# Investigating Hate Speech Beyond Detection and Classification: Uncovering Complex Intensities and Targets

**Abinew Ali Ayele**
Universität Hamburg,
LT Group
Bahir Dar University, FC

**Esubalew Alemneh Jalew**
Bahir Dar University
ICT4D-BiT

**Adem Chanie Ali**
Bahir Dar University
Humanities Faculty

**Seid Muhie Yimam**
Universität Hamburg
HCDS

abinew.ali.ayele@uni-hamburg.de

## Abstract

Despite the complex nature of hate speech, studies focus primarily on detecting its binary categories, often overlooking the continuous spectrum of offensiveness and hatefulness inherent in the message. This study presents benchmark datasets for Amharic, comprising 8,258 tweets annotated for three distinct tasks: category classification, identification of hate targets, and rating of offensiveness and hatefulness intensities. Our study highlights that a considerable majority of tweets belong to the less offensive and less hateful 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. This study revealed that hate and offensive speech cannot be addressed by simplistic binary classification methods. Instead, they manifest themselves as variables across a continuous range of values. The Afro-XLMR-large model exhibits the best performance, achieving F1 scores of 75. 30%, and 70. 59% for the category and target classification tasks, respectively. The 80.22% correlation coefficient of the Afro-XLMR-large and Afro-XLMR-large-with-active-learning models exhibits strong alignments in the regression tasks.

## 1 Introduction

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. However, 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, this limitation has been recognized, prompting a shift towards the adoption of multifaceted methodologies to better understand 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 address 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. (2023, 2022a), predominantly focused on the classification of hate speech as a binary concept, overlooking its varying levels of intensities and targeted groups. In this study, we go beyond the traditional binary classification by examining the varying intensities of hate and offensive speech, as well as the specific communities targeted by such hatred. 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 toward a particular topic, where respondents are asked to choose the options that best reflect their point of view for each item (Subedi, 2016). Likert rating scale provides a quantitative measurement of qualitative data, which helps researchers analyze attitudes or opinions in a structured and comparable manner (Joshi et al., 2015).

The study addresses the following research questions:

- Do hate and offensive speech represent discrete binary categories, or exist on a continuous spectrum of varying intensities?
- What is the extent to which hate speech specifically targets certain groups?
- How frequently do tweets containing hate speech targeting multiple groups appear?

## 2 Data Collection and Annotation

The dataset has been collected from **X/formerly Twitter**. We used different data selection strategies such as hate and offensive lexicon entries. The dataset comprised 8.3k tweets, each annotated by 5 native speakers covering three distinct types of tasks namely; **category**, **target**, and **intensity level**. The category annotation includes **hate**, **offensive**, **normal**, and **indeterminate** classes. In addition, the annotators were requested with identifying the targets of hateful tweets, such as ethnicity, politics, religion, gender, and disability. They were also asked to rate the intensity of hatefulness and offensiveness of each tweet on a 5-point Likert scale, with ratings ranging from 1 to 5. The entire annotation consists of a pilot and five subsequent main annotation batches and achieved a Fleis kappa agreement score of 0.49. More than 83% of the hateful tweets in the target dataset exhibited overlapping occurrences.

Within the annotated dataset, there have been large considerable of hateful tweets targeting people based on their political, ethnic, and religious identities. Politics and ethnic identities mainly appeared together within hateful tweets in the dataset as indicated in Figure 1.

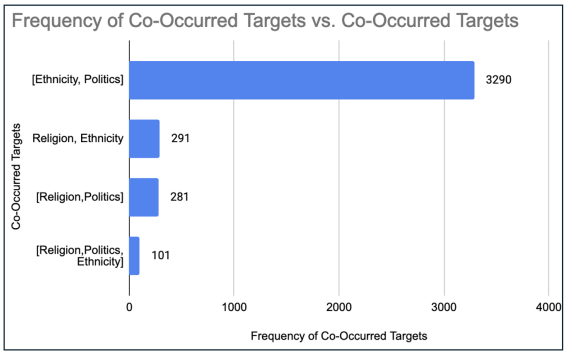


Figure 1: Major overlapping hatred target occurrences across hateful tweets in the dataset.

We mapped offensiveness and hatefulness intensities of messages, representing in a continuum of ranging between 0 and 10, where 0 denotes a normal speech. Offensiveness intensities have been represented in a range of 1-5 while hatefulness mapped with 6-10 intensities, perfectly creating an 11-scale intensity dataset as depicted in Figure 2.

## 3 Results and Discussion

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 situations in Ethiopia. Notably, these two targets often co-exist in hateful tweets, underscoring the intricate nature of Ethiopia's sociopolitical dynamics, especially within ethnic contexts. Our findings also showed variations in toxic intensities of tweets, emphasizing the need to develop regression models capable of predicting the level of hatefulness and offensiveness in tweets. The majority of hateful (69%) and offensive (72%) tweets fall into less hate and less offensive categories, respectively. Although severe offensive tweets constitute 8%, extreme hateful tweets that could call for violence and genocide accounts 11% of the hateful category. These results signify the need for early interventions from stakeholders to mitigate hate speech in Amharic.

This study employed a 70:15:15 data-split strategy for train, development, and test sets construction across all tasks and models. We conducted a comprehensive exploration of various models for the detection of hate speech **categories**, their associated **targets**, and their **intensity levels**. The study employed models such as **AmRoBERTa** (Yimam et al., 2021), **XLMR-Large-fintuned** (Conneau et al., 2019), **AfroXLMR-large** (Alabi et al., 2022), and variants of **AfriBERTa; small, base, large** (Ogueji et al., 2021) and **AfroLM-Large-with-active-learning** (Dossou et al., 2022) for all experiments. Afro-XLMR-large demonstrated superior performance across all tasks **category classification**, **target classification** and **intensity prediction**. It achieved 75.30% and 70.59% F1 scores on both tweet category and hatred target classification tasks, respectively. We performed regression tasks using the intensity rating scale data, where the models achieved Pearson's r correlation coefficients ranging from 74.94% to 80.22%, indicating strong correlations, as shown in Figure 3. These findings denote a robust relationship between the predicted values and the actual observations, underscoring promising performance outcomes across all models. The Afro-XLMR-large and AfroLM-Large (w/ AL) models presented the best results in the regression tasks, which is 80.22%.

As presented in Figure 4, the majority of errors, 47.84%, within the predicted intensities showed only 1 scale variation with the actual annotation scores. The second majority presented a 2 scale differences between the actual and predicted intensities, which accounts 28.36% of the errors.

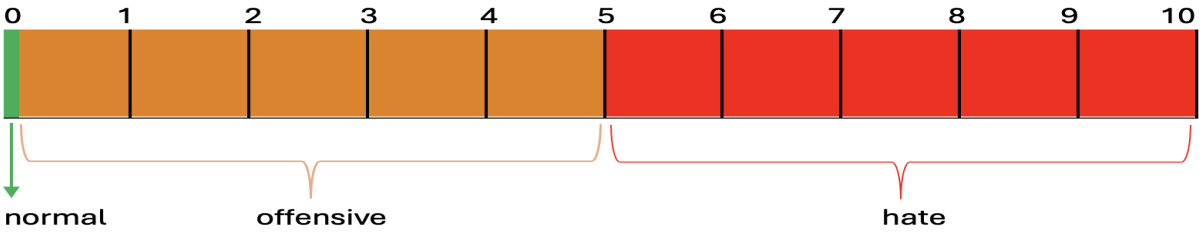


Figure 2: Distributions of 0-10 rating scales

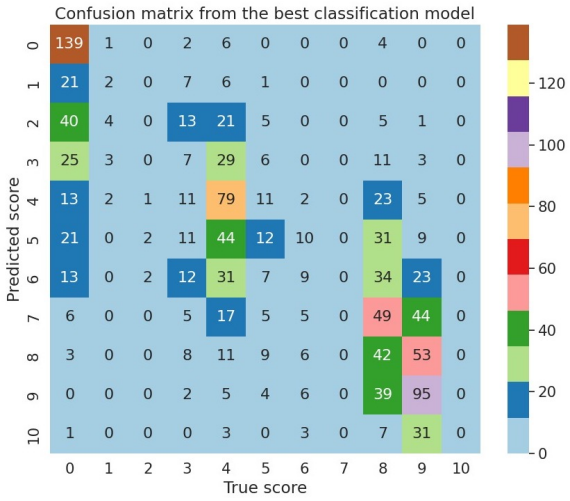


Figure 3: Confusion matrix from Afro-XLMR-large.

Over 76% of the predictions are closer to the actual values, with 1 or 2 intensity scale differences. Such small variations are also common experiences among human experts due to subjectivity.

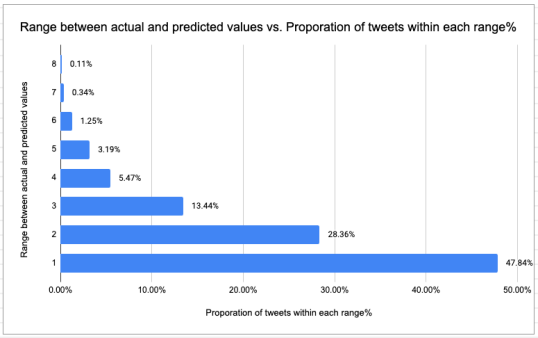


Figure 4: Variations within actual and predicted intensity ratings of tweets.

## 4 Conclusion

This paper introduces datasets comprising 8,258 annotated tweets for categorizing hate speech, identifying targets such as ethnicity, politics, and religion, and assigning intensity levels using Likert scales. With five annotators, a Fleiss kappa score of 0.49 was achieved. The analysis indicates that political and ethnic targets frequently co-occur within Ethiopia's sociopolitical landscape, necessitating regression models to predict intensity levels. The Afro-XLMR-large model performed exceptionally well across all tasks, illustrating that offensiveness and hatefulness can be treated as continuous variables. Future research could focus on refining intensity levels and leveraging the dataset for conflict monitoring and peace-building efforts. The datasets, guidelines, models, and source code will be released under a permissive license.

## 5 Limitations

The research study faces several limitations affecting its findings. The small dataset of 8,258 tweets limits the robustness and generalizability of the results. Furthermore, the low availability of normal and offensive class instances may hinder accurate detection of these categories. The dataset's extreme imbalance, primarily focused on political and ethnic targets, could overlook other types of hate speech. The pre-selection of tweets through dictionaries also distorts the true distribution of hateful content. Lastly, the underrepresentation of certain intensity levels may impair the performance of both classification and regression models. These issues underscore the necessity for future research using larger and more balanced datasets.

## 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.

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 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 *2022 International Conference on Information and Communication Technology for Development for Africa (ICT4DA)*, pages 114–120, Bahir Dar, Ethiopia.

Abinew Ali Ayele, Seid Muhie Yimam, Tadesse Destaw Belay, Tesfa Asfaw, and Chris Biemann. 2023. 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 detection system. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4177–4185, Marseille, France. European Language Resources Association.

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.

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, 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.

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.

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*, volume 35, pages 14867–14875, Palo Alto, California, USA.

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.

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.

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.

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.

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, California, USA.

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).