Resources for Annotating Hate Speech in Social Media Platforms Used in Ethiopia: A Novel Lexicon and Labelling Scheme

Nuhu Ibrahim[†], **Felicity Mulford**[†], **Matt Lawrence**[†] **and Riza Batista-Navarro**[†,‡]
[†]Centre for Information Resilience, London, UK
[‡]Department of Computer Science, The University of Manchester, UK
hi@nuhuibrahim.com, {felicitym, mattl}@info-res.org, riza.batista@manchester.ac.uk

Abstract

Hate speech on social media has proliferated in Ethiopia. To support studies aimed at investigating the targets and types of hate speech circulating in the Ethiopian context, we developed a new fine-grained annotation scheme that captures three elements of hate speech: the target (i.e., any groups with protected characteristics), type (i.e., the method of abuse) and nature (i.e., the style of the language used). We also developed a new lexicon of hate speech-related keywords in the four most prominent languages found in Ethiopian social media: Amharic, Afaan Oromo, English and Tigrigna. These keywords enabled us to retrieve social media posts (also in the same four languages) from three platforms (i.e., X, Telegram and Facebook), that are likely to contain hate speech. Experts in the Ethiopian context then manually annotated a sample of those retrieved posts, obtaining fair to moderate inter-annotator agreement. The resulting annotations formed the basis of a case study of which groups tend to be targeted by particular types of hate speech or by particular styles of hate speech language.

Keywords: Hate speech, Ethiopian languages, Social media, Annotation scheme, Lexicon development

1. Introduction

Social media platforms have emerged as potent communication tools, empowering individuals to voice opinions, exchange information and participate in diverse discussions (Poell and Van Dijck, 2015). Nevertheless, the unrestricted environment of these platforms has also fostered the spread of hate speech, presenting notable hurdles to societal cohesion, particularly in culturally diverse settings like Ethiopia. With the surge of digital communication that has encouraged the intertwining of personal and public life online, hate speech has discovered novel channels for propagation, frequently targeting marginalised communities or minority groups (Kovács et al., 2021) and intensifying social divides (Targema and Lucas, 2018).

Ethiopia, a country known for its rich linguistic and cultural diversity, has witnessed the rapid spread of hate speech on social media platforms such as Twitter, Telegram and Facebook (Delelegn, 2021). Recent events, including inter-ethnic violence and political unrest, have underscored the destructive impact of online hate speech on Ethiopian society. For instance, the escalation of tensions between ethnic groups in various regions has been fuelled, in part, by the dissemination of hate speech and incendiary rhetoric on social media platforms (Delelegn, 2021).

Minority languages continue to face scarcity in computational resources for gathering and analysing extensive textual datasets, resulting in minimal to no resources for automatically detecting hate speech on social media (El-Haj et al., 2015; Kovács et al., 2021). This research aims to develop a fine-grained labelling scheme for annotating hate speech. The labelling scheme helps in producing a richly annotated hate speech dataset that does not only identify hate but also the targeted groups with protected characteristics, and the type and nature of hate speech. In addition, this research aims to develop a lexicon across four languages (Amharic, Afaan Oromo, English and Tigrigna) which are indicative of hate speech along gendered, ethnic and religious lines, which to the best of our knowledge is currently the most comprehensive one for the Ethiopian context.

This research builds upon an earlier study conducted by the Centre for Information Resilience (CIR) that considered the lived experiences and lasting impacts of online abuse through a review of existing literature and interviews with 14 women who hold prominent positions in media, civil society and other public roles in Ethiopia (Centre for Information Resilience, 2023). Their findings highlight the toxicity of online environments, and interviewees revealed that the online abuse and harassment they received have had real-world impacts, including psychological harm, damaged professional reputations, disrupted family life and the silencing of women both online and offline. Considering the gravity of hate speech proliferating on the internet in minority languages and its impact on events in Ethiopia, we argue that there is a pressing need to develop resources that will enable the development of natural language processing (NLP) methods that can aid in automatically detecting such hate speech. To this end, we present: (1) a fine-grained annotation scheme for labelling hate speech circulating in social media platforms used in Ethiopia; (2) a new

The Fifth Workshop on Resources for African Indigenous Languages @LREC-COLING-2024 (RAIL), pages 115–123
25 May, 2024.

lexicon of hate speech-related keywords, covering inflammatory terms used in Amharic, Afaan Oromo, English and Tigrigna;[1] and (3) a corpus of social media posts annotated based on the fine-grained annotation scheme.

2. The Ethiopian Context

This research used the Ethiopian Government's definition of hate speech, as set out within the Hate Speech and Disinformation Prevention and Suppression Proclamation (No.1185/2020) (Federal Democratic Republic of Ethiopia, 2020), which defines it as "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."

In our study, we explored hate speech in social media platforms commonly used in Ethiopia. We meticulously adhered to the definition of hate speech provided by the Ethiopian Government, as stated above. Ethiopia is highly diverse in terms of languages that are in use, with over 80 languages spoken (Leyew, 2020). For reasons of feasibility and resource constraints, we decided to focus on analysing content in only four predominant languages—Amharic, Afaan Oromo, English and Tigrigna. The selection of these languages was informed by their prominence in social media platforms in the country (Zelalem, 2010).

Meanwhile, three different online platforms were chosen as the source of the content for analysis, namely X (formerly Twitter), Telegram and Facebook. These platforms were selected based on their widespread usage in Ethiopia (Daracho, 2020; Asale, 2020), coupled with the affordances provided by their policies regarding data collection and processing (Sosa and Sharoff, 2022; Giglietto et al., 2012). Additionally, these were three sites that were reported by interviewees (in CIR's earlier research) as the environments in which they faced online abuse (Centre for Information Resilience, 2023).

3. Related Work

The development of hate speech labelling schemes and lexicons for Ethiopian languages within the realm of NLP has gained increasing attention in recent years, driven by the growing recognition of the linguistic diversity and cultural richness of Ethiopia. While there is a scarcity of literature specifically dedicated to this topic, several related efforts have provided valuable insights into the challenges, methodologies and approaches relevant to hate speech labelling schemes and lexicon development for Ethiopian languages.

Peace Tech Lab (2023) reported around 21 inflammatory terms, their related spellings and associated terms, their meanings and the reasons why these terms are inflammatory. They also provided an additional 16 that are offensive and should be looked out for. Minale (2022) curated hateful keywords in Amharic and their translation in English, and then grouped the keywords into categories, namely, 'Ethiopian nation', 'gender', 'hate-related', 'offensive' and 'religious' keywords. They used these keywords to automatically collect Amharic data from three social media sites: Facebook, Twitter and YouTube. These datasets were then categorised by human annotators into four categories: 'normal speech', 'racial hate speech', 'religious hate speech', 'gender gate speech' and 'disability hate speech'.

Meanwhile, Jha and Mamidi (2017) collected sexist English posts from Twitter by matching terms or hashtags that are generally used when exhibiting what they refer to as "benevolent sexism". Some of these terms and hashtags were: *"as good as a man"*, *"like a man"*, *"for a girl"*, *"smart for a girl"*, *"love of a woman"*, *"#adaywithoutwomen"*, *"#womensday"*, *"#everydaysexism"* and *"#weareequal"*. The collected posts were manually annotated and were used to train a machine learning-based model to classify posts into three categories ('Hostile', 'Benevolent', 'Others') depending on the kind of sexism they exhibit. Similar to Minale (2022), the work by Jha and Mamidi (2017) curated Amharic sexist keywords and used them to collect posts from Twitter; they also built various classification models.

Some previous work focussed on hate speech analysis for Ethiopian languages. For instance, Getachew (2020) and Ayele et al. (2022) investigated Amharic hate speech. Kanessa and Tulu (2021) and Defersha and Tune (2021) focussed on hate speech in Afaan Oromo while Bahre (2022) studied hate speech in Tigrigna.

We found that most researchers in the Ethiopian context have concentrated on curating hate speech lexicons for Amharic, and only limited efforts have attempted to curate hate speech lexicons for other Ethiopian languages, e.g., Afaan Oromo, Tigrigna and English (as used in the Ethiopian context). In contrast, our work curated hate speech lexicons for multiple Ethiopian languages: Amharic, Afaan Oromo, English and Tigrigna. These languages are the most prominently used in social media platforms in the country (Zelalem, 2010).

Additionally, most labelling schemes developed for Ethiopian languages only classified hate speech as either 'hate' or 'no hate'. Some studies such as that by Minale (2022) went further to define cate-

[1] https://github.com/Centre-for-Information-Resilience/ethiopia-hate-speech-lexicon

gories of hate: ‘normal speech’, ‘racial hate speech’, ‘religious hate speech’, ‘gender gate speech’ and ‘disability hate speech’. Jha and Mamidi (2017) also categorised hate/sexist posts into ‘Hostile’, ‘Benevolent’ or ‘Others’, however, they concentrated only on identifying sexism. Our research goes beyond existing work in developing a fine-grained labelling scheme that identifies three elements in hate posts: the target, type and nature of hate speech.

4. Annotation Scheme

Annotation schemes typically contain a set of guidelines or rules used to annotate or label data with specific information or attributes (Bird et al., 2009). This section discusses the development of a fine-grained labelling scheme for labelling hate speech on social media platforms in the Ethiopian context.

In line with the definition of hate speech in the Ethiopian context, stated in Section 2, for a post to be considered as containing hate speech, it has to be targeted towards an individual or group with a protected characteristic. When a post contains hate speech, our labelling scheme requires annotators to label three elements in the post: the target, type and the nature of hate speech. We refer the reader to Figure 1) for a diagram that provides an overview of the labels in our annotation scheme. In our work, the type of hate speech refers to the method of abuse (such as threats), while nature refers to the style used in the language that expresses abuse (such as irony or stereotyping).

To capture the information about the target of hate speech, the words that convey which individual/group is being targeted should be assigned any of the following labels:

- `Gender`: An individual or group of people of a particular gender.[2]
- `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.

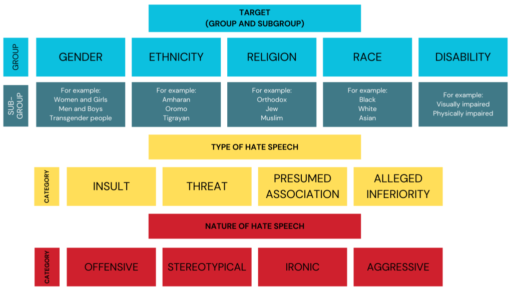

Figure 1: Overview of the categories in our annotation scheme.

Furthermore, to capture information about the type of speech, the words that convey the method of abuse should be assigned any of the following labels:

- `Insult`: Insults or denigrating expressions against an individual/group due to protected characteristics.
- `Threat`: Intimidation, threats or incitement to hatred, violence or violation of individuals’ rights, due to protected characteristics.
- `Presumed Association`: Presumed association of protected characteristics with negative connotations.
- `Alleged Inferiority`: References to the alleged inferiority (or superiority) of an individual/group with a protected characteristic.

Lastly, the nature or style of hate speech often varies from one post to another. While not essential for the classification of hate speech, collecting information on style captures the nuances in the language used in expressing hate. To capture this information, hate speech-containing language needs to be labelled as any of:

- `Aggressive`: Includes strong language that seeks to physically intimidate, threaten or incite physical violence against the recipient, or which requests, suggests or promotes a violation of the recipient’s rights.
- `Offensive`: 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.
- `Ironic`: Includes jokes, satire or sarcastic messaging which targets a protected characteristic of the recipient and could be harmful. Hateful content is sometimes conveyed using

[2] Although the Ethiopian Government’s hate speech definition (Federal Democratic Republic of Ethiopia, 2020) does not explicitly reference sexual identity, we incorporated sexual identity within this category to capture hate based on sexual orientation.

nuances in language, such as sarcasm, humour or satire.

- Stereotypical: Corresponds to implicit or explicit references to stereotypical beliefs or prejudices about an individual/group with protected characteristics.

The labelling scheme ensures that a multi-lingual team of annotators have a shared understanding of what labels constitute hate speech. In addition, when used, the labelling scheme produces a rich hate speech dataset that will not only tell whether a post contains hate but also the target category of the protected characteristics receiving the hate, and the type and nature of the hate received. This will help to answer substantive questions like:

- To what extent do groups with particular protected characteristics, e.g., gender, religion, ethnicity, race, etc, receive hate on social media?
- What type and nature of hate speech are prevalent?
- Do certain protected characteristics receive more hate on social media, compared to others?
- How does hate speech vary across target subgroups, i.e., women, men, homosexuals for the gender category?
- How does hate speech vary when multiple protected characteristics are targeted (i.e., hate speech that targets individuals/groups along multiple identity lines)?

5. Case Study

Our annotation scheme was applied to a case study aimed at investigating which groups with protected characteristics have often been targeted by hate speech in the Ethiopian context, as well as the type and nature of language addressed to them. This section outlines the steps we took to collect and annotate data from various social media platforms in support of the case study.

5.1. Data Collection

We collected two types of data: keywords that form a new Ethiopian hate speech lexicon, and social media posts forming a new hate speech corpus.

5.1.1. Lexicon Development

Considering the huge volume of social media posts that get published on a daily basis, we developed a lexicon of keywords to aid in the collection of posts that are likely to contain hate speech. Specifically, we collected keywords across four languages—Amharic, Afaan Oromo, English and Tigrigna—that are indicative of hate speech along gendered, ethnic and religious lines.

The lexicon was developed through desk-based research that employed both identification and refinement of existing hate speech lexicons (Minale, 2022; Degu, 2022; Getachew, 2020; James, 1998; Jha and Mamidi, 2017; Gashe, 2022; Gao et al., 2017; Peace Tech Lab, 2023; Hatebase.org, 2023; Thalikir, 2016; Centre, 2021; Shariatmadari, 2016; Center for the Advancement of Rights and Democracy, 2023), the identification of other keywords and narratives during in-person, semi-structured interviews carried out during CIR's earlier study (Centre for Information Resilience, 2023) and a roundtable discussion that brought together 21 individuals from an array of civil society organisations, UN agencies, and women and girls' rights advocacy groups.

A first draft of the lexicon was shared with 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. It was clarified to the stakeholders that the keywords will be leveraged only for collecting as many posts as possible (a high-recall but low-precision approach), and manual inspection is still necessary, as we recognise that—as with any dictionary-based approach—many keywords are ambiguous and thus their presence in a post does not necessarily mean that the post contains hate speech. Hence, human annotators will analyse whether the content indeed contains hate speech, according to the developed labelling scheme.

The resulting lexicon consists of 2,058 inflammatory keywords across the four languages within the scope of this study. We believe that, to date, this is the most comprehensive lexicon for the Ethiopian context. Figure 2 and 3 respectively show the number and distribution of keywords curated for each protected characteristic.

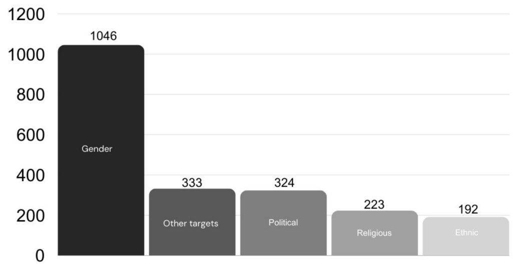

Figure 2: Number of keywords curated for each protected characteristic

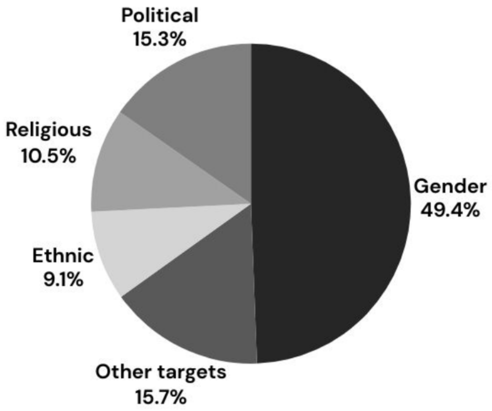

Figure 3: Distribution of keywords curated for each protected characteristic

5.1.2. Data Collection

Social media posts were collected from the platforms of interest, namely, X (formerly Twitter), Telegram and Facebook. To collect data from X, the Meltwater social media analysis tool[3] was employed. Meltwater supports the use of keyword search for tweets posted no longer than 18 months from the date of search. To ensure a relevant sample was obtained, English posts were only retrieved if they originated from Ethiopia.

In collecting data from Telegram, the official Telegram APIs[4] were used. As Telegram supports only searches within Telegram Channels to which a user belongs, social media experts from Ethiopia were engaged to meticulously curate a list of widely popular and influential public Telegram Channels in Ethiopia. Subsequently, we joined a total of 285 Telegram channels; the Telegram posts in these channels that contain keywords in our lexicon were collected.

To extract data from Facebook, social media experts from Ethiopia were again engaged in selecting a list of prominent and influential public Ethiopian Facebook groups and profiles. As an outcome of this engagement, a list of 300 Facebook profiles or groups was curated, and posts from these groups containing keywords in our lexicon were collected.

5.1.3. Data Pre-processing

Data pre-processing is a crucial step that involves cleaning, transforming and organising raw textual data to make it suitable for analysis (Tabassum and Patil, 2020).

Textual data collected from social media often contain irrelevant or erroneous information that complicates analysis or interpretation. To mitigate this issue, we carried out the following tasks on the datasets from X, Telegram and Facebook:

- Removal of HTML tags and special characters.
- Case-folding of text (i.e., making all characters lowercase) to ensure case insensitivity.
- Removal or replacement of punctuation.
- Removal of duplicate posts.

The following cleaning tasks were done only on the datasets in the English language:

- Removal of numerical values, dates and other non-textual information.
- Removal of stop words that do not carry any significant meaning (e.g., *"and"*, *"the"*, *"in"*).
- Normalisation of abbreviations and acronyms.

5.1.4. Data Anonymisation

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

5.1.5. Sampling for Further Analysis

The data collection process resulted in the collection of tens of millions of posts, as illustrated in Table 1. 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, we selected a random sample to obtain more manageable datasets. Table 1 shows the number of posts resulting from each step of the data preparation process and the number of posts chosen for subsequent analysis.

5.2. Annotation Task

The annotation task entails enlisting proficient human annotators who are familiar with the domain of interest to employ the developed labelling scheme for determining whether the posts in the collected dataset contain hate speech. The annotators used Doccano (Nakayama et al., 2018), an open-source annotation tool that we employed to label the posts

[3] https://explore.meltwater.com/meltwater-media-monitoring

[4] https://core.telegram.org/

	X	Telegram	Facebook
Posts collected	865,224	326,471,094	7,230
Posts after preprocessing	527,522	906,471	7,230
Random sample for annotation	2634	2107	2264

Table 1: The number of posts obtained in each step of the data preparation process.

in our datasets according to our annotation scheme, i.e., to annotate the hate speech targets (protected characteristics), the type and nature of hate speech.

6. Annotation Results

To ensure consistency in the application of the fine-grained labelling scheme, it was essential to calculate inter-annotator agreement (IAA) scores.

Two human annotators were enlisted to annotate the randomly chosen English posts. The primary annotator who participated in the development of the fine-grained labelling scheme and is knowledgeable of the Ethiopian context, was responsible for annotating the entire selection of English posts. To allow for estimation of IAA, a secondary annotator was assigned to annotate 10% of the dataset annotated by the primary annotator.

For Amharic, 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 assigned with the Tigrigna and Afaan Oromo datasets) 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.

For Afaan Oromo and Tigrigna, an annotator was enlisted per language. IAA was considered unnecessary for the annotators, as these same annotators had previously worked on annotating the Amharic dataset, and the results of IAA agreement on the Amharic dataset indicated their competence in identifying hate speech, labelling its target, categorising speech types and assessing the sentiment of hate.

IAA was calculated using Cohen's Kappa (k) and Fleiss' Kappa metrics. The IAA scores, presented in Table 2, showed fair to moderate agreement between annotators (Landis and Koch, 1977).

Language	Annotators	Kappa	Agreement
English	E1 & E2	0.46	Moderate
Amharic	A1 & A2	0.38	Fair
Amharic	A1 & A3	0.46	Moderate
Amharic	A2 & A3	0.32	Fair
Amharic	A1, A2 & A3	0.39	Fair

Table 2: Result and interpretation of estimating inter-annotator agreement between annotators in terms of Kappa scores.

7. Discussion

The resulting lexicon covers a higher percentage of gender-related keywords (49.4%) compared to those related to ethnicity (9.1%) or religion (10.5%); see Figures 2 and 3. Despite this imbalance, it is worth noting that the corpus of social media posts constructed based on our lexicon nevertheless revealed a greater prevalence of hate speech targeting other identity groups, as illustrated in Figure 4. For example, out of all the posts in the full dataset, ethnic hate speech comprised 44.5%, whereas gendered hate speech and religious hate speech represented 30.2% and 17.5%, respectively. Racial hate and hate speech targeting people with disabilities made up a smaller proportion of the dataset (4.6% and 0.3%, respectively).

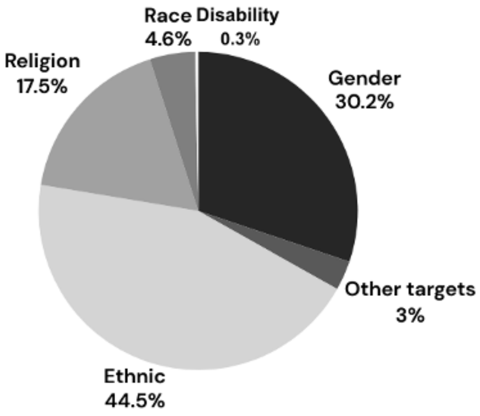

Figure 4: Distribution of hate-containing posts according to hate target.

As can be seen in Figure 5, when the protected characteristic identity groups are broken down into individual hate targets, other interesting trends become visible. The more targeted groups are women and girls (21% of the dataset), closely followed by Oromos (19.1%) and Amharans (16.7%). As the lexicon comprised more gender-related keywords, this is not surprising. Other targets of hate speech within the dataset, albeit in smaller proportions, include Orthodox Christians (8.7%), men (5.9%) and Tigrayans (5.5%).

The 'additional hate targets' category (in Figure 5) is comprised of all the other target groups outside of the top 7 most prevalent targets; this includes Protestants, white people, transgender people, atheists, Arabs, multiracial people and Jews. The 'other target' category was selected in cases where hate speech targeting a protected characteristic was present, but it does not fall under any of the categories.

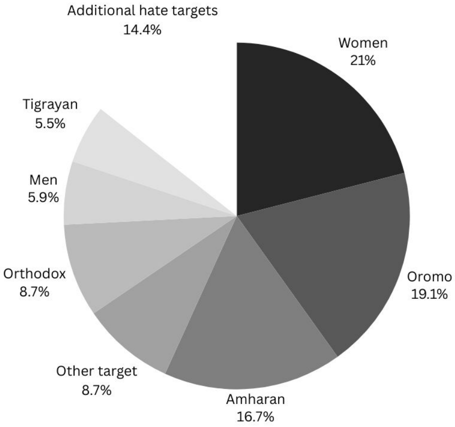

Figure 5: Distribution of hate-containing posts according to specific hate target.

Table 3 shows that women and girls receive proportionally more insulting hate speech (36.55%) than Amharans (28.31%), Muslims (29.51%) and Oromos (22.51%). They receive less insulting hate speech compared to homosexual people (38.96%) and Tigrayans (45.21%). Additionally, women and girls receive proportionally more hate containing alleged inferiority (22.2%), followed by Muslims (16.39%), Amharans (11.90%), homosexuals and Tigrayans (12.99% and 13.01, respectively), and Oromos (9.98%). Conversely, women and girls receive (proportionally) the least threats (13.51%) compared to Oromos (26.11%), Amharans (22.22%), Tigrayans (22.6%), homosexual people (20.78%) and Muslims (18.03%). Women and girls are also among the hate target subgroups which receive proportionally less hate containing presumed association (27.74%), compared to Amharans (37.57%), Muslims and Oromos (36.07% and 41.4%, respectively).

Interestingly, it was identified that offensive language is the more prevalent nature or style of hate speech across all hate targets analysed (see Table 4). Contrary to the pattern observed in offensive language, women and girls receive the highest proportion of stereotypical language (26.17%), closely followed by homosexuals (22.81%), then Muslims (12.5%), Amharans (9.27%), Oromos (4.98%) and Tigrayans (2.65%). Only Muslims receive a higher proportion of ironic language (21.25%) than Women and girls (20.37%). Even so, women and girls are considerably more targeted by ironic language than Tigrayans (11.5%), homosexuals (8.77%), Amharans (6.85%) and Oromos (5.32%).

Target	Type	%
Women	Insult	36.55
	Presumed Association	27.74
	Threat	13.51
	Alleged Inferiority	22.20
Amharan	Insult	28.31
	Presumed Association	37.57
	Threat	22.22
	Alleged Inferiority	11.90
Oromo	Insult	22.51
	Presumed Association	41.40
	Threat	26.11
	Alleged Inferiority	9.98
Muslim	Insult	29.51
	Presumed Association	36.07
	Threat	18.03
	Alleged Inferiority	16.39
Tigrayan	Insult	45.21
	Presumed Association	19.18
	Threat	22.60
	Alleged Inferiority	13.01
Homosexual	Insult	38.96
	Presumed Association	27.27
	Threat	20.78
	Alleged Inferiority	12.99
Orthodox	Insult	32.78
	Presumed Association	36.11
	Threat	24.44
	Alleged Inferiority	6.67

Table 3: Number of hate-containing posts according to target (top 7) and type of hate speech.

8. Conclusion

In our research, we developed a fine-grained annotation scheme for labelling hate speech in posts published in social media platforms used in Ethiopia. The annotation scheme formed the basis of producing a richly annotated hate speech corpus that does not only identify hate-containing posts but also the targeted protected characteristics, the type of hate, and the nature of the language used in hate speech.

In addition, this research produced a lexicon covering four languages used in Ethiopia, i.e., Amharic, Afaan Oromo, English and Tigrigna, that contains keywords that are indicative of hate speech along gendered, ethnic and religious lines. To the best of our knowledge, this lexicon is currently the most

Target	Nature	%
Women	Aggressive	7.29
	Ironic	20.37
	Offensive	46.17
	Stereotypical	26.17
Amharan	Aggressive	23.39
	Ironic	6.85
	Offensive	60.48
	Stereotypical	9.27
Oromo	Aggressive	39.20
	Ironic	5.32
	Offensive	50.50
	Stereotypical	4.98
Muslim	Aggressive	17.50
	Ironic	21.25
	Offensive	48.75
	Stereotypical	12.50
Tigrayan	Aggressive	29.20
	Ironic	11.50
	Offensive	56.64
	Stereotypical	2.65
Homosexual	Aggressive	12.28
	Ironic	8.77
	Offensive	56.14
	Stereotypical	22.81
Orthodox	Aggressive	27.13
	Ironic	6.20
	Offensive	58.91
	Stereotypical	7.75

Table 4: Number of hate-containing posts according to target (top 7) and nature of hate speech.

comprehensive one for the Ethiopian context. Our future work will be focussed on investigating how the annotated corpus resulting from this study, can enable the development of machine learning-based models that can automatically detect and categorise hate speech, as well as automatically identify the specific targets of hate speech.

Ethics Statement

The Centre for Information Resilience (CIR) follows the Berkeley Protocol on Digital Open-Source Investigations. For this study, care was taken to anonymise data and to comply with the terms and conditions of the platforms. To mitigate the impact of vicarious trauma, annotators were offered one-to-one support from the CIR Research Coordinator (the second author of this paper). This was to ensure that the annotators were not directly impacted by exposure to hate speech. Annotators were also made aware that they have access to appropriate resources should professional help become necessary.

Acknowledgements

We would like to thank Adyam Solomon Tesfay, Alemu Teshome Baki and Fasika Tadesse for their hard work as annotators of our datasets.

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.

Bibliographical References

Moges Ayele Asale. 2020. The Tributes and Perils of Social Media Use Practices in Ethiopian Socio-political Landscape. In *Proceedings of the 22nd HCI International Conference*, volume 12427, page 199, Copenhagen, Denmark. Springer Nature.

Abinew Ali Ayele, Skadi Dinter, Tadesse Destaw Belay, Tesfa Tegegne Asfaw, Seid Muhie Yimam, and Chris Biemann. 2022. The 5Js in Ethiopia: Amharic Hate Speech Data Annotation using Toloka Crowdsourcing Platform. In *Proceedings of the 2022 International Conference on Information and Communication Technology for Development for Africa*, pages 114–120. IEEE.

Weldemariam Bahre. 2022. *Hate speech detection from Facebook social media posts and comments in Tigrigna language*. Ph.D. thesis, St. Mary's University.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc.

Center for the Advancement of Rights and Democracy. 2023. CARD's Bi-weekly Social Media Conversation Sensitivity Report.

The Wilson Centre. 2021. Malign Creativity: How Gender, Sex, and Lies are Weaponized Against Women and girls Online.

Centre for Information Resilience. 2023. Silenced, shamed, and threatened: The online abuse of women who participate in Ethiopian public life.

Lisanu Damene Daracho. 2020. Social Media Impact on Social Life of Public Servant in Mari Mansa District, Dawuro Zone, Southern Region, Ethiopia. *New Media and Mass Communication*, 93:1–7.

NB Defersha and KK Tune. 2021. Detection of Hate Speech Text in Afan Oromo Social Media Using Machine Learning Approach. *Indian J Sci Technol*, 14(31):2567–78.

Mekuanent Degu. 2022. Amharic dataset for hate speech detection. Mendeley Data.

Misganaw Delelegn. 2021. *Hate Speech Regulation in Ethiopia: Lessons to Be Learned From Other Jurisdictions*. Ph.D. thesis, Bahir Dar University.

Mahmoud El-Haj, Udo Kruschwitz, and Chris Fox. 2015. Creating language resources for under-resourced languages: methodologies, and experiments with Arabic. *Language Resources and Evaluation*, 49:549–580.

Federal Democratic Republic of Ethiopia. 2020. Hate Speech and Disinformation Prevention and Suppression Proclamation (No. 1185/2020).

Lei Gao, Alexis Kuppersmith, and Ruihong Huang. 2017. Recognizing Explicit and Implicit Hate Speech Using a Weakly Supervised Two-path Bootstrapping Approach. In *Proceedings of the 8th International Joint Conference on Natural Language Processing*, pages 774–782, Taipei, Taiwan. Asian Federation of Natural Language Processing.

S. M. Gashe. 2022. Hate Speech Detection and Classification System in Amharic Text with Deep Learning. List of Amharic Hate Speech Keywords (Lexicons).

Surafel Getachew. 2020. Amharic Facebook Dataset for Hate Speech detection. Mendeley Data.

Fabio Giglietto, Luca Rossi, and Davide Bennato. 2012. The Open Laboratory: Limits and Possibilities of Using Facebook, Twitter, and YouTube as a Research Data Source. *Journal of Technology in Human Services*, 30(3-4):145–159.

Hatebase.org. 2023. Hatebase.org.

Deborah James. 1998. Gender-linked derogatory terms and their use by women and men. *American Speech*, 73(4):399–420.

Akshita Jha and Radhika Mamidi. 2017. When does a compliment become sexist? Analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the 2nd Workshop on NLP and Computational Social Science*, pages 7–16.

Lata Guta Kanessa and Solomon Gizaw Tulu. 2021. Automatic Hate and Offensive speech detection framework from social media: the case of Afaan Oromoo language. In *Proceedings of the 2021 International Conference on Information and Communication Technology for Development for Africa*, pages 42–47. IEEE.

György Kovács, Pedro Alonso, and Rajkumar Saini. 2021. Challenges of hate speech detection in social media: Data scarcity, and leveraging external resources. *SN Computer Science*, 2:1–15.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Zelealem Leyew. 2020. Language and society in Ethiopia. *Bulletin of the Department of Linguistics and Philology 40 years*, page 64.

Samuel Minale. 2022. Amharic Social Media Dataset for Hate Speech Detection and Classification in Amharic Text with Deep Learning. Mendeley Data.

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. doccano: Text Annotation Tool for Human.

Peace Tech Lab. 2023. Hateful Speech and Conflict in the Federal Democratic Republic of Ethiopia: A lexicon of hateful of inflammatory words and Phrases.

Thomas Poell and José Van Dijck. 2015. Social media and activist communication. *The Routledge companion to alternative and community media*, pages 527–537.

David Shariatmadari. 2016. Eight words that reveal the sexism at the heart of the English language.

Jose Sosa and Serge Sharoff. 2022. Multimodal Pipeline for Collection of Misinformation Data from Telegram. In *Proceedings of the 13th Language Resources and Evaluation Conference*, pages 1480–1489, Marseille, France. European Language Resources Association.

Ayisha Tabassum and Rajendra R Patil. 2020. A Survey on Text Pre-Processing & Feature Extraction Techniques in Natural Language Processing. *International Research Journal of Engineering and Technology*, 7(06):4864–4867.

Tordue Simon Targema and Joseph M Lucas. 2018. Hate speech in readers' comments and the challenge of democratic consolidation in Nigeria: A critical analysis. *Jurnal Pengajian Media Malaysia*, 20(2):23–38.

Thalikir. 2016. Everyday misogyny: 122 subtly sexist words about women and girls (and what to do about them).

Amsale Zelalem. 2010. *Design and Implementation of Multilanguage Electronic Dictionary for Smart Phones: A Dictionary of Amharic, Afaan Oromo, English and Tigrigna Languages*. Ph.D. thesis, Addis Ababa University.