledge including slang terms was crucial in identifying harmful posts, but there was still high disagreement amongst experts with the same level of granular knowledge. Our study participants resolved these disagreements through deliberation meetings where each post they disagreed on was annotated by the reason of disagreement, and final labels decided on after discussion. While our experts started out disagreeing 71% of the time, this dropped to 40% after 5 deliberation sessions.

[1]https://data.who.int/countries/231

To find out what types of expertise social media platforms recruit for vs. the expertise we found to be important, and how they handle disagreements amongst content moderators, we furthermore interviewed 15 commercial content moderators with a minimum of 1-year experience. While these moderators valued in-depth familiarity with dialects, cultural practices, and broader social contexts, platforms preferred superficial cultural awareness and language skills of dominant languages in a region. In spite of the high disagreements amongst experts found by our study and the need to have clear processes for resolving them, we found that content moderators are prevented from raising such disagreements by organizational hierarchies, exploitative working conditions and inflexible platform policies. Thus, improving the working conditions of content moderators, and providing them with space to have disagreements they deliberate on, are imperative for appropriate moderation of harmful content.

In summary, the main contributions of this work are: (1) We provide insights into the types of expertise and skills needed to appropriately moderate harmful content, and how those differ from the expertise prioritized by social media platforms. (2) We find high disagreement in labeling harmful content even amongst expert groups, that is best resolved in deliberation meetings, whereas commercial content moderators are hindered from expressing their disagreements during the moderation process. (3) We illustrate how current working conditions and processes restrain moderators from exercising their expertise and raising disagreement, and provide suggestions for better content moderation practices to curb the dissemination of harmful social media content.

2 RELATED WORK

2.1 Disagreement in Data Annotation

There has been extensive research on annotator disagreement when labeling hate speech [85], toxicity [31, 81], and misinformation [95]. Disagreement can be caused by problems with defining and designing annotation tasks, such as task difficulty or unclear annotation description [40]. Another well-studied source of disagreement is annotators' subjectivity, as their race [58, 61], age [24, 66], education level, personality [37, 66], language [66], personal beliefs and values [16, 68], political orientation [37], and knowledge of hate speech [84] can affect annotation practices and thereby the annotated data. Annotators' judgments about different social groups are also subject to normative social stereotypes, resulting in a biased understanding of language directed toward marginalized groups [19]. Davani et al. [20] have further examined the influence of individual and geo-cultural variations on understanding offensive language and highlighted the significance of including diverse perspectives in annotation. An additional source of difficulty with annotator reliability and consistency is language's ambiguous and contextual nature [10]. Thus, ambivalence in words and phrases that can allow multiple interpretations, the lack of context to interpret them, and the inherent subjectivity in their interpretation can all result in variations in annotation [8, 90].

Researchers and practitioners have suggested various methodological and technical solutions to effectively deal with annotation disagreements, although majority voting is still the primary choice

[28, 60]. For instance, Gordon et al. [34] and Davani et al. [21] propose methods to predict each annotator's labels before modeling or aiming to resolve annotator disagreement. These methods are useful when getting a representative pool of annotators is impossible [34]. Others propose different methodologies to take expertise into account in the data annotation process [10, 22, 28, 56, 77, 88, 91]. Although these works studied the process of annotating harmful content and called for involving expert annotators, the consistency of annotations and causes of variation among experts remain an open question. We add to this line of work by exploring the extent to which experts agree in their annotations and examining the main sources of their disagreements. While experts are best positioned to provide well-informed decisions, we still find high levels of disagreements amongst them, which deliberation meetings are crucial in resolving.

2.2 Expertise in Content Moderation

Content moderation is, in a way, data annotation put into practice in real-time or near real-time. Social media platforms, for instance, rely on content moderators to improve the safety of their platforms [32], with human moderators annotating content to decide whether a post should stay online or not [64]. Social media platforms also use content moderators' annotations to train hate speech detection and other models [30]. As moderators' decisions greatly influence the content delivered to users of the platforms and society at large [35], researchers have advocated for designing better moderation strategies that account for diverse perspectives. Fleisig et al. [28] recommend, for instance, recruiting representative groups to capture better labels. Vaccaro et al. [80] suggest that platforms involve diverse groups in designing their policies, facilitate effective communication, and provide emotional support to moderators. Others have called for investing in a workforce of moderators with a deep understanding of the moderation policy as well as the cultural contexts of the content to be moderated [88], and involving native language speakers and local experts in crafting context-aware content moderation policies [91].

Prior studies on content moderation primarily focus on Western contexts that do not pertain to genocide during armed conflict. The few studies that have focused on a genocidal context discuss the 2016-17 genocide of Rohingya communities in Myanmar [14, 27, 62]. Rio [62] documents the spread of hateful and violent language targeting the Rohignya, and Brooten [14] and Fink [27] describe the weaponization of platforms such as Facebook to "stoke fear, normalize hateful views, and facilitate acts of violence against the Rohingya and other Muslim communities in Myanmar." Stecklow [75] notes commercial social media platforms' insufficient allocation of resources, such as content moderators fluent in local languages. Nkemelu et al. [54] stress the need to involve what they call context experts, people with "deep and personal knowledge of the context resulting from their lived experience," in the creation of automated tools to detect hate speech targeting Rohignya communities. Similar to those focused on Myanmar, the very few works analyzing the impact of social media platforms on the Tigray war either study campaigns spread by specific networks [15], the shortcomings of automated hate speech detection tools [78], or the companies' failure to adequately resource content moderation in the relevant languages and enforce their own policies during conflicts [18].

Our paper goes beyond calls to increase content moderators speaking various languages, presenting a granular analysis of the expertise and procedures needed to moderate genocidal content. Moderators participating in the Data Workers' Inquiry project noted that social media platforms disregard their perspectives, devalue their expertise, and penalize and fire them when they organize for better working conditions [2]. We further detail how these working conditions impact content moderators' ability to exercise the necessary expertise to effectively moderate harmful content, especially during armed conflict where the speed with which genocidal content is curbed is more crucial than other contexts.

3 METHODS

In this section, we outline our methodologies for designing and executing two studies. First, we discuss the design and execution of our expert annotation study, in which 7 study participants spent 4 months annotating hateful social media posts pertaining to the Tigray war. [2] We then describe the interview study with 15 commercial content moderators. The data annotation study gave us insight into the expertise needed to identify and classify hateful social media posts, quantified the level to which even people with necessary expertise disagree, and identified effective processes to address these disagreements. The interview study allowed us to compare our findings to the practices of commercial social media platforms, and to provide recommendations for better moderation practices given the input of commercial content moderators.

3.1 Expert Annotation Study

3.1.1 A Brief Background on the Tigray War. We anchored our study on the 2020-2022 Tigray war, given the documented consequences of inadequate moderation of genocidal social media posts pertaining to this war, the staggering number of lives lost, and the different languages and contexts needed to identify hateful language in this context [6, 94]. Tigray is a region in the northern part of Ethiopia, the second most populous country in Africa and the most populous landlocked country in the world.[3] The region has a population of approximately 7 million ethnic Tigrayans and is estimated to comprise 6% of Ethiopia's population [53]. The Tigray war is reportedly the deadliest armed conflict of the 21^{st} century [39, 83], with an estimated 600,000-800,000 casualties [59]. The warring parties were the Ethiopian National Defense Forces (ENDF), Eritrean Defense Forces (EDF), Amhara Regional Forces (ARF) and Amhara Militia (Fano) on one side, and the Tigray Defense Forces (TDF) on the other [26, 46, 87]. One of the longest recorded blockades in history was instituted during this war, with the Ethiopian government blocking access to phone and internet, and preventing the entry of food, fuel and other essentials to Tigray for two years, resulting in the mass starvation of Tigrayans [29, 36, 82, 89]. A number of human rights organizations have documented ethnic cleansing against Tigrayans perpetuated during the war, with at

[2] This 4 months includes the time spent in discussions to understand the categories and their respective explanation of the platform's policies. In addition, some deliberation meetings took longer than an hour. Hence, the remaining disagreed posts are discussed in the following week.

[3] https://achpr.au.int/en/member-states/ethiopia

least 100,000 Tigrayan women estimated to be victims of rape as a weapon of war [38]. The most comprehensive report on the topic was published by Newlines Institute in 2024, and asserts that a genocide has been committed against the people of Tigray [63].

3.1.2 Participant Recruitment. We recruited participants with the following characteristics:

(1) Most are native speakers of the dominant languages spoken by the warring parties (Tigrinya by EDF and TDF, Amharic by Fano, ARF, and ENDF), with some of our participants also speaking Arabic and Tigre (other dominant Eritrean languages). All participants are fluent in English, the language in which a significant amount of social media content is generated by the warring parties.
(2) Our participants are a mix of journalists exiled by the warring parties for their critical coverage, activists who have been independently archiving social media data pertaining to the war, dissidents who are targeted by the warring parties and are faced with social media harassment campaigns that result in physical attacks, and former content moderators whose full time job was to moderate this content for major social media platforms.

Table 1 summarizes the relevant backgrounds our annotators have, in addition to their linguistic knowledge.

3.1.3 Codebook Design. We based our codebook on X's policies pertaining to hate and violence on the platform.[4] Specifically, the codebook follows X's policies to classify posts under *Hateful Conduct*, *Violent Speech*, *Abuse and Harassment*, *Violent and Hateful Entities*, and *Glorification of Violence* (accessed between September and December 2023). Posts can violate multiple policies: for instance, there are insults and slurs that can fall under both *Abuse and Harassment* and *Hateful Conduct*. After studying the initial codebook, our annotators held discussions to identify what they believed was missing to flag hateful content that was spreading on X during the conflict. This led to adding subcategories that distinguish between *verified misinformation* and *suspected misinformation* under *misinformation*. We added categories like *I do not understand* and *I can not read the language* to ensure that annotators only label posts they understand. Labelers could also skip posts they did not believe they had adequate context to classify.

After this step, the codebook contained 10 categories: *misinformation (verified or unverified)*, *violent speech*, *abuse and harassment*, *violent event denial*, *dehumanization*, *neutral*, *irrelevant*, *I do not understand (lacks context)*, and *I can not read the language*.

3.1.4 Gathering Social Media Posts. We used the now-defunct Twitter Academic API to extract all posts that could pertain to the Tigray war. To do this, we first gathered 209 keywords and key phrases,[5] including known slurs and slangs used during the war, words that can be neutral like "Ethiopia," "Eritrea" and "Tigray," and the words' variants like "Tigrai" and "Tgrai." Our keywords consisted of terms in English, Amharic, Tigrinya, and Arabic and were created in collaboration with our expert annotators. We retrieved X content posted between January 1, 2018 and January 15, 2023 that included these keywords, and further cleaned the data to remove noisy keywords and terms for which most posts are unrelated to Ethiopia, Eritrea or the Tigray war. For example, we removed tweets in Spanish, Portuguese, and Catalan containing the keyword "junta," a term that was used by the Ethiopian government and its allies to describe Tigrayans and their supporters [63] but is a common word used in contexts unrelated to ours in these languages. The final dataset comprises 5.5 million posts, including original tweets, quoted tweets, and replies.

3.1.5 Annotating Social Media Posts. Each week, we randomly selected 55 tweets from our dataset and assigned them to our annotators to label according to the codebook. [6] We decided on 55 posts per week to restrict the participants' exposure to harmful content to a maximum of one hour per day, based on feedback from participants and the psychological impacts this content can cause [11, 73, 74]. After each round of annotation, our expert annotators spent a minimum of 60 minutes discussing their disagreements and arrived at a final agreed-upon label for each post after deliberation. They also documented the sources of their disagreements and the processes by which they resolved them. They performed 5 rounds of annotations via LabelStudio[7] and classified a total of 340 X posts.

Table 5 in the appendix gives examples of annotated posts, their translations to English, and the categories they were classified to. For instance, one post reads: *"#Eritrea #Ethiopia #HOA appreciate @[anony]@[anony] rebuffing Z @[anony] called @[anony] meeting. #TPLF spewed #FakeAxumMassacre needs investigation. But, lies mustn't be basis to harass sovereign nations. #TPLFstartedTheWar #EritreaPrevails #StopScapeoatingEritrea."* The post includes the hashtag #FakeAxumMassacre which emerged shortly after the hashtag #AxumMassacre which was used to highlight the killing of hundreds of unarmed civilians inside a church in the city of Axum, Tigray in November of 2020 [4, 7, 86]. Even though the event was recounted by survivors and investigated and corroborated by the likes of Amnesty International, Human Rights Watch and Associated Press, all of whom attributed the killings to Eritrean troops (EDF), this post claims that the massacre is "fake". Our annotators thus classified this post into the "Violent Event Denial" category.

3.1.6 Analysis. We used multiple quantitative methods to assess disagreements between annotators. We used *Krippendorf's Alpha* to measure disagreements between annotators' initial labels (rather than their agreed-upon annotation after deliberation). *Krippendorf's Alpha* is widely used to measure annotator variations and accounts for missing data where each instance is not labeled by all annotators [43]. In addition to *Krippendorf's Alpha*, we used metrics for *"complete agreement,"* measuring whether annotators agree on all labels assigned to a particular post (e.g., whether a post should be labeled both as *Abuse* and *Dehumanization*), and *"at least one class agreement"* which measures if annotators agree on at least one label assigned to a post. A *complete agreement* of 60%, for instance, indicates that our annotators agreed on all of their labels for 60% of posts. Whereas a score of 60% for *at least one class agreement* indicates that the annotators agreed on at least one label on 60% of the posts they labeled.

[4] https://help.x.com/en/rules-and-policies
[5] https://github.com/nuredinali/Social-Media-Harms-Keywords
[6] During the second round, annotators were given 2-3 weeks as the round contained 120 posts.
[7] https://labelstud.io/

Table 1: The expert annotators involved in the annotation process. Each of these experts worked for at least two years in their respective areas.

Area of expertise	Description
Journalists	Reporters who cover the East African political landscape, focusing on Ethiopia and Eritrea. We incorporated both freelance journalists and those who work full time at specific news media agencies.
Activists	Activists focusing on human rights in the region. We incorporated both independent activists and former employees of human rights organizations. These activists' advocacy ranges from highlighting government repression to providing firsthand accounts of the atrocities committed during the Tigray war.
Content Moderators	Individuals who have spent time working as content moderators in a particular market/region and are also from the specific society where they moderate.
Data Archivists	Individuals who spent considerable time independently archiving hateful posts on various social media platforms during the Tigray war. The archived posts violated platform policies but were not moderated. Data archivists have a large possible overlap with activists.

We held deliberation meetings and documented instances where annotators discussed why they did not have the necessary expertise to label a particular post, or when some annotators explained the context of a post to others who did not have the same understanding as them. This allowed us to identify the additional expertise needed to understand the specific context relevant to a post that others did not understand. We also used the notes accompanying each deliberation meeting to identify the main themes of disagreement amongst the annotators. To achieve this, we conducted a thematic analysis [13] of the notes to identify core themes.

3.2 Interview Study with Content Moderators

To compare the findings from our expert annotation study to the practices of major social media platforms, we conducted semi structured interviews with 15 content moderators who are or were involved in moderating harmful content.

Our interviews aimed to answer these questions:

(1) What expertise is most valued by social media companies in recruiting content moderators, and how does this compare to the expertise most valued by commercial content moderators?
(2) What processes do social media platforms follow to resolve disagreements amongst content moderators, and how do they compare to the procedures content moderators would like to have?

3.2.1 Participant Recruitment. We recruited participants by disseminating flyers and interest forms on multiple social media platform groups for former content moderators in Africa. The flyers and interest forms included the purpose of the study and asked interested participants to briefly describe their experience in moderation, the market they worked in (e.g., Ethiopia, South Africa, Nigeria), and the type of content they moderated. From the larger sample, we selected participants primarily involved in moderating harmful content, including violent speech, hate speech, abuse and harassment, misinformation, graphic content, nudity, and child abuse content. In addition to Ethiopia, we recruited participants focused on Nigeria, South Africa, Kenya, Uganda, and Somalia. But we oversampled the moderators who worked in the Ethiopian market as we wanted to compare the interview results with our expert annotation study and given the large scale of hate and violence spread on social media in the last few years in that market relative to the other markets.[8] Within Ethiopia, we selected moderators working in 4 markets: Afaan Oromo (a language spoken by the largest ethnic group in Ethiopia), English, Amharic and Tigrinya.

In addition to content moderators, our sample included quality analysts (QA) and subject matter experts (SME) to incorporate insights from different roles. All participants worked for one of the largest social media platforms (which we pseudonymized as InstantChat here) through a particular business process outsourcing company (henceforth BPO) which we pseudonymize as OutModeration. BPOs are third-party organizations to which social media companies often contract out their content moderation or data annotation. On average, our participants had 2-3 years of experience working as moderators. Refer to table 4 in the Appendix for details on the region/market of the participants.

3.2.2 Interview Procedure. Each interview began with a participant describing their background and professional experience as a content moderator. Participants were asked about their roles, the types of content they moderated, InstantChat's policies, and OutModeration's organizational structure. Our initial discussions focused on participants' onboarding, recruitment and training, as well as their perceptions of the work in each of these stages. Next, participants described how disagreements over flagged content were handled within their teams, including the role of power dynamics and any challenges they faced. Finally, participants reflected on how their background knowledge and lived experiences affect their work, and shared their preferred approaches to content moderation. All interviews were conducted online on Zoom, with the average length of the interviews being 64.8 minutes. Participants were compensated with 35 USD in their local currency.

3.2.3 Qualitative Coding and Analysis. We conducted a thematic analysis [13] on the interview data using the software MAXQDA.

[8] https://www.c-span.org/video/?515042-1/facebook-whistleblower-testifies-protecting-children-online

We started our analysis by conducting open coding on five interviews to closely examine the data. We then discussed our preliminary codes and potential themes, and completed open coding for the rest of the interviews. We conducted focused coding to identify main themes and organize and revise codes around them, having 2 rounds of discussions to exchange our understanding of the data, refine the coding system, and structure the findings.

3.3 Positionality Statement

More than half of the authors, including the first author, are either Tigrayans whose friends and families were killed and had to flee during the genocide enacted on Tigrayans during the 2020-2022 Tigray war, or Eritreans and other Ethiopians targeted (through online bullying and harassment and offline physical attacks) for their anti-war and anti-genocide stance. These authors saw firsthand the scale at which genocidal social media content was spreading unabated, which motivated them to investigate this content moderation failure and the expertise needed for effective moderation of harmful content. Both our data annotation and interview studies are from the standpoint that a genocide was enacted on Tigrayans, and do not lend credence to perspectives that may say otherwise, which we find to engage in genocide denial. The author team's expertise includes professional work in content moderation, research in hate speech detection and data labor, social media analysis, and ethical development of AI systems. Our experience advocating for data workers additionally motivated us to center content moderators in our critical examination of the commercial content moderation process.

4 FINDINGS

In this section, we outline the types of expertise necessary for moderating harmful content, as found through our expert annotation study and articulated by commercial content moderators in our interview study (Section 4.1). We then measure the level of disagreement amongst expert annotators during our annotation study, provide a descriptive analysis of the prevalence and causes of disagreements amongst experts who annotate and moderate harmful content, and describe processes for resolving these disagreements (Section 4.2). Sections 4.3 and 4.4 discuss the expertise prioritized by social media platforms and the processes they use to resolve disagreements amongst content moderators, and compare them to the expertise and processes our interviewees found to be necessary for effective moderation of harmful content. We have included additional interview quotes in Table 6 in the Appendix A.1 for further reference. In summary, we answer the following questions: what expertise is needed to moderate hateful social media content effectively? To what extent do expert annotators disagree with annotating harmful content? what are their sources of disagreement? What expertise do social media platforms prioritize instead? What is the current disagreement resolution process and how does this compare with those deemed valuable by the content moderators and expert annotators?

4.1 What expertise is needed to moderate harmful content?

Our data annotation and interview studies uncovered common skills and expertise that are necessary to appropriately moderate harmful content. In addition to linguistic understanding, moderators and annotators need to have specific knowledge of dialects, in-depth knowledge of cultural contexts, and other domain expertise to identify and classify harmful content. In-depth cultural knowledge is also required to grasp rich cultural variations and understand nuances of terms across social contexts. Our interview participants further described mental resilience, concentration and patience, and ability to perform teamwork to be important skills to moderate harmful content. These latter themes were not as salient in our annotation study, given that annotators were not under time pressure to moderate content and were not exposed to harmful posts for more than an hour per day. Finally, our annotation study showed that thoroughly understanding and regularly consulting the codebook was important, and our interview participants further stressed the importance of rigorous enforcement of platforms' policies.

4.1.1 Specific Knowledge of Dialects. Although it is clear that people need to understand the language a particular content is in to gauge whether it is harmful, our study shows that knowledge and familiarity with dialects are important to classify harmful content even in the same language. For example, while most of our annotators took the term "ሰው· በላ" to mean "one who eats people" (its literal translation), one of our annotators noted that this is a term commonly applied to describe leaders, with its more accurate meaning in this context being "brutal." This importance of dialectical knowledge was further stressed by our interviewees, where content moderators noted that OutModeration often expects them to moderate content in dialects they do not understand.

4.1.2 In-Depth Knowledge of Cultural Contexts. During our annotation study, we found that in-depth cultural knowledge was important to discern the social and cultural connotations and contexts of particular content and to interpret them accordingly. For example, the word "ወያነ" in Tigrinya means revolutionary, but among Eritreans, "ወያነ" is used to describe the Tigray People's Liberation Front (TPLF), whereas this is not the case among Tigrayans. And during the Tigray war, the Ethiopian government and its allies often used the Amharic equivalent of the term (ወያኔ) to imply that all Tigrayans are members of the TPLF [63]. Our annotators learned about the differences in the term's connotations among different communities when discussing why they disagreed on labeling a post containing this word. The importance of this level of cultural knowledge was also echoed by content moderators in our interview study.

4.1.3 Domain Expertise. In addition to dialectical and cultural knowledge, specific types of domain expertise can be important to correctly classify harmful posts. For example, journalists may be better at identifying verified misinformation compared to activists or data archivists. Conversely, long-time refugee advocates working with specific populations might be better at identifying slurs targeting those populations than journalists. As an example, our data annotators came across #There'sNoFanoinOromia a social media campaign, which claimed that the Amhara militia force

known as "Fano" was not present in the Oromia region of Ethiopia. Only an Ethiopian journalist amongst the annotators identified this post as "verified misinformation" because they worked on a report investigating that particular campaign.

4.1.4 Familiarity with Specific Networks on Social Media Platforms. Through our data annotation study, we found that even those who had appropriate linguistic and cultural knowledge did not identify harmful content on social media if they did not spend enough time on these platforms to understand the type of content that was disseminated across different networks. For example, one of our annotators was in Tigray during the 2020-2022 war when there was a 2-year siege without Internet or phone communication.[9] In spite of being a primary target of hateful social media content, this annotator could not identify many genocidal terms targeting Tigrayans.

An example of a post this annotator didn't identify as hate speech is an Amharic post translated to "Clean the cockroaches." In order to identify this post as genocidal content, one would need to know that the Ethiopian government and its allies regularly described Tigrayans as cockroaches and that the post was from an account belonging to a prominent activist who mostly disseminated anti-Tigrayan content.

In contrast, another Tigrayan annotator in the diaspora who spent an extended period of time independently studying and archiving harmful social media posts during the genocide quickly identified this post as genocidal content. Their familiarity with the type of content disseminated on social media by various networks in the context of the Tigray war enables this annotator to quickly detect implicit calls for violence, recognize troll accounts, and understand their targets.

4.1.5 Rigorous Understanding of Policy. In our annotation study, participants often consulted the codebook to remind themselves of how a particular post should be categorized. Sometimes, disagreements were resolved when annotators jointly read the codebook together during the deliberation meetings. This need for a robust understanding of platform policies was also echoed by our interviewees. Additionally, moderators further highlighted the importance of applying these policies rigorously because some of their colleagues aligned moderation practices with their political stances and ethnic backgrounds, retaining content that clearly violates platform policies but supports moderators' views or opposes their enemies (discussed further in Section 4.2.2).

4.1.6 Mental Resilience. Our interview participants pointed mental resilience is a crucial skill for commercial content moderation. All participants highlighted the severe psychological distress they suffer due to their constant exposure to disturbing social media content, which is still present even after they stopped working as content moderators. Moderators' working conditions, such as having to moderate a large volume of harmful content under strict time constraints, workplace surveillance and control, and the inability to seek emotional support from friends and family members due to nondisclosure agreements, exacerbated the mental toll of commercial content moderation work. Therefore, our interviewees perceived the ability to be emotionally detached from work and stay mentally strong as crucial.

This theme was not as salient during our data annotation study, given that our annotators were exposed to harmful content for a maximum of 1 hour per day as part of the study. However, some of our annotators were concurrently independently archiving harmful social media posts, and discussed the mental toll of being exposed to these posts for an extended period of time. This led us to start a wellness program for our team designed by a psychotherapist who treats secondary trauma.

4.1.7 Concentration and Patience. Commercial content moderators work under strict time limits, during which they quickly have to understand and interpret flagged content while ensuring that they patiently analyze the entire content before making decisions. Given the volume of content they have to analyze, moderators have to furthermore ensure that they don't lose concentration as the day progresses. Teka notes:

> "when you read [a piece of content,] maybe you're tired [and] you might miss some words. (...) Sometimes you find very, very long tickets and maybe [the words that align with the policy and validate your moderation decision are] at the bottom of that long paragraph. So if you do not read the whole [content], you might not capture what is validating [your decision]. (...) [So] you have to be very patient [and] have to read keenly."

Our annotators also noted the need to sometimes further analyze the context of a post, even going back to the original post and analyzing its responses before making decisions. While they noted the importance of thoroughly analyzing posts to ensure that they are correctly interpreted, the need for concentration wasn't highlighted by our annotators as they were not met with the volume of content that commercial moderators have to handle.

4.1.8 Effective Teamwork. Our data annotators had regular deliberation meetings and discussed their perspectives on how posts should be annotated before jointly making decisions on posts they disagreed on. Our interviewees found this type of teamwork to be ideal: while moderators are expected to work individually on the tickets, they often collaborate in practice by, for example, helping each other understand unfamiliar content or contexts. Commercial content moderators might also informally discuss platform policies and controversial tickets so that they can reach a consensus and prevent their disagreements from reaching quality analysts.

4.2 How much and why do experts disagree on flagging harmful content?

Our expert annotation study allowed us to investigate to what extent annotators with the type of contextual expertise detailed in Section 4.1 disagree with each other in categorizing harmful social media posts. Our measurements are summarized in Table 2. We see that the Krippendorf's Alpha coefficient increases from 0.2 after the first round of annotation to 0.55, the level of disagreement amongst all labels decreases from 71% to 40%, and disagreement on at least one label decreases from a high of 60% to 33% by the 5^{th} round. The consistent increase in the level of annotator agreement

[9] https://www.accessnow.org/press-release/two-years-internet-shutdowns-tigray/

shows the importance of deliberation meetings to have a joint understanding of the codebook, to debate annotations, and to share contextual knowledge. Not only is this type of teamwork preferred by commercial content moderators as echoed by our interviewees (Section 4.1.8), but it could lead to more consistent moderation practices as moderators share their understanding of moderation policies and provide missing context.

We further compared labels created through majority voting to those that were obtained through deliberation, and found that 51% of the labels reached through majority voting were different from those created through deliberation meetings. In some cases, each annotator had a different label for a particular post and all annotators reached consensus after deliberation. In other cases, one annotator convinced the majority of annotators to change their labels after discussion, highlighting the importance of deliberation meetings. The biggest disagreement stemmed from whether a post was neutral or not, and the least disagreement was in identifying posts that were suspected of misinformation. Table 3 in the appendix further details the distribution of labels and disagreements across them.

We also analyzed the deliberation notes from our annotators to understand the main sources of their disagreements and compared them to the sources of disagreements mentioned by the commercial content moderators in our interview study, which we list below.

4.2.1 Cultural Variability. As discussed in Section 4.1.2, the same term can have differing connotations in various cultures. Thus, annotators with different cultural backgrounds can disagree on how a particular post using a term known to all of them should be annotated if there is ambiguity as to which cultural context the term is being used in. For example, terms deemed abusive but not dehumanizing in one culture may be perceived as extremely dehumanizing in another, which leads to disagreement on whether certain posts should be labeled as "Abuse and Harassment" or "Dehumanization".

The word "ሳሕሳሕ" in Tigrinya, for instance, has a number of meanings, one of which can be translated to "extremely dirty." However, the Eritrean annotators noted that amongst urban Eritreans this term is a slur that carries implicit discrimination and is mostly used to dehumanize Tigrayans. The Eritrean annotators then unanimously decided that the term should be labeled as dehumanizing, while the Tigrayans did not believe this to be true before the deliberation meeting during which additional context was provided by the Eritreans.

Our interviewees provided similar examples during commercial content moderation. For example, Aya notes that similar disagreements often arose between Somali Kenyans and Somali Somalians who both speak the same language (Somali):

> "Maybe the quality analyst (...) speaks stronger Somali from Somalia, and we are here in Kenya. So he always used to take us down (...), because they'll be like, no you (...) Somali Kenyans. You think maybe this is (...) dehumanizing, but (...) according to Somalis and from Somalia, this is not (...) this is just like a joke. This is not even like bullying (...) So you see, (...) we used to disagree a lot."

4.2.2 Differing Political Stances. People with the same amount of contextual knowledge can have differing political stances. Most of the annotators we recruited for our annotation study were targeted by all the warring parties at one point or another and did not support any of the political parties involved in the Tigray war. However, we ensured that we recruited people who did not deny the genocide of Tigrayans as was widely being done during the war in spite of overwhelming evidence [63]. Hence, disagreements on posts did not arise from differing views of what was transpiring on the ground.

On the contrary, content moderators in our interview study noted instances of moderators clearly violating platform policies to support a particular warring party's stance, such as deciding not to delete violating posts by a party they support. This led to increased disagreement among moderators or between moderators and quality analysts, resulting in considerable tension at the workplace and in some cases even physical confrontations, as reported by the participants. Fiona says:

> "Since we were having problems [wars] at our country, we were also having problems in the office because we couldn't agree on the policies. (...) It was hellish for the Tigrinya content moderators. That's my experience from every content moderator. Some of the Amharic speaking content moderators were rational and unbiased (...) [but others were very biased. Even sometimes] there were huge fights."

4.2.3 Ambiguity of Political and Societal Terms. In times of conflict and especially genocide, there can be ambiguity on who specific terms are targeting, often because warring parties intentionally conflate entire groups of people with particular parties they are at war with [45]. Coming back to the example we discussed in Section 4.1.2, a post containing the term "ወያነ" (in Tigrinya) can be used in its true meaning ("revolutionary"), to refer to the Tigray People's Liberation Front (TPLF: a political party), or intentionally conflated with the people of Tigray and anyone opposing the Tigray war. Similarly, the term TPLF could be used to refer to the political party or intentionally conflated to imply that Tigrayans and all those who oppose the war are members of the TPLF. Given that platform policies do not afford political parties the same protections as specific ethnic groups, how posts containing the terms above should be classified depends on whether they refer to Tigrayans or their political parties. While this implication can sometimes be clear given contextual knowledge surrounding specific posts, there are times when it is ambiguous, creating disagreements amongst annotators even after deliberation meetings to resolve them.

4.2.4 Posts missing context. Related to the point above, posts with missing contexts create ambiguity that results in disagreements on how they should be classified [72]. When this occurred, our annotators attempted to jointly find the original posts and discuss their contexts in order to arrive at agreed-upon labels. However, this is not always possible during commercial content moderation, where workers are under strict time limits (e.g., spending a maximum of 90 seconds per content) and intensive surveillance. So they have to make decisions on posts with missing contexts or posts

Table 2: The level of disagreement within the seven experts. N=7

Round	Sample size	Krippendorf's Alpha	Complete disagreement (%)	At least one class disagreement (%)
1	55	*0.20*	0.71	0.40
2	120	*0.41*	0.65	0.60
3	55	*0.44*	0.56	0.47
4	55	*0.46*	0.44	0.38
5	55	*0.55*	0.40	0.33

they did not have time to read/watch completely, which results in disagreements. As Jacob notes:

> "It might look like someone is dead to you. [But to me] it looks like no, this person (...) just fainted, collapsed (...), he's still breathing. So yeah, (...) a lot of those tickets [are] very (...) ambiguous."

4.2.5 Unclear and Out-of-Date Moderation Policies. Our annotators sometimes disagreed with the codebook we followed, which was derived from X's policies as discussed in Section 3.1.3. Our interview participants further noted that it is important to frequently update social media platform policies and ensure that they apply to regions with rapidly developing contexts. For instance, new dehumanizing words can quickly emerge without platform policies being updated to reflect them. This can lead to inconsistencies where some moderators do not label posts containing these words as dehumanizing to ensure that they follow (out of date) platform policies, whereas others flag them as dehumanizing given their contextual knowledge. Teka explains:

> "There was a word crisis in the Tigray region in Ethiopia (...) people were referring (...) [to the Tigrayan people as] locusts (...) during that word crisis, we have been seeing that word used a lot of times (...) It was creating a lot of inconsistencies in the way that we are handling the job. Because for some of us, we don't have the context, we just ignore it (...) we have Tigrinya speaker content moderators (...), they start deleting that content."

Some moderators also noted that unclear moderation policies are compounded by the short training (3 weeks) they were given to understand and follow these policies, making it difficult to remember every detail. This causes labeling inconsistencies between moderators who rigorously follow platform policies vs those who do not.

4.3 What skills or expertise are prioritized by social media platforms?

Based on our interview study, we identify four forms of expertise valued by social media platforms for content moderation: language skills, superficial cultural awareness, robust understanding of platform policies, and mental strength. Notably, in addition to language skills, only 2 out of the 8 forms of expertise we identified in Section 4.1, robust understanding of platform policies and mental strength (similar to mental resilience), are valued by social media platforms.

4.3.1 Language Skills. There was a strong consensus among participants that having language skills specific to certain regional markets was the most important expertise needed to be hired as a content moderator. Although Moderators are recruited to work on content in specific regional markets that they come from, they can also take on additional work in languages they speak but are not native to. Sabrin notes:

> "It's a language-based job. So (...) the main requirement was just to know the language that is required."

Participants recalled that language proficiency is a key focus during the recruitment process. Notably, OutModeration conducts language tests during the interviews to ensure that candidates are native speakers.

4.3.2 Superficial Cultural Awareness. OutModeration expects moderators to have superficial cultural awareness, such as knowledge of public figures and cultural and political events, most of which can be easily searched and found on the Internet. Such knowledge is typically tested during the recruitment phase through questions such as "Do you [know] the current affairs in Ethiopia, [and] current affairs in the Oromo region?" or "Which singers sing about songs of violence?" Moderators also undergo training to stay up-to-date on the latest cultural events and keep themselves informed through informal discussions amongst themselves or by following local news.

As we discussed in Section 4.1.2, in-depth cultural knowledge is required to grasp cultural variations and nuances of terms across social contexts, which goes beyond platforms' focus on superficial cultural awareness. For instance, quality analysts use moderators' in-depth cultural and dialectical knowledge to improve content moderation in the corresponding markets, even though the recruitment process does not specifically test for that knowledge. Tarik, a quality analyst, notes:

> "When we hire you (...), the first thing I need to do as a quality analyst is to learn your background, [e.g. the] types of words which are violating [in your area] is not the types of words [that are violating] in our country, but [they are] in the same language. So you teach me things like that (...) and in turn, I teach the rest of the team, how to work on content like that."

4.3.3 Robust Understanding of Platform Policy. InstantChat establishes moderation policies, while OutModeration trains moderators and tests their understanding of these policies through worker training exams during the onboarding process. Moderators are expected to learn the policy by heart and seek clarification and guidance in cases of ambiguity. Policies are also regularly updated to adapt

to the latest trends in regional markets, albeit not at a frequency matching the rapid developments in some regions (see Section 4.2.5). Once moderators are notified of these policy changes, they are expected to immediately adopt them and change their moderation practices. Tarik notes:

> "you need to understand [the policy] as much as you understand your Bible. This is how close the policy should be with you, in order for you not to make mistakes."

4.3.4 Mental Strength. OutModeration's recruitment process attempts to evaluate moderators' mental strength. But our interviewees found the process to be misleading and unable to assess the mental resilience that they identified to be crucial for the job (see Section 4.1.6).

For instance, OutModeration's job advertisement might give the impression that they were recruiting administrators or translators rather than content moderators. During the interview process, participants did not get clarity on the specific tasks required for the job. Notably, our interviewees were not even told that the position they were to fill was that of a content moderator, until they were hired. And once they started their hands-on training, the social media content participants processed was significantly less disturbing than the content they worked on as employees. Sabrin described the confusion caused by this lack of transparency during the recruitment process:

> "in the whole interview process, (...) they didn't even tell me what the job really entails. (...) Unfortunately, when I came here [and worked as a content moderator], that's [when] I then realized that (...) it's really not what I thought it is."

OutModeration's recruitment process tested our interviewees' mental strength through a psychometric test, a conversation with a psychological counselor, or a combination of both. However, participants perceived such evaluations to be shallow and unable to assess the mental resilience that they identified to be crucial for the job (see Section 4.1.6). Interviewees described the psychological assessment as a "light conversation" with the counselor covering "general questions." Fiona recalls:

> "The counselor would ask you about your strengths and weakness? And how do you deal with stressful situations or contents (...)."

4.4 How do social media platforms resolve disagreements amongst content moderators?

As we found in Section 4.2, open discussion amongst annotators of harmful social media content allows them to understand different cultural contexts and resolve disagreements, which leads to more consistent labels. This need for teamwork in order to exchange contextual knowledge was also echoed by commercial content moderators we interviewed (Section 4.1.8). Nevertheless, these moderators point out that they do not have the agency to voice their disagreements with how posts are labeled and that their views are often devalued or completely disregarded.

Figure 1 summarizes the disagreement resolution process at OutModeration. Commercial moderators refer to the social media content they process as "tickets." Multiple moderators work on each ticket, and if disagreements arise amongst them, quality analysts review the tickets. To address daily disagreements among moderators and between moderators and quality analysts, quality analysts hold regular calibration sessions with moderators to discuss controversial tickets. The mechanisms to reach consensus vary across regional markets, but two stand out: taking the quality analysts' judgment or majority voting.

These mechanisms often do not result in full consensus. In majority votes on content involving ethnic conflicts or gender-specific understandings, the larger ethnic or gender group consistently wins regardless of whether or not their judgments are correct. This is particularly harmful when the content relates to the genocide of minority populations whose ethnic groups are often outnumbered among commercial content moderators. In contrast, our annotation study showed that when there is space for open conversation, even one person with the appropriate domain expertise (e.g., a journalist who reported on verified misinformation as noted in Section 4.1.3) can change the others' views, and provide them with missing contextual, cultural or dialectical information to arrive at the correct label.

In most of the cases involving disagreements at OutModeration, the views of the quality analysts are prioritized. However, moderators can escalate disagreements to team leaders, who forward the questions and seek clarification from market specialists. We uncovered 3 social media platform practices that prevent moderators from exercising their expertise and agency and voicing their disagreements: rigid organizational hierarchies, exploitative working conditions, and inability to influence platform policies.

4.4.1 Rigid Organizational Hierarchies. Our data shows a hierarchical organizational structure within OutModeration where content moderators are positioned at the bottom. **Team leaders** perform administrative tasks like ensuring moderators start work on time and monitoring their pace. Though this position doesn't exist in all regional markets, **subject matter experts** usually sit and work with moderators and provide guidance in case of questions about tickets and policies, and **quality analysts (QAs)** approve or flag moderators' works. Quality analysts can disagree amongst themselves on how to handle specific tickets, and these disputes are concealed from content moderators and resolved by leaders of the QA team or majority votes amongst QAs.

When disagreements on tickets aren't resolved within the "top members" of the QA team and require a majority vote among QAs, they are forwarded to **market specialists** for resolution, who are InstantChat's staff instead of OutModeration and have the final say over moderation decisions and their explanations. Content moderators can also escalate their disagreements to market specialists. However, unlike quality analysts, content moderators are at the bottom of the organizational hierarchy. Thus, although moderators can escalate their disagreement, in most cases, the market specialists align with quality analysts' judgments. Tarik, a QA, described his interaction with market specialists:

> "because to a market specialist, (...) Quality analyst is real special team, because those are the best of the

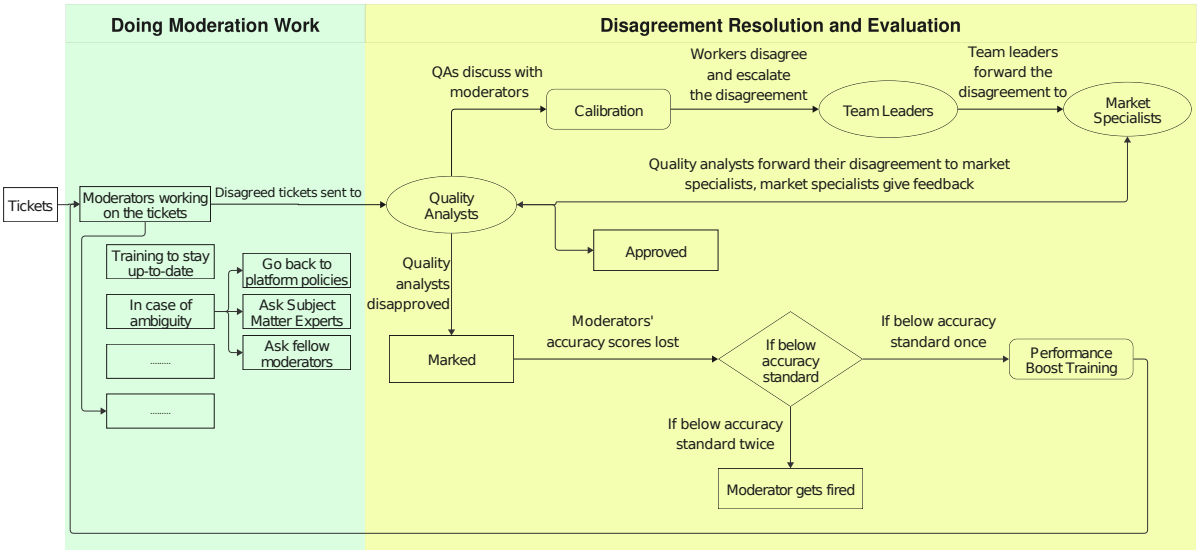

Figure 1: Summary of the content moderation, disagreement resolution, and evaluation workflows at OutModeration.

> best that they've chosen to put in a team together. (...) And so I had a feeling that even though we weren't correct, the market specialist would agree with us. So the moderator would always end up wrong."

Such rigid organizational hierarchy imposes higher social costs on moderators who want to express their views and challenge the judgments of quality analysts.

4.4.2 Exploitative working conditions. Moderators' performances are constantly measured, monitored, and evaluated, resulting in them being under pressure to adhere to strict time limits and keep up their accuracy rates. Disagreeing with the judgments of quality analysts can risk moderators' losing accuracy scores, which can lead to wage cuts and even job loss. When quality analysts mark tickets as incorrect, the corresponding moderators' accuracy rates drop. If the accuracy rate falls below a fixed standard for the first time, a moderator receives a performance boost training before returning to work. But moderators for whom this drop occurs twice are summarily fired.

Commercial content moderators in precarious conditions are particularly vulnerable to wage reductions and job insecurity. For example, those working on content related to conflicts might become targeted by extremists or warring parties, leading to concerns about their personal safety and fear of returning to their home countries. Moderators under these types of conditions avoid expressing disagreement in order to secure their pay and jobs, even though they know they are being exploited and discriminated against across regional markets as Ernest explains: " [for people from] South Africa I think they [get] full payment [like 65,000 Kenyan shillings] (...) [and] Ethiopians including me (...)[are] paid like 35,000 Kenyan shillings."

4.4.3 Inability to Influence Platform Policies. As we discussed in Sections 4.2.5, annotators and content moderators often disagree with the labeling guidelines enacted by social media platform policies, one of the reasons being that these policies often fail to address region-specific contexts. Despite these perceived gaps, moderators have limited agency to challenge the policies and initiate changes. Tarik notes:

> "We turn [moderators] into robots, we force them to understand policy (...) So it's not about what you know, what you think you know, it's what policy says (...) eventually everyone just became auto pilots on the job."

Some moderators try to convince people in higher positions to report policy issues to market specialists and suggest changes, but content moderators' opinions "weren't really valued as they should have," with moderators even receiving delayed replies or non-responses from market specialists. Mary says:

> "sometimes these specialists will just tell you that, hey, stop arguing, (...) please follow the policy (...) sometimes we could argue (...) and then after months, you will find that InstantChat has updated something (...), but it has to come from them. Not necessarily from our opinions. And sometimes you (...) find that some updates are (...) not enough. And you know, there's nothing you can do about it. Just follow the policy, do your job. Go home, that's it."

5 DISCUSSION

Overall, our findings highlight the unique insights that can be gained by studying the under-researched contexts of social media platform moderation in non-Western languages, particularly in the context of genocide. Our work found the superficial cultural awareness and basic language skills prioritized by a BPO moderating content for one of the biggest commercial social media platforms,

to be inadequate. Our annotation study further showed that annotators have differing interpretations of posts depending on their contextual, dialectical, and domain-specific knowledge, even when they are native speakers of the same language. Prior work has found that social media users who are from marginalized communities also find the aforementioned expertise to be important in moderating content pertaining to them. For instance, similar to our findings, various works have highlighted the need for minority voices [23, 25], dialectical knowledge [3], and in-depth, localized cultural knowledge [44, 55, 71]. Our work supports these findings and shows the extent to which those minoritized perspectives are even more important to consider in fast changing contexts of armed conflicts and genocide that affect a minority of content moderators and social media users who are uniquely at risk of political violence. In addition, we present new insights that emerge due to our contextual focus on the Tigray war.

For instance, valuing domain expertise could mean involving journalists who cover specific regions and topics [57]. However, our work shows that this in itself is insufficient. Even amongst journalists with in-depth knowledge of the topics, regions and languages being covered, our annotation study showed that only journalists who investigated specific campaigns were able to accurately flag some posts related to those campaigns as verified misinformation, while others covering the same region and with understanding of the same context did not flag these campaigns as such. Similarly, we found that familiarity with specific social media platforms and how certain networks of political actors operate on them is important in our context, because studying specific platforms allows people to quickly see new campaigns and networks that evolve in response to the fast-changing reality on the ground. Moderating genocidal content while the genocide is being enacted requires urgency and attunement to fast-changing realities, much more so than other contexts of moderation that have been more widely studied in the CHI literature.

Our findings also note that commercial content moderators value collaborative work, a procedure which is not highlighted in prior work studying content moderation. In the context of genocide and armed conflict, terms vilifying the target population quickly emerge, with the meanings of previously known phrases changing rapidly. In this context, it makes sense that ambiguity of societal terms, where some moderators understand how those terms are evolving whereas others may not, can be a source of disagreement amongst them. As moderators gain knowledge of these campaigns, networks, and platforms, it becomes even more important for them to share this knowledge amongst themselves to more effectively hamper genocidal content from spreading.

Another major contribution of our work is highlighting the extent to which the expertise, processes, and procedures prioritized by commercial social media platforms differ from those we identify to be crucial in moderating harmful content, especially in contexts that are only known to a minority of content moderators. We elucidate how commercial social media platforms de-prioritize the forms of expertise that both content moderators (according to our work) and social media users (according to prior work [71]) find to be crucial in moderating harmful content, during the content moderator recruiting process.

We further show that even if content moderators with the necessary expertise make it through the recruiting process, the organizational practices of BPOs and commercial social media platforms make it difficult for these moderators to exercise their expertise. Moderators regularly identify gaps in platform policies, including the lack of updates that reflect fast-changing contexts in specific regions, such as the emergence of new genocidal words. However, they are unable to effectively influence policy changes to incorporate these updates due to their location at the bottom of the organizational hierarchy and their exploitative working conditions, which severely punish them for deviating from the status quo. Bridging research on expertise in content moderation and worker exploitation, our work shows that exploitative labor conditions and organizational hierarchies enacted by commercial social media platforms result in lack of appropriate moderation of genocidal content.

These exploitative working conditions are also linked to the expertise valued by content moderators. While prior work has extensively documented the mental toll faced by commercial content moderators [2, 50, 64], we show that the moderators themselves believe that mental resilience should be a form of expertise prioritized by social media platforms. Similarly, our work highlights concentration and patience as forms of expertise that are important for all content moderators, rather than solely those in specific social groups involved in data work, such as women [12] and disabled people [92]. Nevertheless, the content moderator recruiting process misleads job applicants on what their tasks and working conditions will look like: moderators note the importance of mental resilience, while employers seem to downplay the severity of the content they evaluate.

Given our results, we present a number of actionable recommendations to improve current moderation procedures to more effectively curb the dissemination of harmful content on social media. These considerations are not meant to serve as an exhaustive checklist but instead offer a starting point for responsible content moderation practices. We note that while many of the issues pertaining to the working conditions of the commercial moderators who participated in our interview study can be attributed to Out-Moderation's organizational structure, Muldoon et al. [49] point out a number of positive working conditions at BPOs that stemmed from requirements instituted by clients. These results show that requesters' involvement can drive positive changes. Thus, echoing existing initiatives on ethical and responsible sourcing [1], we recommend that social media platforms take a proactive role in establishing more inclusive and effective disagreement resolution mechanisms and providing a better working environment.

5.1 Accuracy as the Main Measure for Moderation Performance

Our findings show that data annotation and content moderation are interpretative tasks deeply influenced by the expertise and perspectives of the annotators and moderators [10, 41, 67], emphasizing the situatedness of interpreting harmful content and the ambiguity of such content itself. This calls into question the common practice of equating accuracy with "correctness" because accuracy in these settings mainly reflects how closely workers' interpretations

align with annotation guidelines, moderation policies, or expert judgment, all of which are inherently partial. Relying primarily, if not solely, on accuracy as the measure of moderation performance poses a further concern in the contexts of low-resource languages and non-Western cultures, as moderation policies are often written in English and grounded in Western cultures, and are not well adapted to the target language and culture. At times, some policies are outdated, unclear, and contextually irrelevant. In this regard, relying heavily on accuracy can reinforce the norms of Western cultures [71] and be counter-productive to effective moderation, because it can lead workers to rigidly follow prescribed interpretations of data without striving for contextual understanding. We therefore suggest that social media platforms request BPOs to take the following actions.

- Ensure dialectical diversity amongst moderators speaking the same language and hire moderators with in-depth cultural and contextual knowledge, especially in regions with a multiplicity of languages and dialects and a diversity of cultures. These expertise go beyond the superficial cultural knowledge (like knowledge of public figures) that BPOs test for, and are needed to discern nuances in the harmful content spread during times of conflict and genocide.
- Remove overly punitive measures against moderators with lower than standard rates twice in a row. Platform policies and QA teams may lack the necessary contextual, cultural and dialectical understanding of the content to be moderated. The focus on accuracy rates coupled with overly punitive measures discourages moderators who have this knowledge from raising disagreement and sharing the knowledge necessary to keep policies up to date.

5.2 Rigid Organizational Hierarchies and Exploitative Working Conditions

Commercial moderators face wage cuts or, in the worst case, job termination if they exceed the time limit stipulated for each ticket or fall below a set accuracy rate. This can happen in cases where moderators' contextual knowledge leads them to make different moderation decisions than instructed by QAs. Quality analysts' power over moderation decisions, accuracy scores, and the support they gain from market specialists make it very unlikely for an individual moderator to express disagreement. Additionally, the lack of mental health support and inadequate breaks from being constantly bombarded by disturbing content can lead to lasting trauma and severe mental health issues [2, 50, 64, 76]. These conditions make moderators likely to follow the interpretation instructed by platform policies, quality analysts, and market specialists and prioritize income and job security instead of their own expertise. We therefore make the following recommendations to improve moderators' working conditions.

- Policymakers should establish clear, enforceable standards for ensuring responsible and ethical sourcing of data services. This includes fair wages, better working conditions, access to benefits, and equitable treatment for in-house and outsourced moderators. Platforms must be required to publicly disclose detailed information about labor conditions, including work hours, payment, benefits, and the availability of psychological support. Such transparency would empower worker advocacy groups by providing information which can be used to effectively advocate for fair treatment and strengthen collective worker power.
- Platforms must go beyond superficial mental health initiatives, such as simply providing on-site counselors, and prioritize care that addresses the specific challenges of content moderation. Participants in our study noted that on-site counselors often provided generic advice and did not understand the severe psychological harms tied to moderation work. Platforms and BPOs should support proactive approaches, including peer support networks and regular check-ins with mental health professionals trained in moderation induced stress and trauma. They should also redesign workflows to ensure structured, frequent breaks and implement workload caps to avoid burnout and limit trauma exposure. Moderators should be able to take these breaks without their wages suffering as a result, and without seeking approval for time off from their managers.

5.3 Ineffective Mechanism for Raising Disagreement and Influencing Platform Policy

We found that moderation teams achieve consensus either through imposing the judgments of quality analysts or market specialists on moderators, or through majority voting. But as our findings show, the varying extents of disagreement among experts (Section 4.2) challenge the epistemic authority of quality analysts or market specialists as definitive decision-makers whose judgment is believed to ensure objectivity and accuracy [47]. Our annotation study also found that using majority voting risks silencing minority voices, as the composition of groups can significantly impact the voting result: only 49% of the majority voting results agreed with the annotations after deliberation. Our findings echo prior work on the importance of disagreement as a valuable source of information [9, 10, 22], and extend previous work on the significance of deliberation meetings [42, 69] by showing that they are even more vital while moderating hateful content during active genocides targeting minority populations who are often underrepresented amongst content moderators. We therefore recommend the following actions to encourage moderators to raise disagreement without repercussions.

- Social media platforms should establish processes that enable content moderators to co-design and periodically update moderation policies and improve the adaptation of these policies to the target languages and cultures. This includes establishing a process that involves not only market specialists in the co-designing and updating of policies but also moderators who do the day-to-day moderation.
- Social media platforms should have frequent deliberation meetings that allow for the exchange of varying viewpoints without fear of negative repercussions for their differing views. Although deliberation meetings currently exist, the potential for negative repercussions often discourages moderators from sharing their genuine views, thereby undermining the fundamental purpose of these discussions. It is

essential to provide sufficient time and foster an environment where moderators feel comfortable exchanging their views with their peers.
- Social media platforms should provide detailed annual reports that disclose their moderation practices, as they do for the European Union through the Digital Services Act. While our work provides initial insights into the moderation procedures of platforms, such reports are needed to obtain a more detailed picture, especially in regions where social media platforms are used to fuel conflicts.

6 LIMITATIONS AND FUTURE WORK

Our data annotation study was performed with 7 annotators. While our annotators had a range of domain expertise and cultural backgrounds, incorporating annotators with additional perspectives would strengthen our study. An estimated 100 languages are spoken in Ethiopia and 15 in Eritrea [17, 51]. Future work could expand our work by covering more of these languages and cultures.

Our interview study focuses primarily on the perspectives of commercial content moderators working at social media platforms via BPOs since they are at the front lines of flagging harmful content on social media. Future work should explore the perspectives of quality analysts, market specialists and policy designers who can provide additional insights into the organizational practices and processes of social media platforms.

7 CONCLUSION

Through two studies focused on the annotation and moderation of harmful social media posts targeting Tigrayans during the 2020-2022 Tigray war, we investigated the expertise needed to effectively moderate harmful social media content during times of conflict and genocide. Our findings show stark disparities between the forms of expertise content moderators deem crucial and those prioritized by social media platforms. While social media platforms prioritize basic linguistic skills and superficial cultural knowledge, our study found in-depth cultural knowledge and granular dialectical knowledge to be crucial to effectively flag harmful posts. And while social media platforms resolve disagreements through majority voting or using experts higher in the organizational hierarchy as tie-breakers, we find open discussions and deliberation meetings where moderators can exchange contextual information and resolve disagreements to be more effective ways of ensuring that harmful content is appropriately flagged. Based on our findings, we provide 7 recommendations to change content moderation practices to more effectively curb harmful social media posts. These include removing overly punitive measures based on accuracy rates, providing mental health care that addresses moderators' needs, involving moderators in co-designing and periodically updating policies, providing detailed annual reports about moderation practices in the majority world, and allowing frequent deliberation meetings to effectively resolve disagreements.

Acknowledgments

This research is partially supported by a grant from the Internet Society Foundation to the Distributed AI Research Institute (DAIR). We thank Asmelash Teka for his feedback, our other contributors who chose to be anonymized, and all research participants who took part in this study. We also thank our anonymous reviewers for their helpful comments in revising this paper.

References

[1] 2021. *Responsible Sourcing of Data Enrichment Services.* Technical Report. Partnership on AI. partnershiponai.org/responsible-sourcing
[2] 2024. The Data Workers' Inquiry. https://data-workers.org/berlin Edited by M. Miceli, A. Dinika, K. Kauffman, C. Salim Wagner, and L. Sachenbacher.
[3] Alham Fikri Aji, Genta Indra Winata, Fajri Koto, Samuel Cahyawijaya, Ade Romadhony, Rahmad Mahendra, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Timothy Baldwin, Jey Han Lau, and Sebastian Ruder. 2022. One Country, 700+ Languages: NLP Challenges for Underrepresented Languages and Dialects in Indonesia. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 7226–7249. https://doi.org/10.18653/v1/2022.acl-long.500
[4] International Amnesty. 2021. Ethiopia: The Massacre in Axum. https://www.amnesty.org/en/documents/afr25/3730/2021/en/
[5] International Amnesty. 2022. Kenya: Meta sued for 1.6 billion USD for fueling Ethiopia ethnic violence. https://www.amnesty.org/en/latest/news/2022/12/kenya-meta-sued-for-1-6-billion-usd-for-fueling-ethiopia-ethnic-violence/
[6] International Amnesty. 2023. Ethiopia: Eritrean soldiers committed war crimes and possible crimes against humanity after signing of agreement to end hostilities – new report. https://www.amnesty.org/en/latest/news/2023/09/eritrean-soldiers-committed-war-crimes-and-possible-crimes-against-humanity-in-the-tigray-region-after-signing-of-agreement-to-end-hostilities/
[7] Cara Anna. 2021. 'Horrible': Witnesses recall massacre in Ethiopian holy city. https://apnews.com/article/witnesses-recall-massacre-axum-ethiopia-fa1b531fea069aed6768409bd1d20bfa
[8] Kofi Arhin, Ioana Baldini, Dennis Wei, Karthikeyan Natesan Ramamurthy, and Moninder Singh. 2021. Ground-Truth, Whose Truth? – Examining the Challenges with Annotating Toxic Text Datasets. http://arxiv.org/abs/2112.03529 arXiv:2112.03529 [cs].
[9] Lora Aroyo and Chris Welty. 2013. Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. *WebSci2013. ACM* 2013, 2013 (2013).
[10] Lora Aroyo and Chris Welty. 2015. Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation. *AI Magazine* 36, 1 (March 2015), 15–24. https://doi.org/10.1609/aimag.v36i1.2564 Number: 1.
[11] Andrew Arsht and Daniel Etcovitch. 2018. The human cost of online content moderation. *Harvard Journal of Law and Technology* 2 (2018).
[12] Asmita Bhutani Vij. 2023. *Women Workers Behind the AI Revolution: The Production and Reproduction of Data Annotation Platforms.* Doctoral Theses. University of Toronto, Toronto, Canada. http://hdl.handle.net/1807/130603
[13] Virginia Braun and Victoria Clarke. 2012. *Thematic analysis.* American Psychological Association.
[14] Lisa Brooten. 2020. When media fuel the crisis: Fighting hate speech and communal violence in Myanmar. *Media, journalism and disaster communities* (2020), 215–230.
[15] Ellen Tveteraas Claire Wilmot. 2021. Dueling Information Campaigns: The War Over the Narrative in Tigray. https://mediamanipulation.org/case-studies/dueling-information-campaigns-war-over-narrative-tigray
[16] Gloria Cowan, Miriam Resendez, Elizabeth Marshall, and Ryan Quist. 2002. Hate Speech and Constitutional Protection: Priming Values of Equality and Freedom. *Journal of Social Issues* 58, 2 (Jan. 2002), 247–263. https://doi.org/10.1111/1540-4560.00259
[17] Donald Edward Crummey. 2024. Ethiopia - Rural, Urban, Highlands | Britannica. https://www.britannica.com/place/Ethiopia/Ethnic-groups-and-languages
[18] Caroline crystal. 2023. Facebook, Telegram, and the Ongoing Struggle Against Online Hate Speech. https://carnegieendowment.orgundefined?lang=en
[19] Aida Mostafazadeh Davani, Mohammad Atari, Brendan Kennedy, and Morteza Dehghani. 2023. Hate Speech Classifiers Learn Normative Social Stereotypes. *Transactions of the Association for Computational Linguistics* 11 (March 2023), 300–319. https://doi.org/10.1162/tacl_a_00550
[20] Aida Mostafazadeh Davani, Mark Díaz, Dylan Baker, and Vinodkumar Prabhakaran. 2024. D3CODE: Disentangling Disagreements in Data across Cultures on Offensiveness Detection and Evaluation. http://arxiv.org/abs/2404.10857 arXiv:2404.10857 [cs].
[21] Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations. *Transactions of the Association for Computational Linguistics* 10 (Jan. 2022), 92–110. https://doi.org/10.1162/tacl_a_00449
[22] Remi Denton, Mark Díaz, Ian Kivlichan, Vinodkumar Prabhakaran, and Rachel Rosen. 2021. Whose Ground Truth? Accounting for Individual and Collective Identities Underlying Dataset Annotation. http://arxiv.org/abs/2112.04554

arXiv:2112.04554 [cs].

[23] Thiago Dias Oliva, Dennys Marcelo Antonialli, and Alessandra Gomes. 2021. Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online. *Sexuality & Culture* 25, 2 (April 2021), 700–732. https://doi.org/10.1007/s12119-020-09790-w

[24] Mark Diaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. 2018. Addressing Age-Related Bias in Sentiment Analysis. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3173986

[25] Rebecca Dorn, Lee Kezar, Fred Morstatter, and Kristina Lerman. 2024. Harmful Speech Detection by Language Models Exhibits Gender-Queer Dialect Bias. In *Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (San Luis Potosi, Mexico) *(EAAMO '24)*. Association for Computing Machinery, New York, NY, USA, Article 6, 12 pages. https://doi.org/10.1145/3689904.3694704

[26] Think Thank European Parliament. 2022. Ethiopia: War in Tigray - Background and state of play | Think Tank | European Parliament. https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2022)739244

[27] Christina Fink. 2018. Dangerous speech, anti-Muslim violence, and Facebook in Myanmar. *Journal of International Affairs* 71, 1.5 (2018), 43–52.

[28] Eve Fleisig, Su Lin Blodgett, Dan Klein, and Zeerak Talat. 2024. The Perspectivist Paradigm Shift: Assumptions and Challenges of Capturing Human Labels. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 2279–2292. https://doi.org/10.18653/v1/2024.naacl-long.126

[29] Zecharias Zelalem Foundation, Thomson Reuters. 2022. FEATURE-Six million silenced: A two-year internet outage in Ethiopia. *Reuters* (Sept. 2022). https://www.reuters.com/article/markets/commodities/feature-six-million-silenced-a-two-year-internet-outage-in-ethiopia-idUSL8N2ZM09X/

[30] Fasica Berhane Gebrekidan. 2024. Content moderation: The harrowing, traumatizing job that left many African data workers with mental health issues and drug dependency. https://data-workers.org/fasica/ In: M. Miceli, A. Dinika, K. Kauffman, C. Salim Wagner, and L. Sachenbacher (eds.) The Data Workers' Inquiry.

[31] Spiros V. Georgakopoulos, Sotiris K. Tasoulis, Aristidis G. Vrahatis, and Vassilis P. Plagianakos. 2018. Convolutional Neural Networks for Toxic Comment Classification. In *Proceedings of the 10th Hellenic Conference on Artificial Intelligence*. 1–6. https://doi.org/10.1145/3200947.3208069 arXiv:1802.09957 [cs].

[32] Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. Yale University Press. Google-Books-ID: cOJgDwAAQBAJ.

[33] GlobalWitness. 2022. Facebook continues to approve hate speech inciting violence and genocide during civil war in Ethiopia. https:///en/campaigns/digital-threats/ethiopia-hate-speech/

[34] Mitchell L. Gordon, Michelle S. Lam, Joon Sung Park, Kayur Patel, Jeffrey T. Hancock, Tatsunori Hashimoto, and Michael S. Bernstein. 2022. Jury Learning: Integrating Dissenting Voices into Machine Learning Models. In *CHI Conference on Human Factors in Computing Systems*. 1–19. https://doi.org/10.1145/3491102.3502004 arXiv:2202.02950 [cs].

[35] James Grimmelmann. 2015. The Virtues of Moderation. *Cornell Law Faculty Publications* (April 2015). https://scholarship.law.cornell.edu/facpub/1486

[36] Million Haileselassie. 2023. Ethiopia's Tigray region ravaged by deadly famine – DW – 12/12/2023. https://www.dw.com/en/ethiopia-tigray-famine-crisis-residents-call-for-aid/a-67701012

[37] Danula Hettiachchi, Indigo Holcombe-James, Stephanie Livingstone, Anjalee De Silva, Matthew Lease, Flora D. Salim, and Mark Sanderson. 2023. How Crowd Worker Factors Influence Subjective Annotations: A Study of Tagging Misogynistic Hate Speech in Tweets. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 11, 1 (Nov. 2023), 38–50. https://doi.org/10.1609/hcomp.v11i1.27546

[38] Katharine Houreld. 2023. Raped during Ethiopia's war, survivors now rejected by their families. *Washington Post* (Nov. 2023). https://www.washingtonpost.com/world/2023/11/26/ethiopia-tigray-rape-survivors-stigma/

[39] Azeem Ibrahim. 2024. The Tigray War May Be One of the Deadliest Conflicts of This Century. https://nationalinterest.org/feature/tigray-war-may-be-one-deadliest-conflicts-century-211281 Publisher: The Center for the National Interest.

[40] Oana Inel, Khalid Khamkham, Tatiana Cristea, Anca Dumitrache, Arne Rutjes, Jelle van der Ploeg, Lukasz Romaszko, Lora Aroyo, and Robert-Jan Sips. 2014. CrowdTruth: Machine-Human Computation Framework for Harnessing Disagreement in Gathering Annotated Data. In *The Semantic Web – ISWC 2014*, Peter Mika, Tania Tudorache, Abraham Bernstein, Chris Welty, Craig Knoblock, Denny Vrandečić, Paul Groth, Natasha Noy, Krzysztof Janowicz, and Carole Goble (Eds.). Springer International Publishing, Cham, 486–504. https://doi.org/10.1007/978-3-319-11915-1_31

[41] Sanjay Kairam and Jeffrey Heer. 2016. Parting crowds: Characterizing divergent interpretations in crowdsourced annotation tasks. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 1637–1648.

[42] Shivani Kapania, Alex S Taylor, and Ding Wang. 2023. A hunt for the Snark: Annotator Diversity in Data Practices. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–15. https://doi.org/10.1145/3544548.3580645

[43] Klaus Krippendorff. 2011. Computing Krippendorff's alpha-reliability.

[44] Nayeon Lee, Chani Jung, and Alice Oh. 2023. Hate Speech Classifiers are Culturally Insensitive. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, Sunipa Dev, Vinodkumar Prabhakaran, David Adelani, Dirk Hovy, and Luciana Benotti (Eds.). Association for Computational Linguistics, Dubrovnik, Croatia, 35–46. https://doi.org/10.18653/v1/2023.c3nlp-1.5

[45] Tirrell Lynne. 2012. Genocidal Language Games1. In *Speech and Harm: Controversies Over Free Speech*, Ishani Maitra and Mary Kate McGowan (Eds.). Oxford University Press, 0. https://doi.org/10.1093/acprof:oso/9780199236282.003.0008

[46] Eliza Mackintosh. 2021. Ethiopia is at war with itself. Here's what you need to know. https://www.cnn.com/2021/11/03/africa/ethiopia-tigray-explainer-2-intl/index.html

[47] Milagros Miceli, Martin Schuessler, and Tianling Yang. 2020. Between Subjectivity and Imposition: Power Dynamics in Data Annotation for Computer Vision. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (Oct. 2020), 115:1–115:25. https://doi.org/10.1145/3415186

[48] Dan Milmo and Dan Milmo Global technology correspondent. 2021. Rohingya sue Facebook for £150bn over Myanmar genocide. *The Guardian* (Dec. 2021). https://www.theguardian.com/technology/2021/dec/06/rohingya-sue-facebook-myanmar-genocide-us-uk-legal-action-social-media-violence

[49] James Muldoon, Callum Cant, Mark Graham, and Funda Ustek Spilda. 2023. The poverty of ethical AI: impact sourcing and AI supply chains. *AI & SOCIETY* (Dec. 2023). https://doi.org/10.1007/s00146-023-01824-9

[50] James Muldoon, Mark Graham, and Callum Cant. 2024. *Feeding the Machine: The Hidden Human Labor Powering A.I.* Canongate Books, Edinburgh.

[51] Unicef Namibia. 2017. The impact of language policy and practice on children's learning: Evidence from Eastern and Southern Africa. *Country Review* (2017).

[52] United Nations. 2022. Statement by the UN Special Adviser on the Prevention of Genocide condemning the recent escalation of fighting in Ethiopia. https://www.globalr2p.org/resources/statement-by-the-un-special-adviser-on-the-prevention-of-genocide-condemning-the-recent-escalation-of-fighting-in-ethiopia/

[53] Scott Neuman. 2021. 9 Things To Know About The Unfolding Crisis In Ethiopia's Tigray Region. *NPR* (March 2021). https://www.npr.org/2021/03/05/973624991/9-things-to-know-about-the-unfolding-crisis-in-ethiopias-tigray-region

[54] Daniel Nkemelu, Harshil Shah, Michael Best, and Irfan Essa. 2023. Tackling Hate Speech in Low-resource Languages with Context Experts. In *Proceedings of the 2022 International Conference on Information and Communication Technologies and Development* (Seattle, WA, USA) *(ICTD '22)*. Association for Computing Machinery, New York, NY, USA, Article 5, 11 pages. https://doi.org/10.1145/3572334.3572372

[55] Daniel Nkemelu, Harshil Shah, Michael Best, and Irfan Essa. 2023. Tackling Hate Speech in Low-resource Languages with Context Experts. In *Proceedings of the 2022 International Conference on Information and Communication Technologies and Development* (Seattle, WA, USA) *(ICTD '22)*. Association for Computing Machinery, New York, NY, USA, Article 5, 11 pages. https://doi.org/10.1145/3572334.3572372

[56] Christina A. Pan, Sahil Yakhmi, Tara P. Iyer, Evan Strasnick, Amy X. Zhang, and Michael S. Bernstein. 2022. Comparing the Perceived Legitimacy of Content Moderation Processes: Contractors, Algorithms, Expert Panels, and Digital Juries. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW1 (April 2022), 82:1–82:31. https://doi.org/10.1145/3512929

[57] Christina A. Pan, Sahil Yakhmi, Tara P. Iyer, Evan Strasnick, Amy X. Zhang, and Michael S. Bernstein. 2022. Comparing the Perceived Legitimacy of Content Moderation Processes: Contractors, Algorithms, Expert Panels, and Digital Juries. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW1, Article 82 (April 2022), 31 pages. https://doi.org/10.1145/3512929

[58] Jiaxin Pei and David Jurgens. 2023. When Do Annotator Demographics Matter? Measuring the Influence of Annotator Demographics with the POPQUORN Dataset. http://arxiv.org/abs/2306.06826 arXiv:2306.06826 [cs].

[59] David Pilling. 2024. The young people sifting through the internet's worst horrors. *Financial Times* (Jan. 2024). https://www.ft.com/content/ef42e78f-e578-450b-9e43-36fbd1e20d01

[60] Barbara Plank. 2022. The "Problem" of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 10671–10682. https://doi.org/10.18653/v1/2022.emnlp-main.731

[61] Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Díaz. 2021. On Releasing Annotator-Level Labels and Information in Datasets. http://arxiv.org/abs/2110.05699 arXiv:2110.05699 [cs].
[62] Victoire Rio. 2020. The role of social media in fomenting violence: Myanmar. *Toda Peace Institute* (2020).
[63] rkremzner@newlinesinstitute.org. 2024. Genocide in Tigray: Serious breaches of international law in the Tigray conflict, Ethiopia, and paths to accountability. https://newlinesinstitute.org/rules-based-international-order/genocide-in-tigray-serious-breaches-of-international-law-in-the-tigray-conflict-ethiopia-and-paths-to-accountability-2/
[64] Sarah T Roberts. 2014. *Behind the screen: The hidden digital labor of commercial content moderation.* University of Illinois at Urbana-Champaign.
[65] Claire Sanford. 2021. Facebook Whistleblower Frances Haugen Testifies on Children & Social Media Use: Full Senate Hearing Transcript. https://www.rev.com/blog/transcripts/facebook-whistleblower-frances-haugen-testifies-on-children-social-media-use-full-senate-hearing-transcript
[66] Yisi Sang and Jeffrey Stanton. 2022. The Origin and Value of Disagreement Among Data Labelers: A Case Study of Individual Differences in Hate Speech Annotation. In *Information for a Better World: Shaping the Global Future*, Malte Smits (Ed.). Springer International Publishing, Cham, 425–444. https://doi.org/10.1007/978-3-030-96957-8_36
[67] Sebastin Santy, Jenny T Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. 2023. NLPositionality: Characterizing design biases of datasets and models. *arXiv preprint arXiv:2306.01943* (2023).
[68] Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection. http://arxiv.org/abs/2111.07997 arXiv:2111.07997 [cs].
[69] Mike Schaekermann, Joslin Goh, Kate Larson, and Edith Law. 2018. Resolvable vs. Irresolvable Disagreement: A Study on Worker Deliberation in Crowd Work. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (Nov. 2018), 1–19. https://doi.org/10.1145/3274423
[70] Liam Scott. 2021. How Social Media Became a Battleground in the Tigray Conflict. https://www.voanews.com/a/how-social-media-became-a-battleground-in-the-tigray-conflict-/6272834.html
[71] Farhana Shahid and Aditya Vashistha. 2023. Decolonizing Content Moderation: Does Uniform Global Community Standard Resemble Utopian Equality or Western Power Hegemony?. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 391, 18 pages. https://doi.org/10.1145/3544548.3581538
[72] Clemencia Siro, Mohammad Aliannejadi, and Maarten Rijke. 2024. Context Does Matter: Implications for Crowdsourced Evaluation Labels in Task-Oriented Dialogue Systems. In *Findings of the Association for Computational Linguistics: NAACL 2024*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 1258–1273. https://doi.org/10.18653/v1/2024.findings-naacl.80
[73] Ruth Spence, Antonia Bifulco, Paula Bradbury, Elena Martellozzo, and Jeffrey DeMarco. 2023. The psychological impacts of content moderation on content moderators: A qualitative study. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace* 17, 4 (2023).
[74] Ruth Spence, Antonia Bifulco, Paula Bradbury, Elena Martellozzo, and Jeffrey DeMarco. 2024. Content moderator mental health, secondary trauma, and well-being: A cross-sectional study. *Cyberpsychology, Behavior, and Social Networking* 27, 2 (2024), 149–155.
[75] Steve Stecklow. 2018. Why Facebook is losing the war on hate speech in Myanmar. *Reuters* (Aug. 2018). https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/
[76] Miriah Steiger, Timir J Bharucha, Sukrit Venkatagiri, Martin J. Riedl, and Matthew Lease. 2021. The Psychological Well-Being of Content Moderators: The Emotional Labor of Commercial Moderation and Avenues for Improving Support. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 341, 14 pages. https://doi.org/10.1145/3411764.3445092
[77] Elizabeth Stewart. 2021. Detecting Fake News: Two Problems for Content Moderation. *Philosophy & Technology* 34, 4 (Dec. 2021), 923–940. https://doi.org/10.1007/s13347-021-00442-x
[78] Kirubel Tadesse. [n. d.]. Overselling AI: Facebook's Content Moderation Issues in Ethiopia | Annenberg. https://www.asc.upenn.edu/research/centers/milton-wolf-seminar-media-and-diplomacy-5
[79] Moges Teshome. [n. d.]. "The Road to Hell is Paved with Good Intentions": the Role of Facebook in Fuelling Ethnic Violence | Annenberg. https://www.asc.upenn.edu/research/centers/milton-wolf-seminar-media-and-diplomacy/blog/road-hell-paved-good-intentions-role-facebook-fuelling-ethnic-violence
[80] Kristen Vaccaro, Ziang Xiao, Kevin Hamilton, and Karrie Karahalios. 2021. Contestability For Content Moderation. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2 (Oct. 2021), 318:1–318:28. https://doi.org/10.1145/3476059
[81] Betty van Aken, Julian Risch, Ralf Krestel, and Alexander Löser. 2018. Challenges for Toxic Comment Classification: An In-Depth Error Analysis. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, Darja Fišer, Ruihong Huang, Vinodkumar Prabhakaran, Rob Voigt, Zeerak Waseem, and Jacqueline Wernimont (Eds.). Association for Computational Linguistics, Brussels, Belgium, 33–42. https://doi.org/10.18653/v1/W18-5105
[82] David Volodzko. 2022. There's Genocide in Tigray, but Nobody's Talking About it. (May 2022). https://www.thenation.com/article/world/genocide-in-tigray/
[83] Alex de Waal and Mulugeta Gebrehiwot Berhe. 2024. Ethiopia Back on the Brink. *Foreign Affairs* (April 2024). https://www.foreignaffairs.com/ethiopia/ethiopia-back-brink
[84] Zeerak Waseem. 2016. Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, David Bamman, A. Seza Doğruöz, Jacob Eisenstein, Dirk Hovy, David Jurgens, Brendan O'Connor, Alice Oh, Oren Tsur, and Svitlana Volkova (Eds.). Association for Computational Linguistics, Austin, Texas, 138–142. https://doi.org/10.18653/v1/W16-5618
[85] Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, Jacob Andreas, Eunsol Choi, and Angeliki Lazaridou (Eds.). Association for Computational Linguistics, San Diego, California, 88–93. https://doi.org/10.18653/v1/N16-2013
[86] Human Rights Watch. 2021. Ethiopia: Eritrean Forces Massacre Tigray Civilians | Human Rights Watch. https://www.hrw.org/news/2021/03/05/ethiopia-eritrean-forces-massacre-tigray-civilians
[87] Human Rights Watch. 2022. "We Will Erase You from This Land". *Human Rights Watch* (April 2022). https://www.hrw.org/report/2022/04/06/we-will-erase-you-land/crimes-against-humanity-and-ethnic-cleansing-ethiopias
[88] Dylan Weber. 2024. Scaling Expertise: A Note on Homophily in Online Discourse and Content Moderation. *New England Journal of Public Policy* 36, 1 (June 2024). https://scholarworks.umb.edu/nejpp/vol36/iss1/13
[89] WHO. 2022. Crisis in Northern Ethiopia. https://www.who.int/emergencies/situations/crisis-in-tigray-ethiopia
[90] Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating Expressions of Opinions and Emotions in Language. *Language Resources and Evaluation* 39, 2 (May 2005), 165–210. https://doi.org/10.1007/s10579-005-7880-9
[91] Richard Ashby Wilson and Molly K. Land. 2020. Hate Speech on Social Media: Content Moderation in Context. https://papers.ssrn.com/abstract=3690616
[92] Di Wu. 2023. Good for tech: Disability expertise and labor in China's artificial intelligence sector. *First Monday* (Jan. 2023). https://doi.org/10.5210/fm.v28i1.12887
[93] Daniel Zaleznik. 2021. Facebook and Genocide: How Facebook contributed to genocide in Myanmar and why it will not be held accountable. https://systemicjustice.org/article/facebook-and-genocide-how-facebook-contributed-to-genocide-in-myanmar-and-why-it-will-not-be-held-accountable/
[94] Zecharias Zelalem. 2021. Why Facebook keeps failing in Ethiopia. https://restofworld.org/2021/why-facebook-keeps-failing-in-ethiopia/
[95] Xinyi Zhou and Reza Zafarani. 2021. A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. *Comput. Surveys* 53, 5 (Sept. 2021), 1–40. https://doi.org/10.1145/3395046 arXiv:1812.00315 [cs].

A Appendix

A.1 Data Annotation Details

Table 3 provides details on the class distribution of the annotated data, including the percentage representation of each class and the annotation disagreement among the annotators when labeling posts into the class. Since each post can be assigned multiple classes, the percentages do not sum up to 100%. Notably, 42.3% of the dataset is classified as neutral, and this class also generates the highest level of disagreement (34.7%) among annotators.

A.2 Examples of Annotated Posts

Trigger warning: Table 5 presents examples of user posts that contain potentially offensive and upsetting terms. we present some examples of posts from the annotation study. We randomly selected posts from different categories. We include the original posts, their translations to English, the categories they were classified into, and provide the context and reasoning for their

categorization. Some of the posts are slightly paraphrased to protect the anonymity of the posters.

Class	Distribution	Disagreement in labeling the class
Neutral	42.3%	34.7%
Abuse and Harassment	38.8%	31.5%
Dehumanization	14.4%	20.8%
Violent Speech	8.2%	9.7%
Violent Event Denial	8.5%	4.4%
Verified Misinformation	2.9%	8.2%
Suspected Misinformation	1.8%	4.1%

Table 3: Table outlining the class distribution and levels of disagreement in labeling posts into the class.

A.3 Interview Participant Details

Pseudonym	Market/Language
Hermela	Afaan Oromo, English
Belay	Tigrinya, Amharic, English
Jacob	English
Elene	Amharic, Afaan Oromo, English
Henry	Swahili, Kirundi, English
Sabrin	Somali, English
Teka	Amharic, Afaan Oromo, English
Fiona	Amharic, English
Aya	Somali, English
Maya	Hausa, Pidgin, English
Tarik	All Quality Support
Mary	Swahili, English
Ezana	Amharic, Tigrinya, English
Ernest	Afaan Oromo, English
Adam	Amharic, Tigrinya, English

Table 4: Interview participants' pseudonym and market/region. N=15. Gender (8 = male, 7 = female)

Original post	Translated to English	Category	Context
ትርፍራፊውን ማጽዳት ግን ይቀጥል። አንዷ ያል-ተነቀለች አረም ናት በኋላ የምትራባው።	But let the cleanup continue. It's one weed that has not been pulled out that will grow later.	"Violent", "dehumanization"	Weed in this case is referring to the Tigrayan people or sometimes to the TPLF political party. It was posted during the conflict. Regardless of the ambiguity in its reference, the post is 'violent" as it calls for cleaning up the people and 'dehumanizing" for referring to human beings as weeds.
ወዲዞም ርሹናት ዓጋመ፡ ረሳሕ ዓሌት ታሪኽ ኣልቦ	Son of these dead Agame, dirty race without history	"Dehumanization", "Abuse"	Agame, is a word used to refer to a specific cultural group from the Eastern Zone of Tigray, particularly Adigrat. It is insulting the people in this cultural group as a whole by calling them a 'dirty race" and 'without history".
ኣረሙን ነቅለን ማስወገድ የግድ ንው። የ ኢት-ዮጲያ ህዝብ ሁሉ መተባበር ኣለበት። Posted: March, 2020	It is necessary to remove the weeds. All the people of Ethiopia must cooperate.	"neutral"	The context and time matters. This post was made in March 2020. This was before the war started and there was another weed problem in the Tana river of Amhara region. So, the annotator needs to understand and take into account these contexts.
#Eritrea #Ethiopia #HOA appreciate @[anony]@[anony] rebuffing Z @[anony] called @[anony] meeting. #TPLF spewed #FakeAxumMassacre needs investigation. But, lies must n't be basis to harass sovereign nations. #TPLFstartedTheWar #EritreaPrevails #StopScapegoatingEritrea		"Violent Event Denial"	The post mentions that the Axum Massacre was fake and needs investigation. However, the massacre has been confirmed and well documented by an Amnesty International report where Eritrean fighting forces systematically killed hundreds of unarmed civilians in the northern city of Axum [4]. Hence denying the tragic massacre.
ሰማዕ እንዶ ኢድካን ሂውተካን ሓውስኩ ቆልማጽ ረሳሕ ዓጋመ። ለመጭ ዓጋመ።	Listen, put your hands and your life in it, you dirty bastard Agame.	"Abuse"	Abusing a person based on their cultural group (Agame). This is a common insult used against Tigrayans to attack them based on their cultural identity.
ትግሬ የሚባል መጥፎ አረም ማጥፋት ብቸኛው መፍትሄ ነው!!!	Wiping out the bad weed known as Tigre is the only solution!!	"Violent", "dehumanization"	Referring to Tigrayans as weeds which is dehumanizing and calling for them to be wiped out which is violent.

Table 5: Example posts from the annotation study, along with the context and reasoning used to classify them into specific categories.

A.4 Additional Data from Findings

Themes	Selected Quotes
4.1.1 Specific Knowledge of Dialects	Jacob: "There are 11 dialects in the Oromo language. (...) We have a person from each dialect here in our team of like 30 people. The Oromo team here comprises like 30 people, but the market specialist is on his own, one person in Dublin. So if this person from this dialect says, this is violating in our dialect (...) Most of the times he sticks with what he knows. and it is hard to convince him that in another dialect this word is a contradiction, so it has to be removed. And he says, don't remove that word."
4.1.2 In-Depth Knowledge of Cultural Contexts	Hermela:"Some content that we know culturally, (...) it's social values of where we come from. We actually know it's violating. But it's not covered in policies."
4.1.4 Familiarity with Specific Networks on Social Media Platforms	Fiona: "they [are] just social media goons, or something like that. (...) they're familiar, familiar people on the social media, especially on Twitter."
4.1.5 Rigorous Understanding of Policy	Mary: "You have to follow the policy. Whether it's what in your region, what happened, it doesn't matter. But you follow the policy, what they told you, what they trained you, that's what your action, when you see any content is fitted, violating." Ezana: "Even if your enemy is being attacked, and you want that [content] to stay in the [InstantChat] platform, you have to follow policy, you have to delete." Jacob: "If (...) you are an activist against the government. The government has been brutalizing people. (...) You've been criticizing them. You come into content moderation from that kind of background. (...) It will be very difficult for you to separate that profession of activism from now dealing with the policy because these government soldiers are also being attacked. The government security forces are being slaughtered (...) How will you deal with that situation?"
4.1.6 Mental Resilience	Adam: "the resilience is the first skill required to moderate content. Because it's not easy for not everyone can or has the heart to look at all the content and [moderate] in difficulties." Elene: "I think also being emotionally strong. (...) So you have to be very strong to understand that this is my job. (...) You have to have a way of making that thing not get to you because if for example, someone was bullying someone for like being sick [on social media content]. For example, someone has cancer, or this person is being bullied, and you're there thinking about your family member who has cancer, and you imagine whatever bad words that I've been told on that person, you imagine this could be my sick aunt, or this could be my sick grandmother. You have to be able to differentiate. Let me just not think about this for now. Let me just work on this. This is not directed to my people. This is just a bad person who decided to bully other people and doesn't. So you have to be resilient, strong."
4.1.7 Concentration and Patience	Teka: "So being very attentive when you are reading especially in our language." Interviewer: "what kind of like skills or knowledge or quality that you find important to do content moderation? (...)" Belay: "(...) Secondly, you need to focus."
4.1.8 Effective Teamwork	Adam: "(...) ability to work in a team setting (...) some of the content you can do alone, you need like a expertise (...) most of the content moderation, it comes in a team and not as an individual." Henry: "it's an open office. And so you can chat with someone who is next to you. (...) How did you action this? Does it go for dangerous organization? Does it go for sexual activity? you will have a discussion. And it comes when someone is not sure of what is the [right moderation]. So there is a group, which is Oromo group, you posted the ticket there. And someone will go and open it. And someone is writing very sure that this ticket goes for this. And that's how we manage to help each other. So every market has its own group." Belay: "Mostly for the Hausa, we have the more of unity among us, we always kind of do more of like team bonding with meets, it makes people feel at home and try to enlighten you even not at work, even at home, we can decide to come home teach you more about how the policies work, because, definitely, when you [as a newcomer]'re in the training room, all things have been rushed, you just be rushed. So, when you come in, we try to offer you (...) by assisting you (...) that's the way we always try to come together, maybe be it at work and at home, when you don't understand things, we come, we help you. This is how we Hausa team always do it in order to keep up with our team performance."
4.2.1 Cultural Variability	Teka: "I feel like they [the policies] try to generalize, like every culture, every language in one bucket, that's the big problem.'

4.2.2 Differing Political Stances	Aya: "Others feel attached, they take things personal. That was the major issue. Like, if you see a ticket that maybe is attacking a certain tribe or ethnicity, and there's someone [moderators] from that ethnicity, you'll see someone fighting, oh, you guys are like this. (...) At some point, the quality analysts were also taking things pass on. (...) You see this ethnicity, religion, and such things (...) It's very partial when it comes to religion, so people could even go weeks without talking to each other."
4.2.4 Posts missing context	Aya: "You are supposed to take down the content using, like, 40 seconds per the content (...) you can't afford to stay too long on one ticket. (...) the ticket will disappear, and you'll be called by your supervisor: Why haven't you acted on this ticket? You stayed with this ticket for two minutes." [This quote this connected to the finding that workers are under strict time limits and have to make decisions on posts with missing contexts or posts they didn't have time to read/watch completely.] Sabrin: "[For] some some posts, you would need to get a little bit of context, maybe more details, especially where there's no picture of video. It's just someone writing something (...) [and] it can be a bit tricky. (...) It could be a song, or it could be a poem."
4.2.5 Unclear and Out-of-Date Moderation Policies	Fiona: "The policies were made a long time ago, like two years before we joined. (...) so the policies were already there, because there were workers from the Amharic team who were working starting from 2019. So we only joined in 2021. So the policies were all there already. They were made. So like, we couldn't do anything. Even after we flagged them, (...) they wouldn't accept." Hermela: "You know, it's a lot of contents with different kind of violations. And then every day you can see, you can face, (...) you encounter a new kind of violation that's not covered in the police."
4.3.1 Language Skills	Fiona: "the skills they were looking for [are] language, [the] ability to understand and speak the language to Tigrinya or Amharic." Ezana: "After that we will go for the language interview (...) just to confirm if you are a native speaker of that language."
4.3.2 Superficial Cultural Awareness	Jacob: "So now the quality guy is the one who calls you, and he has a set of quiz that you have to answer. Do you [have] the market knowledge? Do you [know] the current affairs in Ethiopia, [and] current affairs in the Oromo region? (...) Which singers sing about songs of violence? (...) So he's just testing if you are aware of what is going on right now in the Oromo region." Mary: "Who is the president, who's the minister, who's the public figures, or who's the artists and all those things, of course, you should have also some knowledge somehow Yes. You [have] to put [it] under the policy to action. Sometimes now if it is a public figure and even if he's been bullied or something, someone is saying something about him, it's not violated, so you ignore it. If you don't know [that this person is a public figure], you actually need for delete."
4.3.3 Robust Understanding of Platform Policy	Aya: "The most hard part was that this policies could change every week. (...) Sometimes even in three days, the policy is completely changed, and you are no longer going to take the actions you used to take before. (...) Sometimes when there's a major change, they call for emergency training for like two hours or an hour, and then you go back to work, and they expect you to now change completely according to how the policy has changed." Tarik: "We had to make sure that you understand it is not what you think, not what you say, it's not how you feel about what we are moderating. It's about what the policy says. So whether you see it violating or not, if a policy doesn't cover it, or it doesn't follow the policy indicators, you will leave it on their platform."
4.3.4 Mental Strength	Sabrin: "The second interview was psychometric test something (...) and then (...) one of their counsellors from their wellness team called me. And we just had a sort of like a light conversation, asking me if I'm comfortable to relocate to Kenya, am I okay with it, and all those things. And yeah, after those 3 calls. Then we're done." Henry: "while in the recruiting recruitment process, they told me it's something related with translating my language into the English language, or the English language to my language. (...) as I was going through the training session, I just realized, this is a far different job. A far different task from interpreting and translating." Fiona: "let me speak for myself. I was not prepared to see graphic contents, or I wasn't prepared to watch so many insults. And bad words that consume them everyday, everyday, everyday. (...) I never knew the job was the content moderation was to do this. I just thought, Okay, it's just InstantChat, it could be a cool job and stuff until they realized it. (...) The next day that we finished training, and then you'd be like, boom, there's very disturbing content."

4.4.1 Rigid Organizational Hierarchies	Tarik: "A lot of disagreements [among quality analysts]. You know, we all want to be right. And we want to think that we know the policy better. (..) But (...) we couldn't make it noticeable for the moderators. They should not know that we debate and fight in that room." [This quote this connected to the finding that quality analysts can disagree amongst themselves, but these disputes are concealed from content moderators.] Mary: "at [the] end of the day we [moderators] have to follow the experts. whatever they say we have to follow." Teka: "like the [market] specialist or the SME said that this should be marked as this word, you just agree with it, because even if you don't feel like it's right, or even if you have opinion, it does not matter. Because at the end of the day, we are just there to make money, and no one wants to lose their money. So we just agree with everyone."
4.4.2 Exploitative working conditions	Aya: "40 seconds you need to have taken down everything. So if you haven't mastered the policy, again there's something called AHT, the average time you take on one content, yeah, which was crazy." Fiona: "But InstantChat was, I think hiring vulnerable people. Like me, I'm a refugee or immigrant. (...) They needed vulnerable people in general. It's not majority, but most of the content moderators especially from South Africa or Uganda." Jacob: "He [market specialist] says don't remove that word (...). if you go against him. (...) We have a percentage like your accuracy. (...) We have quality score, and this quality score you have to keep a quality of 90 and above. But if you keep deleting what the guy said don't delete, your quality will go down, and you'll face punishment here and there."
4.4.3 Inability to Influence Platform Policies	Aya: "Sometimes, to get to the policy team is quite a hassle. Because by the time you raise your issue with the quality analyst, by the time they take it, sometimes they even just don't raise your issue to the manager or to the quality team. So they just leave it hanging, you keep on following them, oh what did you guys say about this ticket you guys marked me down? They'll be like, ah, you know, I asked the manager, but the manager hasn't gotten back to me. He said, we are waiting for the quality team. It will take weeks. You might even never get back the response of a certain thing until (...) maybe a whole majority of content moderators complain." Maya: "To be honest, sometimes you just had to do it because the specialist who is saying that and if you keep arguing with them, [they] may be going to report you to the manager that this person is maybe hard headed. So sometimes you had to just agree with a specialist not because they're right, just because they're kind of your boss." Jacob: "Most of the times [the market specialist] disagrees with us, and when he disagrees with us we cannot do anything to push the the policy to be changed or updated. So it remains at that. And if actually, you meet a market specialist who doesn't want to be corrected, who doesn't want to to learn new things (...) you would be in trouble."

Table 6: Additional data from the interviews supporting the findings.