Centre for Information Resilience

Normalised and invisible:

An analysis of gendered hate speech on social media in Ethiopia

The Centre for Information Resilience

February 2024

1. Executive summary

As people's personal and public lives are increasingly played out on the internet and through social media, a new frontier in the fight against gender-based violence has emerged. While the internet serves as a conduit for information dissemination, social connection, and the facilitation of activism and political mobilisation, it concurrently serves as a platform for the perpetuation of technology-facilitated gender-based violence (TFGBV) and discrimination.

The Centre for Information Resilience's (CIR) research into TFGBV in Ethiopia, roundtables and workshops in Addis Ababa, and this study into gendered hate speech, all signal that women and girls in Ethiopia are targeted with several different types of TFGBV, including hate speech and harassment.

TFGBV does not occur in a vacuum, but rather is rooted in and reinforces historical, religious, political, and cultural attitudes (figure 1). Combatting TFGBV is essential to protecting women and girls online and empowering their safe and meaningful participation in all forms of public life. This project aims to:

- Strengthen the evidence base on TFGBV in Ethiopia.
- Better inform government institutions, civil society organisations (CSOs), social media companies, and the general public about TFGBV in Ethiopia.
- Empower CSOs and government institutions in Ethiopia with practical recommendations on how to address TFGBV.

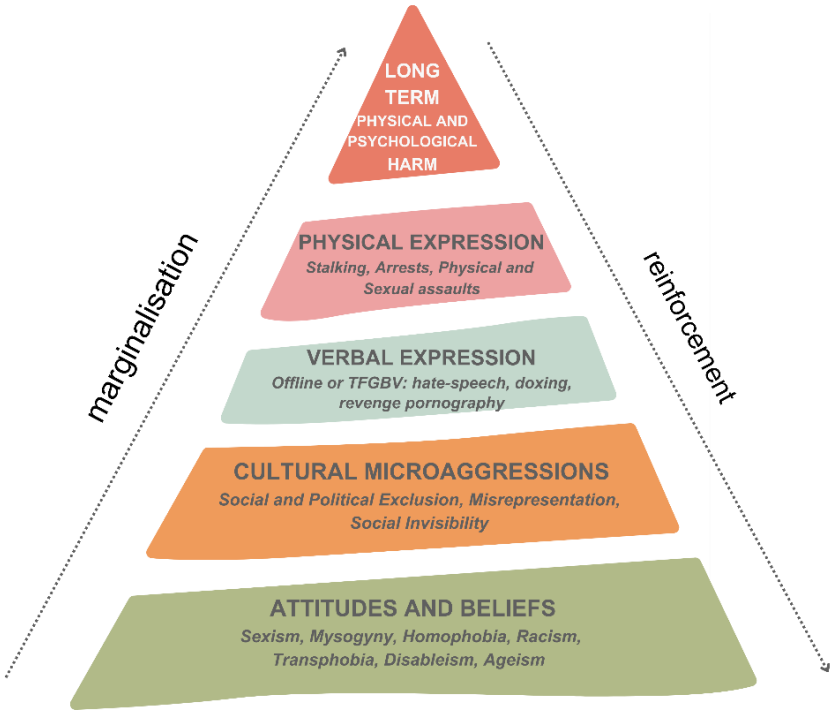


Figure 1: Diagram to demonstrate the cycle of marginalisation within which TFGBV contributes.

Key findings

This report both complements and builds on CIR's earlier study 'Silenced, shamed and threatened: TFGBV targeting women who participate in Ethiopian public life' by conducting a quantitative study of gendered hate speech (one form of TFGBV) in Ethiopia on three social media platforms (Facebook, Telegram, and X).[1]

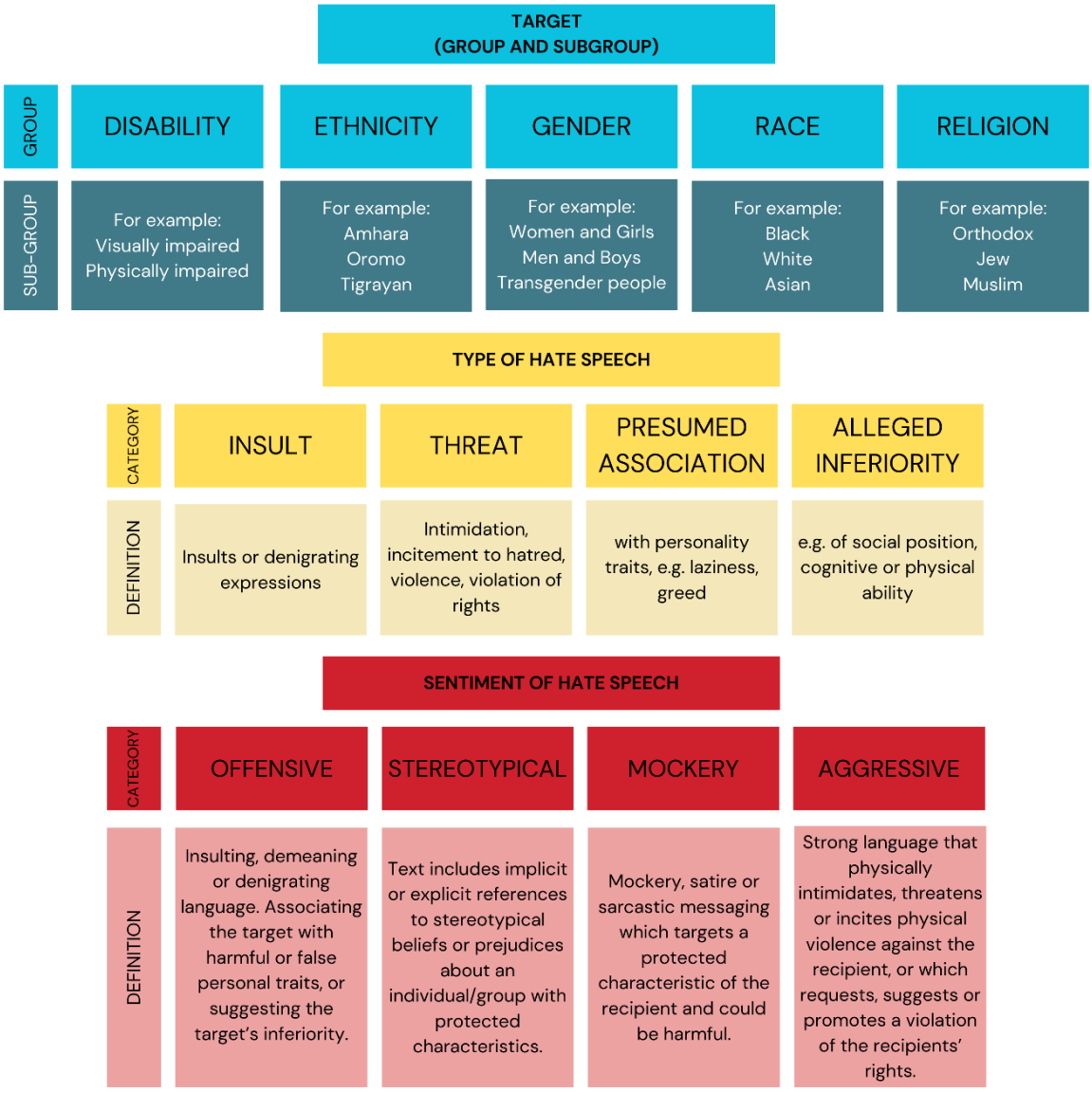


Figure 2: CIR's Conceptual Framework. This underpins the research design and annotation schema, as found in the Annotation Protocol in section 7.2.

[1] The methodology and conceptual framework developed for this research will feature in the "Resources for African Indigenous Languages" academic publication, see: 'Resources for African Indigenous Languages' (2024) Available at: https://sadilar.org/en/the-fifth-workshop-on-resources-for-african-indigenous-languages-rail/

Through workshops and roundtables in Addis Ababa, CIR developed a comprehensive lexicon of 2,058 inflammatory keywords to guide the collection of hate speech data across four languages (Amharic, Afaan Oromo, Tigrigna, and English). CIR identified 327,343,548 social media posts containing at least one inflammatory keyword, of which 7005 were manually annotated according to three dimensions: hate speech target, type, and sentiment (as shown in the conceptual framework above).

Both this and CIR's earlier study found that women and girls receive different types of online abuse than men and boys do. Interviewees in the first study reported that women and girls often face stereotypical abuse, centred around gender roles, and laced with misogyny, while abuse against men and boys often focuses on expressed views or politics. The social media analysis in this study found that women and girls receive more hate speech which includes gendered stereotypes and mockery or irony, than men and boys.[2] Additionally, this study found that the risks associated with being female online can be compounded when other protected characteristics are also targeted, including ethnicity and religion.

Discussions during roundtables and workshops revealed that hate speech is often misunderstood, leading to certain forms of TFGBV being overlooked. For example, certain forms of hate speech, such as the use of gendered stereotypes, insults, demeaning language, or speech implying women and girls' inferiority to other genders, are often considered less harmful than hate speech that is threatening or aggressive. In many conversations, these forms of speech were not considered to be hate speech at all. Better education on what constitutes hate speech, and its impact, is therefore essential in Ethiopia.

Many of the interviewees in CIR's earlier study reported that they saw TFGBV on all social media platforms that they engaged with. While this study found similarities in the sentiment of the hate speech targeting women and girls, there were variations in the types of hate speech seen across the three platforms. For example, proportionally, more insults were found on X than the other platforms, while Telegram had more threats. Facebook had more instances of hate speech which associated gender with certain personality traits (e.g. greed) or suggested the inferiority of women and girls' social position, or cognitive or physical ability due to their gender. Understanding platform variations can inform more targeted solutions, as well as more tailored resources for women and girls to protect themselves in the meantime. One approach might not suit all platforms, or its users.

This study also found that hate speech targeting women and girls differs from hate speech directed against other identity groups. Women and girls were more likely than ethnic or religious hate targets to receive abuse which suggests their inferiority, contains gendered stereotypes, or irony and mockery, and less likely to receive aggressive hate speech. Additionally, the findings from this study underscore that current events offline impact online debate and hate speech. This can be seen by the relatively high prevalence of intersectional abuse targeting women and girls of Amhara and Oromo ethnicities, compared to other ethnicities, in the context of active conflict in these areas of Ethiopia during the data collection time frame. Hate speech that is reactive to

[2] The different categories are defined within the conceptual framework.

political events and uses inflammatory rhetoric may be more apparent than gendered hate speech. Gendered abuse, in the form of stereotypes and the suggestion of inferiority, appears to almost go under the radar. Workshop participants expressed a belief that gendered abuse is so endemic that it has become normalised to the point of invisibility.

This report seeks to highlight forms of gendered hate speech on social media platforms, that actively contribute to the further marginalisation of women and girls in Ethiopia. In CIR's earlier study, Ethiopian women interviewed reported that the online abuse they faced left them feeling silenced, with many withdrawing from online and offline public spaces as a result. Cultivating safe online environments for women and girls is essential to empowering their full and meaningful participation in public life. To have a lasting effect, any strategies to combat TFGBV must address its root causes. This includes countering gender stereotypes and gender-based discrimination and promoting women and girls' representation in all public spaces.

Recommendations

Research into the nature of TFGBV - including who is targeted, where, and how - is essential to crafting effective policy solutions tailored to the specific context. To accompany this report, CIR has worked with stakeholders in Ethiopia to create a policy and community-led recommendations whitepaper.

This study reveals that hate speech differs depending on the target. Understanding these differences provides an entry point for targeted policy solutions to better safeguard women and girls online. CIR hopes that government institutions can use the findings to inform decision making, that social media companies can use them to inform their content moderation efforts, that civil society can use them in their advocacy, and that the public can use them to call for action.

Table of Contents

Acronyms

Centre for Information Resilience	CIR
Civil Society Organisations	CSOs
Gender-Based Violence	GBV
Natural Language Processing	NLP
Technology-Facilitated Gender-Based Violence	TFGBV

2. Introduction

Internet access in Ethiopia presents a paradoxical phenomenon. While it serves as a conduit for information dissemination, social connection, and the facilitation of activism and political mobilisation, it concurrently serves as a platform for the perpetuation of technology-facilitated gender-based violence (TFGBV) and discrimination.

There is no doubt that women, girls, men, and boys in Ethiopia are all victims of online abuse.[3] CIR's first study on this topic in 2023, 'Silenced, Shamed and threatened: the online abuse of women and girls who participate in Ethiopian public life',[4] researched the lived experiences of survivors and lasting impacts of online abuse through a review of existing literature and interviews with 14 women who hold prominent positions in Ethiopian public life, including in the media and civil society. The interviewees reported that there are differences in the sentiment, purpose, and impact of that abuse depending on the gender of the target. The findings revealed the toxicity of online environments in Ethiopia and how online abuse directed against women and girls reflects existing societal divisions around the role of women and girls in society, as well as in relation to ethnicity, politics, and religion.

While TFGBV covers a range of harmful behaviours, in this study, CIR second study narrowed the research data pool to focus solely on hate speech found on social media, through a quantitative analysis of social media posts drawn from X (formerly Twitter), Telegram, and Facebook. A comprehensive lexicon of inflammatory keywords was developed in four languages (Amharic, Afaan Oromo, Tigrigna, and English) to guide data collection. Posts were sampled and subsequently annotated according to three dimensions: hate speech target, type, and sentiment.[5] Furthermore, this study has expanded the types of hate speech being investigated by examining hate speech targeting the protected characteristics found in the Ethiopian Government's Hate Speech and Disinformation Proclamation (disability, ethnicity, gender, race, and religion[6]), not just gender. This addresses the need to investigate TFGBV along intersectional lines, as signalled by CIR's first study into online abuse in Ethiopia.

During roundtables and workshops in Addis Ababa conducted throughout this study, a shared belief was expressed among participants that women and girls in Ethiopia face high levels of TFGBV, including gendered hate speech. Many participants cited the lack of data as a key issue preventing TFGBV from being addressed. This study aims to help fill this data gap regarding the types and sentiment of hate speech that women and girls face on three prominent social media platforms in Ethiopia (X, Telegram, and Facebook).

[3] CIR (2023) Silenced, shamed and threatened: the online abuse of women and girls who participate in Ethiopian public life. Available on the CIR website.
[4] Ibid.
[5] Hate speech type and sentiment are defined in the Annotation Protocol, section 7.1, and explained in subsequent sections.
[6] Proclamation No.1185/2020, As seen in Federal Negarit Gazette (2020) Available at: https://ethionab.org/wp-content/uploads/2022/09/1185_2020_HATE_SPEECH_AND_DISINFORMATION_PREVENTION_AND_SUPPRESSION_PROCLAMATION_.pdf

Why research TFGBV?

TFGBV represents a distinct and harmful form of gender-based violence that actively exacerbates the marginalisation of women and girls, both within the digital and physical space. This can lead to long-term psychological and physical harm to survivors.

TFGBV can create an atmosphere of fear and intimidation. If women and girls do not feel safe online, they may be deterred from fully and meaningfully participating in public life, both online and offline. This exclusion can have devastating impacts, leading to less representative public spaces and democratic processes.

TFGBV reinforces cycles of marginalisation. It is a product of attitudes and beliefs, including sexism and misogyny, and can be rooted in historical, religious, political, and cultural attitudes. These attitudes and beliefs provide the foundation for cultural microaggressions: actions or speech that signal indirect, unconscious, or unintentional prejudice or discrimination towards marginalised groups. These more subtle forms of marginalisation often pave the way for verbal expression, including insults, harassment, and threats, before escalating to physical expression in the form of stalking, arrests, or even physical and sexual assaults. Unsurprisingly, once it has reached this stage, there is a high risk of long-term physical and psychological harm. Finally, silencing and ostracising women and girls from public spaces hinders informed decision-making. The absence of diverse perspectives may even undermine the legitimacy of governing bodies.

Gendered online abuse can have a lasting impact on the women and girls targeted, and societies at large. In an earlier study of gendered online abuse in Ethiopia, CIR interviewed 14 women and girls active in public life who had experienced gendered abuse. Many of the interviewees reported that the abuse they received online silenced them, leading them to withdraw from public spaces, both online and offline.

Combatting TFGBV is essential to protecting women and girls online and ensuring their safe and meaningful participation in all forms of public life. As outlined above, TFGBV does not occur in a vacuum, but rather is rooted in and reinforces historical, religious, political, or cultural attitudes. Research into the specific forms and locations of TFGBV is therefore essential to crafting effective policy solutions that are tailored to the specific contexts in which TFGBV occurs.

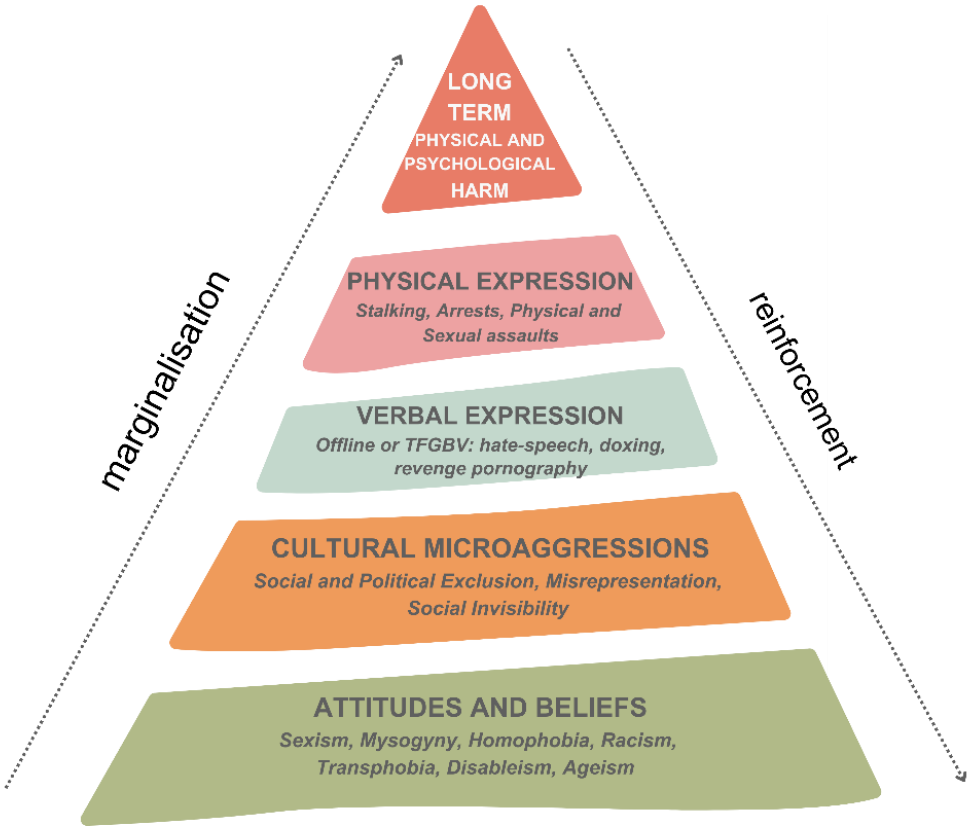


Figure 3: Diagram to demonstrate the cycle of marginalisation within which TFGBV contributes.

Project goals

- To strengthen the evidence base on TFGBV in Ethiopia.
- To inform government institutions, civil society organisations (CSOs), social media companies, and the general public about TFGBV in Ethiopia.
- To empower CSOs and government institutions in Ethiopia with practical recommendations on how to address TFGBV, including policies and community-led initiatives.

Research questions

- **Forms (type and sentiment):** What types of gendered hate speech are prevalent? What is the sentiment of this hate speech? How does hate speech vary by gender subgroup?
- **Intersectionality:** How does hate speech vary when gender is being targeted alongside another protected characteristic, such as ethnic or religious identity?
- **Location:** Does gendered hate speech vary by social media platform?

Research scope

To answer the research questions, CIR has undertaken an analysis of social media content from X (formerly Twitter), Telegram, and Facebook to determine the forms and locations of gendered hate speech and the sentiment of intersectional hate speech (hate speech that targets multiple protected characteristics). To enable the retrieval and analysis of large quantity comparable of social media data, CIR narrowed the scope of the research from all forms of TFGBV to only gendered hate speech. This research used the Ethiopian Government's definition of hate speech, as set out within the 'Hate Speech and Disinformation Prevention and Suppression Proclamation', to ensure relevance to the Ethiopian context as much as possible.[7] Future research could investigate other forms of TFGBV, including the use of imagery to spread hate, revenge pornography and case studies on prominent individuals.

To inform the methodology for data collection, CIR began desk-based research into existing hate speech lexicons in May 2023. During a visit to Addis Ababa in July 2023, CIR refined a comprehensive lexicon in Amharic, Afaan Oromo, Tigrigna, and English, comprised of over 2,000 inflammatory terms, through meetings and a roundtable with representatives from CSO and human rights institutions.

CIR used keyword matching and hate speech detection models to sample posts from the three platforms in August 2023. Between September and November 2023, a team of researchers annotated the dataset for X and Telegram according to three dimensions: hate speech target, type, and sentiment (as specified in the annotation protocol). As CIR's earlier study signalled the need to investigate intersectional hate speech (hate along multiple identity lines), this study was broadened in scope to investigate multiple protected characteristics covered by the Ethiopian Government's Hate Speech Proclamation (disability, ethnicity, gender, race, and religion).[8]

CIR presented preliminary findings to stakeholders during workshops in Addis Ababa in late November 2023. Due to differences in platform policies, Facebook data took longer to acquire, and it was annotated in December 2023. During January, the team analysed the findings and collated them into this report before presenting the findings in early February 2024 to stakeholders in Addis Ababa.

[7] Proclamation No.1185/2020, As seen in Federal Negarit Gazette (2020) Available at: https://ethionab.org/wp-content/uploads/2022/09/1185_2020_HATE_SPEECH_AND_DISINFORMATION_PREVENTION_AND_SUPPRESSION_PROCLAMATION_.pdf

[8] Proclamation No.1185/2020, As seen in Federal Negarit Gazette (2020) Available at: https://ethionab.org/wp-content/uploads/2022/09/1185_2020_HATE_SPEECH_AND_DISINFORMATION_PREVENTION_AND_SUPPRESSION_PROCLAMATION_.pdf

3. Methodology

In this study, CIR investigated the different types of hate speech targeting women and girls on social media in Ethiopia, as well as differences in the sentiment of that hate across platforms and target sub-groups. In this context, 'type' of hate speech refers to the method of abuse (such as 'threats'), while 'sentiment' refers to the emotive or linguistic qualities of the abuse (such as 'irony' or 'stereotyping'). A conceptual framework outlining the different types and sentiments of hate speech assessed is included below in section 3.1. This conceptual framework underpins the study's broader research methodology, including the data annotation protocol, and therefore is essential to interpreting the research findings. CIR annotators only classified content as 'hate speech' if the target of the hate speech (a protected identity group, such as gender) and the type of hate speech could be clearly identified.

CIR designed the methodology for this study to be uniform across the three platforms analysed (X, Telegram, and Facebook) to maximise the comparability of the findings across the platforms. Natural Language Processing (NLP) techniques comprised a core part of the methodology used to analyse and interpret textual social media data and enable automated processing. This included developing a lexicon, collecting data from social media platforms, and processing the data before it could be analysed, which included steps such as removing duplicates and anonymising posts. CIR also used NLP methods to partially automate the identification of hate speech, reducing the manual effort required for annotation. Finally, CIR used NLP techniques during the manual data annotation process. The following sections outline each of these steps.

The methodology developed for this research has been reviewed by leading academics in the field and will feature in the 'Resources for African Indigenous Languages' academic publication.[9]

Conceptual framework

To investigate TFGBV in the form of online gendered hate speech in Ethiopia, CIR developed a conceptual framework to outline the key components of hate speech: target, type, and sentiment. While having a clear target and 'type' is essential for a piece of content to classify as hate speech, and therefore essential to investigate, CIR included an additional assessment of 'sentiment' for a richer analysis.

CIR selected the hate targets for the study using those found within the Ethiopian Government's Hate Speech and Disinformation Proclamation.[10] As hate speech might target women and girls, alongside another protected characteristic, the conceptual framework includes other forms of hate

[9] 'Resources for African Indigenous Languages' (2024) Available at: https://sadilar.org/en/the-fifth-workshop-on-resources-for-african-indigenous-languages-rail/

[10] Proclamation No.1185/2020, As seen in Federal Negarit Gazette (2020) Available at: https://ethionab.org/wp-content/uploads/2022/09/1185_2020_HATE_SPEECH_AND_DISINFORMATION_PREVENTION_AND_SUPPRESSION_PROCLAMATION_.pdf

speech (disability, ethnicity, gender, race, and religion). This allows for an understanding of intersectionality, i.e. when hate speech has multiple targets.

In this context, 'type' of hate speech refers to the method of abuse (such as 'threats'), while 'sentiment' refers to the emotive or linguistic qualities of the abuse (such as 'irony' or 'stereotyping'). For example, hate may be conveyed using nuances in language, such as sarcasm, mockery, or satire. A full overview of this typology can be seen in the figure below.

Having a clear understanding of the building blocks of hate speech is essential to identifying the ways in which hate speech can vary (by target, type, and sentiment), and can therefore inform targeted solutions to preventing and countering its occurrence.

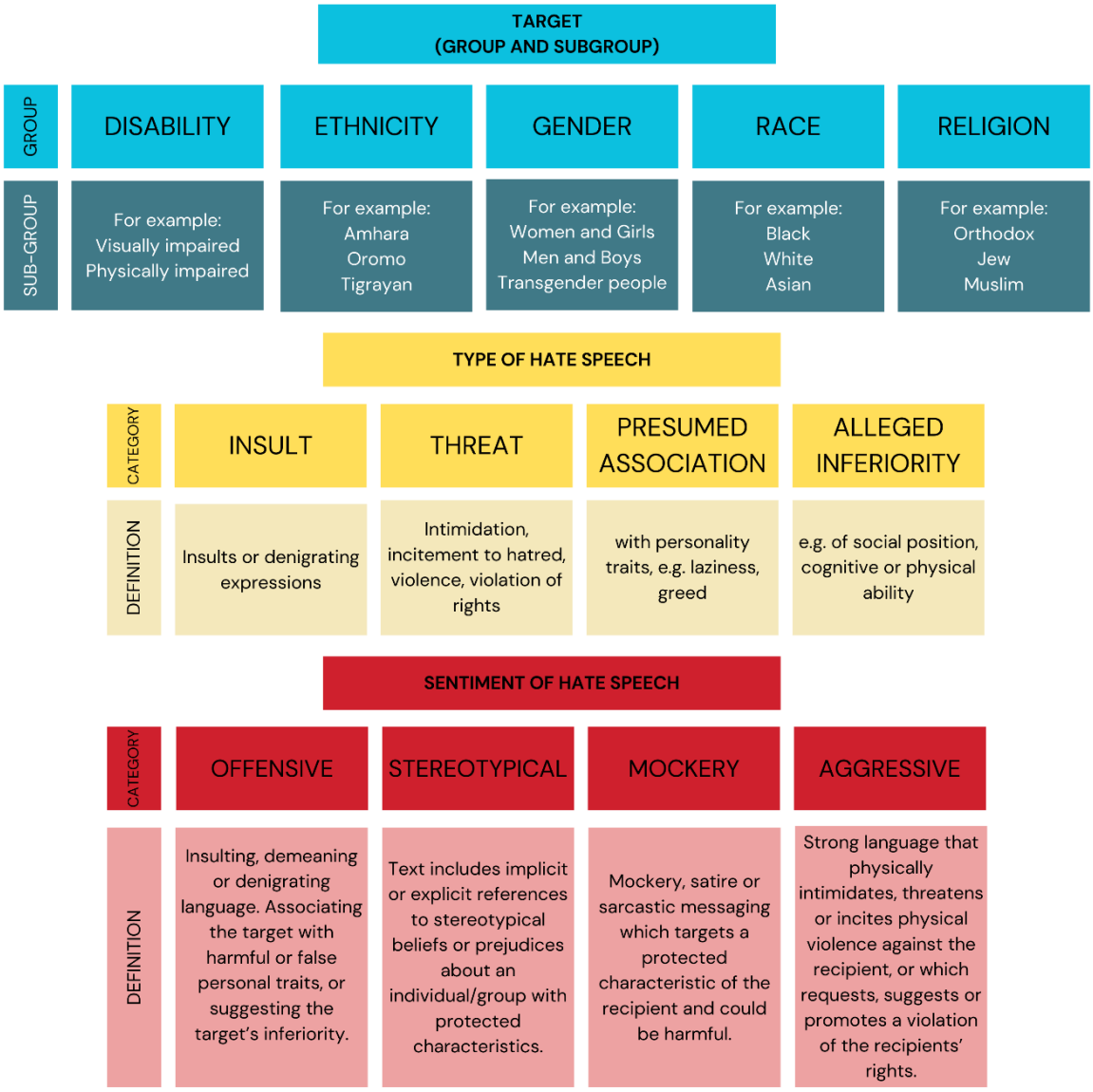


Figure 4: CIR's Conceptual Framework. This underpins the research design and annotation schema, as found in the Annotation Protocol in section 7.2.

Lexicon development

CIR developed a lexicon comprised of 2,058 inflammatory keywords across four languages (Amharic, Afaan Oromo, Tigrigna, and English) which may be indicative of hate speech along gendered, ethnic, and religious lines. CIR believes that this is the most comprehensive lexicon at present for the Ethiopian context. Figure 3 shows the number of keywords curated for each protected characteristic.

The lexicon was developed through desk-based research (identification and refinement of existing hate speech lexicons),[11] the identification of keywords and narratives during the in-person, semi-structured interviews carried out for CIR's first report,[12] and a roundtable of experts in Addis Ababa in July 2023. The roundtable brought together 21 individuals from an array of CSOs, UN agencies, and women and girls' rights advocacy groups. During the session, CIR presented the findings of the first report and facilitated a discussion on terms indicative of hate speech in Ethiopia across the four languages selected. This session yielded a significant number of additional keywords and insight as to why specific terms constitute hate speech in Ethiopia. This discussion guided the development of the hate speech lexicon. Although political affiliations or viewpoints are not protected characteristics under the Ethiopian Government's hate speech definition, additional categories for 'political' hate and 'other terms' were also included in the lexicon, as many terms were raised which could be used in hate speech but do not fall explicitly under a protected characteristic.

A first draft of the lexicon was shared with CIR's partners, stakeholders, and roundtable attendees in Ethiopia for feedback. It became apparent at this stage that there was confusion about why some terms had been included in the lexicon, as they may not, on their own, constitute hate speech. CIR reassured those concerned that the collected keywords would be used to obtain content from social media which could contain hate speech; however, human annotators would then analyse whether the content was/wasn't hate speech, as per the detailed annotation protocol. The presence of a keyword on its own does not necessarily mean that a piece of content is hate speech.

Once all feedback was incorporated, errors corrected, and duplications removed, three Ethiopian social media research experts with Amharic, Afaan Oromo, Tigrigna, and English language skills then reviewed the lexicon to ensure that the keywords were correctly represented in each of the four languages selected for the study. As this study is primarily focussed on gender-based violence, the number of keywords for gender is significantly higher than other protected characteristics. Despite this, CIR believes the keyword lists for the other categories are still more comprehensive than other studies to date.

[11] See section 8, bibliography, for a full list of the resources consulted.
[12] CIR (2023) Silenced, shamed and threatened: the online abuse of women and girls who participate in Ethiopian public life.

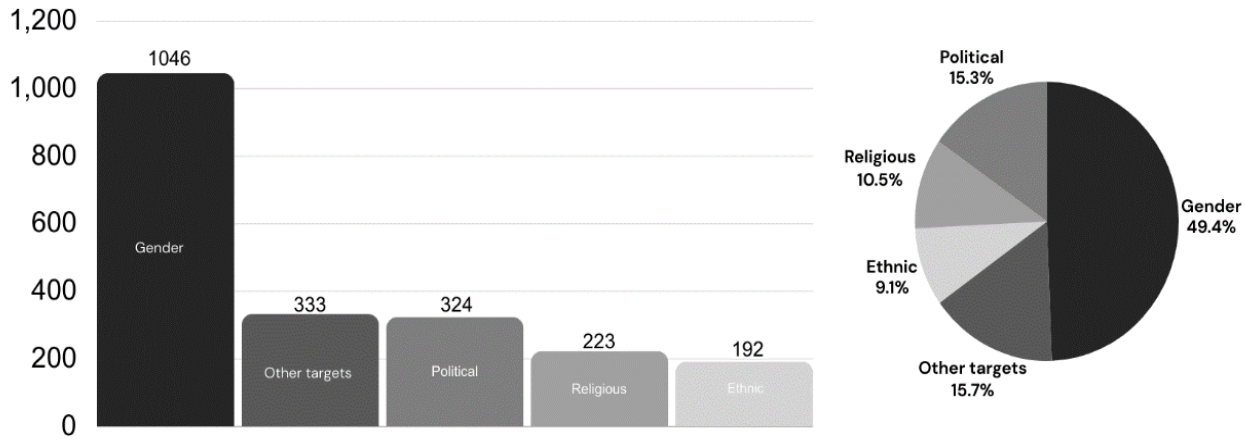


Figure 5: Pie chart and bar graph showing the distribution of inflammatory keywords used in the collection phase.

Data collection

CIR selected X, Telegram, and Facebook for investigation. Interviewees in CIR's first report noted that these three platforms were sites where women and girls face online abuse. Research by Statista, published in 2023, suggests that Facebook is the most popular social media platform in Ethiopia, with 6.5 million users, followed by Messenger, LinkedIn, and Instagram.[13] In early 2023, on Android (which is overwhelmingly the most common smartphone operating system in Ethiopia[14]) Telegram was the most downloaded social media platform.[15] The lack of research into the use of Telegram for perpetrating online abuse signals a significant gap in the literature.

Due to the textual nature of X, Facebook and Telegram, the same methodology could be used to investigate gendered hate speech across all three platforms. While CIR is also interested in gendered hate speech on YouTube and TikTok, the different format (video and related comments) requires a different methodology for analysis. Future study could investigate these platforms.

As different data collection methodologies were needed to collect data from the three platforms, in accordance with their terms and conditions, the raw number of hate speech posts cannot be compared. The study thus analyses the composition of hate speech, by platform. The results are not juxtaposed based on count but based on the proportion of different 'types' 'targets' or 'sentiments' of hate speech and the number of posts isn't significant to the outcomes of the study.

[13] Statista (2023) Social media users by Platform in Ethiopia, Available at: https://www.statista.com/statistics/1312554/social-media-users-by-platform-in-ethiopia/

[14] StatCounter (2023) Mobile Operating System Market Share Ethiopia, Available at: https://gs.statcounter.com/os-market-share/mobile/ethiopia

[15] This was reported by Addis insight, citing data from the Google Play store: Addis Insight (2022) Available at: https://addisinsight.net/why-telegram-is-so-popular-in-ethiopia/; however, this is consistent with reporting from previous years, for example, see: Quartz (2018) Available at: https://qz.com/africa/1214381/in-a-continent-dominated-by-whatsapp-ethiopia-says-yes-to-telegram

3.3.1. X (formerly Twitter)

To collect data from X, CIR used the social media analysis tool **Meltwater**.[16] Meltwater supports the use of keyword search for posts on X within 18 months maximum of the date of search. Meltwater provides metadata for each post, including URL, content (i.e., the text), name and username of the author, country, language, number of users reached, number of engagements, sentiment, the X client used (phone, web, android, iPhone, etc), profile URL, user profile description (bio), number of X followers, number of X following, post views, likes, number of replies, number of reposts, number of shares, number of reactions, and whether the author is verified.

Table 1 contains information on the number of posts extracted from X. To ensure a relevant sample, CIR investigators only retrieved English language posts that originated in Ethiopia.

3.3.2. Telegram

To collect data from Telegram, CIR used the official Telegram API.[17] As Telegram only supports searches within Telegram Channels to which a user belongs, CIR engaged social media experts from Ethiopia to curate a list of widely popular and influential Telegram Channels in Ethiopia where posts could be extracted from. CIR ensured that only publicly accessible Telegram Channels were selected for collecting data in this research, adhering to ethical guidelines.

This resulted in the selection of 310 Telegram Channels. Of these, 285 were successfully joined and used as the sites of Telegram data collection for this research work. The metadata collected from Telegram posts included username and ID of author, text of the post, link to telegram post, channel ID, first name, and last name of author. Table 1 contains information on the number of posts extracted from Telegram.

3.3.3. Facebook

Due to differences in Facebook's policies regarding data extraction, compared to the policies of X and Telegram, CIR was unable to use Meltwater to extract data or write code to scrape data. Instead, CIR engaged social media experts from Ethiopia to select a list of prominent and influential Ethiopian Facebook groups. This engagement generated a list of over 300 Facebook groups. The metadata extracted from Facebook included post URL, username and ID of poster, number of shares, number of likes, text, and number of views.

The different methodology resulted in a far smaller number of posts collected for analysis. This could have had implications on the findings, as a smaller number of true hate posts may have

[16] Meltwater, Available at: https://explore.meltwater.com/uk/all-in-one
[17] Other scraping tools could be used, such as Tom Jarvis' 'Telegram Snowball Sampling' tool, available on GitHub. Available at: https://github.com/thomasjjj/Telegram-Snowball-Sampling

been identified, as the filtering steps used on the datasets for the other two platforms could not be applied. Table 1 contains information on the number of posts extracted from Facebook.

Data preparation

3.4.1. Data pre-processing

Data pre-processing is a crucial step in any natural language processing (NLP) methodology. It involves cleaning, transforming, and organising raw text data to make it suitable for analysis and/or modelling. Data pre-processing is essential to improve the quality and reliability of language models and to extract meaningful insights from textual data. Below is an overview of some of the key tasks involved in the pre-processing of the data collected in the four different languages:

Text cleaning:

Text data collected from social media often contain irrelevant or erroneous information that complicates analysis or interpretation. Cleaning tasks completed for the datasets included:

i. Removing Hypertext Markup Language (HTML) tags and special characters. HTML tags are labels that tell browsers how to display information and special characters are extra symbols in text. HTML tags and special characters are useful for formatting webpages or changing the visual representation of text but are not needed when processing text.

ii. Converting texts to lowercase to ensure case insensitivity. Lowercasing text ensures uniform treatment of words regardless of case. For instance, "apple", "APPLE", and "APPle" become equivalent as "apple" after conversion.

iii. Removing or replacing punctuation.

The following cleaning tasks were done for English language datasets only. CIR carried out these steps to make the textual posts more suitable to the models used for classifying the dataset as hate or not hate:

i. Handling or removing numerical values, dates, and other non-textual information.

ii. Removing stop words (common words like "and," "the," "in" that don't carry significant meaning).

iii. Handling and normalising abbreviations and acronyms. Abbreviations and acronyms are often used to conceal or disguise harmful content. Thus, converting all variations of popular words to a single form helps improve the performance of hate speech detection models.

Removing duplicates:

Following text cleaning, CIR detected and removed duplicate posts. This is essential as duplicates can distort statistical analyses or model training. The number of posts reduced drastically in the X dataset following this stage. This is because multiple keywords often appeared in the same post, resulting in duplicate retrievals of the said posts.

For Telegram, many of the duplicates were posts that stated: *"This group can't be displayed because it violated local laws".* The table below outlines the number of posts removed by Telegram, per language, for "violating local laws". These posts would have contained a keyword that was in CIR's lexicon prior to its removal.

	Number of Posts Removed
English	94,261,111
Amharic	43,356
Afaan Oromo	301,460
Tigrigna	34,508

Table 1: Number of posts removed by Telegram from Telegram groups for violating local laws.

Due to the different methodology, duplicates did not exist in Facebook, so this step was not taken. It was observed that both Facebook and X do not have duplicate posts resulting from posts replaced with the same text for "violating local laws". This may be because both platforms completely remove posts found to be violating their laws.

Data anonymisation:

CIR anonymised the usernames in posts in line with ethical requirements to protect the privacy and confidentiality of individuals whose data was used for research and analysis. This was done by replacing all usernames (i.e., any word appearing after the "@" symbol) with the word "USERNAME".

3.4.2. Classification of dataset as hate or not hate

The aim of this research is to detect and analyse instances of hate speech in posts on social media platforms. Typically, identifying these posts involves a manual review process, where each post is assessed by an expert to determine if it contains hate speech. To expedite the process and reduce the number of non-offensive posts that annotators must review before encountering hate speech, CIR introduced an automated classification step in the data processing pipeline.

This step includes the fine-tuning and deployment of a machine learning model based on transformers to classify posts as either containing hate speech or not. Subsequent analyses were conducted exclusively on posts classified as containing hate speech by the machine learning model. This approach removed posts which were considered likely not to contain hate speech.

It is important to note that due to the lack of pre-trained models in Tigrigna and Afaan Oromo, this pre-classification step was applied exclusively to posts in English and Amharic. As a result, the English and Amharic posts have undergone one extra filtering step compared to the Tigrigna and Afaan Oromo datasets, potentially resulting in the easier identification of hate speech.

3.4.3. Data sample selection for further analysis

The data collection process described above resulted in CIR acquiring of tens of millions of posts (as illustrated in Table 2). Even after extensive data pre-processing, which involved removing duplicates and excessively short posts, over 5 million posts remained.

Due to the constraints posed by limited human resources available for manual annotation to determine hate content, CIR chose to randomly select a manageable sample from the gathered data. Table 2 shows the number of posts chosen for subsequent analysis.

CIR obtained datasets for three platforms and in four languages. CIR randomly selected posts for analysis from the different datasets in order to mitigate potential biases in the selection process.

3.4.4. Data processing steps in numbers by platform

The differences in data collection methods can account for the difference in the number of posts collected across X, Telegram, and Facebook. For example, only 7,230 Facebook posts were collected, compared to 865,224 X posts and 326,471,094 Telegram posts.

After pre-processing, the number of X posts decreased to 527,522, or 60.97% of the posts initially collected. The number of Telegram posts decreased to 906,471, or 0.28% of the posts initially collected. The number of Facebook posts after the pre-processing remained the same (see section 3.3.1 for more information).

Next, CIR used machine learning classifiers to help filter out posts which most likely did not contain hate speech. The classifiers for both English and Amharic language posts predicted a higher prevalence of hate content on X and Facebook, compared to Telegram. This discrepancy can be partly attributed to the fact that the classifiers were primarily trained on X posts, where the writing style differs slightly from that of Telegram or Facebook. For instance, Telegram posts tend to have a longer average length, and abbreviations are more commonly used compared to X. As mentioned above, due to a lack of pre-trained models, this stage was not carried out on the Afaan Oromo or Tigrigna datasets.

Finally, CIR randomly sampled 2,634 X posts, 2,107 Telegram posts and 2,264 Facebook posts for annotation.

	X	Telegram	Facebook
Number of posts collected	865,224	326,471,094	7,230
Number of posts after pre-processing	527,522	906,471	7,230
Number of posts after Machine Learning Hate classification	244,221	239,558	4,135
Random sample for annotation	2634	2107	2264

Table 2: The number of posts at each step of the data preparation process, by platform.

Data annotation

Manual annotation played a pivotal role in the data preparation process, involving human annotators well-acquainted with the domain of interest and specific annotation protocol (see section 7.2, the appendix, for the full annotation protocol).

Ensuring that the multi-lingual team of annotators used the same methodology and had a shared understanding of what constitutes hate speech was essential. CIR's conceptual framework formed the basis of the annotation protocol (see section 3.1). It provided the annotators with clear definitions of hate speech, its targets, and the type and sentiment categories of hate speech (see section 7.2, the appendix, for the full annotation protocol). The protocol included several examples to highlight how the targets, type, and sentiment of hate speech could be identified in text, and how to use the annotation tool.

Following the identification of hate speech, the annotators tagged essential details, including the target of the hate (gender, ethnicity, religion, race, and disability), type and sentiment or the Annotation Protocol in section 7.2). The categories are not mutually exclusive; multiple targets, types, or sentiment categories could be selected. For example, hate speech could be aggressive, yet also use mockery.

The annotators used Doccano, a popular open-source tool, to annotate the posts and tag the hate speech targets, types, and sentiment. The annotations facilitated the subsequent extraction and in-depth analysis of the data.

This exercise also revealed several limitations to the study's methodology (see section 3.6 on limitations). This dialogue between the annotators and data engineers led to the creation of a series of rules to ensure that a consistent approach was applied during annotation (see section 7.3).

3.5.1. The annotators

English

Two human annotators were enlisted to manually annotate the randomly chosen English posts. The primary annotator, who designed the annotation protocol and is knowledgeable of the Ethiopian context, was responsible for annotating the entire selection of English posts. The secondary annotator was assigned to annotate 10% of the dataset annotated by the primary annotator. This approach allows for the estimation of inter-annotator agreement (IAA), enabling the measurement of agreement and consistency between the two annotators.

Amharic

CIR enlisted three people to manually annotate randomly selected Amharic posts. The primary annotator, a native Amharic speaker with experience in social media analysis, undertook the annotation of the entire Amharic dataset, while the other two annotators (who were also assigned with the Tigrigna and Afaan Oromo datasets, respectively) were tasked with annotating approximately 10% of the dataset annotated by the primary annotator. IAA was subsequently estimated using the posts annotated by all three annotators to assess the level of agreement and consistency among them.

Tigrigna

CIR enlisted one person to manually annotate randomly chosen Tigrigna posts. IAA was considered unnecessary for the Tigrigna annotator, as this same annotator had previously worked on annotating the Amharic dataset, and the results of the IAA in the Amharic dataset indicated their competence in identifying hate speech, labelling its target, categorizing speech types, and assessing the sentiment of the hate.

Afaan Oromo

Similarly, CIR enlisted one person to manually annotate randomly chosen Afaan Oromo posts. IAA was considered unnecessary for the Afaan Oromo annotator, as this same annotator had previously worked on annotating the Amharic dataset, and the results of IAA agreement in the Amharic dataset indicated their competence in identifying hate speech, labelling its target, categorizing speech types, and assessing the sentiment of the hate.

3.5.2. Inter-annotator agreement (IAA)

To ensure consistency in the application of the guidelines, IAA[18] scores were calculated using Cohen's Kappa (k) and Fleiss' Kappa metrics. The IAA scores showed 'Fair' to 'Moderate' agreement between annotators (see section 7.4 for more information).

[18] See section 7: Glossary for the full definition.

Roundtables and workshops

CIR held a series of roundtables, workshops, and meetings with stakeholders in July and November 2023 in Addis Ababa, and virtual workshops in December. These were attended by individuals who work in the human rights field. The roundtable and meetings in July provided a forum to debate and develop the lexicon of inflammatory keywords that were used in the data collection. During this period, CIR consulted 42 individuals. While many attended in an independent capacity, others collectively represented 11 different Ethiopian human rights organisations. These discussions resulted in the identification of 2,058 keywords across the four languages.

During November 2023, CIR facilitated a series of meetings and four workshops to present preliminary findings and discuss possible recommendations. CIR conducted four additional virtual workshops in December, allowing more people to engage with the research. In total, CIR consulted 24 people, representing 18 Ethiopian organisations. The aim of these workshops was to speak to people who work across multiple fields, including women and girls' rights, gender-based violence (online and offline), and online safety. CIR wanted to engage with subject matter experts and hear different perspectives about how to research and respond to these issues in Ethiopia. As each context is entirely different, CIR wanted to ensure that a broad range of voices were consulted, an essential step for balanced and well-informed research.

In February 2024, CIR held a series of final meetings with stakeholders to present the findings and provided a two-week feedback window to allow as many voices to be incorporated as possible.

Limitations

Where possible, CIR took measures to reduce the limitations of the methodology and the findings. However, several resource, context, or result limitations must be considered, which have been outlined below.

3.7.1. Languages

Over 80 languages are spoken in Ethiopia. While not all these languages are represented on social media, it was only possible to select four languages for this study due to resource and time constraints. As a result, four languages that are prevalent on social media were selected: Amharic, Afaan Oromo, Tigrigna and English. The findings therefore only represent a sample of speakers of each of the four languages on social media. Despite this, carrying out the study in four languages was resource intensive, requiring skilled researchers who were familiar with the languages, context, and social media analysis. Future studies could adopt CIR's methodologies to conduct similar studies on different languages.

3.7.2. Data collection

Tailored data collection methods were made for each platform to ensure that CIR complied with each platform's unique set of terms and conditions. While necessary, this resulted in the collection of different amounts of data from each platform. For example, CIR collected far fewer posts from Facebook than from X or Telegram. As a result, CIR could not compare how prevalent gendered hate speech is across the three platforms. Instead, proportional data was used to assess the type and sentiment of hate speech on each platform.

During the identification of Facebook pages to collect data from, the researchers noted several trends which could impact the study. For example, the researchers noted that abuse on Facebook is often targeted, such as in comments on posts by prominent people. By only capturing content from public groups, CIR did not obtain these data. Case studies on specific people and the hate levelled against them could be conducted during future study.

CIR researchers also found that many Facebook posts within the groups contained images with text overlayed on top, as well as graphic or erotic imagery. Hate speech in this format may be more likely to evade detection by platforms. Other posts contained manipulated imagery alongside emojis, or even screenshots of other posts (again, images rather than text). CIR also noticed posts with images of young girls looking sad, alongside captions such as "I'm 18 years old and don't have a boyfriend" and "For those of you looking for long lasting marriage and love friends good cash payers and temporary ones, the right channel (down arrow) suggestion below" [sic] (see figure 4 below). As this study focussed on text for annotation in line with the annotation protocol, these posts were not captured in the data collection. Future research could investigate the use of pictures and manipulated imagery in the online abuse of women and girls.

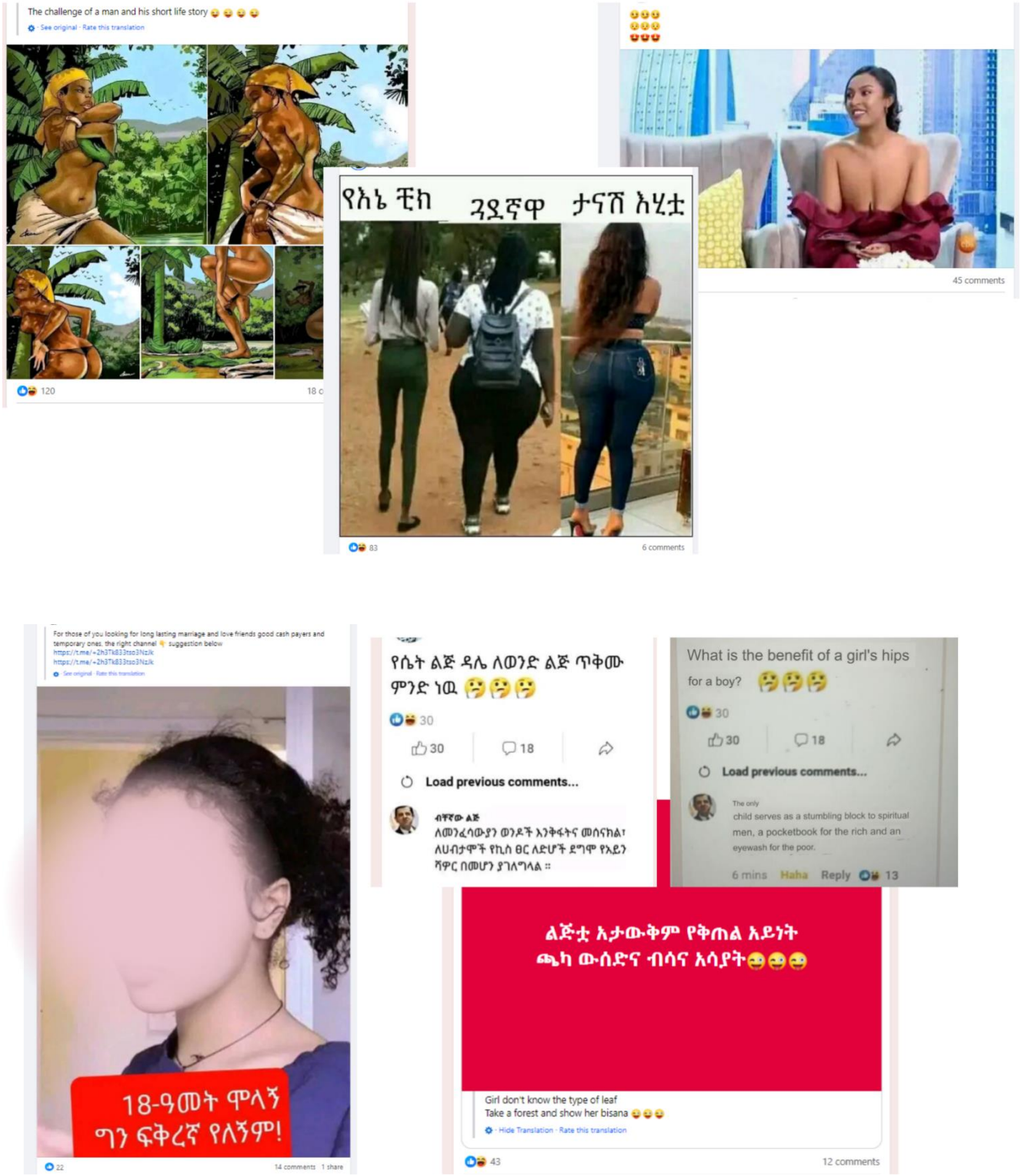


Figure 6: A selection of examples from Facebook showing different styles of posts that would not have been included in the study. The young girls face [bottom left] has been blurred by CIR. The channel and username have been obscured to prevent identification.

3.7.3 Data processing and sampling

Prior to sampling the dataset, CIR used a pre-trained machine learning model to filter out posts that the model classified as 'not hate', thus increasing the chances that CIR would find hate speech to analyse in the dataset. It is important to note that, due to the lack of pre-trained models in Tigrigna and Afaan Oromo, this pre-classification step was applied exclusively to posts in English and Amharic. The absence of pre-trained models is partly due to the scarcity of annotated datasets available for model training in these languages. However, CIR's annotated datasets in these languages could be leveraged to train machine learning models for future studies. In addition, pre-trained models have the potential to automate the identification of hate speech on social media platforms. This could be used for platform content moderation and may lead to a reduction in hate speech across the languages.

The large volume of data CIR retrieved, and therefore the requirement to analyse only a random sample, created several analytical limitations. CIR retrieved 865,224 posts from X, 326,471,094 posts from Telegram, and 7,230 posts from Facebook. This posed several challenges, including slowing down the speed of the pre-processing stage. However, even after this, the dataset was still too large for manual annotation. Given the limited human resources and timeframe, CIR selected a random sample from each social media platform. As the sample was random, CIR was unable to investigate changes in the incidence of hate on the social media platforms over time, or whether there were any spikes in hate speech at certain times (perhaps in relation to significant events offline). Additionally, the random sample prevented CIR from assessing connections between social media users that proliferate hate speech online. As CIR has retained the original dataset, future research could establish if there are certain accounts responsible for large volumes of hate speech.

3.7.4. Intersectionality

CIR expanded the scope of the research significantly in order to investigate intersectional hate speech (i.e., hate speech that targets individuals/groups along multiple identity lines). Originally, the focus was solely on gendered hate speech. To investigate intersectionality, CIR also had to collect data on other hate targets. This meant that CIR had to expand the lexicon to include keywords associated with other protected characteristics (e.g., ethnicity and religion). Using the full lexicon, CIR created a large database of not only gendered hate speech, but also ethnic, religious, racial, and disability-based hate speech. To allow for the analysis of intersectional hate speech, these data were annotated in the same way and utilised in the analysis. A future study could utilise more of the data collected on the other hate targets.

3.7.5. Annotation

CIR felt it was essential to carry out an IAA assessment on a portion of the annotated sample. Although this stage slowed down the speed of annotation, the commendable IAA scores instilled

confidence in the annotators' comprehension of the task and the consistency of the application of the annotation protocol.

When the annotation protocol was put into practice, the difficulties of assessing whether a post constituted hate speech became prominent. Further guidance was needed on how to deal with specific posts. The team consulted with one another where necessary. The rigorous application of the annotation protocol by researchers, however, had the potential to limit CIR's dataset, as posts that may have been hate speech – but could not be confirmed to be hate speech without further context – were excluded. For example, some posts lacked a clear hate target, and were therefore excluded from the study.

Additionally, as the annotation process relied upon textual analysis without context or imagery (as mentioned in relation to Facebook data above), the annotators agreed that the text should be taken at face value. Similarly, annotators noted that references to black skin, including derogatory terms like 'nigga' or 'nigger', were often used alongside gendered hate speech. In fact, one workshop participant suggested that this might be seen as endearing to some, rather than hate. If this was the case, then, due to the rigorous application of the annotation protocol, and the rule to take the text at 'face value', there is a risk that some of these posts were mislabelled. Conversely, taking the text at face value may have led to a smaller sample size. For example, one workshop participant noted that the pronoun 'she' may be used in reference to a man, as an insult. This would not have been identified during the research, as the context will have been lost. As a result, CIR assessed that the results represented just the tip of the iceberg. For more information on the challenges during annotation, see the appendix in section 7.2.

4. Results and discussion

This research has sought to strengthen the evidence base on TFGBV in Ethiopia by investigating the types of hate speech (i.e. the written methods employed by the abuser within the hate speech) that women and girls face on three social media platforms (X, Telegram, and Facebook). CIR also assessed the 'sentiment' of the hate speech to determine the nature of the abuse; for example, whether it was offensive, stereotypical, aggressive, or contained mockery. Additionally, CIR examined variations by gender subgroup across the three platforms. The different types, sentiments and targets are defined in the typology found in section 3.1 and in the Annotation Protocol in section 7.2.

CIR also investigated intersectional hate speech (i.e. when hate has multiple targets) to determine how hate speech varies when women and girls are targeted for their gender alongside another protected characteristic, such as ethnic or religious identity. By collecting data on other hate targets, CIR was also able to compare the composition of hate speech directed at women and girls, to hate speech directed towards people for their ethnic, religious, or racial identities. Where appropriate and relevant, insights from the interviews (CIR's earlier study), roundtables, and workshops have been included in the discussions.

4.1 Gendered hate speech targeting women and girls

In combination, CIR's earlier research into online abuse in Ethiopia, the roundtables, workshops, and this follow up study into social media trends, all signal that women and girls in Ethiopia suffer from several different types of TFGBV. These include hate speech, revenge pornography, and harassment. This study was narrowed in scope to investigate the types, sentiments, and targets of online hate speech as a form of TFGBV.

This study found that 'insults' were the most prevalent 'type' of gendered hate speech targeting women and girls, while 'offensive' speech was the most common 'sentiment' or tone of speech. Using these findings, decision makers can tailor policies to respond to, and counter, gendered hate speech online; for example, through targeted education campaigns that seek to breakdown gendered stereotypes.

4.1.1. Type of hate speech targeting women and girls

CIR investigated the 'type' of hate speech targeting women and girls present within the dataset (insults, threats, presumed association, and alleged inferiority).

	TYPE OF HATE SPEECH			
CATEGORY	INSULT	THREAT	PRESUMED ASSOCIATION	ALLEGED INFERIORITY
DEFINITION	Insults or denigrating expressions	Intimidation, incitement to hatred, violence, violation of rights	with personality traits, e.g. laziness, greed	e.g. of social position, cognitive or physical ability

Figure 7: the definitions of each category, as denoted in the schema in the methodology.

'Insults', a category that also encompasses denigrating language, was the most prominent type of hate speech identified across all platforms, accounting for 36.6% of the total. 'Presumed association' of women and girls with specific characteristics, including greed or laziness (see full list in the annotation protocol, section 7.2), was the second most prominent type of hate speech across all platforms, representing 27.7% of the total.

Women and girls received proportionally less hate in the form of 'alleged inferiority' (22.2% of the total), or speech that references the inferiority of women and girls in relation to their social position, cognitive, or physical ability, amongst other characteristics (see full list in the annotation protocol, section 7.2). 'Threats' – which encompass intimidation, threats or incitement to hatred, violence, or violation of individuals' rights – accounted for 13.4% of the hate speech detected.

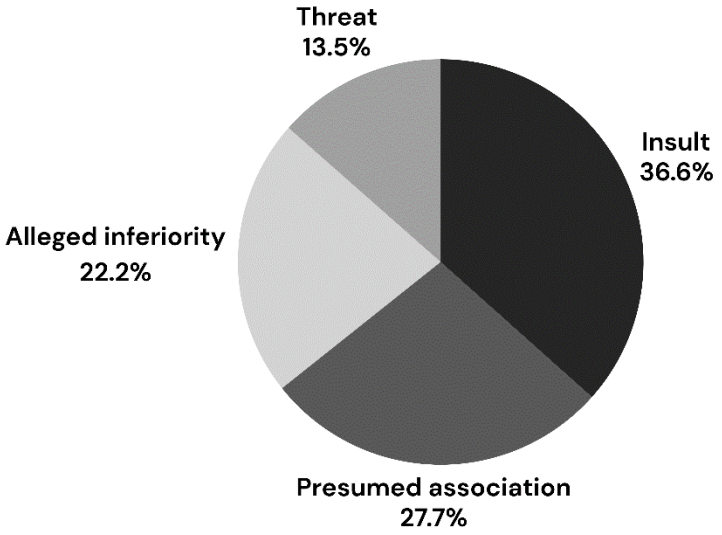


Figure 8: Pie chart showing the distribution of the type of hate speech targeting women and girls.

4.1.2. Sentiment of hate speech targeting women and girls

To gain further insight into the 'sentiment' of hate speech targeting women and girls on the three social media platforms, CIR also looked at whether hate speech contained 'aggressive' language, 'offensive' language, 'irony' (including mockery, at the expense of the target), or 'stereotypes'.[19]

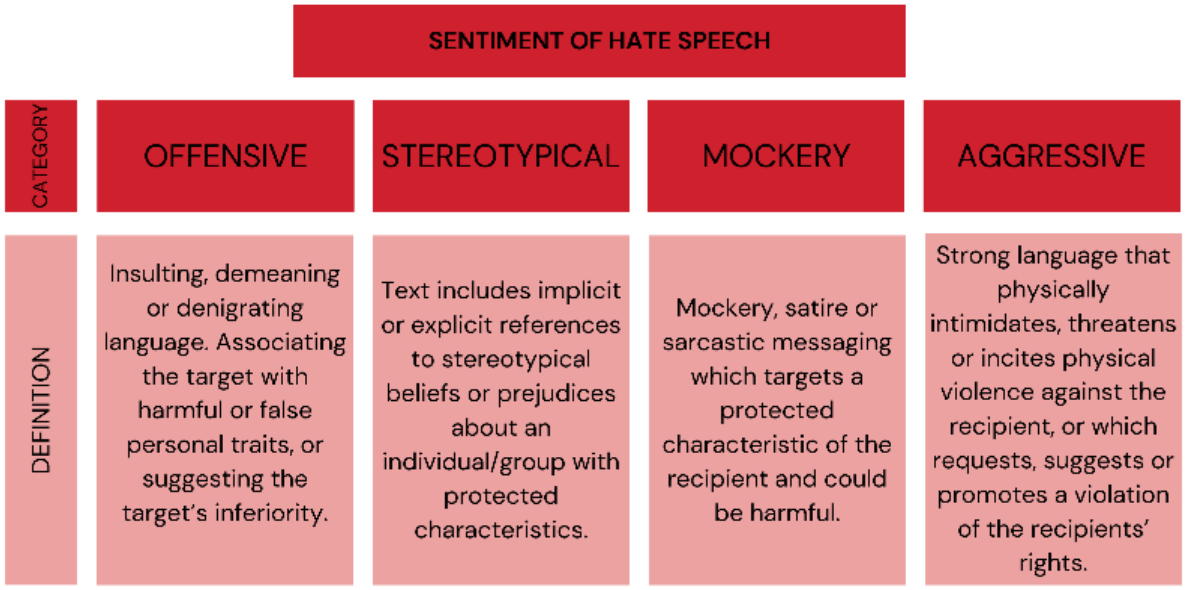


Figure 9: the definitions of each category, as denoted in the schema in the methodology.

Across all platforms, 'offensive' language was the most common sentiment of hate speech found targeting women and girls (46.2% of the dataset). This encompassed several different forms of speech, from insulting, demeaning, or denigrating speech, to associating the target (an individual or group) with harmful or false personal traits, or suggesting the target's inferiority.

The second most prevalent category was 'stereotypical' speech (26.2% of the dataset), or posts that contain implicit or explicit references to stereotypical beliefs or prejudices about an individual or group due to their gender. This was followed by 'irony' (20.3% of the dataset), or posts that include jokes, satire or sarcastic messaging targeting an individual/group due to their gender.

Lastly, the study identified significantly less 'aggressive' hate speech against women and girls than the other categories, with 'aggressive' hate comprising only 7.3% of the dataset. 'Aggressive' language includes speech that intimidates, threatens, or incites physical violence against the recipient, or promotes the violation of their human rights.

[19] Full definitions of each of these categories can be found in section 3.1 and the annotation protocol in section 7.2

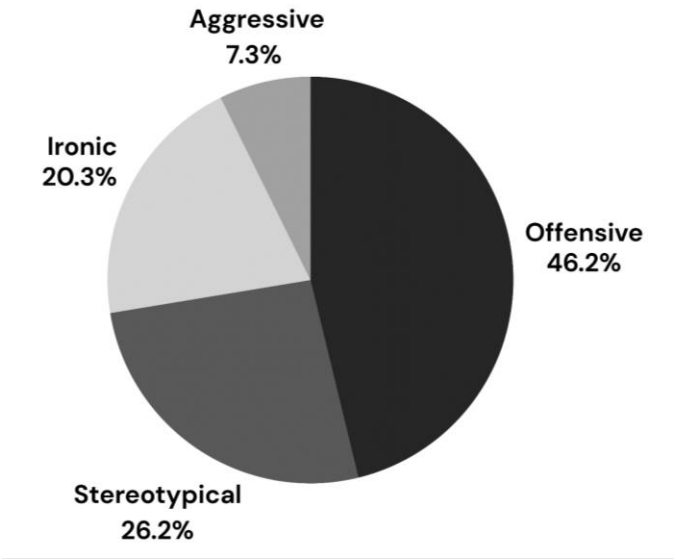


Figure 10: Pie chart showing the distribution of the sentiment of hate speech targeting women and girls.

4.1.3. Discussion: hate speech targeting women and girls

The most prevalent forms of gendered hate speech within the dataset were 'insults' and 'offensive' speech. When the initial findings were presented during eight separate workshops in November and December 2023 several participants expressed shock at the low proportion of 'threats' and 'aggressive' speech. They expressed a belief that 'threats' and 'aggressive' language were far more common than CIR's figures suggest. Discussions on this topic ensued, resulting in other participants outlining how aggressive or threatening hate speech was more likely to be reported and removed, perhaps resulting in a smaller sample in the dataset than reality. Additionally, the absence of contextual information and the strict definition of hate speech employed in CIR's annotation protocol means that not all threats or aggressive language may have classified as hate speech; for example, data was excluded if no clear protected characteristic was targeted.

While several stakeholders in Ethiopia were shocked at these findings, their reactions aligned with discussions CIR held with CSOs and human rights defenders in Ethiopia regarding the normalisation of, and desensitisation to, 'less severe' forms of TFGBV in Ethiopia. For example, many workshop participants expressed a shared belief that hate speech composed of insults, presumed association, or alleged inferiority are often overlooked. One participant suggested that these forms of hate speech are considered less harmful, and thus may be less often reported or recognised. Despite working in the human rights sector, several participants indicated that they would not have classified 'insults', 'presumed association', or 'alleged inferiority' as hate speech, and therefore suggested the need for more targeted, in-depth educational campaigns.

During workshop discussions, a key finding was raised from CIR's first study of TFGBV in Ethiopia which interviewed survivors of gendered abuse; namely that almost all TFGBV survivors interviewed believed that gendered hate speech on social media was so widespread that it had

become normalised and, as a result, society had become desensitised to the issue. There was a shared view among many workshop participants that this normalisation, coupled with the lack of understanding about what can constitute hate speech, results in a lack of reporting of less aggressive or threatening hate speech. These findings reaffirm the need for education on the various forms of hate speech, including those that seem less harmful at first glance.

The findings also reveal the need for broader education on the role of women and girls and women and girls' rights. For example, women and girls faced a significant proportion of stereotypical, gendered hate speech. Similarly, survivors interviewed in CIR's first study reported that they received gendered abuse, including misogynistic insults related to their physical appearance and their role within society, while they claimed that men were more likely to receive abuse related to their political views. Improved education about gender roles and gendered abuse in schools, workplaces, and religious or political institutions is crucial to challenging gendered stereotypes in Ethiopia, and ensuring the full and meaningful participation of women and girls in public life, both online and offline.

4.2 Women and girls only, by platform

All three platforms included in this study have policies forbidding users from engaging in gendered hate speech. However, each platform has different mechanisms for content moderation, different numbers of moderators, and, therefore, a different approach to combatting the issue. Each platform also has a different demographic of users. According to the Digital Ethiopia report, only 43.1% of social media users in Ethiopia in 2023 were female. Data on the gender split by platform is not readily available for Ethiopian users, however ad viewership data can be used as a rough, proxy indicator. Using these data, women and girls make up only 34.2% of Facebook and 16.7% of X users. Similar data is not available for Telegram.

In addition, users interact with each social media platform differently, according to the functions they serve in Ethiopian society. In CIR's first study, interviewees reported using different platforms for distinct reasons, often without uniform answers. This included using platforms as an information source, sites to share their own opinions, locations to share their work, sources of entertainment, locations to engage in advocacy, or mediums to connect with friends and family. Understanding which platforms women and girls engage with, and why, can inform tailored recommendations to combat TFGBV on social media platforms.

This research found variations in the 'type' of hate speech targeting women and girls across the three platforms. Proportionally, more insults were found on X than the other platforms, while Telegram had more threats, and Facebook had more cases of 'presumed association' and 'alleged inferiority' (proportionally; see Figure 7). Across all three platforms, there was proportionally more 'offensive' hate speech than other 'sentiment' categories identified.

4.2.1. Type of hate speech targeting women and girls, by platform

Hate speech targeting women and girls presented itself differently across the three platforms investigated. X had the largest proportion of ‘insults' (48.0% of the posts identified as containing hate speech), compared to Telegram (32.5%) and Facebook (25%).

Facebook had proportionally more hate speech containing ‘presumed association’ of the protected characteristic (here, the female gender) with certain personality traits or characteristics (for example, greed) (31.7%), compared to Telegram (28.3%) and X (25.3%). Similarly, Facebook had proportionally more hate speech containing ‘alleged inferiority’ of the female gender (29.8%) compared to Telegram (23.7%) and X (16.7%).

Finally, Telegram had more threats (15.6%) compared to Facebook (13.5%) and X (10.0%). This could be due to the closed nature of Telegram, compared to Facebook and X, making it harder for content moderation. The channel owner is responsible for moderation, rather than Telegram itself.

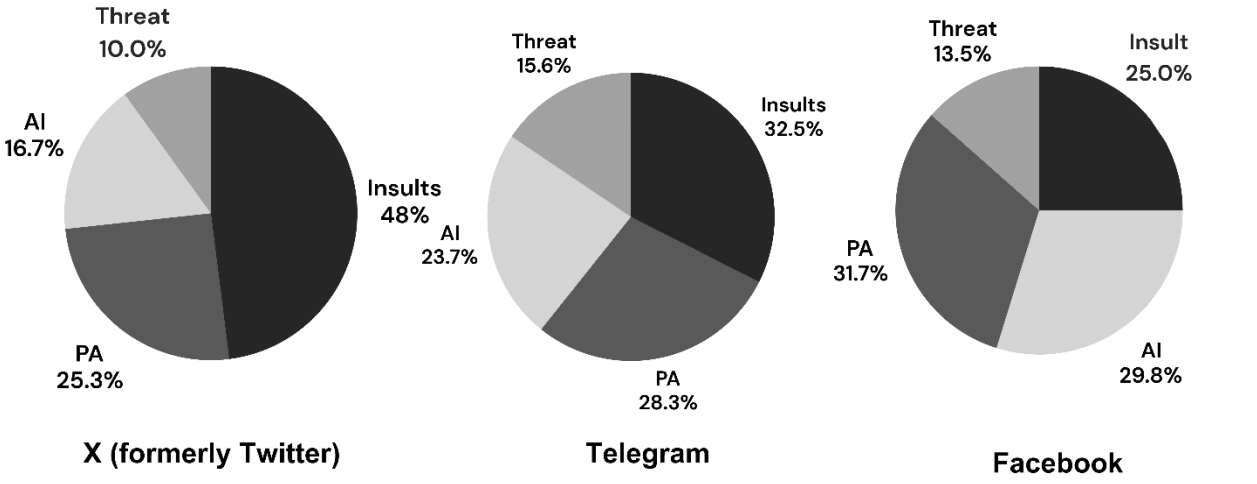


Figure 11: Pie charts showing the distribution of the type of hate speech targeting women and girls in X, Telegram, and Facebook. (PA: Presumed association. AI: Alleged inferiority).

4.2.2. Sentiment of hate speech targeting women and girls, by platform

The ‘sentiment’ of the hate speech targeting women and girls also presented itself differently across the three platforms investigated (see Figure 8). Across all three platforms, ‘offensive’ hate speech was more common than other ‘sentiment’ categories.

X had the largest proportion of stereotypical hate speech (31.6% of all hate speech posts), compared to Telegram (26.8%) and Facebook (12.5%). Telegram proportionally had more hate speech containing irony (29.1%), compared to Facebook (19.3%) and X (9.3%). Facebook had proportionally more ‘aggressive’ hate speech (12.5%) compared to X (6.7%) and Telegram (5.9%).

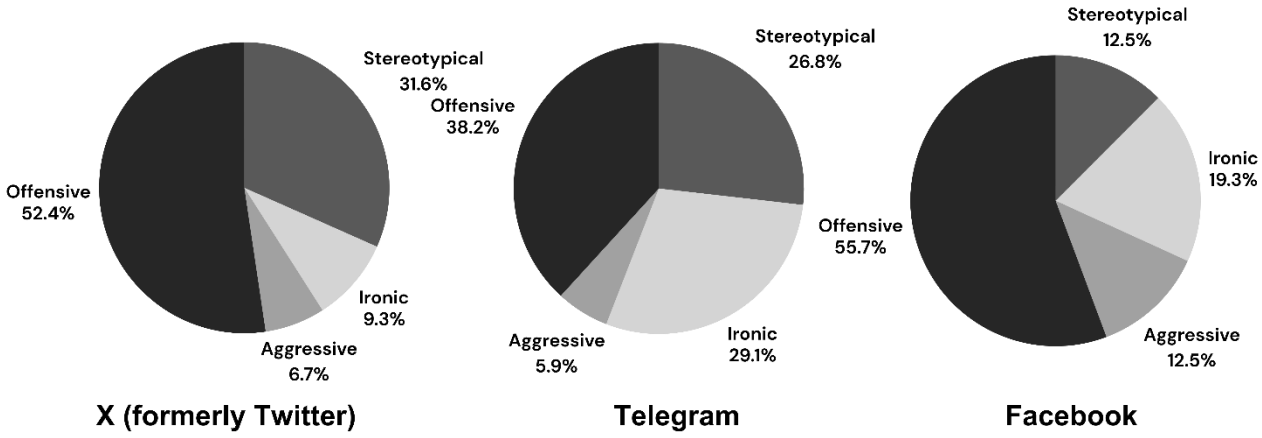


Figure 12: Pie charts showing the distribution of sentiment of hate speech targeting women and girls across, Telegram, and Facebook.

4.2.3. Discussion: hate speech targeting women and girls, by platform

In CIR's earlier study, interviewees reported that TFGBV was prevalent across the social media platforms that they engaged with. CIR's social media analysis revealed that, although gendered hate speech was prevalent on all three platforms, the sentiment and types of gendered hate speech varied by platform. Better understandings of these variations can inform more targeted solutions, as well as more tailored resources for women and girls to protect themselves in the meantime, such as bespoke digital security training for women and girls on each platform. One approach might not suit all platforms, or its users.

During the second study, workshop participants expressed concerns that, even when harmful content on social media platforms is reported, it is rarely removed. There was a shared belief that Facebook was more effective than X at removing reported content. One participant said that they had received threats on X and, despite getting many people to report it, months later it was still visible on the platform. Other participants stated that Ethiopian CSOs provided Facebook with information on trending terms and events, which feeds into the platform's monitoring efforts. While a step in the right direction, more resources are required to combat the volume of hate speech on Facebook, across the many Ethiopian languages.

Other platforms appear to be more active in removing harmful content. CIR identified 94,640,435 posts that Telegram had removed from the groups monitored for this study, each containing at least one keyword from CIR's lexicon, predominantly in the English language dataset.[20] This signals that there is an active removal of harmful content. Interestingly, after analysing the posts that had not been removed, CIR found Telegram to have proportionally more hate speech containing mockery than the other platforms. As suggested during the workshops, many

[20] See section 3.3.1.

participants did not realise that irony could classify as hate speech. This finding could suggest that hate speech containing mockery is less likely to be reported and/or removed on Telegram. Alternatively, this finding could reflect attempts by perpetrators to use irony or mockery to avoid content removal. Targeted campaigns to inform Telegram users that irony or mockery can sometimes amount to hate speech could be effective in raising awareness about the issue. For example, a community of Telegram users could be trained to report hate speech containing mockery, or even to create content to counter mockery. Tailored approaches such as this could complement broader policy solutions.

Reliable data on the types and sentiment of gendered hate speech on social media could empower platforms to improve their understanding of the issue, and adapt their content moderation and removal policies and practices accordingly. Additionally, social media community guidelines could be strengthened by reflecting a richer understanding of the different types of speech prohibited on their respective platforms. Increased public education on what constitutes hate speech, its harmful impacts, and how to report it to social media platforms could also feed into improved social media monitoring efforts. Future study could broaden the scope of this research to additional social media platforms, such as TikTok and YouTube.

4.3 Comparison of gender subgroups

Although the focus of this research was TFGBV against women and girls specifically, where hate speech directed against men was identified, it was also annotated and included in the study. Due to the methodology, CIR cannot confirm the true gender of the victim of abuse. For example, as workshop participants highlighted, abusers sometimes use female pronouns in abuse directed towards men. This must be considered when interrogating the results.

While inexhaustive, this data allows a preliminary comparison of the different type and sentiment of the hate speech received by the different subgroups. Although there were variations in the forms of hate speech received, across both gender subgroups, ‘insults’ were the most common type of hate speech identified. The data for ‘sentiment’ was far more varied.

4.3.1. Hate speech by gender subgroup

Of both the gender subgroups, women and girls received more hate (77.8%) than men (22.2%), see Figure 13). This research was designed to investigate hate speech directed against women and girls, as can be seen by the focus on gendered term in the lexicon. Resultantly, it is unsurprising that women and girls received more hate speech than men within the dataset.

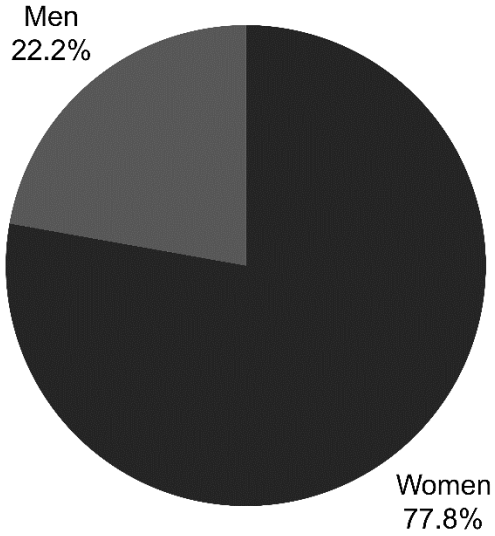


Figure 13: Pie chart showing the distribution of gender-based hate in specific gender targets.

4.3.2. Type of hate speech, by gender subgroups

Interestingly, the 'type' of hate speech identified presents itself differently across the gender subgroups (see Figure 14). For both men and boys and women and girls, there were more 'insults' proportionally, than other types of hate speech identified. Men and boys received the highest proportion of insults, approximately half of the total hate speech count. However, women received proportionally more of each other type of hate speech than men and boys. For example, Women and girls received more hate containing alleged inferiority (22.2%), proportionally, than men (13.8%). Women and girls received more threats (13.5%) than men and boys (9.2%). Women and girls also received slightly more hate (27.7%) that contains presumed association of their gender with specific characteristics, than men and boys (26.3%).

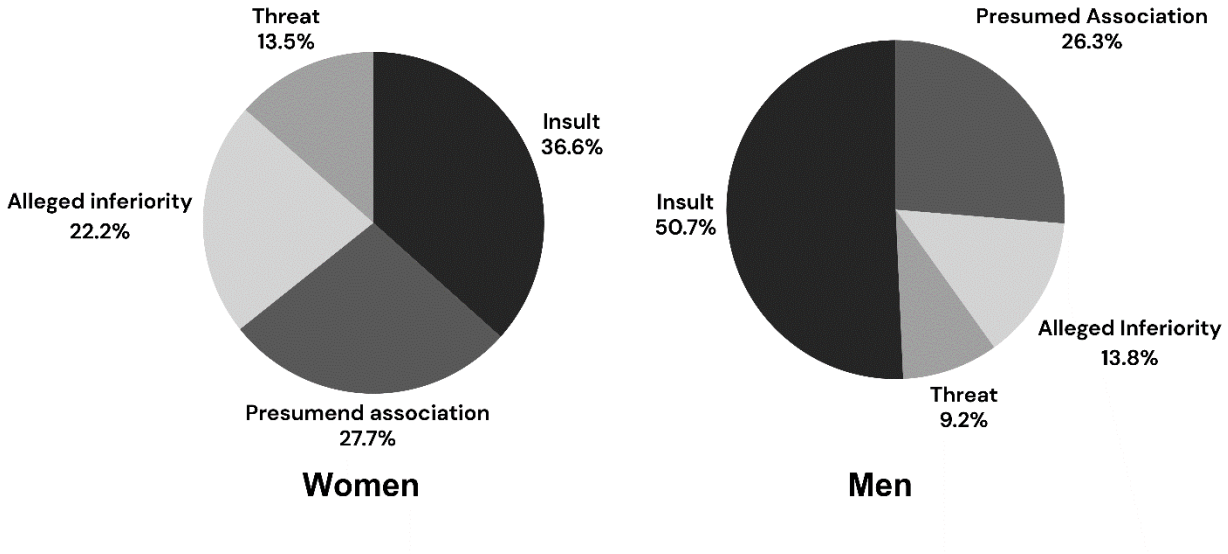


Figure 14: Pie charts showing the distribution of the type of hate speech targeting across gender subgroups.

4.3.3. Sentiment of hate speech, by gender subgroups

CIR also compared the sentiment (aggressive, offensive, mockery, stereotypical) of the hate speech by gender subgroup. This revealed large variations in the sentiment of the hate speech received by both subgroups, as can be seen by the pie charts below. For example, Men and boys receive more offensive hate speech (72.7%), than women and girls (46.2%).

Women and girls receive more stereotypical hate speech (26.2%). Although men and boys receive, proportionally, less stereotypical hate (11.7%) than women and girls, it is still the second most dominant 'sentiment' category for men and boys. There is also notable difference in the amount of abuse containing mockery that the gender subgroups face. 20.3% of the abuse women and girls receive contains mockery, compared to just 7.8% for men and boys.

Finally, aggressive hate speech is the least prevalent sentiment category identified across both subgroups. Men and boys and women and girls receive similar levels of aggressive hate (7.8% and 7.3% respectively).

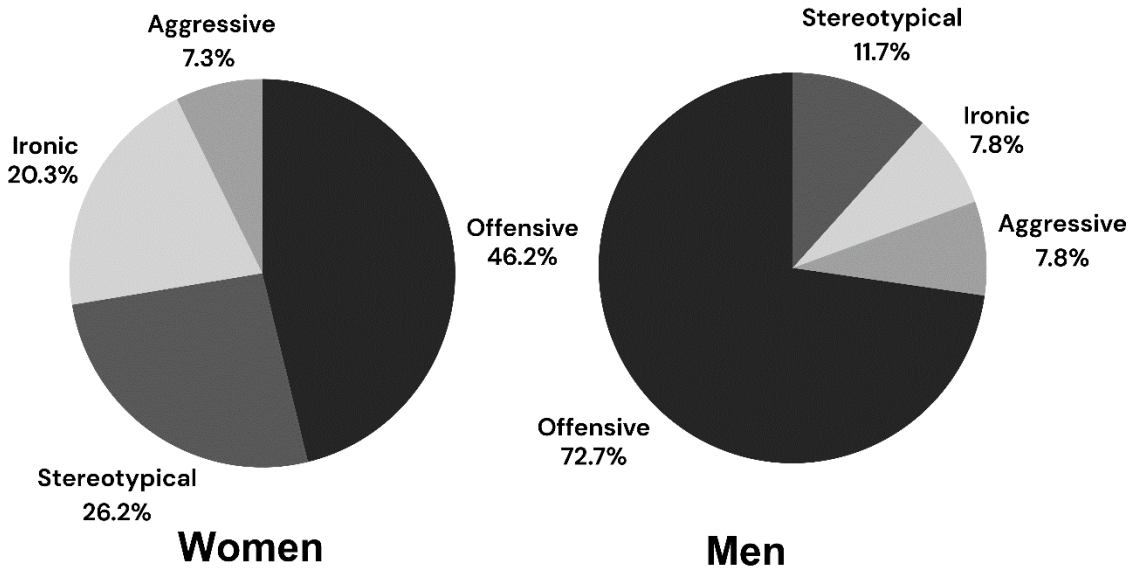


Figure 15: Pie charts showing the distribution of the sentiment of hate speech across gender subgroups.

4.3.4. Hate gender subgroup comparison, by platform

As aforementioned, across all platforms investigated, women and girls received more hate speech than men and boys. However, there are significant differences in the amount of hate speech across the platforms. Interestingly, the proportion of hate speech targeting men was much higher on Facebook (40.1%) than X (17.7%) or Telegram (3.1%).

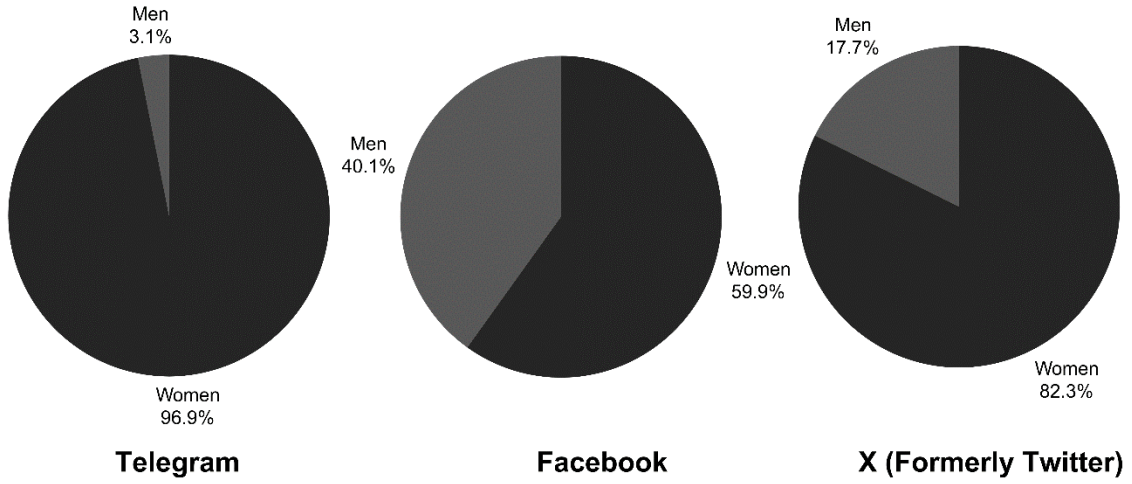


Figure 16: Pie charts showing the distribution of Gender-based Hate in X, Telegram, and Facebook Posts by Specific Gender Targets.

4.3.6. Discussion: hate speech by gender subgroups

Comparing the trends between gender subgroups is revealing. For example, men received less stereotypical hate speech, and hate speech containing mockery, than women and girls, but more aggressive and offensive hate speech. Similarly, interviewees in CIR's earlier study reported that women and girls and men received different types of hate, and that woman specifically faced stereotypical abuse centred around gender roles. Workshop participants reiterated this, stating that men are often attacked for their political views, while women and girls in the same position are more likely to receive abuse related to their appearance or role in society.

After presenting these findings during a workshop, one participant suggested that a commonly held belief in Ethiopia is that women and girls' rights and notions of feminism are often considered a Western concept, at odds with, or even degrading, Ethiopian values. Thus, abusers may criticise, abuse, or ridicule women and girls who appear supportive of women's rights or feminism online for challenging traditional roles in society. Interestingly, a discussion ensued between participants *about 'what constitutes Ethiopian values? And how does feminism degrade them?'* The consensus was that Ethiopian values differ greatly throughout the country, and gender roles differ. One participant suggested that a showcase of the broad variation in Ethiopian gender norms could be useful for breaking down the perception of women and girls' role in society.

CIR found a much larger proportion of hate speech directed at men on Facebook than any other platform. There are several reasons why this could be, none conclusive, including that Facebook has more female users than Twitter does, perhaps leading to less polarised hate environment.

4.4 Comparison with other hate speech targets

To answer questions on intersectional hate speech (when hate speech targets multiple protected characteristics), CIR investigated other identity-based hate speech, in line with the Ethiopian Government's definition of hate speech.[21] Interestingly, ethnicity was the most prevalent protected characteristic within the entire dataset, while gender was the second most targeted. A comparison of the type and sentiment of hate speech received by hate target reveals differences in the forms of abuse women and girls receive. While the proportion of 'threats' and 'aggressive' speech are lower for women and girls than other hate targets, 'alleged inferiority' and 'stereotypical' hate speech or hate speech containing 'mockery' were more prevalent for women and girls.

4.4.1. Hate speech targeting groups with protected characteristics

Although the research focused on TFGBV, and the keyword list contained more gender-related keywords (49.4 %) than ethnic (9.1%) or religious (9.3%), CIR found hate speech targeting other identity groups in the dataset. The keyword list contained terms that are deemed 'inflammatory' or could be used within hate speech. After the annotation step, only posts that classified as hate

[21] See Glossary.

speech using the annotation protocol were included in the sample for analysis. This included hate speech that targeted other protected characteristics (e.g. ethnicity, race…) Of the full dataset, ethnic hate speech comprised 44.5%, while gendered hate speech represented 30.2%, and religious hate speech was 17.5%. Racial hate and hate speech targeting people with disabilities made up a smaller proportion of the dataset (4.6% and 0.3%, respectively).

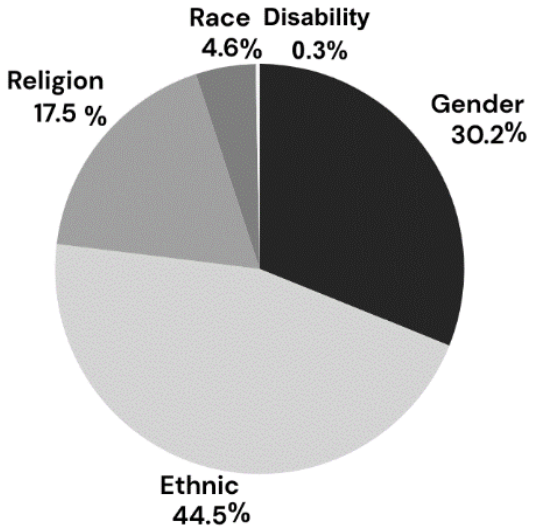


Figure 17: Pie chart showing the distribution of hate in posts by hate target.

4.4.2. Hate speech by hate target subgroups

When the protected characteristic identity groups are broken down into individual hate targets, other interesting trends become visible. Women and girls (21.1% of the dataset) received more hate speech than any other hate target subgroup, closely followed by Oromos (19.2%) and Amharas (16.8%). Again, due to the focus on gender, this is not surprising.

Other targets of hate speech within the dataset, albeit in smaller proportions, include Orthodox Christians (8.8%), men (5.9%) and Tigrayans (5.5%), followed by Muslims and black people (4.5% and 1.3% respectively).

The 'additional hate targets' category is comprised of all the other target groups outside of the top 10 targets, this includes: protestants, white people, transgender people, atheists, Arabs, multiracial people and jews. The 'other targets' category was selected in cases where hate speech targeting a protected characteristic was present, but it was harder to place within one of the categories.

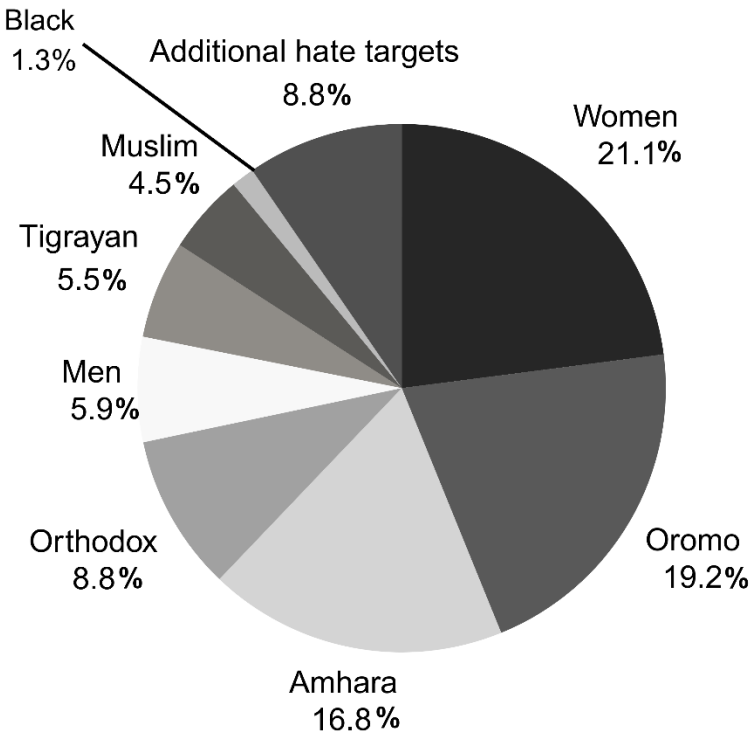


Figure 18: Pie chart showing the distribution of hate in posts by the eight top hate targets. Additional hate targets represent all other targets grouped together.

4.4.3. Comparison of type of hate speech

In addition to identifying the trends in gendered hate speech, CIR analysed how abuse targeting women and girls differs from abuse targeting other identity groups. The six most prevalent hate target subgroups have been included in this comparison.

Women and girls receive proportionally more insulting hate speech (36.6%) than Amharas (31.4%), Muslims (29.5%) and Oromos (22.5%), they receive less insulting hate speech compared to Tigrayans (45.2%). Additionally, women and girls receive proportionally more hate containing alleged inferiority (22.2%), followed by Muslims (16.4%), Amharas (13.2%), Tigrayans (13.0%) and Oromos (9.9%).

Conversely, women and girls receive (proportionally) less threats (13.5%) compared to Oromos (26.1%), Amharas (24.6%), Tigrayans (22.6%), and Muslims (18.0%). Women and girls are also among the hate target subgroups which receive proportionally less hate containing presumed association (27.7%), compared to Amharas (30.8%), Muslims and Oromos (36.1% and 41.4% respectively).

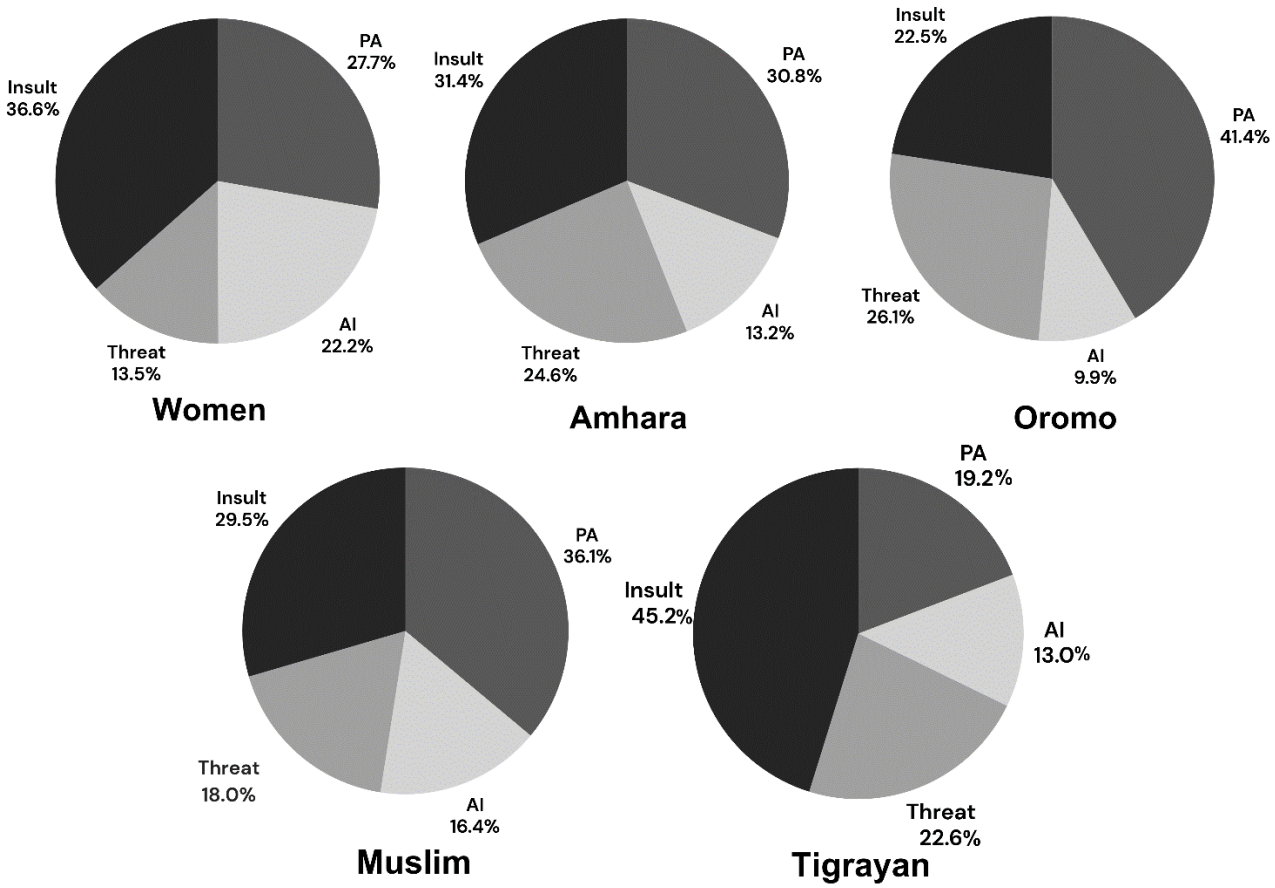


Figure 19: Pie charts showing the distribution of hate in posts by specific target (top 5) and type of hate speech.

4.4.4. Comparison of sentiment of hate speech

The 'sentiment' of hate speech received by women and girls also presents itself differently, when compared to the other hate target subgroups analysed.

Interestingly, CIR identified that offensive hate speech is the most prevalent sentiment across all hate targets analysed. Although women and girls receive a high proportion of offensive hate speech (46.2%), compared to other hate targets they receive less offensive hate speech.

Contrary to the pattern observed in offensive hate speech, women and girls receive the highest proportion of stereotypical hate speech (26.2%), followed by Muslims (12.5%), Amharas (9.3%), Oromos (5.0%) and Tigrayans (2.7%).

Only Muslims receive a higher proportion of hate speech containing mockery (21.3%) than Women and girls (20.4%). Even so, women and girls receive considerably more hate speech containing mockery than Tigrayans (11.5%), Amharas (6.9%), and Oromos (5.0%).

Lastly, CIR also observed that women and girls receive the lowest proportion of aggressive hate speech (7.3%) compared to the other target subgroups.

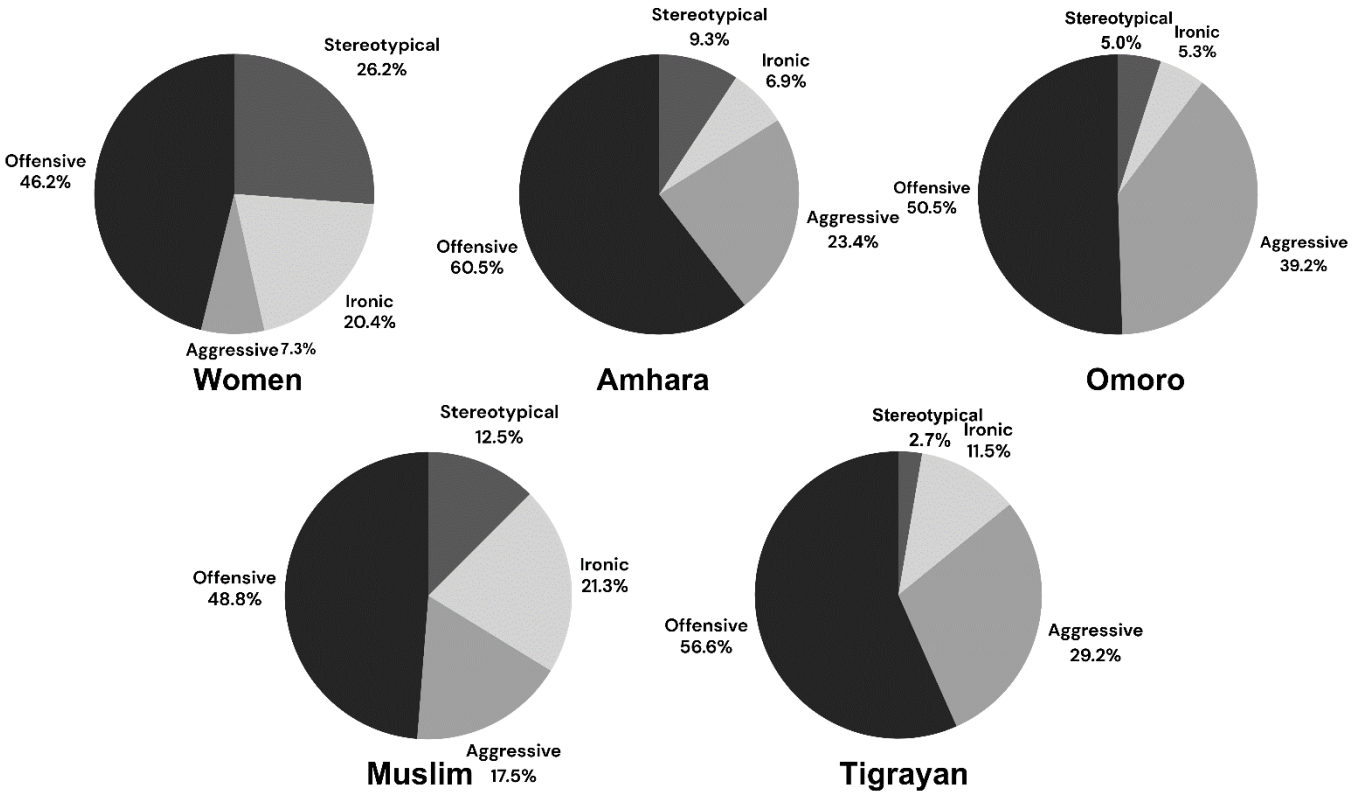


Figure 20: Pie charts showing the distribution of hate in posts by specific hate targets (top 5) and hate sentiment.

4.4.5. Discussion: comparison with other hate targets

This study reveals the differences in the type and sentiment of hate speech when along gendered lines, compared to other identity factors. Gendered hate speech presents itself differently to other forms of hate speech, often in subtle, less threatening ways. For example, women and girls receive more 'alleged inferiority' and 'stereotypical' hate speech or hate speech containing 'mockery' and proportionally less 'aggressive' or 'offensive' hate than the other subgroups, as outlined above. These themes align with those raised by the interviewees in CIR's earlier study, who claimed that women and girls receive misogynistic abuse characterised by gendered stereotypes, related to their appearance and their role in society. For example, interviewees outlined how women and girls in public positions have their hair, clothing, weight, marital status, suspected lovers, and number of children debated on social media and acknowledged that these

things are not discussed in relation to men in public roles. While this might not appear as damaging as 'threats' or 'aggressive' speech, the reinforcement of gender stereotypes can prevent access to opportunities for women and girls and provide a barrier to gender equality. The interviewees also reported that the abuse was so widespread it was normalised.

It's clear that gendered hate speech, as one form of TFGBV, is not the only type of hate speech prevalent in Ethiopia. CIR investigators were aware that ethnic and religious hate speech is a real concern to those working in the human rights sector within the country, due to its prominence. This study did not seek to disrepute this, and in fact, demonstrates that there is a high prevalence of ethnic and religious hate speech. The differences in the types of hate each group receives could provide an entry point for targeted policy solutions. Conversations during workshops in Addis Ababa cemented the need to include these trends within the analysis, despite broadening the scope beyond purely gendered hate.

Current events also influenced the study. While annotating the dataset, it became apparent that the conflict in Amhara was a key driver of online dialogue that contains inflammatory terms from CIR's lexicon. Similarly, CIR identified hate targeting Prime Minister Abiy Ahmed, much of which cited his Oromo ethnicity and Oromo allegiances. These forms of online abuse, which are reactive to political events and inflammatory in nature may be more apparent than gendered hate speech, due to the different nature. Gendered abuse, in the form of stereotypes and the suggestion of inferiority appears to almost go under the radar. Workshop participants expressed a belief that it is so endemic that it has become normalised to the point of invisibility.

All forms of hate speech can have a lasting impact on the recipient, at the individual or group level. While this research specifically focusses on gender and seeks to provide recommendations to combat TFGBV, some of the recommendations may also serve to improve the online environment for other hate targets.

4.5 Intersectional, gendered hate speech

To understand the sentiment of hate speech levelled at women and girls along intersectional lines, CIR investigated hate speech that targeted women and girls, alongside another protected characteristic. The most common combinations of hate targets, as well as the type and sentiment of hate speech women and girls received are discussed below. Unsurprisingly, the type and sentiment of hate speech presents itself differently depending on which protected characteristic appears alongside gender in the same post.

4.5.1. Intersectional gendered hate speech

Within this dataset, women and girls faced intersectional hate speech along six different intersectional identity lines, as can be seen in the pie chart below. This included two ethnic subgroups (Oromo and Amhara), one racial subgroup (Black), and one religious subgroup (Orthodox).

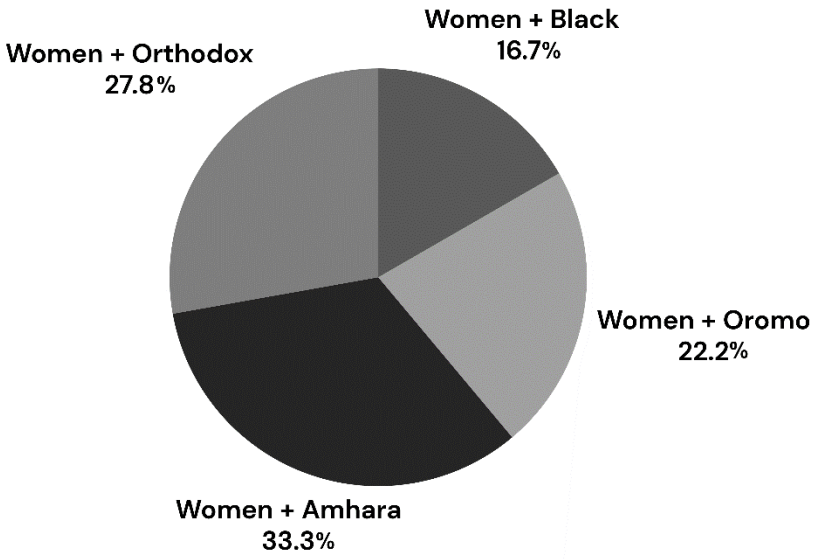


Figure 21: Pie chart showing the distribution of intersectional hate speech faced by women and girls alongside other protected categories.

4.5.2. Type of intersectional, gendered hate speech

When the type of hate speech is broken down, there are clear variations in the composition of the hate speech the women and girls received. Unlike in the data for 'Women and girls' only (Section 4.1 above), 'insult' was not always the most prevalent type of hate speech.

For the two ethnic and gendered hate speech subgroups (Amhara and Oromo), 'alleged inferiority' was the most prevalent type of hate speech. i.e. the women and girls were perceived as not equal, or subordinate, due to their gender and ethnicity, in relation to either their social position, credibility, cognitive or physical ability, or desirability.

For the Orthodox and gendered hate speech subgroup, 'presumed association' was the most prominent type of hate speech. i.e. assumptions were made about the women and girls' personal characteristics and integrity due to their gender and religious affiliation.

Unlike other types of hate speech identified in this study, 'threats' were prominent across a few of the intersectional, gendered hate speech subgroups. For example, 'threats' made up 25.0% of the 'women and girls + orthodox' category, and 22.5% of the 'women and girls + black' category. When women and girls alone are the targets of hate speech, as seen in section 4.1 above, 'threats' only comprise 13.5% of the hate speech received. Thus, women and girls who are labelled or identify as Orthodox or black, may be at higher risk of receiving threats, than other identity groups. However, these findings are of course subject to the caveats of the research methodology.

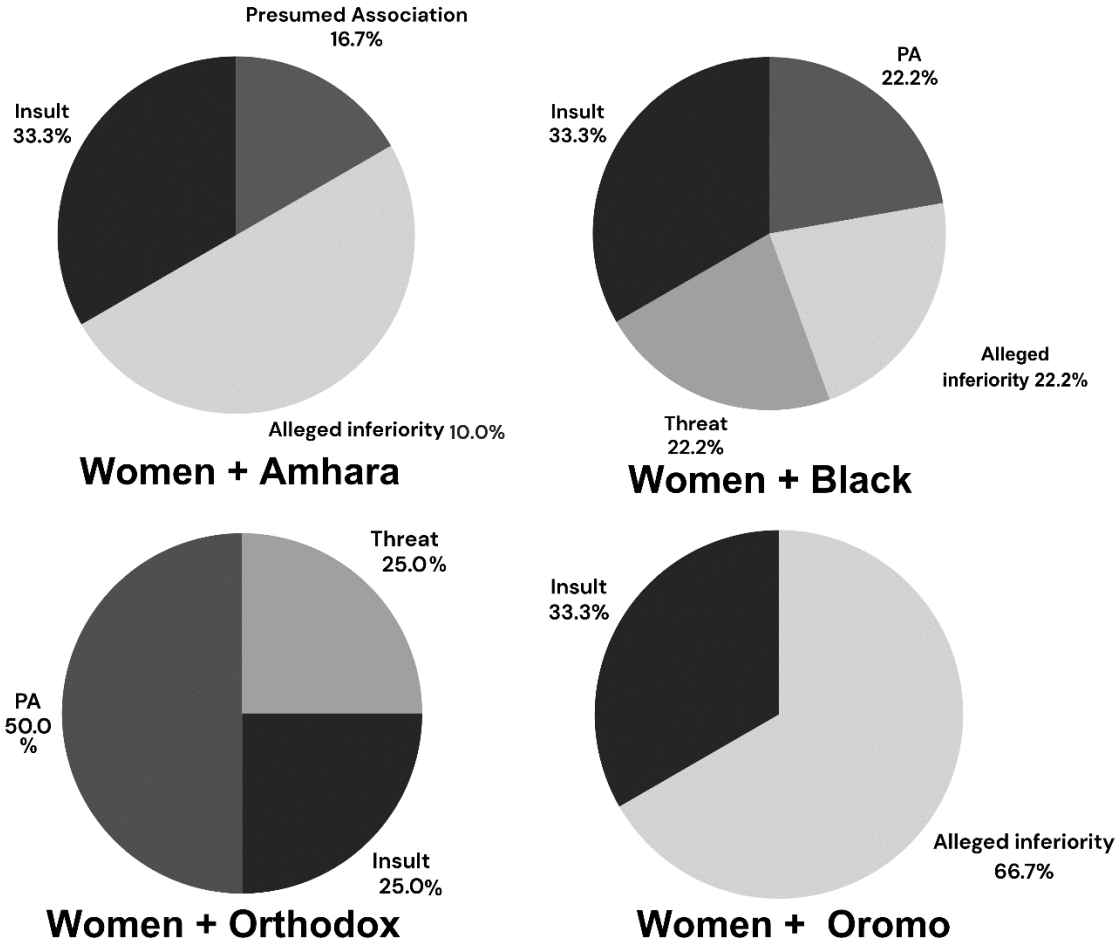


Figure 22: Pie charts showing the distribution of the type of hate speech faced by women and girls alongside other protected categories.

4.5.3. Sentiment of intersectional, gendered hate speech

Similarly, when the 'sentiment' of the intersectional, gendered hate speech is investigated, clear variations are also present. In line with the findings for hate speech targeting women and girls (Section 4.1.2), 'offensive' speech is the predominant category.

'Offensive' speech was the only 'sentiment' category for the 'Women + Oromo' subgroup. In contrast, it accounted for only 57.1% of hate speech directed towards the 'Women + Amhara' subgroup. This disparity is noteworthy, considering there are conflicts in both the Oromia and Amhara regions.

There are, however, other notable findings. For example, the proportions of stereotypical hate speech varied across the different subgroups, with 'Women + Oromo' subgroup receiving none, while it constituted one third of the hate speech targeting the 'Women + Black' subgroup. Aside from the 'Women + Orthodox' subgroup, the subgroups contained less hate speech containing

mockery proportionally, compared to the women and girls only data. Finally, there was a higher proportion of aggressive hate speech for the 'Women + Amhara' and subgroup (14.3%) than the women and girls only data.

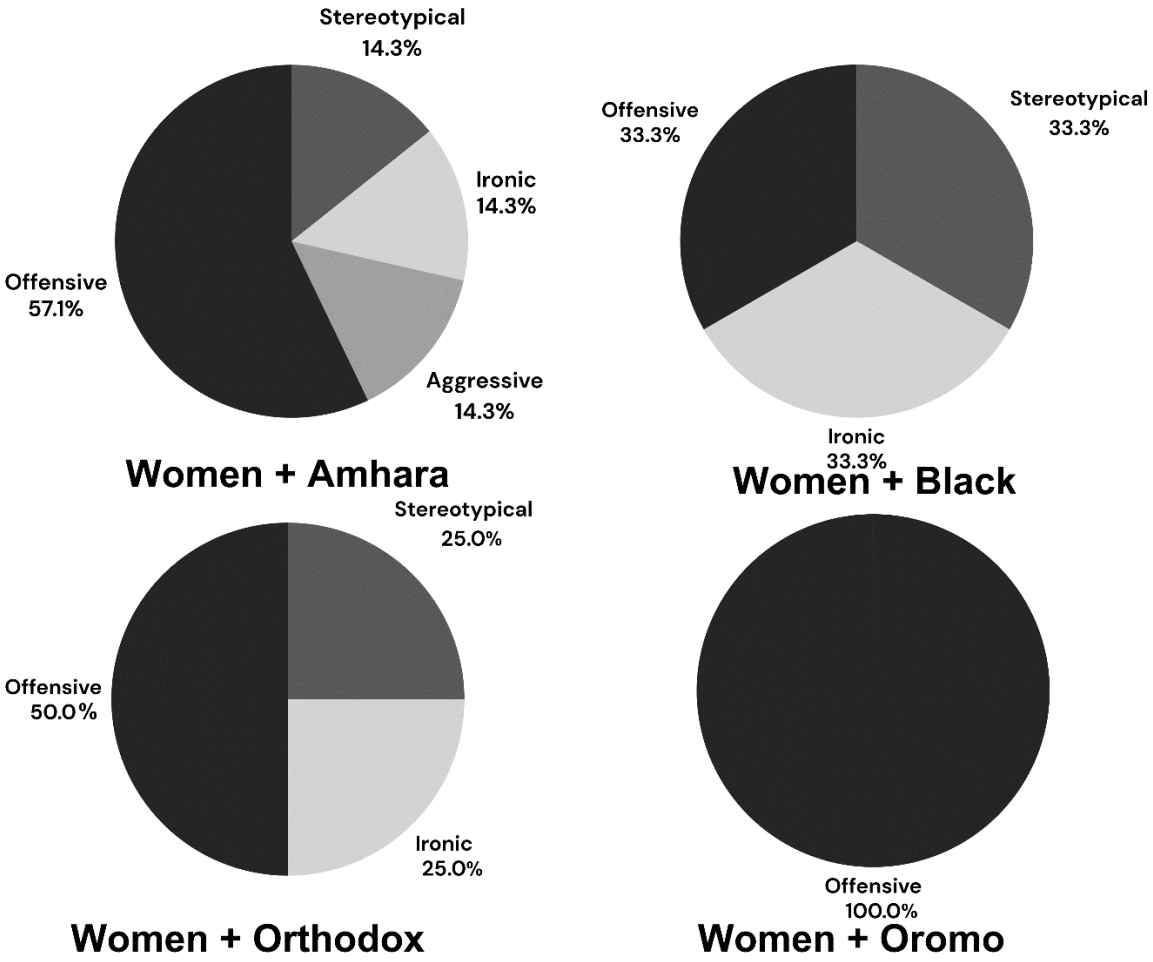


Figure 23: Pie charts showing the distribution of the sentiment of hate speech faced by women and girls alongside other protected categories.

4.5.4. Discussion

Interestingly, there are differences between hate speech targeting 'women and girls' alone and intersectional hate speech. Political events have a considerable impact on the topics of discussion online and the rhetoric used. As aforementioned, during the research period, conflict in Amhara was dominating online debate. Abuse levelled against Amhara women and girls may well be influenced by the conflict. A similar comment applies for Oromo women and girls: conflict and political events may be driving this online trend. Despite this, there were notable differences in the sentiment of the abuse levelled against Oromo and Amhara women, signalling that conflict or political events may not be the sole explanation. Amhara women received more aggressive, or stereotypical hate and hate speech containing mockery than Oromo women in this study. This

signals the complexity of the information environment. Future study could investigate the factors at play that result in these differences. Similarly, it would be interesting to assess whether there was a rise in the abuse levelled against Tigrayan women and girls during the Tigray conflict.

Similarly, religious hate speech is prominent in the Ethiopian context. Not only was there a schism within the Orthodox church during the research period, but much of the rhetoric seen online has religious undertones. This reinforces the belief that current events, offline, have a direct impact on abuse online. As such, when political, ethnic or religious tensions flare up offline, policy makers should be prepared to combat spikes in abuse against specific groups, online. Thus, both preventative and responsive measures are needed to combat hate speech. During one workshop, a participant noted that during times of conflict or unrest, rather than combatting the rise in hate speech the current policy is to shutdown the internet. The participant believed that this was detrimental to peace, as it inflates the voices of the diaspora, who are often criticised for stoking tension. In almost all workshops, participants asked if CIR could see which posts originated within or outside Ethiopia and thus, if the role of the diaspora could be assessed. Unfortunately, this was not possible due to the methodology selected, however future study should take this consideration into account.

4.6 Accusations of homosexuality

In roundtables, workshops and CIR's earlier study, individuals and interviewees reported being accused of being homosexual if they spoke about women's rights or feminism. For example, just under 29.0% of CIR's interviewees in an earlier study reported being labelled as a homosexual by their abusers. Multiple interview participants who advocate for women and girls' rights in online forums reported being accused of homosexuality. They believe that the abusers used the accusation of homosexuality as not only an insult, but a method to impose significant threats to their safety, freedom, and ability to speak in public forums. Homosexuality is a taboo subject within Ethiopia; it is illegal and carries a penalty of one year imprisonment. As such, these claims are not without risk to the individual target.

During this study, CIR identified hate speech targeting women and girls that also included references to, or accusations of, homosexuality. When this data was interrogated, it became clear that hate speech which includes accusations of homosexuality differs from hate speech targeting women and girls alone. For example, hate speech targeting women and girls that includes references to homosexuality contained far more 'threats' (50%) than hate speech targeting just women and girls (13.5%). Additionally, hate speech targeting women and girls that includes references to homosexuality contained far more 'aggressive' speech (40%) than hate speech targeting just women and girls (7.3%).

These findings provide further evidence of a trend reported by the interviewees in CIR's earlier research and subsequent roundtables and workshops in Addis Ababa. Women and girls who are labelled or accused of, homosexuality, appear to be at higher risk of receiving 'threats' and 'aggressive' hate speech than 'women and girls' alone.

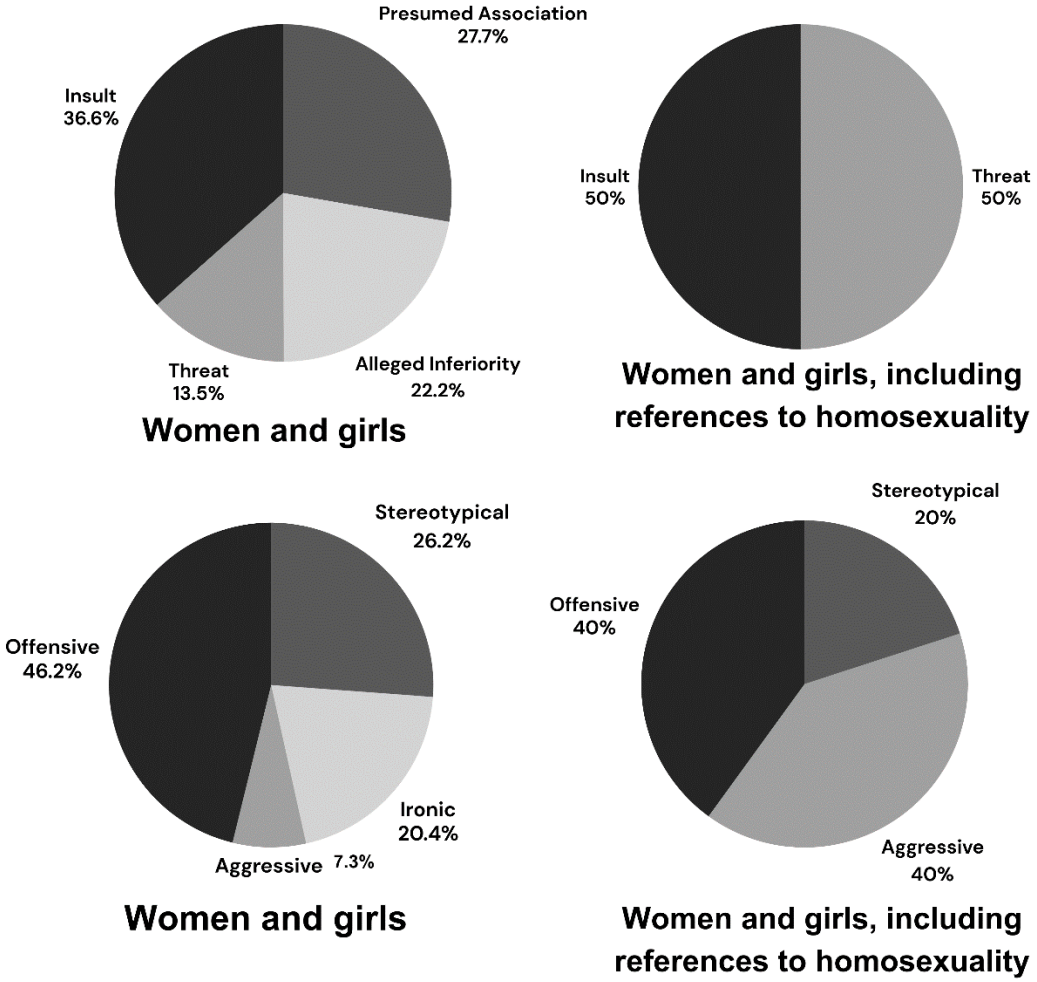


Figure 24: Pie charts showing the distribution of the type and sentiment of hate speech faced by women and girls, as well as women and girls when the text also includes references to, or accusations of, homosexuality.

5. Conclusion

As people's personal and public lives are increasingly played out, in part, on the internet and through social media, a new frontier in the fight against gender-based violence has emerged. This research has sought to strengthen the evidence base on TFGBV in Ethiopia, by investigating the types (insults, threats, presumed association, alleged inferiority) and sentiment (offensive, stereotypical, mockery or aggressive) of hate speech targeting women and girls and girls on three social media platforms: X, Telegram and Facebook. CIR also identified hate speech targeting other gender subgroups, allowing a comparison of the forms of hate speech each subgroup receives. To answer questions on intersectional, gendered hate speech (when hate speech targets gender alongside another protected characteristic), CIR also investigated hate speech along ethnic, religious and racial lines.

Both this and CIR's earlier study (Silenced, shamed and threatened: TFGBV targeting women who participate in Ethiopian public life) found that women and girls receive different types of online abuse than men and boys do. Interviewees in the first study reported that women and girls often face stereotypical abuse, centred around gender roles and laced with misogyny, while abuse against men and boys often focuses on expressed views or politics. Similarly, the social media analysis in this study found that women and girls receive more hate speech which includes gendered stereotypes and mockery or irony, than men and boys, and less aggressive hate speech.

Discussions during roundtables and workshops revealed that hate speech is often misunderstood, leading to certain forms of TFGBV being overlooked. Additionally, certain forms, such as the use of gendered stereotypes, insults, demeaning language, or speech implying women and girls' inferiority to other genders, are often considered less harmful than 'threats' or 'aggressive' speech, or may not even be considered hate speech at all. Better education on what constitutes hate speech, and its impact, is therefore essential in Ethiopia.

Many of the interviewees in CIR's earlier study reported that they saw TFGBV on all social media platforms that they engaged with. While this study found similarities in the sentiment of the hate speech targeting women and girls, there were variations in the types of hate speech seen across the three platforms. For example, proportionally, more insults were found on X than the other platforms, while Telegram had more threats. Facebook had more instances of hate speech which associated gender with certain personality traits (e.g. greed) or suggested the inferiority of women and girls' social position, or cognitive or physical ability due to their gender.

Despite variations, the findings reveal that TFGBV is present on all three platforms investigated, targeting not only women and girls, but other gender subgroups also. As interviewees told CIR, different platforms are used for different purposes. Understanding platform variations (both in their function to the user and the hate speech present) can inform more targeted solutions, as well as more tailored resources for women and girls to protect themselves in the meantime. One approach might not suit all platforms, or its users. Future study could investigate the targets, types and sentiments of hate speech on other popular platforms, including YouTube and TikTok; sites that interviewees also reported frequent TFGBV.

Analysis of intersectional abuse reveals that the risks associated with being female online may be compounded when other protected characteristics are also targeted. For example, women and girls who receive gendered hate, as well as ethnic hate, may be at higher risk of receiving threatening and aggressive hate speech than women and girls who are targeted solely for their gender. The findings from this study also support the view that current events offline impact online debate and hate speech. This can be seen by the relatively high prevalence of intersectional abuse targeting women and girls of Amhara and Oromo ethnicities, compared to other ethnicities, in the context of active conflict in these areas of Ethiopia during the data collection time frame.

The findings provide a unique snapshot of the composition of hate speech when targeting not only women and girls, but also other identity groups. Women and girls were more likely to receive abuse which suggests their inferiority, contains gendered stereotypes, or irony and mockery than ethnic or religious hate targets, and less aggressive hate. While this study focuses on gendered hate speech, ethnicity was the most targeted protected characteristic within the entire dataset, while gender was the second. By collecting data on intersectional hate speech (targeting gender and at least one other protected characteristic), CIR was able to analyse how the type and sentiment of hate speech targeting women and girls varies when it also targets another identity they hold, such as ethnicity or religion. Hate speech targeting ethnic or religious groups, which are reactive to political events and use inflammatory rhetoric, may be more apparent than gendered hate speech. Gendered abuse, in the form of stereotypes and the suggestion of inferiority appears to almost go under the radar. Workshop participants expressed a belief that it is so endemic that it has become normalised to the point of invisibility.

This report seeks to highlight the forms of gendered hate speech on social media platforms, which actively contribute to the further marginalisation of women and girls in Ethiopia. In CIR's earlier study, Ethiopian women interviewed by CIR reported that the online abuse they faced left them feeling silenced, with many withdrawing from public spaces, both online and offline, as a result. Cultivating safe online environments for women and girls is essential to empowering their full and meaningful participation in public life, both online and offline. To have a lasting effect, any strategies to combat TFGBV must address its root causes. This includes countering gender stereotypes and gender-based discrimination and promoting women and girls' representation in all public spaces.

6. Recommendations

To accompany this report, CIR has worked with stakeholders in Ethiopia to create a policy and community-led recommendations whitepaper.

This study reveals that hate speech differs depending on the target. Understanding these differences provides an entry point for targeted policy solutions to better safeguard women and girls online. CIR hopes that government institutions can use the findings to inform decision making, that social media companies can use them to inform their content moderation efforts, that civil society can use them in their advocacy, and that the public can use them to call for action.

7. Appendices

Glossary

Natural language processing, inter-annotator agreement

Term	Definition
Hate Speech (Ethiopian Government's definition)	Speech that deliberately promotes hatred, discrimination or attacks against a person or a discernible group of identity, based on ethnicity, religion, race, **gender,** or disability.[22]
Online abuse	Online abuse is a broad term which encompasses many types of harmful behaviours that occur on the internet. The 'Online Harassment Field Manual' published by PEN America, defines online abuse as "pervasive or severe targeting of an individual or group online through harmful behaviour." This includes, and is not limited to, acts such as hate speech, doxing, and sexual harassment.[23]
Gender-based violence	"[H]armful acts directed at any individual or a group of individuals based on their gender. It is rooted in gender inequality, the abuse of power and harmful norms".[24]
Technology-facilitated gender-based violence (TFGBV)	"[A]n act of violence perpetrated by one or more individuals that is committed, assisted, aggravated and amplified in part or fully by the use of information and communication technologies or digital media against a person on the basis of their gender."[25]
Inter-annotator agreement (IAA)	"A measure of the level of agreement or consistency between two or more human annotators or coders when assessing or labelling the same data or making judgments about the same set of items".[26] It is also known as inter-rater reliability or inter-coder reliability.

Annotation protocol

To limit the impact of individual biases during the data annotation process, CIR asks each annotator to follow these annotation protocol, which outline the different variables that should be considered during the

[22] Federal Negarit Gazette (2020) Available at: https://ethionab.org/wp-content/uploads/2022/09/1185_2020_HATE_SPEECH_AND_DISINFORMATION_PREVENTION_AND_SUPPRESSION_PROCLAMATION_.pdf

[23] PEN America (n.d.) Online Harassment Field Manual, Available at: https://onlineharassmentfieldmanual.pen.org/defining-online-harassment-a-glossary-of-terms/

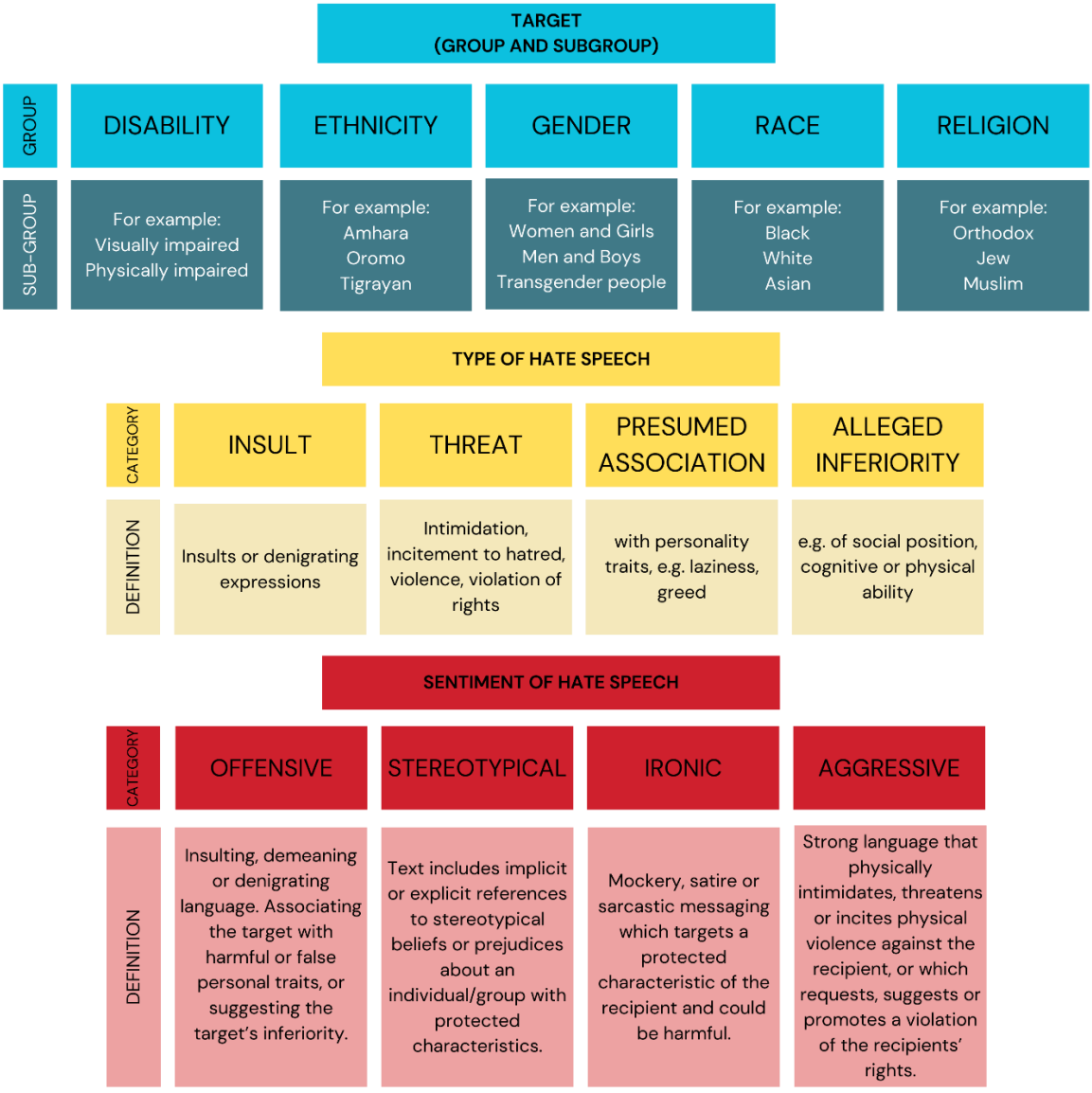
[24] UN Women and girls in GBV Doc

[25] UNFPA (n.d.) Available at: https://www.unfpa.org/sites/default/files/resource-pdf/TFGBV_Brochure-1000x560.pdf; UNFPA (2021) Technology-facilitated Gender-based Violence: Making All Spaces Safe, Available at: https://www.unfpa.org/publications/technology-facilitated-gender-based-violence-making-all-spaces-safe

[26] Kits.ai,"Inter-annotator Agreement IAA". Available at: https://www.kits.ai/glossary/inter-annotator-agreement-iaa

labelling of online posts. Please read these guidelines carefully and refer back to them frequently during the study.

If a post contains hate speech (as defined above), three elements need to be labelled using the annotation tool: the target, the type of speech and the sentiment of speech. The categories pertaining to each of these elements are visually depicted in the diagram below.



NOTE TO ANNOTATORS:

The views expressed during this research are not those of the investigators and the research team do not condone the opinions expressed. As the dataset contains real messages/posts sent by users on Telegram, X (formerly Twitter) and Facebook, this research may expose the annotators to offensive and harmful

content. Please take steps to protect yourself from the impacts of vicarious trauma. For example, please take regular breaks. There are resources to help you saved in Egnyte, in a folder titled 'Vicarious Trauma'.

Target of hate speech

A post that contains hate speech is targeted towards an individual or group with a protected characteristic. To capture this information, the words that convey which individual/group is being targeted should be assigned any of the following labels (in line with the Ethiopian Government's definition of hate speech):

- **Gender:** An individual or group of people of a particular gender.
- **Ethnicity:** An individual or group of people who come from a particular place of origin and culture.
- **Religion:** An individual or group of people belonging to a particular religious group.
- **Race:** An individual or group of people possessing distinctive physical traits associated with a particular race.
- **Disability*:*** An individual or group of people possessing a particular disability.

Example A1: Stupid women and girls shouldn't talk about political issues. Someone should throw acid at her face, maybe then she will shut up.

In the above example, the target should be labelled as **Gender** since the words "women and girls", "her" and "she" indicate that women and girls are the target of the hate speech.

Example A2: @airline how about donating flights to deport the Invaders back to their homeland. #DeportThem

In the above example, the target should be labelled as **Ethnicity** since the word "Invaders" indicates that the hate speech is intended for a group of people whose country of origin is different from that of the speaker (even if the specific group is not explicitly mentioned).

Type of hate speech

A post that contains hate speech uses language that spreads, incites, promotes, or justifies hatred, discrimination, dehumanisation, intimidation, or violence towards the target (individual/group). To capture this information, the words that convey such language should be assigned any of the following labels.

2.1 Insult: Insults or denigrating expressions against an individual/group due to protected characteristics.

Example B1: *Fucking clueless* women and girls should stay in the kitchen and not ruin a good man's name.

2.2 Threat: Intimidation, threats or incitement to hatred, violence or violation of individuals' rights, due to protected characteristics, such as:

- bodily harm and threats to physical safety (to individual or family members)
- rape threats or sexual harassment
- image-based sexual abuse (also referred to as 'revenge pornography')

- death threats
- arbitrary arrests
- restriction of access to services
- doxing
- online harassment
- creation of videos/memes

Example B2: I'll fucking *kill* the RELIGIOUS_GROUP pig.

2.3 Presumed Association: Presumed association of protected characteristics with any of the following negative connotations:

- propensity to crime or violence
- laziness or other vices such as greed, alcoholism, cleanliness level
- fundamentalism or terrorism
- invasion or conquest of another location
- threat to national security
- threat to welfare/competitors in distribution of resources

Example B3: Fucking ETHNIC_GROUP *always take more than they deserve*. Scroungers!

2.4 Alleged Inferiority: References to the alleged inferiority (or superiority) of individual/group with a protected characteristic, for example, in relation to:

- social position
- credibility (e.g. defamation)
- cognitive ability
- physical ability
- ability to engage in societal activities, e.g. politics
- undesirable or inferior lifestyle/cultural practises
- dehumanisation or association with animals or entities considered inferior

Example B4: They ETHNIC_GROUP shouldn't be allowed to vote, *they don't contribute to our society*. They must be silenced.

Sentiment of hate speech

The sentiment of hate speech must be labelled as any of:

3.1 Aggressive: This includes strong language that physically intimidates, threatens or incites physical violence against the recipient, or which requests, suggests or promotes a violation of the recipients' rights.

Example C1.a: I'll fucking *kill* the RELIGIOUS_GROUP pig.
Example C1.b: Stupid women and girls shouldn't talk about political issues. *Someone should throw acid at her face*, maybe then she will shut up.

3.2 Offensive: This includes several different forms of speech, from insulting, demeaning or denigrating language, to associating the target (individual or group) with harmful or false personal traits, or suggesting the target's inferiority.

Example C2: *Stupid immigrants... they come and take our resources.* They make us weak. They are the enemy.

3.3 Mockery: This includes jokes, satire or sarcastic messaging which targets a protected characteristic of the recipient and could be harmful. Hateful content is sometimes conveyed using nuances in language, such as sarcasm, mockery, or satire. Previous studies have expressed the importance of not overlooking this form of hate speech.

Example C3: @xxxxx Ok hoe or whore *you choose sweetie?*

3.4 Stereotypical: Text includes implicit or explicit references to stereotypical beliefs or prejudices about an individual/group with protected characteristics.

Example C4: Fucking clueless women and girls *should stay in the kitchen* and not ruin a good man's name.

Please note, the labels are not mutually exclusive. Something can be both offensive and stereotypical.

Implementing the annotation protocol: rules

Dialogue between the annotators and data engineers led to the creation of a series of rules, outlined below, to ensure consistency during annotation.

Rule 1: Take the text at face value

Investigators should not infer meaning from the text; the text must be taken at face value. For example, although the following messages from the dataset contained offensive language, the investigators were advised not to annotate the text as they do not make sense.

> *Pitchfork Dirty van bitches with all that his bagpipes again, and, to give a toast is pain Surviving on toast is your skin*

> *The highest high, I'm posed to a dyke bitch 'til she wanna Mac go I don't give me scarred I wake*

Similarly, if the target is unclear, investigators were advised not to annotate. The following examples exemplify this challenge.

> *It's just so hard to let you dusty pieces of crap to do the bare minimum when my bitches show out every fucking time….*
>
> In the above example, it is unclear whether the individuals targeted here are being targeted because of a protected characteristic.

You are not muslim so crime lenesu normal new

The above example translates roughly to “you are not muslim so crime is normal for them”. While ‘crime is normal for them’ could fall under ‘presumed association’ of a protected characteristic with crime, it is unclear who the target is. This may have been meant as hate speech, albeit written with a number of grammatical errors, however, no inferences were allowed.

Yehen aswged ant 666 new

The above example translates roughly to: ‘Terminate/finish this one, it is 666’. While ‘666’ implies satanism, and it is inciting violence, the target is unclear.

As this investigation relies solely on textual information, context is lost. As a result, the sentiment of the post may not be captured, such as irony. Additionally, abusive terms may have multiple meanings. For example, the term ‘public toilet’ is used in gendered abuse. According to CIR workshop participants, this term is used against women and girls who people consider sexually promiscuous. Everyone can use a public toilet. Thus, the use of this word against women and girls implies that anyone can use them and that they are unclean/impure. This is both offensive and derogatory. Without context, however, it is hard to tell if the term ‘public toilet’ is being used against a woman, or in a different context altogether. As a result, the following example could be gendered abuse, complaining about a woman moaning during intercourse, or it could be talking about a public toilet:

They should make a public toilet that isn’t so miserably loud when it flushes

This rule may have resulted in hate speech being excluded from the study. While this could be seen as a limitation, it prevented any personal biases from impacting the annotation process. This ensures that the hate speech dataset that is created is robust and reflects the current definition as set out by the Ethiopian Government.

Rule 2: The importance of protected characteristics

For a post to be classified as hate speech, it should target a protected characteristic (gender, race, religion, ethnicity or disability) under the current Ethiopian Government definition.

Sometimes the text was highly offensive; however, if no protected characteristics were targeted, then this does not classify as hate speech under the current definition. For example, the following is not hate-speech:

@USERNAME graduated from cunt university with honors with a pHd in serving

The annotation task revealed a lot of hate, threats and incitement to violence directed against President Abiy and his political party, or other political organisations. However, political views and affiliations are not protected characteristics under the Ethiopian Government’s definition of hate speech. This only classifies as hate speech when a protected characteristic is mentioned. For example, the following is not hate speech:

@USERNAME: #AmharaGenocide by the facist #AbiyAhmedAli #JusticeForEthiopia

However, if the above example made reference to Abiy's ethnicity, then this would classify as hate speech. For example:

> *@USERNAME: #AmharaGenocide by the facist Oromo #AbiyAhmedAli #JusticeForEthiopia #Oromofacism*

Similarly, a Telegram/X (formerly Twitter) user may be targeted for their views. 'Viewpoints' (like political affiliations) are not protected characteristics. This does not classify as hate speech under the current definition unless a protected characteristic is targeted.

Sometimes posts contained information and misinformation about war crimes and human rights violations inflicted on an ethnic/religious/gender group and/or committed by certain armed/political groups. While this could amount to incitement to hatred against a particular entity, this is not hate speech unless there is also a protected characteristic targeted within the text.

Rule 3: Dealing with multiple languages

When posts contained multiple languages, the investigators consulted the wider annotation team which included four hate speech experts spanning four languages: Amharic, Afaan Oromo, Tigrigna, and English. If the text included languages that were not included in this study, the investigators were advised not to annotate the text. If the text included any combination of this investigation's four languages, the team annotated the text together. This ensured that the study remained focussed on the four languages being analysed, however also means that hate speech may have been excluded from the study. Another implication of this is that text may have been pulled within, for example, the English dataset, but it may have contained Afaan Oromo. When analysing the results of the study it is important to bear this in mind.

Rule 4: Copypasta

During the exercise, a number of 'copypasta' texts were identified. Copypasta is a block of text shared on the internet which is literally copy and pasted by multiple users. It often reveals coordination in information sharing. If the copypasta text contained hate speech, as per the guidelines, each instance was annotated. If they didn't conform with the annotation protocol, they were not annotated. Interestingly, this led annotators to identify a number of copypasta websites including: https://ethiopiantruth.com/jawsa-is-a-gang-of-blindly-driven-idiots-who-aim-to-rob-kill-and-terrorize-2/

Rule 5: Additional hate speech terms

When terms which could be indicative of hate speech were identified that were not in the study's hate speech lexicon, the investigators were advised to write these down. These will be added to the final hate speech lexicon that is published alongside the report, to ensure that the lexicon is as comprehensive as possible to aid future research. For example, the terms "flour ranger" and "kufr" were identified within the English dataset and added to the lexicon.

Inter-annotator agreement

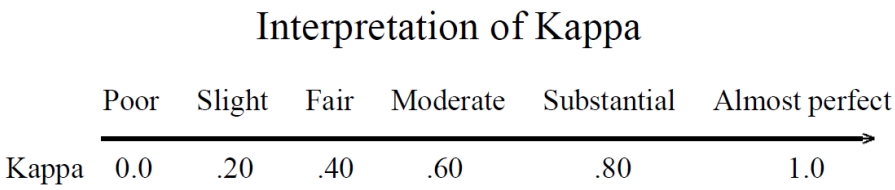
Cohen's Kappa

Cohen's Kappa (κ): Cohen's Kappa is a widely used measure of IAA that takes into account both the observed agreement between annotators and the agreement that would be expected by chance. It provides a value between -1 and 1, where 1 represents perfect agreement, 0 represents agreement no better than chance, and negative values indicate disagreement worse than chance. The diagram below provides the interpretation of Cohen's Kappa scores:

Fleiss' Kappa

Fleiss' Kappa: Fleiss' Kappa is an extension of Cohen's Kappa for more than two annotators. It is suitable for cases where there are multiple annotators providing judgments on the same items.

The diagram below provides the interpretation of Fleiss' and Cohen's Kappa scores:



Kappa	Agreement
< 0	Less than chance agreement
0.01–0.20	Slight agreement
0.21– 0.40	Fair agreement
0.41–0.60	Moderate agreement
0.61–0.80	Substantial agreement
0.81–0.99	Almost perfect agreement

The results of estimating inter-annotator agreement are shown below. The last comparison for Amharic is Fleiss's Kappa (comparing three annotators); all other comparisons are Cohen's kappa.

S/N	LANGUAGE	ANNOTATORS	KAPPA SCORE (k)	AGREEMENT
1	English	Eng Annotator A & Eng Annotator B	0.46	Moderate Agreement
2	Amharic	Amh Annotator A & Amh Annotator B	0.38	Fair Agreement
		Amh Annotator A & Amh Annotator C	0.46	Moderate Agreement
		Amh Annotator B & Amh Annotator C	0.32	Fair Agreement
		Amh Annotator A, Amh Annotator B & Amh Annotator C	0.39	Fair Agreement

Table X: IAA Agreement scores.

8. Bibliography

Akshita Jha and R. Mamidi (2017) 'When does a compliment become sexist? Analysis and classification of ambivalent sexism using X (formerly Twitter) data', NLP+CSS@ACL, Available at: https://api.semanticscholar.org/CorpusID:1570443.

Deborah James (1998) 'Gender-Linked Derogatory Terms and Their Use by Women and girls and Men', American Speech, Vol. 73, No. 4 (Winter, 1998), Duke University Press, pp. 399-420

Gao, L., Kuppersmith, A., & Huang, R. (2017). Recognizing explicit and implicit hate speech using a weakly supervised two-path bootstrapping approach. arXiv preprint arXiv:1710.07394.

Gashe, S. M. (2022). Hate Speech Detection and Classification System in Amharic Text with Deep Learning. List of Amharic Hate Speech Keywords (Lexicons). Retrieved from http://annotate.shegerapps.com/hs_keywords.php

Mekuanent Degu (2022), "Amharic dataset for hate speech detection", Mendeley Data, V3, doi: 10.17632/fhvsvsbvtg.3

Samuel Minale (2022), "Amharic Social Media Dataset for Hate Speech Detection and Classification in Amharic Text with Deep Learning", Mendeley Data, V1, doi: 10.17632/p74pfhz3yx.1

Surafel Getachew (2020), "Amharic Facebook Dataset for Hate Speech detection", Mendeley Data, V1, doi: 10.17632/ymtmxx385m.1

Reports and web articles:

CARD's Bi-weekly Social Media Conversation Sensitivity Report, see: https://www.cardeth.org/

David Shariatmadari (2016) 'Eight words that reveal the sexism at the heart of the English language' Available at: https://www.theguardian.com/commentisfree/2016/jan/27/eight-words-sexism-heart-english-language [last accessed 9 Nov 2023].

Hatebase.org (2023) Available at: https://hatebase.org/ [last accessed 9 Nov 2023].

Hate Speech Dataset Catalogue, Available at https://hatespeechdata.com/ [last accessed 9 Nov 2023].

Peace Tech Lab (n.d.) 'Hateful Speech and Conflict in the Federal Democratic Republic of Ethiopia: A lexicon of hateful of inflammatory words and Phrases' Available at: https://static1.squarespace.com/static/54257189e4b0ac0d5fca1566/t/60bfaa27a19f0752ecd1426d/1623173770820/EthiopiaLexicon2021_web.pdf [last accessed 9 Nov 2023].

Thalikir 2016 'Everyday misogyny: 122 subtly sexist words about women and girls (and what to do about them)' Available at: http://sacraparental.com/2016/05/14/everyday-misogyny-122-subtly-sexist-words-women and girls/ [last accessed 9 Nov 2023].

The Wilson Centre (2021) 'Malign Creativity: How Gender, Sex, and Lies are Weaponized Against Women and girls Online' Available at: https://www.wilsoncenter.org/publication/malign-creativity-how-gender-sex-and-lies-are-weaponized-against-women and girls-online [last accessed 9 Nov 2023].

Existing CIR publications on hate speech

Afghan Witness (2023), 'Violence behind a screen: rising online abuse silences Afghan women' Available at: https://www.afghanwitness.org/reports/violence-behind-a-screen%3A-rising-online-abuse-silences-afghan-women--

Eyes on Russia (2023), 'Incitement to Kill: Tracking hate speech targeting Ukrainians during Russia's war in Ukraine' Available at: https://www.info-res.org/post/incitement-to-kill-tracking-hate-speech-targeting-ukrainians-during-russia-s-war-in-ukraine

Myanmar Witness (2023), 'Digital Battlegrounds: Gendered hate speech report on the politically motivated abuse of Myanmar women and girls online' Available at: https://www.myanmarwitness.org/reports/digital-battlegrounds

9. Funding

This material has been funded by UK International Development from the UK government; however, the views expressed do not necessarily reflect the UK government's official policies.


Centre for
Information
Resilience