The 5Js in Ethiopia: Amharic Hate Speech Data Annotation Using Toloka Crowdsourcing Platform

Abinew Ali Ayele
Faculty of Computing
Bahir Dar Institute of Technology
Bahir Dar University
Bahir Dar, Ethiopia
abinewaliayele@gmail.com

Tadesse Destaw Belay
College of Informatics
Wollo University
Kombolcha, Ethiopia
tadesseit@gmail.com

Skadi Dinter
Dept. of Informatics
Universität Hamburg
Hamburg, Germany
skadi.dinter@studium.uni-hamburg.de

Tesfa Tegegne Asfaw
ICT4D Research Center
Bahir Dar Institute of Technology
Bahir Dar University
Bahir Dar, Ethiopia
tesfat@gmail.com

Seid Muhie Yimam
Dept. of Informatics
Universität Hamburg
Hamburg, Germany
seid.muhie.yimam@uni-hamburg.de

Chris Biemann
Dept. of Informatics
Universität Hamburg
Hamburg, Germany
christian.biemann@uni-hamburg.de

***Abstract*—This paper presents an Amharic hate speech annotation using the Yandex Toloka crowdsourcing platform. The dataset for this paper is collected from 5 consecutive years in Ethiopia (5Js), namely from 2018-2022, following the 'June' events in Ethiopia. Every June for the last five years, some events put the country in violence. Accordingly, we annotate 5,267 tweets, nearly 1k tweets every year. We explore the main challenges of crowdsourcing annotation for Amharic hate speech data collection using Toloka. We attain a Fliess kappa score of 0.34 using three independent annotators that annotate the tweets and the gold label is determined using majority voting. Using the datasets, we build different classification models using classical machine learning and deep learning approaches. The classical machine learning algorithms, LR and SVM, achieve an F1 score of 0.49, while NB archives an F1 score of 0.46. The deep learning algorithms (LSTM, BiLSTM, and CNN) achieve a similar F1 score, that is 0.44, which is the lowest performance of all models. The contextual embedding models, Am-FLAIR and Am-RoBERTa achieve F1 scores of 0.48 and 0.50 respectively. We publicly release[1] the dataset, source code, and models with a permissive license.**

***Index Terms*—hate speech annotation, low-resource languages, crowdsourcing annotation, Yandex Toloka, Twitter data, Amharic**

I. Introduction

Crowdsourcing in machine learning entails facilitating recruiting of data annotators on a large scale. Data annotation is the prerequisite task for many NLP tasks such as sentiment analysis, entity extraction, and text categorization. Various crowdsourcing annotation platforms have been proposed for different applications and they have their advantages and disadvantages [1]. The following are among the advantages of crowdsourcing: 1) It allows numerous anonymous end-users to participate within a short period from different backgrounds. 2) It provides an opening, realistic test environment including time, location, end-user, multimedia content (e.g. image, audio/video stream), and network access. 3) It reduces time and cost related to experimental facilities, in-lab personnel, and traditional participant recruitment schemes [2].

Although crowdsourcing has several advantages and makes it possible for researchers to reach a wide audience for evaluation, there still exist some challenges that remain unresolved. Some of them are 1) trustworthiness regarding the dataset, 2) unreliable annotation accuracy, 3) not supporting low-resourced languages, 4) end-users perform without supervision (untrained participants), they may give erroneous feedback, carelessly or dishonestly, and 5) a high number of malicious annotators who cheats with Google Translate to get more rewards [1].

The majority of Natural Language Processing (NLP) researches focus on only 20 out of the 7000 languages of the world. African and Asian Languages are among the low-resource languages which are still understudied [3, 4]. The reason behind these unaddressed languages might be the limitations of crowdsourcing platforms, lack of online infrastructures such as payment methods, internet connection problems, shortage of online native performers, and the lack of awareness of online jobs [5].

There are some attempts to build hate speech classification for the Amharic language [6, 7, 8]. In previous studies, data annotation for Amharic hate speech was conducted in the laboratory with a few personnel and limited contexts of user opinions. This paper proposed a crowdsourcing annotation scheme for hate speech data collection and explored the challenges associated with low-resource language in general and Amharic in particular. The dataset of this research is taken from the 5J (5 consecutive Ethiopian June months) where there

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

were controversial issues in the country and have been major topics both main stream and social media platforms.

This study addresses the following research questions:

1) **RQ1**) What are the main challenges of crowdsourcing annotation for Amharic hate speech data collection?
2) **RQ2**) To what extent do crowdsourcing annotation agreements differ from in-lab annotation agreements for Amharic?

In this paper, we present the analysis of the Toloka crowdsourcing platform for low-resource languages like Amharic. The paper has the following main contributions:

1) Explore the challenges of the Yandex Toloka crowdsourcing platform for low-resource languages.
2) Proposes different techniques of crowdsourcing annotations for low-resource language researchers like Amharic.
3) Build classification models for datasets collected using crowdsourcing platforms.

The remaining sections of the paper are organized as follows. The study provides introductory information about the Amharic language in Section II. The related works are presented in Section III. Data collection strategies and techniques are presented in Section IV. Data annotation strategies are described in Section V while error analysis of the annotation task are presented in Section VI. We present classification models in Section VII, and the result and discussion part of the paper in Section VIII. Conclusion and future work are presented in Section IX.

II. Amharic Language

Amharic is the second-largest Semitic language family widely spoken next to Arabic. Amharic scripts originated from the Ge'ez alphabet which is called Fidäl or 'Ethiopic script'. Among the languages spoken in Ethiopia, Amharic is the most widely spoken language. It is the official working language of the Federal Democratic Republic of Ethiopia (FDRE) and many regional states within the country [9, 10]. Moreover, it is used in governmental administration, public media, mass communication, and nationally used for commercial transactions. In Amharic, there are 34 core characters each having seven different variations to represent vowels. Amharic is a morphologically complex language. At the sentence level, Amharic follows Subject-Object-Verb (SOV) word order. Models built for English and even for Semitic languages, could not work without extensive rework.

III. Related Work

Casanovas and Oboler [11] defined hate speech as a speech that targets identity in terms of ethnicity, gender, disability, or political and religious ideology, which indirectly or directly focuses on their group identity and has the potential to incite violence while the offensive speech usually targets individuals to be offended based on their personal behaviors but do not target their group identity. We have adopted these definitions for this research project.

The ever-increasing of social media platforms and the expansion of the internet aggravated the spread of hateful content globally. The task of hate speech detection and classification has drawn the attention of many researchers for more than a decade. Despite most of the works being conducted for English and some European languages, there is a raising attempt among the low-resource languages like Amharic and other African and Asian languages [12].

We reviewed the works conducted for the Amharic hate speech classification task. The work by Mossie and Wang [7, 8] collected an Amharic hate speech dataset from the Facebook pages of individuals and organizations. The annotation was done with a few annotators and achieved a kappa score of 0.57 using in-lab annotation. Similarly, Abebaw et al. [6] collected Facebook comments and posts in the same manner as Mossie and Wang [7] and achieved Cohen's kappa score of 0.8 with two annotators. Amharic hate speech researchers such as [3, 6, 7, 13] conducted annotation using in-lab annotation environment. They all annotated their dataset using in-lab personnel with a minimal number of annotators and the dataset represents the perceptions of only limited users.

Whereas studies by [14, 15, 16] have employed crowdsourcing annotation techniques to annotate Hate Speech data for English, Arabic, Germany, and French languages. However, low inter-annotator agreements of less than 0.25 were reported except for the work by Mathew et al. [14], which is 0.46.

Nowadays, crowdsourcing is getting more attention for data annotation due to its lower cost, higher speed, and diversity of opinions compared to labeling data with experts [17]. Amazon Mechanical Turk (MTurk)[2], Yandex Toloka[3], and Crowdflower[4] are among top crowdsourcing platforms. However, performers at crowdsourcing marketplaces are non-professional and their annotation results are much noisier than that of expert annotations [17]. MTurk is used broadly in the research community because of the 24/7 worldwide workforce. In particular, Amazon MTurk is difficult to use from outside of the United States, Europe and some parts of Asia [18]. It requires funding that can be difficult to obtain for some junior researchers in developing nations [5].

In our research, we have used Yandex Toloka crowdsourcing to annotate and collect the dataset. Yandex Toloka is a rising crowdsourcing platform similar to that of MTurk. Yandex Toloka has more than 25K performers executing around 6M hits in more than 500 different projects every day [2]. We found out that, Toloka is preferable for low-resource languages since it is relatively cheap, supports annotation from developing countries, has a training facility for performers, and allows filtering of performers by language or country.

IV. Data Collection

The source of this research data is taken from the Ethiopian Twitter dataset repository, which has been collected since 2014 [19]. The number of tweets stored in the repository

[2] https://www.mturk.com/
[3] https://toloka.yandex.com/
[4] http://crowdflower.com/

surpasses 12 million tweets. A sample of 5400 tweets was chosen using seed keywords for annotation over 5 years period, including the 400 pilot tweets. The dataset contains a large number of tweets that are written in the 'Fidäl' script, which includes Amharic, Awgni, Guragigna, Ge'ez, Tigrinya, or other Semitic languages that uses the Fidäl script. We used the Pycld2 Python language identification library to select Amharic tweets.

We consider the five Junes' incidences (**5Js**), from June 2018 to June 2022, to compile our research data samples. It is a surprising coincidence that there were incidences during the consecutive five years in June that brought excessive violence in the country, which was aggressively addressed by the mainstream and social media as well. In June 2018, there was an assassination attempt on the newly elected Prime minister of Ethiopia, Dr. Abiy Ahmed Ali while celebrating his first 100 days achievements as a prime minister with the people gathered at Maskel Square, Addis Ababa. In June 2019, the Ethiopian army chief and three Amhara region higher officials including the head of the state were assassinated at the same time in Addis Ababa and Bahir Dar respectively. In June 2020, the well-known Oromo artist, Hachalu Hundessa has been killed near his residence. There was terrible violence across the Oromia region where 86 ethnic Amhara civilians were killed, many wounded, and properties including hotels and buildings were burned and destroyed. The government arrested main Oromo opposition party leaders and accused them of intensifying the violence and conflicts. In June 2021, the 6th Ethiopian national election was held while the Tigray People Liberation Front (TPLF) was at war with the federal government in the northern part of the country and the Oromo Opposition parties abandoned themselves from participating in the election. Finally, in June 2022, hundreds of ethnic Amhara people were massacred in Qellem Wellega, Oromia region by the so-called Terrorist Oromia liberation army, called "Shenie". We collected the tweets for each June incidence starting from the incidence date for about one month. We removed re-tweets, anonymized users, and performed all important preprocessing tasks.

V. Data Annotation

Annotation for hate speech classification task by itself is a complex task. Low-resource languages are particularly challenged by the scarcity of annotators, annotation frameworks, and lack of expert researchers in the area to pursue hate speech research. We have used three annotators for each tweet. We prepared annotation guidelines and upload them to the Yandex Toloka crowdsourcing platform. The annotation interface in Toloka is presented in Figure 1. Besides, we presented 20 tweets as a training task for users to complete before they start the annotation task. We used 50 control tweets with their gold labels to screen out malicious annotators during the annotation process. Each task presented to users contains 15 tweets and one of them is taken from control tweets using the smart mixing technique of Yandex Toloka. We used Fleiss Kappa as an inter-annotator agreement (IAA) measure which is appropriate for three overlapping annotators. Figure 2 shows a sample Toloka interface for one of the pools with performers' basic information. As indicated in Figure 2, the average time to submit an assignment was 5.16 minutes, which means 0.34 minutes were required to annotate a tweet. The figure also showed that 88 users were interested to participate in the task while 85 participated and submitted at least one assignment in the pool. It was also indicated that the average number of assignments submitted by each user was 2.54 (nearly 38 tweets per user).

A. Pilot Annotation

We conducted two rounds of pilot annotations and used 400 tweets for both pilots (200 tweets for each pilot). In the first pilot, 14 annotators participated and an inter-annotator agreement of 0.15 has been achieved. We examine the annotation results manually using the control tweets and identified performers who probably use Google Translate (Amharic to English) to annotate the tweet, and we have blocked such users from participating in the main task. Accordingly, we sent personalized messages to 4 users indicating their negligence during the annotation with examples taken from their annotation result and referred them to read the guidelines and re-complete the training tasks. For the rest of the annotators, we sent a generalized message to remind them to take due care and read the tweets carefully while annotating the main tasks. However, there is no way to completely avoid malicious annotators who use Google Translate before they complete some tasks.

In the second pilot task, a total of 29 users participated in the annotation task where the majority are new users. An inter-annotator agreement of 0.25 has been achieved for the second pilot annotations. We compiled the two pilots together and achieved an overall inter-annotator agreement of 0.20. The possible reasons for the low inter-annotator agreements could be the low price per task, limited or no training for performers, lack of sufficient annotators for low-resource languages, and random annotators seeking more rewards.

B. The Main Annotation Task

For the main annotation task, 5 pools were created where each pool containing 1000 tweets from 5 different years of datasets spanning 2018 - 2022 Ethiopian Junes' controversial and miserable events. In every pool, new users join the task, and some users are banned from the project. Overall, 579 users from 27 different countries participated in the task where 17 users are banned from the Toloka crowdsourcing system and 154 were banned from our project. The majority of performers are only from three countries, 207 from Ethiopia, 197 from Pakistan, and 65 from the United States. Most of the performers participated in only a few tasks. Toloka has the option to choose annotators either based on the country in which they live or based on the language they speak. We choose users who can speak Amharic from all over the world. We have paid 0.1$ for each hit, which contains 15 tweets.

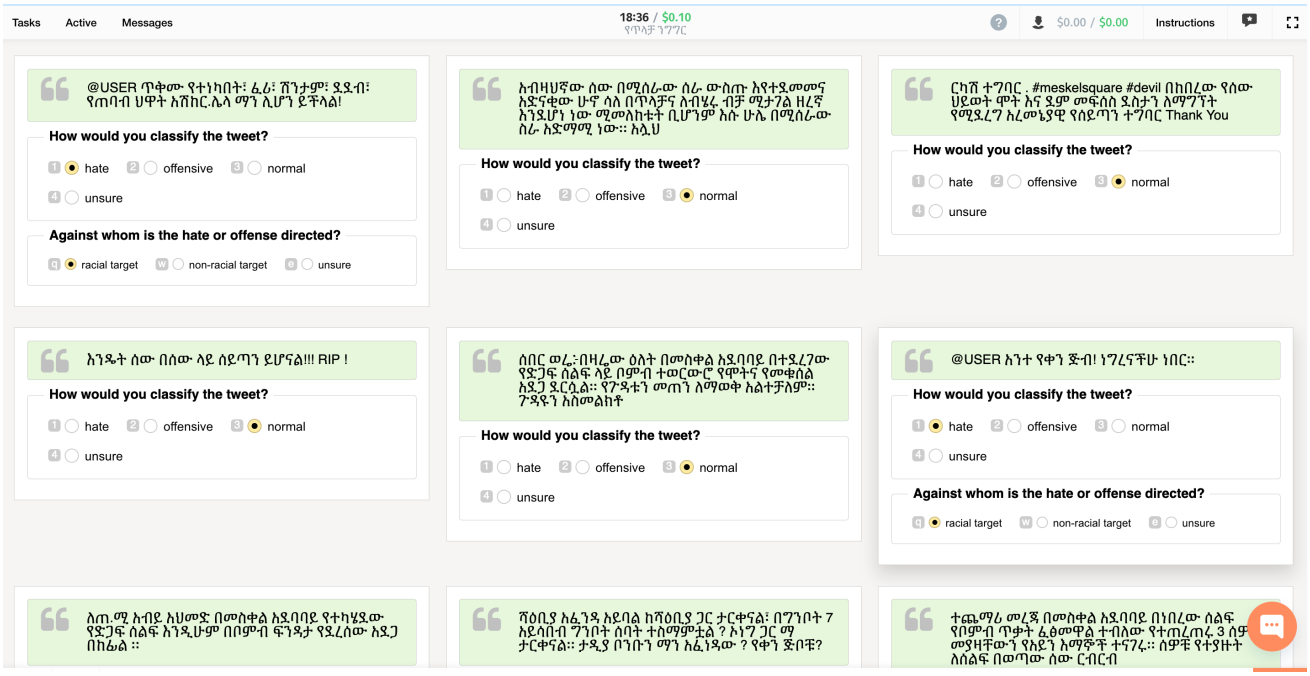

Fig. 1. The Toloka UI for Amharic hate speech annotation.

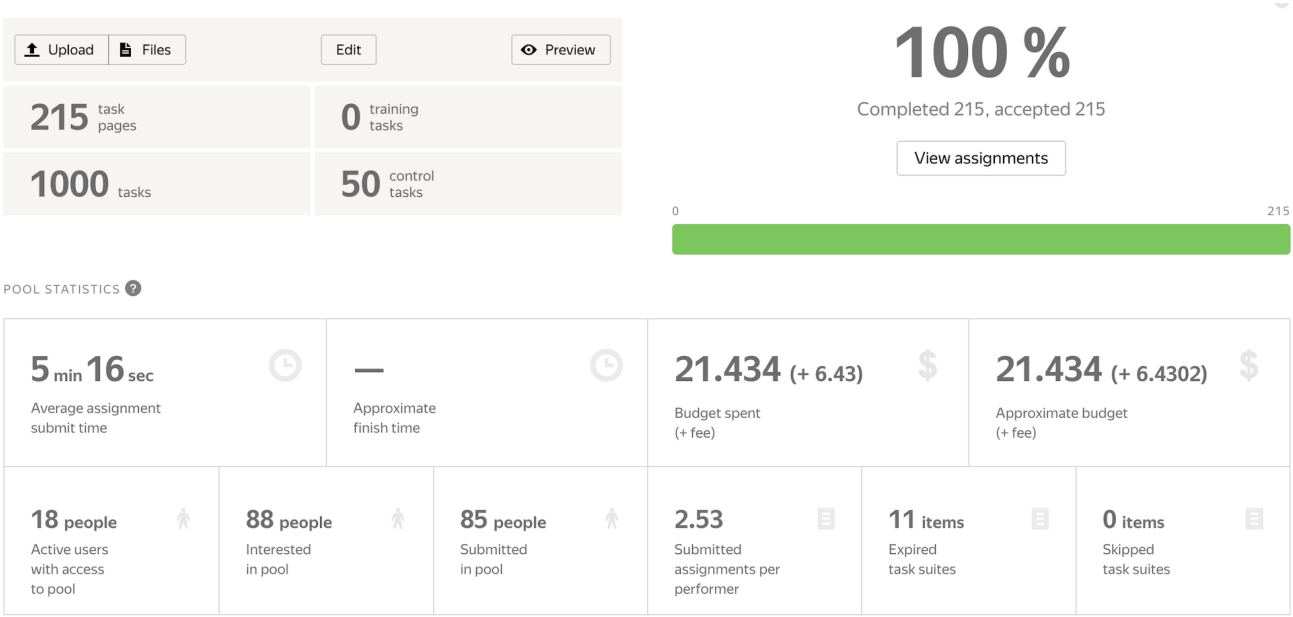

Fig. 2. Sample Toloka interface that shows Performers participated and are interested in one of the pools.

After analyzing the first and the second pools, we suspect many of the annotators are malicious users and used the second option, which is filtering performers by country. We allow annotators from Ethiopia and United Ara Emirates to perform the task but, no one has started the task in two days period. We then allow all Amharic-speaking annotators from all over the world, and the task was completed in an hour and a half. We checked this strategy two times and got the same result. We presented the biographic information of participant users in the annotation task in Table I. We achieved 0.34 IAA for the overall dataset which looks good for crowdsourcing annotation approaches. This indicates that banning suspected malicious users from the project has helped to increase the IAA. When compared to other related studies, for example, the work by Del Vigna12 et al. [15] reported a 0.26 inter-annotator agreement score on the Italian dataset. The work by Ousidhoum et al. [16] reported 0.153, 0.202, and 0.244 IAA scores of kappa coefficient on English, Arabic, and French

parameter	# Count
Total performers participated	579
Number of Countries	27
Performers blocked from projects	154
Performers blocked from the system	17
The average age of performers	30 years

TABLE I
PERFORMERS' BASIC INFORMATION

data sets respectively using Amazon MTurk. Besides, Mathew et al. [14] have reported a 0.46 inter-annotator agreement score on the English data set, which indicates a moderate agreement among annotators. The Fleiss kappa agreement score of 0.34 for our annotation task is moderately similar to other tasks in the literature, hence we decide to use the dataset to build classification models.

The gold labels are determined based on majority voting for each tweet. If two of the annotators agree on a label, the label is determined by taking the majority label. For the tweets in which three annotators choose different category labels, which are 801 out of 5,400 tweets, we have annotated by a fourth performer. However, the performers still choose the fourth class label on 134 tweets, and we then exclude them from the study. Finally, a total of 5,267 annotated tweets were used for the experiment.

VI. Annotation Error Analysis

We choose some of the annotated tweets randomly in the pilot study that contains disputes among the four annotators and become difficult to decide on the gold label even after re-annotating for the fourth time. The annotation errors are presented with examples as depicted in Table II. The gold labels for the three tweets would have been normal, offensive, and hate respectively, but at least two of the four performers would have chosen the labels. The possible reasons might be the performers are malicious users who do not understand the Amharic language or knows the language but choose the labels without reading the tweets to collect as much reward as possible.

As shown in Table III, we also choose some tweets for which performers annotate wrongly either with negligence or some tweets containing idiomatic and poetic expressions that have no direct meaning for readers who might not understand the context. Table III presents the tweets concerning the labels by the three performers and the gold labels. The gold labels for the first two tweets (tweet #1 and tweet #2) would have been 'hate' than 'normal' since the two tweets contain abusive content targeting ethnic groups. Similarly, tweets #3 and #4 did not contain any hate content despite all performers choosing the 'hate' label, and the gold labels are chosen as 'hate' wrongly. However, tweets #5 and #6 contained poetry and idiomatic expressions. Tweet #5 ("the dispersed flour comes as a storm") represents a poetry expression spoken in a specific context. It was spoken in July 2021 when TPLF rebels captured North Wollo after the Ethiopian government announced the complete destruction of TPLF rebels during the law enforcement period in November 2020. Tweet #6 ("@USER you are joking; while fearing the donkey, you deal with what the donkey carries") represents an idiomatic expression to express when people deal with silly challenges than dealing with the real challenge.

VII. Classification Models

A. Classical Machine Learning Approaches

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 [20]. Among these algorithms, we have applied logistic regression (LR), support vector machine (SVM), and Naive Bayes (NB) classification algorithms with bag-of-words (BOW) and n-gram feature extraction methods.

B. Deep Learning Models

In this study, we have employed three deep learning algorithms such as long short-term memory (LSTM), bidirectional long short-term memory (BiLSTM), and Convolutional Neural Network (CNN). We have 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 Ethiopian Tweeter Data - Amharic (ETD-AM) dataset, namely Am-FLAIR and Am-RoBERTa [12]. RoBERTa is a replication of the BERT model which is developed by Facebook [21]. Unlike BERT, RoBERTa allows training on longer sequences and dynamically changes the masking patterns. FLAIR is a very powerful framework that is developed by the Zalando Research group and built on top of PyTorch [22].

C. Experimental Settings

Our annotated tweets are further split into training, development, and test instances using an 80:10:10 split. 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 trainable parameters or weights are computed using the following hyper-parameters as; *embedding dimension = 100, epochs = 10, batch_size = 64, activation = softmax, and optimizer = adam.*

VIII. Results and Discussion

In this section, the different Amharic hate speech classification experiments are presented. We experimented with the classical machine learning models (such as LR, SVM, and NB) and deep learning models (LSTM, BiLSTM, and CNN) as well as the fine-tuned Amharic transformer models like Am-FLAIR and Am-RoBERTa. We use the F1-score, Precision, Recall, and Accuracy for comparing the performance of the models. We used macro F1-score for the comparisons of the accuracy of NLP models – it is a harmonic balance between precision and recall measures. As we can see from Table IV, the hate speech classifier model from RoBERTa contextual embedding outperforms other classifiers, which achieves a

#	Tweet	Anno1	Anno2	Anno3	Anno4
1	አፈጉባዔ ወይዘሮ ሙፈሪያት ካሚል የደኢህዴን ሊቀመንበር ሆነው ተመረጡ (Speaker Mrs. Muferiat Kamil was elected as the chairman of the SPDM Party)	offensive	unsure	normal	hate
2	@USER @USER ወረኛ ፈሳም ነገር ነህ አጉተን ብሎ ተሳዳቢ። (@USER you are talkative and coward what abuser you are)	unsure	offensive	normal	hate
3	@USER ባንዳ ላህያ ነህ:: የባንዳዎች ዘር (@USER You are a renegade ass. The race of renegades)	offensive	normal	unsure	hate

TABLE II
SAMPLE TWEETS WHERE THE GOLD LABEL CAN NOT BE DETERMINED EVEN AFTER FOURTH ROUND ANNOTATIONS.

#	Tweet	Anno1	Anno2	Anno3	Gold
1	@USER አንተ ደደብ ቄሮ የሚባል ከብትህን ግራ። (@USER You idiot. educate your cow called Kero.)	normal	normal	offensive	normal
2	ነፍጠኛ ከኦሮምያ ከኢትዮጵያ ካልጠፋ ሰላም የለም:: (If the musketeer does not disappear from Oromia and Ethiopia, there will be no peace.)	normal	normal	normal	normal
3	አማራነትን መርጦ የዘር ጭፍጨፋ ማድረግ ይቁም!! (Stop genocide by choosing Amhara!!)	hate	hate	hate	hate
4	@USER ተጠያቂነት ካልሰፈነ ጭፍጨፋው ይቀጥላል:: (@USER Without accountability, the massacre will continue.)	hate	hate	hate	hate
5	የተበተነው ዱቄት ኣውሎ ነፋስ ሆኖ መጣ። (The scattered powder came as a storm.)	normal	normal	unsure	normal
6	@USER አንተ ቀልድ፡ አህያውን ፈርቶ ዳውላውን (@USER you are joking; while fearing the donkey, you deal with what the donkey carries)	hate	hate	hate	hate

TABLE III
ANNOTATION ERROR ANALYSIS

0.50 F1 score. As we have made an error analysis from the predicted test file, some tweets like idiomatic expressions are contextually understood from RoBERTa embedding. From the classical classifiers, Logistic Regression and SVM perform better than NB, which both achieve an F1 score of 0.49. Our experimented deep learning classifiers, LSTM, BiLSTM, and CNN achieve an F1 score of 0.44, which is less accurate than classical classifiers. Our low inter-annotator agreement result of the crowdsourcing-based dataset (0.34) compared to in-lab annotation approaches, has an impact on the performance of the models since the quality of the dataset highly affects the performance of the classification models. As we have analyzed from the predicted test file, idiomatic expressions are not handled by the model and remained challenging for the Amharic hate speech classification task.

classifier	Precision	Recall	Accuracy	F1-score
Log.Reg	0.49	0.54	0.54	0.49
SVM	0.48	0.54	0.54	0.49
NB	0.52	0.52	0.52	0.46
LSTM	0.43	0.46	0.46	0.44
BiLSTM	0.43	0.46	0.46	0.44
CNN	0.43	0.48	0.48	0.44
FLAIR	0.46	0.51	0.52	0.48
RoBERTa	**0.49**	**0.51**	**0.51**	**0.50**

TABLE IV
AMHARIC HATE SPEECH CLASSIFIER MODELS RESULT

IX. CONCLUSION AND RECOMMENDATION

In this paper, we presented a crowdsourcing-based hate speech dataset collection approach which is the first of its kind to the knowledge of the researchers for the Amharic Twitter dataset. In this research, 5,267 tweets are annotated into hate, offensive, normal, and unsure classes. The dataset can be a benchmark dataset for the crowdsourcing-based Amharic dataset. Three different supervised machine learning, deep learning models, and contextual embedding models are presented that were trained on the collected dataset. The contextual embedding model, Am-RoBERTa outperformed all the classical and deep learning models with an F1-score of 0.50 performance. The dataset, models, and source codes are publicly released[5] with a permissive license to advance hate speech classification research in Amharic.

In future work, we plan to build a semi-supervised 'active learning and distance supervision approach to select hateful tweets employing the human-in-the-loop annotation approach. We also plan to explore and compare in-lab and crowdsourcing experiments combined with active learning setups.

REFERENCES

[1] Z. Wang, D. Tao, and P. Liu, "Development and challenges of crowdsourcing quality of experience evaluation

[5]https://github.com/uhh-lt/ethiopicmodels

for multimedia," in *International Conference on Big Data Computing and Communications*, 2015, pp. 444–452.

[2] H. Garcia-Molina, M. Joglekar, A. Marcus, A. Parameswaran, and V. Verroios, "Challenges in data crowdsourcing," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 4, pp. 901–911, 2016.

[3] A. Magueresse, V. Carles, and E. Heetderks, "Low-resource languages: A review of past work and future challenges," *arXiv preprint arXiv:2006.07264*, 2020.

[4] A. Wang, C. D. V. Hoang, and M.-Y. Kan, "Perspectives on crowdsourcing annotations for natural language processing," *Language resources and evaluation*, vol. 47, no. 1, pp. 9–31, 2013.

[5] E. Öhman, "Challenges in annotation: annotator experiences from a crowdsourced emotion annotation task," in *Proceedings of the Digital Humanities in the Nordic Countries 5th Conference*. CEUR Workshop Proceedings, 2020, pp. 293–301.

[6] Z. Abebaw, A. Rauber, and S. Atnafu, "Multi-channel convolutional neural network for hate speech detection in social media," in *International Conference on Advances of Science and Technology*, Bahir Dar, Ethiopia, 2021, pp. 603–618.

[7] Z. Mossie and J.-H. Wang, "Social network hate speech detection for amharic language," *Computer Science & Information Technology*, pp. 41–55, 2018.

[8] Z. Mossie and J.-H. Wang, "Vulnerable community identification using hate speech detection on social media," *Information Processing & Management*, vol. 57, no. 3, 2020.

[9] A. Salawu and A. Aseres, "Language policy, ideologies, power and the ethiopian media," *South African Journal for Communication Theory and Research*, vol. 41, no. 1, pp. 71–89, 2015.

[10] A. M. Gezmu, B. E. Seyoum, M. Gasser, and A. Nürnberger, "Contemporary amharic corpus: Automatically morpho-syntactically tagged amharic corpus," in *Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing*, New-Mexico, NM, USA, 2018, pp. 65–70.

[11] P. Casanovas and A. Oboler, "Behavioural compliance and law enforcement in online hate speech." in *TERECOM@ JURIX*, Groningen, The Netherlands, 2018, pp. 125–134. [Online]. Available: http://ceur-ws.org/Vol-2309/11.pdf

[12] S. M. Yimam, A. A. Ayele, G. Venkatesh, I. Gashaw, and C. Biemann, "Introducing various semantic models for amharic: Experimentation and evaluation with multiple tasks and datasets," *Future Internet*, vol. 13, no. 11, 2021.

[13] S. G. Tesfaye and K. Kakeba, "Automated amharic hate speech posts and comments detection model using recurrent neural network," *Preprint paper*, 2020.

[14] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee, "Hatexplain: A benchmark dataset for explainable hate speech detection," in *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence*, vol. 35, no. 17, Palo Alto, California, USA, 2021, pp. 14 867–14 875.

[15] F. Del Vigna12, A. Cimino23, F. Dell'Orletta, M. Petrocchi, and M. Tesconi, "Hate me, hate me not: Hate speech detection on facebook," in *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*. Venice, Italy: ITASEC17, 2017, pp. 86–95.

[16] N. Ousidhoum, Z. Lin, H. Zhang, Y. Song, and D.-Y. Yeung, "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)*, Hong Kong, China, 2019, pp. 4675–4684.

[17] A. Drutsa, V. Farafonova, V. Fedorova, O. Megorskaya, E. Zerminova, and O. Zhilinskaya, "Practice of efficient data collection via crowdsourcing at large-scale," *arXiv preprint arXiv:1912.04444*, 2019.

[18] J. W. Vaughan, "Making better use of the crowd: How crowdsourcing can advance machine learning research." *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 7026–7071, 2017.

[19] S. M. Yimam, A. A. Ayele, and C. Biemann, "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)*, Paris, France, 2019, pp. 1–5.

[20] J. P. Mueller and L. Massaron, *Machine learning for dummies*. John Wiley & Sons, 2021.

[21] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[22] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, and R. Vollgraf, "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)*, Minnesota, USA, 2019, pp. 54–59.