Exploring Amharic Hate Speech Data Collection and Classification Approaches

Abinew Ali Ayele[1,2], **Seid Muhie Yimam**[1], **Tadesse Destaw Belay**[3], **Tesfa Tgegne Asfaw**[2], **Chris Biemann**[1]

[1]Universität Hamburg, Germany, [2]Bahir Dar University, Ethiopia, [3]Wollo University, Ethiopia

{abinew.ali.ayele, seid.muhie.yimam, chris.biemann}@uni-hamburg.de,
tadesseit@gmail.com, tesfat@gmail.com

Abstract

In this paper, we present a study of efficient data selection and annotation strategies for Amharic hate speech. We also build various classification models and investigate the challenges of hate speech data selection, annotation, and classification for the Amharic language. From a total of over 18 million tweets in our Twitter corpus, 15.1k tweets are annotated by two independent native speakers, and a Cohen's kappa score of 0.48 is achieved. A third annotator, a curator, is also employed to decide on the final gold labels. We employ both classical machine learning and deep learning approaches, which include fine-tuning AmFLAIR and AmRoBERTa contextual embedding models. Among all the models, AmFLAIR achieves the best performance with an F1-score of 72%. We publicly release the annotation guidelines, keywords/lexicon entries, datasets, models, and associated scripts with a permissive license[1].

1 Introduction

In this digital era, social media platforms have become an important part of everyday life for people globally. The 2023 Global Digital Report disclosed that nearly 5.16 billion people use the internet and the number of social media users exceeded 4.76 billion worldwide. Over 64.4% of the world's population is already online, and nearly 60% of the people are active users of different social media platforms (Kemp, 2023).

Hateful content targeting minorities is rapidly spreading across social media platforms and becoming a major socio-political and cultural challenge in the world (Williams et al., 2020). To tackle the problem, many countries, like Ethiopia, crafted hate speech regulation laws even though the regulations have limitations for implementation (Ayalew, 2020). Moreover, there has been a rising interest among researchers in hate speech detection to expose and regulate this phenomenon with technological solutions. In this regard, researchers like Mathew et al. (2021); Ousidhoum et al. (2019); Poletto et al. (2017); Davidson et al. (2017); Waseem and Hovy (2016) have proposed several hate speech classification models and datasets for the development of automatic hate speech detection systems. Despite many researchers claiming state-of-the-art performance on their own datasets, the models can not be generalized for all languages and datasets (Gröndahl et al., 2018).

Ethiopia's legal regulations that were designed to counteract hate speech are not very well implemented. This is due to the complex nature of the online community, which is difficult to control by local laws, and the anonymity of online users who spread hateful messages while hiding behind their screens. Moreover, the available hate speech classification models built for high-resource languages such as English could not be used for low-resource languages like Amharic since such tasks incorporate cultural, social, and political variations in addition to language-specific differences. We have compiled Amharic hate speech datasets from Twitter and built classification models using different machine-learning approaches.

In this paper, we addressed the following research questions, which are formulated for Amharic, but also apply to other low-resource languages:

1. How to identify appropriate data collection and selection approaches for constructing hate and offensive speech datasets for Amharic?

2. What are the main challenges in the annotation and classification tasks of Amharic hate speech?

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

The paper presented benchmark hate speech data selection approaches, a dataset consisting of over 15.1k annotated tweets, and various classification models. This work has the following main contributions:

1. A well-defined hate speech data selection and preprocessing pipeline for hate speech annotation,
2. The collection of benchmark hate and offensive speech lexicon entries,
3. The development of hate speech annotation guidelines and strategies for quality data annotations, and
4. Releasing benchmark dataset and classification models for Amharic hate speech task.

We organized the remainder of the paper as follows. The study provides introductory information about the Amharic language in Section 2. Related works are presented in Section 3. Data collection and preprocessing details are discussed in Section 4. Data annotation strategies are described in Section 5. We present classification models in Section 6 and the results and discussion part of the paper in Section 7. An error analysis of model results is described in Section 8. Section 9 concludes and shortly discusses future avenues, limitations are indicated in Section 10.

2 Amharic Language

Amharic is the second-largest widely spoken Semitic language next to Arabic. It is written from left to right with its own unique 'Fidäl' scripts. Fidäl is a syllable-based writing system where the consonants and vowels co-exist within each graphic symbol. Amharic is the working language of the Federal government in Ethiopia and many regional states in the country (Salawu and Aseres, 2015). In Amharic, there are 34 core characters each having seven different variations to represent vowels. Besides, it has 51 labeled characters, 20 numerals, and 8 punctuation marks. Amharic uses more than 310 unique characters and is a morphologically complex and highly inflected language (Gezmu et al., 2018).

3 Related Work

Hate speech refers to language content that targets identity such as ethnicity, gender, disability, or political and religious ideology, which indirectly or directly focuses on their group identity and has the potential to incite violence (Casanovas and Oboler, 2018). In contrast, offensive speech is a speech that usually targets individuals to be offended but not based on their group identity (Casanovas and Oboler, 2018).

Hate speech has been addressed by many researchers using data scraped from online messages on social media. Among the various studies, Waseem and Hovy (2016); Davidson et al. (2017); Founta et al. (2018); ElSherief et al. (2018); Ousidhoum et al. (2019); Founta et al. (2019); Winter and Kern (2019); Mathew et al. (2021); Röttger et al. (2022b); Demus et al. (2022); Röttger et al. (2022a) conducted hate speech research in languages such as English, German, French, Arabic, Spanish, Portuguese and Hindi and published their datasets and models to advance further research.

As indicated in Table 1, among a few studies conducted for Amharic, the work by Mossie and Wang (2018); Tesfaye and Kakeba (2020) and Abebaw et al. (2021) used binary classification (hate or non-hate) labels on Facebook comments using different machine learning algorithms, while Mossie and Wang (2020) further tried to identify vulnerable communities to hate speech among the major ethnic groups in Ethiopia. The studies by Abebaw et al. (2021); Mossie and Wang (2018, 2020) have collected their datasets from the Facebook pages of some media organizations for a few months and from limited users. Ayele et al. (2022b) presented a crowd-sourced Amharic hate speech dataset from Twitter with a kappa score of 0.34 and a model performance of 50% for the F1-score with AmRoBERTa which is fine-tuned for the Amharic. The dataset presented by Ayele et al. (2022b) is a low-quality dataset since it is collected using a crowd-sourcing annotation approach in a low-resource language context and its lower performance score may also be associated with the dataset quality. Even though most of the authors have reported state-of-the-art performance results, we can not reproduce the results since neither the datasets nor the models are published publicly except the one described in Ayele et al. (2022b).

4 Data Collection and Preprocessing

We have been collecting and storing Amharic tweets every day since 2014 and built a Twitter dataset in a relational database using the Twitter API. Our algorithm scrapes large numbers of tweets

Author	Size	Labels	Best Method	Best Score	Resources Available
Mossie and Wang (2018)	6,120	hate, not hate	Naïve Bayes	79.8%: acc	No
Mossie and Wang (2020)	14,266	hate, not hate	CNN-GRU	92.6%: acc	No
Tesfaye and Kakeba (2020)	30,000	hate, free	LSTM	97.9%:acc	No
Abebaw et al. (2021)	2,000	hate, not hate	SVM	92.5% :F1	Dataset only
Abebaw et al. (2022)	2,000	hate, not hate	MC-CNN	68.5% :F1	Same Dataset
Ayele et al. (2022b)	5,267	hate, normal, offensive	RoBERTa	50.0%: F1	Yes

Table 1: Amharic hate speech studies (data size, labels, method, and best score and resource availability)

that are written in Amharic, Awgni, Guragigna, Ge'ez, Tigrinya, or other Semitic languages that use the Fidäl script. Currently, we have collected and stored more than 18 million tweets. As indicated in Figure 1, the number of tweets stored in our repository showed a substantial increase since 2020 due to the evolving economic, social, and political dynamics in Ethiopia. Particularly, in the years 2020, 2021, 2022, and 2023 until April showed a significant increase in the number of tweets collected every day, which might be due to the following reasons:

1. The prevalence of the Covid-19 pandemic and its global impacts,
2. Ethiopia's Tigray region holds a regional election in defiance of the federal government,
3. The escalations of various national socio-political problems in Ethiopia,
4. The conflict between the federal government and the Tigray People's Liberation Front (TPLF) in the Tigray region,
5. The 6th Ethiopian national election,
6. The assassination of artist Hachalu Hundessa and the imprisonment of opposition political party leaders in Oromia region due to the mass demonstrations and violence in the region following the death of the artist, and
7. The Grand Ethiopian Renaissance Dam (GERD) dispute between Ethiopia and Egypt reached a high peak. The GERD case was even taken to the UN security council despite Ethiopia's complaints that it was not a security issue at all.

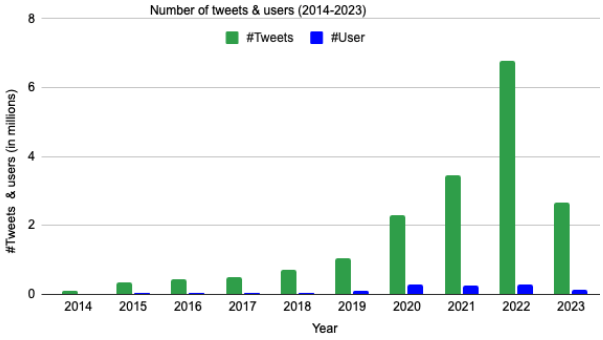

Figure 1: Number of tweets and users scraped per year

For this research, we collected 3.8 million tweets from October 2020 to November 2021 for 14 consecutive months, mainly focusing on tweets that are written during the socio-political dynamics in Ethiopia, mainly related to the reasons mentioned above (#2, #3, #4, #5, and #7).

4.1 Data Sampling

Figure 2 presented the various data collection, preprocessing, and sampling strategies employed in the paper. We removed retweets and filtered out non-Amharic tweets using the Python language detection tool[2] resulting in 902k tweets out of 3.8 million tweets. Through employing hate and offensive lexicon entries, we further filtered the tweets

[2] https://pypi.org/project/langdetect/

and reduced the target dataset to 153k tweets. Figure 3, shows a sample of some hate and offensive keywords that are used to filter the dataset. The keywords were collected from volunteer communities through Google Forms shared via social media platforms. We have also used the keywords listed in Yimam et al. (2019) as an initial query.

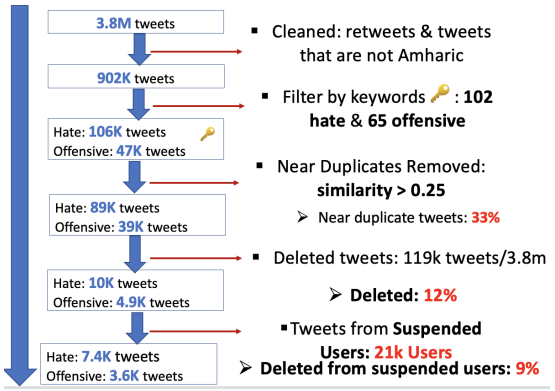

Figure 2: Data selection and preprocessing pipeline

We further examined the filtered tweets for a random number of samples and find out that there are tweets with unique IDs but are duplicates or near-duplicates of each other. This might be due to some users who copy and post others' tweets with some minor modifications. We explored different mechanisms and employed shingling methods to filter the near duplicate tweets using the Jaccard similarity index. The Jaccard similarity measure of all the pairs of tweets was calculated and the near duplicate tweets were obtained. We considered a 25% similarity score as the maximum tolerable threshold value and achieved 130k unique clean tweets by removing all the tweets that have a Jaccard index greater than the threshold value (i.e. with less than 25% similarity). It is indicated that 33% of the tweets are near duplicates in the corpus, and therefore are excluded from being sampled for this study.

Figure 3: Sample hate and offensive keywords

4.2 Dealing with Deleted Tweets

Twitter deletes some tweets that are reported as inappropriate and even suspends some users due to various reasons. We explored many deleted tweets and found out that 12% of the tweets in our repository are deleted from Twitter and are no more available. Among the deleted tweets, around 9% are from suspended users alone. We have annotated some samples of deleted tweets from both active and suspended users for pilot investigations if they contain more hateful content than the accessible tweets.

We have finally created two large pools of unlabelled tweets, one containing keywords and the other without keywords. The keyword-based unlabelled pool consisted of around 113k accessible tweets containing hate and offensive keywords. The second unlabelled pool, which is without keywords, is comprised of accessible tweets that do not contain hate and offensive keywords. The tweets are anonymized by replacing usernames with <USER> tokens and removing URLs from the tweets.

5 Data Annotation

Previous studies on Amharic hate speech classification such as Mossie and Wang (2018, 2020); Abebaw et al. (2021) identified two classification categories (i.e. hate vs non-hate) while studies in English and other languages (Davidson et al., 2017; Mulki et al., 2019) used Hateful, Offensive, and Normal class categories. Recently, the study by Mathew et al. (2021) introduced the "unsure" category and employed four class categories, which are hate, offensive, normal, and unsure. We used the WebAnno[3] annotation tool, which is a web-based annotation framework for all annotations.

5.1 Pilot Annotation

As the first round of pilot annotation, we annotated 3k tweets containing hate and offensive keywords. As indicated in Table 2, the pilot data annotation covered mainly tweets from 3 different categories such as accessible tweets, deleted tweets from suspended users, and deleted tweets from active users.

Each tweet is annotated by three annotators. While the first two annotators labeled each tweet independently, the third annotator who served as

[3] https://www.inf.uni-hamburg.de/en/inst/ab/lt/resources/software/webanno.html

Category	Hate	offensive	Normal	Unsure	# Total
Accessible tweets with keywords	498	198	252	8	959
Deleted tweets from suspended with keywords	490	254	244	14	1000
Deleted tweets from active users with keywords	488	173	387	6	1055
Total number of annotated Tweets	**1,477**	**623**	**885**	**28**	**3013**

Table 2: Pilot annotated tweets by category

a curator or an adjudicator made the decisions on the final gold labels. A total of 5 annotators were involved in the pilot annotation task and each annotator earned 0.5 ETB or $0.01 cents per tweet. The annotators can label 150 tweets per hour and earn 75 ETB or $1.5, which is nearly equivalent to the hourly wage of BSc holders in Ethiopia. We prepare training manuals and annotation guidelines and deliver intensive training to make the task clear for the annotators and the curator.

The pilot annotation result consisted of 1487, 892, 627, and 28 tweets labeled as hate, offensive, normal, and unsure class labels respectively. We employed Cohen's kappa coefficient to compute the inter-annotator agreement (IAA) and achieved a 0.44 agreement score for the pilot annotation. Other related studies, for example, Del Vigna et al. (2017) reported a 0.26 inter-annotator agreement score on the Italian dataset while Ousidhoum et al. (2019) reported 0.153, 0.202, and 0.244 IAA scores of kappa coefficient on English, Arabic, and French datasets respectively. Besides, Mathew et al. (2021) reported a 0.46 inter-annotator agreement score on the English data set, which indicated a moderate agreement among annotators. Therefore, our 0.44 inter-annotator agreement score fell under the moderate category which encouraged us to pursue the main annotation task.

As shown in Table 2, hateful tweets seemed more dominating in the dataset since the pilot annotations in all categories used tweets consisting of keywords only. The deleted tweets were examined and compared with the accessible tweets if they contained more hateful content. No significant differences were found in the distributions of hateful tweets across the three categories (accessible tweets, deleted tweets from suspended users, and deleted tweets from active users). The deleted tweets are excluded from being sampled in the final dataset since they are no more available on Twitter.

5.2 Error Analysis of Pilot Annotations

Hate speech annotation is highly subjective and challenging even for human annotators (Fortuna et al., 2022; Ayele et al., 2022a). During the pilot study, we observed disagreements between annotators on their annotation labels due to the subjective nature of hate speech annotation. In some cases, the curator also deviated from both annotators and selected a different annotation label. Such annotation errors were analyzed with examples as presented in Figure 4. Despite hate speech annotation is a very subjective task, we tried to understand the different views of annotators using expert judgments. Three experts, a lawyer (Assistant professor in Law), a political science expert (Ph.D. student), and a journalism expert (Associate professor of media and communications) were engaged in a focus group discussion to analyze the potential sources of annotation disagreements between the annotators as well as the adjudicator. The experts evaluate the annotation deviations and suggest possible justifications for the source of the disagreements on the labels of those tweets. In general, we observed that hate speech annotation is a highly context-sensitive and challenging task (Ayele et al., 2022a), which usually resulted in lower inter-annotator agreements.

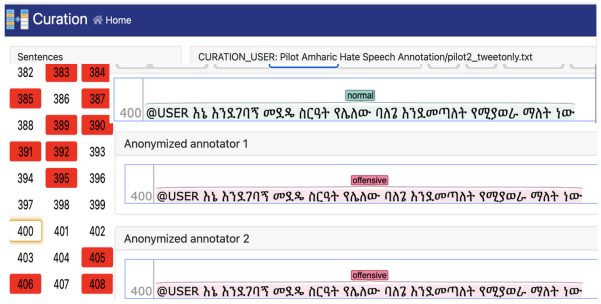

Figure 4: Sample deviations between annotators and the adjudicator taken from WebAnno (Yimam et al., 2013)

As shown in Figure 4, the two annotators agreed that the tweet (translated in English here) "as I understood it, 'Mededе' means a crazy, naughty and disrespectful person who talks randomly" is offen-

sive. The reason was that the annotators might have thought that the tweet targeted the user indicated in the tweet ('@USER') while the curator labeled the tweet as normal since the curator thought that the author of the tweet was defining the word 'Mededе' rather than targeting an individual. The red colored numbers (the left side) in Figure 4 showed that the two annotators disagreed on that item label while the tweets shaded with light red and light cyan colors (right side) represented the annotator's and curator's decisions respectively. In most cases, where annotators faced tweets with mixed languages other than Amharic, they usually annotated the tweet as "Unsure".

5.3 Main Annotation Task

The pilot annotation indicated that the selection from the lexicon-based unlabelled pool suffered from data imbalance problems. Therefore, we mixed the lexicon-based unlabelled pool with the non-lexicon-based pool on a 70/30 proportion. Each batch of annotations comprised 70% from the keyword-based unlabelled pool and 30% from the unlabelled pool with no keywords respectively. The annotation of the dataset including the pilot study took over a year. We performed the pilot annotations in 6 batches and the main annotations in 22 batches, where we analyzed each batch before pursuing the next batch. The annotators were nominated from different cultural, religious, gender, and age categories, and each user annotated from 3,800-4500 tweets. A kappa score of 0.48 is achieved on a dataset of over 15.1k tweets on the main annotation task which is better than the pilot task. The dataset consisted of 6,664, 5,554, 2,283, and 86 hate, normal, offensive, and unsure class label distributions respectively. The 86 tweets annotated as "unsure" were further examined with expert consultations to explore the sources of annotation decisions. Since the majority of the tweets labeled "unsure" contained mixed languages of non-Amharic words that confused annotators, they were excluded from being used in the experiment.

6 Classification Models

Texts on social media platforms are usually unstructured, written in mixed scripts, and lack uniformity in writing styles than texts in the normal context. Moreover, social media texts do not follow spelling/grammar rules as well as other language standards that make hate speech detection tasks a complex problem. Hate speech is linguistically, culturally, and historically dependent on the context of the speech and requires developing classifiers that capture these dependencies (Albadi et al., 2018).

6.1 Classical Machine Learning Approaches

These days, most hate and offensive speech classification studies mainly employ deep learning approaches despite they require large amounts of labeled datasets. In this study, we apply both the classical machine learning and deep learning approaches. We have also employed two contextual embedding approaches from the Amharic Semantic resource repository (Yimam et al., 2021).

The classical machine learning algorithms learned to make predictions through varieties of iterative learning processes from data without being explicitly programmed but only based on patterns and inference on the data (Mueller and Massaron, 2021). Among these algorithms, we have applied logistic regression (LR), support vector machine (SVM), and Naïve Bayes (NB) classification algorithms with bag-of-words (BOW) and n-gram feature extraction methods.

6.2 Deep Learning Models

Most of the current research studies on hate speech detection and classification tasks are based on deep learning approaches with contextual embedding rather than statistical approaches. Deep Learning is a machine learning technique that can be trained to predict outputs from a given set of inputs in a supervised learning approach. It has networks capable of learning in hierarchical layers to understand representations and features from data in increasing levels of complexity and uses these multiple layers to progressively extract higher-level features from the raw inputs (Young et al., 2018).

In this study, we employed recurrent neural networks (RNN), long short-term memory (LSTM), bidirectional long short-term memory (BiLSTM), and convolutional neural networks (CNN). The LSTM network addresses the long-term dependency problem by introducing a memory into the network. RNN is well known in natural language processing applications despite its suffering from vanishing gradient problems. Particularly, the LSTM solves the vanishing gradient problem (Oshikawa et al., 2018). The relative insensitivity to gap length is an advantage of LSTM over RNNs (Glasmachers, 2017; Miedema, 2018), and other sequence learning methods in numerous tasks and

Classifier	Precision	Recall	Accuracy	F1-score
Logistic Regression (LR)	0.68	0.68	0.68	0.67
Linear Support Vector Machine (LSVM)	0.68	0.67	0.67	0.67
Naïve Bayes (NB)	0.68	0.65	0.65	0.63
Recurrent Neural Network (RNN)	0.61	0.62	0.62	0.62
Long Short-Term Memory (LSTM)	0.61	0.62	0.62	0.61
Bidirectional Long Short-Term Memory (BiLSTM)	0.61	0.62	0.62	0.61
Convolutional Neural Network (CNN)	0.62	0.63	0.63	0.62
Framework for state-of-the-art NLP (FLAIR)	**0.72**	**0.72**	**0.72**	**0.72**
Robustly Optimized BERT (RoBERTa)	0.70	0.70	0.70	0.70

Table 3: Performance of the models

applications. The Bi-LSTM neural network learns long-term dependencies without retaining duplicate context information and operates in both directions to incorporate past and future context information through its LSTM units.

We also employed two contextual embedding models, the RoBERTa (A Robustly Optimized BERT Pre-training Approach) and the FLAIR (a very simple framework for state-of-the-art NLP) that are fine-tuned with the Amharic dataset, namely Am-FLAIR and Am-RoBERTa (Yimam et al., 2021). RoBERTa is a replication of the BERT model, which is developed by Facebook (Liu et al., 2019). Unlike BERT, RoBERTa allows training on longer sequences and dynamically changes the masking patterns. FLAIR is a very powerful framework that is developed by Zalando and built on top of PyTorch (Akbik et al., 2019).

7 Results and Discussion

We employed the 80:10:10 data split mechanism for creating the train, development, and test instances. We have used the development dataset to optimize the learning algorithms. All the results reported in the remaining sections are based on the test dataset instances. Deep learning algorithms are computed using the following hyper-parameters, *embedding dimension = 100, epochs = 10, batch_size = 64, activation = softmax, and optimizer = adam*.

F1-score (F1), Precision (P), Recall (R), and Accuracy (Acc) are used to compare the performance of the models. We conducted experiments with the classical machine learning models such as LR, LSVM, and NB; deep learning models like RNN, LSTM, BiLSTM; and CNN, and the fine-tuned Amharic transformer models such as AmFLAIR and AmRoBERTa.

As presented in Table 3, logistic regression (LR) achieved 67% F1-score and 68% performance for precision, recall, and accuracy. LSVM achieved a 68% precision score, and 67% recall, accuracy, and F1-scores. The Naïve Bayes obtained the least F1-score which is 63% from all classical methods. LR and LSVM outperformed the Naïve Bayes in all measures except for precision. LSTM, BiLSTM, RNN, and CNN achieved lower and nearly similar results in all measures of precision, recall, accuracy, and F1 scores. We attribute this to the size of the dataset; while it is common sense that deep learning approaches can achieve higher results by better modeling the properties of large training data, it seems that our dataset was not large enough to leverage their power. The Am-FLAIR contextual embedding model achieved 72% scores for all measures such as precision, recall, accuracy, and F1-scores, which is the overall best result in our experiments. AmRoBERTa also achieved 70% precision, recall, accuracy, and F1 scores, which are the second-best scores. In general, the contextual embedding models such as AmFLAIR and AmRoBERTa outperformed both the deep learning and the classical machine learning methods in all performance measures on the dataset. This confirms the general trend of well-performing transformer-based language models also for the case of Amharic.

8 Error Analysis from Model Outputs

We examined model-predicted tweets against their corresponding gold labels to observe discrepancies. As indicated in Table 5, the model correctly classified 1,034 tweets out of 1,501 test examples. We randomly took 25% of the incorrectly classified instances and conducted extensive investigations in a focus group discussion with three domain experts to explore the potential reasons for the errors.

#	Tweet	English translation	Gold	Predicted
1	እውነትም በእድሜ ትንሹ መሪ?	Oh, truly the youngest leader?	offensive	normal
2	የ *_*ስ ዘ*_ናዬ አያት በቅ_ኝ በኩል የቆመ_ው ባንዳ	M*l** Z*****'s grandfather, the betrayer, standing on right side	hate	normal
3	በኦሮሚያ ክልል የተደረገው የድጋፍ ስልፍ የኦሮሞ ብልፅግና እና ኦነግ ሸኔን አንድነት ያሳየ ነው ተባለ	The rally in Oromia showed the unity of PP and OLF-Shene parties.	hate	normal
4	@ USER ለወራሪ ጋር ሽምግልና የለም። እምሽክ ነው	No mediation with the invaders, just destroy them.	hate	normal

Table 4: Model errors: wrongly predicted tweets against the gold labels

		PREDICTION			
		Hate	Offen.	Normal	Total
GOLD	Hate	**516**	85	101	702
	Offen.	63	**154**	47	264
	Normal	104	67	**364**	535
	Total	683	306	512	**1501**

Table 5: Confusion matrix from FLAIR

63.6% of the errors are mistakes by the model while 28.8% of errors are due to annotator mistakes. The experts found that the remaining 7.6% errors are difficult to judge due to a lack of background contexts. We found out that the main reasons for the errors are annotation bias, association with some keywords, lack of background contexts, informal writing styles in social media, mixed language use, the presence of sarcasm, and idiomatic expressions. Annotation bias, presence of sarcasm, association with some keywords, and the lack of background contexts constituted 29.7%, 13.6%, 11%, and 8.5% of the causes for the errors, respectively. There were also cases where even the experts could not come up with justifications for some errors due lack of background contexts to label some tweets. To showcase the possible justifications for the errors, we to took 5 tweets as presented in Table 4. Tweets with ironic/sarcastic expressions even confused human annotators. For example, **Tweet 1** in Table 4 with the gold label 'offensive', targeted an individual with sarcasm expression and is wrongly predicted as 'normal' by the model. **Tweet 2** annotated as 'hate' is wrongly predicted as 'normal' by the model. This is due to typographic errors in the tweet such as missing characters and unnecessary spaces between characters that we indicated with the '-' symbol. The '*' symbols are used to hide sensitive words from the tweets. Despite **Tweet 3** looking positive news, it contained ironic expressions that the model did not predict correctly. But annotators knew the additional background contexts to understand and label the tweet. **Tweet 4** with gold label 'hate' is wrongly predicted as 'normal' by the model due to the inclusion of informal terms that are not used in the standard Amharic writing system that could confuse the model.

9 Conclusion and Future Work

The paper presented data selection and annotation strategies, and classification models for the Amharic Twitter dataset. A total of 15.1k tweets were annotated into hate, offensive, normal, and unsure classes. We proposed data selection and sampling strategies, a list of hate and offensive lexicon entries, and an annotated dataset for Amharic hate speech research. We also presented both classical and deep learning models trained on a new dataset. The study explored hate speech annotation challenges and revealed that annotation of social media texts for hate speech classification is highly context-dependent. Models that have used contextual embedding models such as Am-FLAIR and Am-RoBERTa outperformed all the models, where Am-FLAIR achieved the best scores of all.

In future work, we plan to use semi-supervised active learning to select hateful tweets employing the human-in-the-loop annotation approach. Exploring the targets of hateful content can also be another future work to deal with. To advance hate speech classification research in Amharic and other low-resource languages; the dataset, hate and offensive keyword lexicons, the best-performing models, annotation guidelines, data selection pipelines, and associated source codes are publicly released with a permissive license [4].

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

10 Limitations

The research study encountered the following limitations. Firstly, the small dataset size could limit the robustness and applicability of the results to be generalized in various contexts. Secondly, the scarcity of the offensive class instances within the dataset might impact the model's ability to accurately detect offensive content. Additionally, the lack of diversity among annotators might have introduced biases in the labeled data, affecting the model's ability to handle inputs from various cultural or linguistic backgrounds. Moreover, the study explored only a few models and embedding approaches and might potentially overlook more effective alternatives. Lastly, the hyperparameters of the models were not extensively fine-tuned to explore opportunities for optimizing performances. These limitations collectively highlight the need for further investigations with larger datasets, diverse annotators, and a broader exploration of models and fine-tuning techniques.

Acknowledgment

The dataset collection of this project is supported by the **Lacuna Fund**. In recognition of this support, we extend our acknowledgments to Lacuna Fund on behalf of the **ICT4D Research Center** at Bahir Dar Institute of Technology, Bahir Dar University, Bahir Dar, Ethiopia.

References

Zeleke Abebaw, Andreas Rauber, and Solomon Atnafu. 2021. Multi-channel convolutional neural network for hate speech detection in social media. In *International Conference on Advances of Science and Technology*, pages 603–618, Bahir Dar, Ethiopia. Springer.

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.

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, Minneapolis, MN, USA.

Nuha Albadi, Maram Kurdi, and Shivakant Mishra. 2018. Are they our brothers? analysis and detection of religious hate speech in the arabic twittersphere. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 69–76, Barcelona, Spain. IEEE.

Yohannes Eneyew Ayalew. 2020. Defining 'Hate Speech'under the Hate Speech Suppression Proclamation in Ethiopia A Sisyphean Exercise? *Ethiopian Human Rights casanovas2018behavioural Series*, 12.

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.

Pompeu Casanovas and Andre Oboler. 2018. Behavioural compliance and casanovas2018behavioural enforcement in online hate speech. In *TERECOM@ JURIX*, pages 125–134, Groningen, The Netherlands.

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.

Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*, pages 86–95, Venice, Italy. ITASEC17.

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, WA, USA. Association for Computational Linguistics.

Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding. 2018. Hate lingo: A target-based linguistic analysis of hate speech in social media. In *Twelfth International AAAI Conference on Web and Social Media*, pages 42–51, Palo Alto, CA, USA.

Paula Fortuna, Monica Dominguez, Leo Wanner, and Zeerak Talat. 2022. Directions for NLP practices

applied to online hate speech detection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11794–11805, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Antigoni Maria Founta, Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Athena Vakali, and Ilias Leontiadis. 2019. A unified deep learning architecture for abuse detection. In *Proceedings of the 10th ACM conference on web science*, pages 105–114, New York City, NY, USA. Association for Computing Machinery.

Antigoni Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Twelfth International AAAI Conference on Web and Social Media*, pages 491–500, Palo Alto, CA, USA.

Andargachew Mekonnen Gezmu, Binyam Ephrem Seyoum, Michael Gasser, and Andreas Nürnberger. 2018. Contemporary amharic corpus: Automatically morpho-syntactically tagged amharic corpus. In *Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing*, pages 65–70, Santa Fe, NM, USA. Association for Computational Linguistics.

Tobias Glasmachers. 2017. Limits of end-to-end learning. In *Proceedings of the Ninth Asian Conference on Machine Learning*, pages 17–32, Seoul, Korea. PMLR.

Tommi Gröndahl, Luca Pajola, Mika Juuti, Mauro Conti, and N Asokan. 2018. All you need is" love" evading hate speech detection. In *Proceedings of the 11th ACM workshop on artificial intelligence and security*, pages 2–12, Toronto, ON, Canada. Association for Computing Machinery.

Simon Kemp. 2023. DIGITAL 2023: GLOBAL OVERVIEW REPORT. Technical report, DataReportal. Last accessed: July 16, 2023.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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.

Fenna Miedema. 2018. Sentiment analysis with long short-term memory networks. *Vrije Universiteit Amsterdam*, 1:1–17.

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.

John Paul Mueller and Luca Massaron. 2021. *Machine learning for dummies*. John Wiley & Sons.

Hala Mulki, Hatem Haddad, Chedi Bechikh Ali, and Halima Alshabani. 2019. L-hsab: A levantine twitter dataset for hate speech and abusive language. In *Proceedings of the third workshop on abusive language online*, pages 111–118, Florence, Italy.

Ray Oshikawa, Jing Qian, and William Yang Wang. 2018. A survey on natural language processing for fake news detection. *arXiv preprint arXiv:1811.00770*.

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.

Fabio Poletto, Marco Stranisci, Manuela Sanguinetti, Viviana Patti, and Cristina Bosco. 2017. Hate speech annotation: Analysis of an italian twitter corpus. In *4th Italian Conference on Computational Linguistics, CLiC-it 2017*, volume 2006, pages 1–6, Rome, Italy. CEUR-WS.

Paul Röttger, Debora Nozza, Federico Bianchi, and Dirk Hovy. 2022a. Data-efficient strategies for expanding hate speech detection into under-resourced languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5674–5691, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Paul Röttger, Haitham Seelawi, Debora Nozza, Zeerak Talat, and Bertie Vidgen. 2022b. Multilingual HateCheck: Functional tests for multilingual hate speech detection models. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 154–169, Seattle, WA, USA. Association for Computational Linguistics.

Abiodun Salawu and Asemahagn Aseres. 2015. Language policy, ideologies, power, and the ethiopian media. *South African Journal for Communication Theory and Research*, 41(1):71–89.

Surafel Getachew Tesfaye and Kula Kakeba. 2020. Automated amharic hate speech posts and comments

detection model using recurrent neural network. *Research Square, DOI: https://doi.org/10.21203/rs.3.rs-114533/v1*, pages 1–14.

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.

Matthew L Williams, Pete Burnap, Amir Javed, Han Liu, and Sefa Ozalp. 2020. Hate in the machine: Anti-black and anti-muslim social media posts as predictors of offline racially and religiously aggravated crime. *The British Journal of Criminology*, 60(1):93–117.

Kevin Winter and Roman Kern. 2019. Know-center at SemEval-2019 task 5: Multilingual hate speech detection on Twitter using CNNs. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 431–435, Minneapolis, MN, 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):1–18.

Seid Muhie Yimam, Iryna Gurevych, Richard Eckart de Castilho, and Chris Biemann. 2013. WebAnno: A flexible, web-based and visually supported system for distributed annotations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 1–6, Sofia, Bulgaria. Association for Computational Linguistics.

Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. 2018. Recent trends in deep learning based natural language processing [review article]. *IEEE Computational Intelligence Magazine*, 13(3):55–75.