Exploring Boundaries and Intensities in Offensive and Hate Speech: Unveiling the Complex Spectrum of Social Media Discourse

Abinew Ali Ayele[1,2], **Esubalew Alemneh Jalew**[2], **Adem Chanie Ali**[2], **Seid Muhie Yimam**[1], **Chris Biemann**[1]

[1] Universität Hamburg, Germany, [2] Bahir Dar University, Ethiopia

Abstract

The prevalence of digital media and evolving sociopolitical dynamics have significantly amplified the dissemination of hateful content. Existing studies mainly focus on classifying texts into binary categories, often overlooking the continuous spectrum of offensiveness and hatefulness inherent in the text. In this research, we present an extensive benchmark dataset for Amharic, comprising 8,258 tweets annotated for three distinct tasks: *category classification*, *identification of hate targets*, and *rating offensiveness and hatefulness intensities*. Our study highlights that a considerable majority of tweets belong to the *less offensive* and *less hate* intensity levels, underscoring the need for early interventions by stakeholders. The prevalence of *ethnic* and *political* hatred targets, with significant overlaps in our dataset, emphasizes the complex relationships within Ethiopia's sociopolitical landscape. We build classification and regression models and investigate the efficacy of models in handling these tasks. Our results reveal that hate and offensive speech can not be addressed by a simplistic binary classification, instead manifesting as variables across a continuous range of values. The Afro-XLMR-large model exhibits the best performances achieving F1-scores of 75.30%, 70.59%, and 29.42% for the category, target, and regression tasks, respectively. The 80.22% correlation coefficient of the Afro-XLMR-large model indicates strong alignments.

Keywords: Intensity, Hatefulness, Offensiveness, Rating scale

1. Introduction

In the world of rapid innovations, the prevalence and influence of social media persistently expand, along with the diverse array of online content crafted by a multitude of contributors, which has become readily available for consumption and engagement (Sazzed, 2023). Remarkably, over 60% of the world's population is actively participating in social media. However, social media platforms have become the main places for the dissemination and proliferation of hate speech (Bran and Hulin, 2023; Mathew et al., 2021; Davidson et al., 2017; Waseem and Hovy, 2016; Ayele et al., 2023b). The ease of communication and the global reach of these platforms have enabled users to spread hateful and offensive content aggressively in wider circles (Zufall et al., 2022). The anonymity of online users on social media granted hateful message propagators to spread toxic content by hiding themselves behind their digital screens (Bran and Hulin, 2023; Kiritchenko et al., 2021; Zufall et al., 2022). Hate speech on social media can take various forms, including discriminatory language, threats, harassment, and the incitement of violence against specific individuals or groups of communities (Mathew et al., 2021; Davidson et al., 2017; Ayele et al., 2023a). This online hate speech can have real-world consequences, contributing to social divisions, fueling hostility, and inciting violence in some circumstances (Abraha, 2017; Yimam et al., 2019). As a result, social media companies, policymakers, and researchers are increasingly focused on developing strategies to detect, combat, and mitigate the impact of hate speech on these platforms without compromising the principles of freedom of speech and user safety (Pavlopoulos et al., 2017; Ayele et al., 2023a).

For the past couple of years, there has been increasing attention and interest in exploring hate speech among researchers from diverse academic disciplines, including social science, psychology, media and communications studies, and computer science (Tontodimamma et al., 2021; Davidson et al., 2017; Mathew et al., 2021; Davidson et al., 2019; Chekol et al., 2023; Ayele et al., 2023b).

Many studies, including those by Davidson et al. (2017); Fortuna et al. (2020); Waseem and Hovy (2016); Mathew et al. (2021); Plaza-del arco et al. (2023); Clarke et al. (2023); Caselli and Veen (2023) and others, adopt a binary approach to hate speech classification. These works aim to distinguish and label content as either hate or non-hate. Nevertheless, this binary viewpoint lacks the capacity to capture the diverse and context-dependent features of hate speech, which resist easy classification. We posit that hate speech classification demonstrates a spectrum of continuity (Bahador, 2023). In contemporary studies, there has been a recognition of this limitation by prompting a shift towards adopting multifaceted methodologies to gain a better understanding of the nature, dimension, and intensity of hate speech (Beyhan et al., 2022; Sachdeva et al., 2022). This further enhances hate speech

detection capabilities and employs more effective mitigation strategies to tackle its propagation on social media and its impact on the physical world.

Studies on hate speech in low-resource languages, particularly Amharic, such as those conducted by Abebaw et al. (2022); Mossie and Wang (2018); Ayele et al. (2022b); Tesfaye and Kakeba (2020); Ayele et al. (2023b), predominantly concentrated on the detection of hate speech as a binary concept, overlooking its varying levels of intensities.

In this study, our focus extends beyond the binary approach to include the varied intensities of hate and offensive speech. For the intensity rating approach, we adopt the Likert rating scale during annotation. Likert rating scale is a commonly used tool to measure attitudes, opinions, or perceptions of respondents towards a particular subject, where respondents are asked to choose the options that best reflects their viewpoint for each item (Subedi, 2016). Likert rating scale provides a quantitative measurement of qualitative data, which helps researchers to analyze attitudes or opinions in a structured and comparable manner (Joshi et al., 2015).

The dataset was collected from X, formerly Twitter and annotated a total of 8.3k tweets. Five native Amharic speakers individually provided annotations for each tweet. Our annotations covered three distinct types: **category**, **target**, and **intensity level**.

In the **category** type of annotations, we requested annotators to classify each tweet into specific categories. These categories include:

1. **Hate**: Tweets that promote prejudice, discrimination, hostility, or violence against individuals or groups targeting their group identities to marginalize or harm them.

2. **Offensive**: Tweets that are likely to cause discomfort, annoyance, or distress to people, but do not target any of their group identities.

3. **Normal**: Tweets that do not contain any hate or offensive language and are considered within the boundaries of acceptable and respectful discourse.

4. **Indeterminate**: This consists of tweets that are challenging to categorize due to various reasons, such as tweets that contain mixed languages, and typographical errors. It also includes tweets that are unclear or incomprehensible to determine its content accurately.

The **target** annotation type involves identifying the specific groups, individuals, or communities who are the recipients of the hate speech within the tweet. This process aids in understanding the intended targets of the harmful content, providing insights into the context and potential impact.

Lastly, the **intensity level** annotation type is a valuable measure for assessing the intensities of hate and offensive speech. It provides a means to measure where a tweet falls along the spectrum of harm, from milder instances to more severe cases. This type of annotation aids in understanding the varying degrees of harm and evaluating the subtle nature of such content.

The following are the main research questions that we address in this paper:

- **RQ-1:** Do hate and offensive speech represent discrete binary categories, or exist on a continuous spectrum of varying intensities?

- **RQ-2:** What is the extent to which hate speech specifically targets certain groups of the population? and,

- **RQ-3:** What is the occurrence and nature of tweets containing hate speech directed towards multiple target groups?

The main contributions of this study include the following but not limited to:

1. Presenting a benchmark dataset for hate speech category and target detection tasks, supplemented with intensity level ratings,

2. Providing comprehensive annotation guidelines for hate speech categories, targets, and approaches to measure the intensity of offensiveness and hatefulness, and

3. Developing classification and regression models for predicting hate intensity levels and detecting hate speech and its targets.

Despite focusing on Amharic, the outlined approach can be further extended to other languages and cultural contexts.

2. Related Works

There is no clear and simple demarcation between hate speech, offensive speech, and protected free speech due to its complex nature. The complexity arises from the subjective nature of the offense, contextual variability, diversity of intent, varying degrees of harm, and variations in legal definitions (Madukwe et al., 2020; Ayele et al., 2022a). Recognizing this complexity is important for balancing the protection of free speech rights with the need to address and mitigate harmful content effectively. This necessitates a holistic approach to be employed in determining the nature and consequences of such speech by considering the intent, impact, cultural context, and legal frameworks (Zufall et al., 2022; Beyhan et al., 2022; Chandra et al., 2020).

Over the past several years, a lot of research attempts have been dedicated to exploring and analyzing hate speech using social media data.

However, the majority of these studies approached hate speech detection and classification tasks as a binary categorization or dissecting it into three or four distinct classes. For instance, Davidson et al. (2017); Mathew et al. (2021); Ousidhoum et al. (2019); Waseem and Hovy (2016); Sigurbergsson and Derczynski (2020); Clarke et al. (2023) are among the studies conducted for resourceful languages that focused on detecting hate speech and its targets. Clarke et al. (2023) and Mathew et al. (2021) attempted a bit deeper study and investigated explainable hate speech detection approaches beyond detecting its presence in a text. Kennedy et al. (2020) studied hate speech by contextualizing classifiers with explanations that encourage models to learn from the context. Ocampo et al. (2023) explored the detection of implicit expressions of hatred, highlighting the complexity of the task and underscoring that hate speech is not yet well studied.

Hate speech detection studies conducted so far in the Amharic language also approach the problem as a binary classification task. For instance, Mossie and Wang (2018); Defersha and Tune (2021); Abebaw et al. (2022); Tesfaye and Kakeba (2020) investigated Amharic hate speech as a binary hate and non-hate class, and Mossie and Wang (2020) identified similar binary label categories, but further explored targeted communities. Ayele et al. (2022b) explored Amharic hate speech in four categories such as hate, offensive, normal, and unsure, and Ayele et al. (2023b) employed similar categories except the exclusion of the unsure class in the latter study. In addition to textual studies, a few multimodal research attempts for Amharic such as Degu et al. (2023); Debele and Woldeyohannis (2022) explored Amharic hate speech using meme text extracts and audio features, treating the task as a discrete binary task.

Recent studies indicated that hate and offensive speeches are not simple binary concepts, rather they exist on a continuum, with varying degrees of intensity, harm, and offensiveness (Bahador, 2023; Sachdeva et al., 2022). In practical scenarios, hate speech exhibits a wide spectrum, encompassing mild stereotyping on one end and explicit calls for violence against a specific group on the other (Beyhan et al., 2022). Demus et al. (2022) explored hate speech categories, targets, and sentiments in two or three discrete categories while analyzing the toxicity of the message using the Likert scale ratings of 1-5 to show the potential of a message to "poison" a conversation.

The study by Chandra et al. (2020) investigated the intensity of online abuse by classifying it into three separate discrete labels, namely 1) biased attitude, 2) act of bias and discrimination, and 3) violence and genocide. The annotators chose among these labels and employed the majority voting scheme for the gold labels. This online abuse intensity study employed the classical categorical approach which is a binary perspective and failed to represent the diverse fine-grained contexts in a spectrum of continuum values.

In this study, we aim to explore the extent of offensiveness and hatefulness intensities of tweets on a rating scale of 1-5, and 0 representing normal tweets.

3. Data Collection and Annotation

This section presented the descriptions of data collection and annotation procedures.

3.1. Data Collection

The dataset has been collected from Twitter/X spanning over 15 months since January 1, 2022. During this time, a multitude of highly controversial dynamics were occurring within the complex sociopolitical landscape of Ethiopia. Over 3.9M tweets that are written in Amharic Fidäl script were crawled, and further filtered by removing retweets, and the tweets that are written in languages other than Amharic. We used different data selection strategies such as hate and offensive lexicon entries, and the inclusion of seasons in which controversial social and political events happened.

3.2. Data Annotation

3.2.1. Overall Annotation Procedures

We customized and employed the Potato-POrtable Text Annotation TOol[1] for the data annotation. Annotators were provided annotation guidelines, took hands-on practical training, completed independent sample test tasks, and participated in group evaluation of independent sample tests they completed. A total of 8.3k tweets are annotated into **hate**, **offensive**, **normal**, and **indeterminate** classes as shown in Table 2. Besides, annotators were requested to identify the targets of hateful tweets and also indicate their ratings of the extent of hatefulness and offensiveness intensities of tweets on a 5-point Likert scale as indicated in Figure 1. The entire annotation process consists of a pilot round and five subsequent batches for the primary task annotations. Each tweet is annotated by 5 independent annotators, and the gold labels are determined with a majority voting scheme. A Fleiss' kappa score of 0.49 is achieved among the five annotators. We compensated annotators with a payment of $0.03 per tweet, roughly 180 ETB per hour on average,

[1] https://github.com/davidjurgens/potato

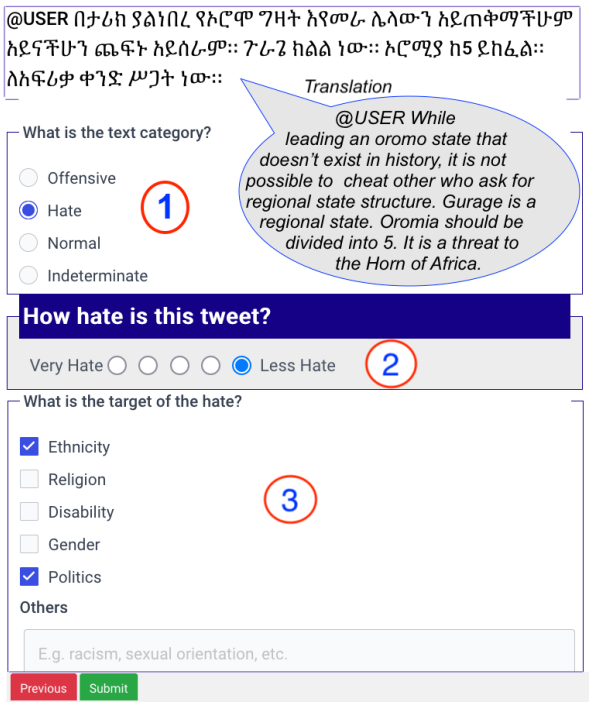

Figure 1: Potato GUI for the three types (1 - category, 2 - intensity, and 3 - target) of annotation tasks.

nearly the same as the hourly wage of a Master's degree holder in Ethiopia.

3.2.2. Backgrounds of Annotators

A total of 11 Amharic native speakers, 5 female and 6 male annotators, were engaged in the annotation task, representing a diverse range of ethnic, religious, gender, and social backgrounds. Annotators comprised of 6 MSc graduates and 5 MSc students from both Natural and Social Science disciplines.

Table 1 presented examples, which showed the structure of the annotated dataset for the three types of annotations; namely category, hatred target and intensity (hatefulness and offensiveness) annotations.

3.2.3. Tweet Category Annotation

As indicated in Table 2, the 5 annotators absolutely agreed on 3.2k tweets out of 8.3k, which is 39% of the total dataset. The absolute agreements on each category label among the annotators consisted of 38% and 31% for hateful and offensive tweets, respectively. The best absolute agreement of 49% per category label is achieved for the normal class. The indeterminate class consisting of only 42 tweets, demonstrated exceptionally infrequent occurrence and is excluded from our experiments. The indeterminate tweets are composed in a language other than Amharic or are unintelligible, thus failing to convey clear messages to the annotators. While determining majority-voted tweets for two labels with equal frequency of 2, we handle ambiguities by giving priority to **hate, offensive, and indeterminate** labels, respectively.

3.2.4. Target Annotation

As indicated in Table 3, a significant majority of the target dataset, totaling 3,249 tweets (53.4%), comprised of instances expressing hatred and hostility towards **political** targets. Political hatred tweets primarily centered on individuals based on their political ideologies, affiliations, or support for specific occasions. While ethnic hatred tweets presented the second majority, 38.8% of hateful tweets, religious and other targets exhibited smaller proportions in the dataset. Annotators achieved better absolute agreements on **ethnic**, **political**, and **religious** hatred targets. Overall, there is complete consensus on 14.3% of the hatred targets, which amounts to 867 instances within the target dataset. However, **gender** and other targets such as **disability** are scarcely represented in this dataset, which addresses **RQ-2**. The **none_hate** represented tweets that do not contain any hateful content.

Table 4 demonstrated the number of times different distinct targets appeared simultaneously across the 5 annotators within the original dataset. It provided a detailed overview of the collective perspectives of these annotators regarding the simultaneous presence of distinct targets. The majority of overlapping occurrences that happened between **ethnic** and **political** targets in the dataset showed how *ethnic and political hatred targets frequently intersect and overlap with one another*, emphasizing the complex relationship between these two targets. This overlap is likely a manifestation of Ethiopia's political landscape, which is primarily structured around ethnic divisions (Mostafa and Meysam, 2023). In Ethiopia, most political parties are established based on ethnic affiliations. This underscores the intricate connection between ethnicity and political tensions in the nation's sociopolitical context, which addresses **RQ-3**.

3.2.5. Intensity Level Annotation

We have organized our intensity level annotation task into three distinct segments. **Normal** texts are assigned a score of **0**, waiving the need for intensity level annotations. The offensiveness scale spans from **less offensive (1)** to **very offensive (5)**, utilizing a 5-point Likert scale for intensity level annotation. Similarly, the intensity of hatefulness is also rated on a 5-point Likert scale, ranging from **less hate (1)** to **very hate (5)**.

Table 5 presented the offensiveness and hatefulness intensities of tweets that appeared at least

Tweet	Category					Hatred Targets					Offensiveness Intensity					Hatefulness Intensity				
አንች ሸርሙጣ ከማያባሽ አትግቢ ይሄ ጭፈራ ቤት አይደለም ግም	off	off	off	off	off	--	--	--	--	--	3	4	5	5	4	--	--	--	--	--
You a whore, don't interfer in matters that doesn't concern you. This is not night club.																				
አሸባሪው የኦሮሙማ መንግስት	hat	hat	hat	hat	hat	['eth', 'pol']	['eth']	['eth']	['eth', 'pol']	['pol']	--	--	--	--	--	4	4	4	4	4
The terrorist Oromo-led government																				
@USER አንተ ደንቆሮ ነህ ስለ አርቶዶክስ አታቅም.	off	off	off	off	hat	--	--	--	--	['rel', 'dis']	4	3	4	4	--	--	--	--	--	3
You are ignorant, you don't know about Orthodox.																				
ቀይ መስቀል ለተፈናቃዮች 5 ሚሊየን ብር ግምት ያለው የዓይነት ድጋፍ አደረገ	nor	nor	nor	nor	nor	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
The Red Cross provided 5 million birr in-kind support to the displaced.																				

Table 1: Dataset examples with 5 annotators for category, hatred target and intensity (hatefulness and offensiveness) annotations. **Keys:** off = offensive, hat = hate, nor = normal, eth = ethnicity, pol = politics, rel = religion, dis = disability

Label	Majority Voted	Fully Agreed	Fully Agreed %
Hate	4,149	1,575	38%
Offensive	2,164	664	31%
Normal	1,945	956	49%
Indeterminate	42	6	14%
Total	**8,300**	**3,201**	**39%**

Table 2: Distribution of majority voted and fully agreed on category labels.

Target	Majority Voted	Fully Agreed	Fully Agreed %
Ethnic	**2,357**	**326**	14%
Politics	**3,249**	**487**	15%
Religion	359	54	15%
Gender	42	0	0%
Other	33	0	0%
None_Hate	2,220	1,620	73%
Total	**8,300**	**2,487**	**30%**

Table 3: Distribution of hatred targets across majority voted and fully agreed tweets.

2 times as offensive and hateful across the 5 annotators, respectively. Average offensiveness and hatefulness intensities on majority-voted tweets are

Coexisted Targets	Frequency	Percent
Ethnic, Politics	**3,290**	**83.0%**
Religion, Ethnic	291	7.3%
Religion, Politics	281	7.1%
Ethnic, Politics, Religion	101	2.6%
Major Co-occurrences	3,963	100%

Table 4: Main overlapping occurrences of targets.

	Majority Voted		Fully Agreed	
Label	Range	G-avg	Range	G-avg
Hate	0.4-5.0	2.48	1.4-5.0	3.56
Offensive	0.4-4.8	2.34	1.6-4.8	3.66

Table 5: Hatefulness and offensiveness intensities. The "range" indicates the intensity ranges per tweet while "G-avg" shows the grand average intensities. **Keys:** G-avg = Grand Average.

lower than the absolutely agreed tweets. The majority voted tweets exhibit wider ranges of intensities for both offensiveness and hatefulness, 0.40-4.80 and 0.40-5.0, respectively. This indicated that hate and offensive annotated tweets in the dataset are represented in a spectrum of wider ranges. Therefore, hatefulness and offensiveness **are not simple binary measures**, rather they exist on **a continuum with varying degrees of intensity**.

In the category of completely agreed tweets, the range of offensiveness intensity spans from a minimum average intensity of 1.60 to a maximum average intensity of 4.80 per tweet. Meanwhile, in the case of hateful tweets, their hatefulness intensity encompasses intensities ranging from a minimum of 1.40 to a maximum of 5.0 across the subset of entirely agreed tweets. The wider intensity ranges and the cumulative average intensity values for offensiveness and hatefulness on the completely agreed tweets highlight the presence of varying degrees of intensity, even among tweets that have absolute agreements.

Label	Average Range	Stage	Tweet Count	%
Offensive	[0.2 - 3.0)	Mild	**2,008**	**69%**
	[3.0 - 4.0)	Moderate	676	23%
	[4.0 - 5.0]	Severe	245	8%
Hate	[0.2 - 3.0)	Early Warning	**3,489**	**72%**
	[3.0 - 4.0)	Dehuman-ization	808	17%
	[4.0 - 5.0]	Voilence & Incitement	528	11%

Table 6: Hatefulness and offensiveness intensity ranges, and distribution of tweets across stages.

3.3. Mapping Hate and Offensive Intensities

Bahador (2023) categorized hate speech into three major **stages**, namely 1) early warning, 2) dehumanization and demonization, and 3) violence and incitement. The **early warning** category starts with targeting **out-groups**[2] to different types of negative speech that have less intensity. **Dehumanization and demonization** involve dehumanizing and demonizing the out-groups and their members, associating with subhuman or superhuman negative characters. The last category, **violence and incitement** starts from the conceptual to the physical attacks and can result in more severe consequences such as incitement to violence and or even death against the out-groups under target.

Similarly, Chandra et al. (2020) classifies online abuse into three labels; 1) **biased attitude**, 2) **acts of bias and discrimination**, and 3) **violence and genocide**; to showcase the mild, moderate, and severe categories of abuse intensity.

The classification categories of Bahador (2023) and Chandra et al. (2020) are employed to represent the hatefulness and offensiveness intensities of tweets as indicated in Table 6. We employed the revised rating scale described in Section 3.2.5 and represent offensiveness into three stage categories (Chandra et al., 2020), mild, moderate, and severe represented by 1-3, 4, and 5 rating scales, respectively. Similarly, the first category of hatefulness, early warning is represented from 1-3 ratings on the 5-point Likert scale. The second, dehumanizing and demonizing, and the third, incitement to violence categories are represented with scale 4 and scale 5, respectively.

As shown in Table 6, we carefully selected tweets labeled offensive at least by two annotators and the remainder labeled normal to explore the offensiveness intensity of tweets. Similarly, we did the same for hatefulness and analyzed the hatefulness and offensiveness intensities separately. Offensive tweets that fall under the mild category, start from 0.2 minimum average intensity when only one of the annotators chooses offensive and rates its' offensiveness 1, and end at 3 maximum average intensity value. Tweets under this category comprised 69% of the offensive tweets and are assumed to be less offending when compared with the other categories. Highly offending tweets constitute 8% of the offensive tweets that present incitement or threats of violence against an individual while the moderate category accounts for 23% of the tweets that dehumanize or demonize individuals.

The majority of hateful tweets comprised of 72% tweets, fall under the less hate, early warning category. The 17% and 11% of tweets that fall under the second and third categories, respectively, require serious attention among different stakeholders such as the government, social media organizations, researchers, and non-governmental organizations (national and international). The mild and early warning stages of offensiveness and hatefulness can be taken as a demarcation point to enforce mitigation strategies by content moderators or other stakeholders. The playground for tackling hate and offensive speech on social media shall be at the first stages of early warning and mild, respectively. For our analysis and experimentation, we transform this scale to a range of 0 to 10, effectively creating an **11-point Likert scale**. In this revised scale, a score of 0 represents **normal** tweets while **offensive** and **hate** categories are scaled from 1 to 5 and 6-10 intensity ranges, respectively. The score of 1 and 5 denotes **less offensive** and **highly offensive** tweets, respectively. Similarly, 6 signifies **less hate**, and 10 represents a tweet characterized by **intense hate**. Figure 2 indicated the transformed dataset on an 11-point Likert rating scale.

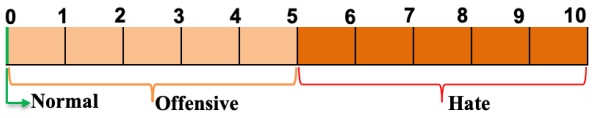

Figure 2: Mapping the dataset in an 11-point Likert rating scale.

3.4. Dataset Summary

A total of 8,258 instances were utilized for building classification and regression models, excluding the 42 indeterminate labeled instances. We presented the distributions of the dataset labels for the category, target, and intensity level classification and regression experiments in Table 2, Table 3, and Figure 3, respectively.

We convert the average values calculated from the input of five annotators into whole numbers,

[2]Out-groups are anyone who does not belong in the group but belongs to another group

resulting in a set of 11 labels spanning from 0 to 10. In this context, a label of 0 represents tweets labeled as **normal** while a label of 10 indicates tweets characterized as **extremely hateful**. Figure 3 illustrates that scale labels 1 and 10 are associated with a relatively smaller number of instances in comparison to the other labels, as these values correspond to the two extremes of the spectrum.

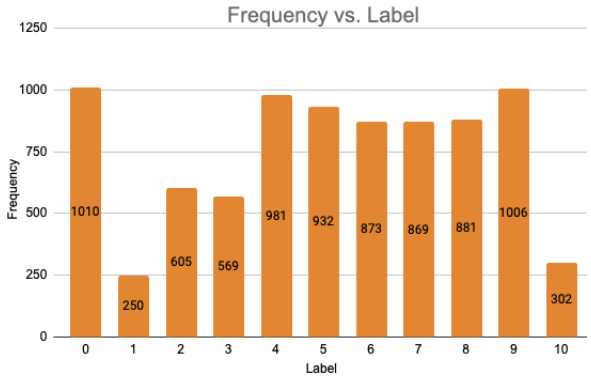

Figure 3: Distributions of 0-10 rating labels.

4. Experimental Setup

We employed a 70:15:15 data-splitting approach to create the training, development, and test sets. This dataset remained consistent across all experiments, including **category classification**, **target classification**, and **intensity scale regression**. The development dataset was instrumental in refining the learning algorithms, and all the results reported in this study are based on data from the test set.

We utilized the transformer models such as **AmRoBERTa**, **XLMR-Large-fintuned**, **AfroXLMR-large**, and **AfriBERTa** variants (small, base, large), and **AfroLM-Large (w/ AL)** for all experiments. AmRoBERTa is a RoBERTa-based language model that has been fine-tuned specifically with the Amharic language dataset, making it well-suited for downstream tasks and applications involving Amharic text (Yimam et al., 2021). We also utilized Afro-XLMR-large (Alabi et al., 2022), a multilingual language model tailored for African languages, including Amharic. This model demonstrated exceptional performance in various natural language processing tasks for African languages. Moreover, we fine-tuned the XLMR-Large (Conneau et al., 2019) model using the same corpus that was utilized to train AmRoBERTa. We also employed the small, base, and large **AfriBERTa** variants (Ogueji et al., 2021), and **AfroLM-Large (w/ AL)**, Pretrained multilingual models on many African languages including Amharic (Dossou et al., 2022). AfroLM Large (w/AL) is a special type of AfroLM Large which is designed with self active learning setups.

Tweet category classification results (in %)			
Classifier	**P**	**R**	**F1**
AmRoBERTa	75.01	75.06	74.82
XLMR-large-finetuned	73.60	73.45	73.50
Afro-XLMR-large	**75.37**	**75.30**	**75.30**
AfriBERTa-large	72.48	72.40	72.43
AfriBERTa-base	73.46	73.20	73.30
AfriBERTa-small	73.05	73.12	73.06
AfroLM-Large (w/ AL)	72.02	71.99	71.98
Hate target classification results (in %)			
AmRoBERTa	66.74	66.42	66.02
XLMR_large_fintuned	65.57	66.18	65.85
Afro_XLMR_large	**70.34**	**70.94**	**70.59**
AfriBERTa_large	66.94	67.47	67.14
AfriBERTa_base	66.04	66.42	66.11
AfriBERTa_small	65.38	66.02	65.68
AfroLM-Large (w/ AL)	64.26	64.57	64.23

Table 7: Performance of models for category and hatred targets classification of tweets. **Keys:** P = Precision, and R = Recall, AfroLM-Large (w/ AL) = AfroLM-Large (with Active Learning).

F1-score variations across tasks (in %)			
Classifier	**Cat.**	**Tar.**	**Diff.**
AmRoBERTa	74.82	66.02	**8.80**
XLMR-large-finetuned	73.50	65.85	7.65
Afro-XLMR-large	**75.30**	**70.59**	**4.71**
AfriBERTa-large	72.43	67.14	5.29
AfriBERTa-base	73.30	66.11	7.19
AfriBERTa-small	73.06	65.68	7.38
AfroLM-Large (w/ AL)	71.98	64.23	7.75

Table 8: F1-score Performance variations across models for category and hatred target classification tasks. **Keys:** Cat = Category, Tar = Target, and Diff = Difference, AfroLM-Large (w/ AL) = AfroLM-Large (with Active Learning).

5. Result and Discussion

As shown in Table 7, the Afro-XLMR-large model outperformed the other 6 models on both tweet category and hatred target classification tasks with 75.30% and 70.59% F1-scores, respectively. In comparison to their performance on target classifications, all models exhibited a pronounced increase in all performance indicators such as precision, recall and F1-scores when undertaking the category classification task. Table 8 indicated the spectrum of F1-score variations across diverse models. The performance variations observed in these two tasks extends from 4.71% for Afro-XLMR-large to 8.80% for AmRoBERTa. This disparity might be due to the class representation variations in the target classification task.

We conducted **regression** experiments on the dataset collected through the utilization of an 11-

Regression results on Likert's 11-scale (in %)	
Classifier	**Pearson's cor. coeff. (r)**
AmRoBERTa	77.23
XLMR-large-fintuned	76.17
Afro-XLMR-large	**80.22**
AfriBERTa_large	75.38
AfriBERTa_base	76.57
AfriBERTa_small	74.94
AfroLM-Large (w/ AL)	**80.22**

Table 9: Performance of models on the regression tasks with Likert's 11-scale data.

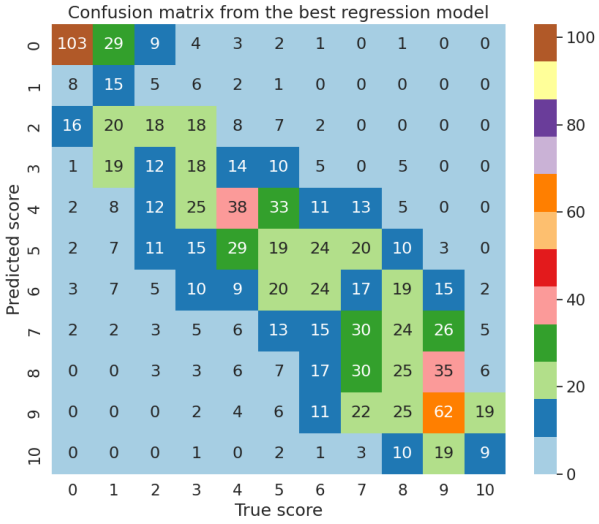

Figure 4: Confusion matrix from Afro-XLMR-large.

point Likert scale, which was employed to measure intensity levels across a broad spectrum of ratings. In these experiments, real-valued scores spanning from 0 to 10 were utilized, and various models were applied for analysis. As part of our methodology, we focused on enhancing the visualization of the regression results for better interpretation. To achieve this goal, we rounded the results and illustrated them with visual representations presented in Figure 4.

Regression experiments were also performed on the 11-point Likert scale data with various models, and their performance was assessed using Pearson's r correlation coefficients. As suggested by Schober et al. (2018), correlation coefficients falling between 0.70 and 0.89 are considered to indicate a strong correlation. Hence, the Pearson's r correlation coefficients achieved in this study, ranging from 74.94% to 80.22% demonstrated strong correlations. These findings denote a robust relationship between the predicted values and the actual observations, underscoring promising performance outcomes across all the models. The Afro-XLMR-large and AfroLM-Large (w/ AL) models presented the best results in the intensity scaling regression tasks, which is 80.22%. Figure 4 reveals that the majority of misclassified instances are clustered along the diagonal within the dark-colored boxes. This suggests that the true labels and their predicted counterparts are closely aligned. For instance, the true label 9 is frequently predicted as 7, 8, or 10, but seldom as 0, 1, 2, 3, or 4, which are considerably distant from 9. Conversely, there are only a few cases where extremely low true labels, such as 0, 1, 2, and 3, are predicted as higher extreme values, such as 7, 8, 9, or 10, and vice versa. In general, the regression model consistently displayed superior and more dependable performance as evidenced by the distribution of predictions in the confusion matrix. The findings indicate that considering hate speech as a continuous variable, rather than adopting a binary classification, is a more suitable approach. Regression-based methods excel at capturing the intricate and evolving characteristics of hate speech, recognizing the subtle variations and intensities within this complex and sensitive domain. This approach aligns with the dynamic and multifaceted nature of hate speech in the real-world situations, where it often exists on a spectrum of varying intensities, defying the usual simple binary categorization approaches. These findings address our research question, **RQ-1**.

6. Conclusion and Future Work

This paper introduced extensive benchmark datasets encompassing 8,258 tweets annotated for three tasks. These tasks included 1) **categorizing** hate speech into labels such as hate, offensive, and normal, 2) identifying the **targets** of hate speech, such as ethnicity, politics, and religion etc, and 3) assigning hate and offensive speech **intensity levels** using **Likert rating scales** to indicate offensiveness and hatefulness. To ensure robust annotation, each tweet is annotated by five annotators, resulting in a Fleiss kappa score of 0.49. Our contribution extended beyond the dataset itself; we provided comprehensive annotation guidelines tailored to each task and offered illustrative examples that effectively outlined the scope and application of these guidelines. After a comprehensive analysis of the dataset, a clear pattern emerged, highlighting the prominence of **political** and **ethnic** targets, which mirrors the complex and unstable sociopolitical environment of Ethiopia. Notably, these two targets often co-occur in hateful tweets, underscoring the intricate nature of Ethiopia's sociopolitical dynamics, especially within ethnic contexts. Furthermore, our findings have demonstrated variations in the intensity of hate speech, emphasizing the necessity to develop regression models capable of gauging the level of toxicity in tweets. We conducted a comprehensive exploration of various models for the de-

tection of hate speech **categories**, their associated **targets**, and their **intensity levels**. Afro-XLMR-large demonstrated superior performance across all tasks **category classification**, **target classification** and **intensity prediction**. Our research illustrated that offensiveness and hatefulness cannot be simply categorized as binary concepts; instead, they manifest as continuous variables that assume diverse values along the continuum of ratings.

In the future, there is potential for a more in-depth examination of hatefulness and offensiveness intensities at finer levels. Moreover, the dataset could be subjected to further analysis to determine whether the predicted hate speech intensity levels can be employed as a valuable tool for monitoring and preventing potential conflicts, which would be particularly beneficial for peace-building efforts. We released our dataset, guidelines, top-performing models, and source code under a permissive license[3].

Limitations

The research study has the following limitations. The small dataset size, 8,258 tweets, could limit the robustness and applicability of the results to be generalized in various contexts. Secondly, the scarcity of the normal and offensive class instances within the dataset might impact the model's ability to accurately detect these categories. The extreme data imbalance in the target dataset, dominated by political and ethnic targets, might have affected the detection of other targets. The pre-selection strategy of tweets with dictionaries also affected the true distribution of hateful tweets in the corpus. Additionally, the smaller representations of label 1 and label 10 in the dataset annotated for rating intensity levels might have affected the performance of classification and regression models. These limitations collectively highlight the need for further investigations with larger datasets, and balanced representations of the examples for all the three types of tasks.

7. Bibliographical References

Zeleke Abebaw, Andreas Rauber, and Solomon Atnafu. 2022. Design and implementation of a multichannel convolutional neural network for hate speech detection in social networks. *Revue d'Intelligence Artificielle*, 36(2):175–183.

Halefom H Abraha. 2017. Examining approaches to Internet regulation in Ethiopia. *Information and Communications Technology Law*, 26(3):293–311.

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Abinew Ali Ayele, Tadesse Destaw Belay, Seid Muhie Yimam, Skadi Dinter, Tesfa Tegegne Asfaw, and Chris Biemann. 2022a. Challenges of Amharic hate speech data annotation using Yandex Toloka crowdsourcing platform. In *Proceedings of the sixth Widening NLP Workshop (WiNLP)*, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Abinew Ali Ayele, Skadi Dinter, Tadesse Destaw Belay, Tesfa Tegegne Asfaw, Seid Muhie Yimam, and Chris Biemann. 2022b. The 5Js in Ethiopia: Amharic hate speech data annotation using Toloka Crowdsourcing Platform. In *Proceedings of the 4th International Conference on Information and Communication Technology for Development for Africa (ICT4DA)*, pages 114–120, Bahir Dar, Ethiopia.

Abinew Ali Ayele, Skadi Dinter, Seid Muhie Yimam, and Chris Biemann. 2023a. Multilingual racial hate speech detection using transfer learning. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 41–48, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Abinew Ali Ayele, Seid Muhie Yimam, Tadesse Destaw Belay, Tesfa Asfaw, and Chris Biemann. 2023b. Exploring Amharic hate speech data collection and classification approaches. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 49–59, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Babak Bahador. 2023. Monitoring hate speech and the limits of current definition. In Christian Strippel, Sünje Paasch-Colberg, Martin Emmer, and Joachim Trebbe, editors, *Challenges and perspectives of hate speech research*, volume 12 of *Digital Communication Research*, pages 291–298. Berlin.

Fatih Beyhan, Buse Çarık, İnanç Arın, Ayşecan Terzioğlu, Berrin Yanikoglu, and Reyyan Yeniterzi. 2022. A Turkish hate speech dataset and

[3] https://github.com/uhh-lt/AmharicHateSpeech

detection system. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4177–4185, Marseille, France. European Language Resources Association.

João Bran and Adeline Hulin. 2023. *Social Media 4 Peace: local lessons for global practices*. Countering hate speech. the United Nations Educational, Scientific and Cultural Organization (UNESCO).

Tommaso Caselli and Hylke Van Der Veen. 2023. Benchmarking offensive and abusive language in Dutch tweets. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, Toronto, Canada. Association for Computational Linguistics.

Mohit Chandra, Ashwin Pathak, Eesha Dutta, Paryul Jain, Manish Gupta, Manish Shrivastava, and Ponnurangam Kumaraguru. 2020. AbuseAnalyzer: Abuse detection, severity and target prediction for gab posts. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6277–6283, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Muluken Asegidew Chekol, Mulatu Alemayehu Moges, and Biset Ayalew Nigatu. 2023. Social media hate speech in the walk of Ethiopian political reform: analysis of hate speech prevalence, severity, and natures. *Information, Communication & Society*, 26(1):218–237.

Christopher Clarke, Matthew Hall, Gaurav Mittal, Ye Yu, Sandra Sajeev, Jason Mars, and Mei Chen. 2023. Rule by example: Harnessing logical rules for explainable hate speech detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 364–376, Toronto, Canada. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the Eleventh International AAAI Conference on Web and Social Media*, volume 11, pages 512–515, Montréal, QC, Canada. Association for Computational Linguistics.

Abreham Gebremedin Debele, Michael Melese and Woldeyohannis. 2022. Multimodal Amharic hate speech detection using deep learning. In *2022 International Conference on Information and Communication Technology for Development for Africa (ICT4DA)*, pages 102–107. IEEE.

Naol Bakala Defersha and Kula Kekeba Tune. 2021. Detection of hate speech text in afan oromo social media using machine learning approach. *Indian Journal of Science Technology*, 14(31):2567–2578.

Mequanent Degu, Abebe Tesfahun, and Haymanot Takele. 2023. Amharic language hate speech detection system from Facebook memes using deep learning system. *Available at SSRN 4389914*.

Christoph Demus, Jonas Pitz, Mina Schütz, Nadine Probol, Melanie Siegel, and Dirk Labudde. 2022. Detox: A comprehensive dataset for German offensive language and conversation analysis. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 143–153, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Bonaventure F. P. Dossou, Atnafu Lambebo Tonja, Oreen Yousuf, Salomey Osei, Abigail Oppong, Iyanuoluwa Shode, Oluwabusayo Olufunke Awoyomi, and Chris Emezue. 2022. AfroLM: A self-active learning-based multilingual pretrained language model for 23 African languages. In *Proceedings of The Third Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*, pages 52–64, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Paula Fortuna, Juan Soler, and Leo Wanner. 2020. Toxic, hateful, offensive or abusive? What are we really classifying? An empirical analysis of hate speech datasets. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6786–6794, Marseille, France. European Language Resources Association.

Ankur Joshi, Saket Kale, Satish Chandel, and D Kumar Pal. 2015. Likert scale: Explored and explained. *British journal of applied science & technology*, 7(4):396–403.

Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani,

and Xiang Ren. 2020. Contextualizing hate speech classifiers with post-hoc explanation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5435–5442, Online. Association for Computational Linguistics.

Svetlana Kiritchenko, Isar Nejadgholi, and Kathleen C. Fraser. 2021. Confronting abusive language online: A survey from the ethical and human rights perspective. *J. Artif. Intell. Res.*, 71:431–478.

Kosisochukwu Madukwe, Xiaoying Gao, and Bing Xue. 2020. In data we trust: A critical analysis of hate speech detection datasets. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 150–161, Online. Association for Computational Linguistics.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. HateXplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence*, pages 14867–14875, Palo Alto, CA, USA. Association for the Advancement of Artificial Intelligence.

Zewdie Mossie and Jenq-Haur Wang. 2018. Social network hate speech detection for Amharic language. In *4th International Conference on Natural Language Computing (NATL2018)*, pages 41–55, Dubai, United Arab Emirates. AIRCC Publishing.

Zewdie Mossie and Jenq-Haur Wang. 2020. Vulnerable community identification using hate speech detection on social media. *Information Processing & Management*, 57(3):1–16.

Ghaderi Hajat Mostafa and Mirzaei Tabar Meysam. 2023. The impact of spatial injustice on ethnic conflict in Ethiopia. *Geopolitics Quarterly*, 19(70):41–65.

Nicolas Ocampo, Ekaterina Sviridova, Elena Cabrio, and Serena Villata. 2023. An in-depth analysis of implicit and subtle hate speech messages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1997–2013, Dubrovnik, Croatia. Association for Computational Linguistics.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. Small data? No problem! Exploring the viability of pretrained multilingual language models for low-resourced Languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4675–4684, Hong Kong, China. Association for Computational Linguistics.

John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2017. Deeper attention to abusive user content moderation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1125–1135, Copenhagen, Denmark. Association for Computational Linguistics.

Flor Miriam Plaza-del arco, Debora Nozza, and Dirk Hovy. 2023. Respectful or toxic? Using zero-shot learning with language models to detect hate speech. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 60–68, Toronto, Canada. Association for Computational Linguistics.

Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia von Vacano, and Chris Kennedy. 2022. The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 83–94, Marseille, France. European Language Resources Association.

Salim Sazzed. 2023. Discourse mode categorization of Bengali social media health text. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 52–57, Toronto, Canada. Association for Computational Linguistics.

Patrick Schober, Christa Boer, and Lothar A Schwarte. 2018. Correlation coefficients: Appropriate use and interpretation. *Anesthesia & Analgesia*, 126(5):1763–1768.

Gudbjartur Ingi Sigurbergsson and Leon Derczynski. 2020. Offensive language and hate speech detection for Danish. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3498–3508, Marseille, France. European Language Resources Association.

Basu Prasad Subedi. 2016. Using Likert type data in social science research: Confusion, issues and challenges. *International journal of contemporary applied sciences*, 3(2):36–49.

Surafel Getachew Tesfaye and Kula Kakeba. 2020. Automated Amharic hate speech Posts and comments detection model using recurrent neural network. *Preprint*. Version 1.

Alice Tontodimamma, Eugenia Nissi, Annalina Sarra, and Lara Fontanella. 2021. Thirty years of research into hate speech: topics of interest and their evolution. *Scientometrics*, 126(1):157–179.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93, San Diego, CA, USA. Association for Computational Linguistics.

Seid Muhie Yimam, Abinew Ali Ayele, and Chris Biemann. 2019. Analysis of the Ethiopic Twitter Dataset for Abusive Speech in Amharic. In *In Proceedings of International Conference On Language Technologies For All: Enabling Linguistic Diversity And Multilingualism Worldwide (LT4ALL 2019)*, pages 210v–214, Paris, France.

Seid Muhie Yimam, Abinew Ali Ayele, Gopalakrishnan Venkatesh, Ibrahim Gashaw, and Chris Biemann. 2021. Introducing various semantic models for amharic: Experimentation and evaluation with multiple tasks and datasets. *Future Internet*, 13(11).

Frederike Zufall, Marius Hamacher, Katharina Kloppenborg, and Torsten Zesch. 2022. A legal approach to hate speech – operationalizing the EU's legal framework against the expression of hatred as an NLP task. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 53–64, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

8. Language Resource References