Challenges of Amharic Hate Speech Data Annotation Using Yandex Toloka Crowdsourcing Platform

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

[1]Faculty of Computing, Bahir Dar University, Bahir Dar, Ethiopia
{abinewaliayele, tesfat}@gmail.com
[2] College of Informatics, Wollo University, Kombolcha, Ethiopia
tadesseit@gmail.com
[3]Dept. of Informatics, Universität Hamburg, Hamburg, Germany
{seid.muhie.yimam, christian.biemann}@uni-hamburg.de, skadi.dinter@studium.uni-hamburg.de

Abstract

This paper presents an Amharic hate speech annotation using the Yandex Toloka crowdsourcing platform. The dataset was compiled over a five-year period (2018-2022), following some contentious events in Ethiopia that sparked violence. We consider one-month data starting from the event date in each year. Accordingly, we annotate 5,400 tweets, which is approximately 1,000 tweets per year. We explore the main challenges of crowdsourcing annotation for Amharic hate speech data collection using Yandex Toloka. We attain a Fleiss-Kappa score of 0.34 with three independent annotators that annotate the tweets where the gold label is determined using majority voting. We built deep learning based classification models with LSTM and BiLSTM and achieved a 0.44 F1-score for both models. We released the dataset, source code, and models under a permissive license[1].

1 Introduction

Crowdsourcing platforms are used to annotate large datasets for machine learning projects (Wang et al., 2015). It allows numerous anonymous end-users to participate within a short period from different backgrounds and provides a realistic test environment including time, location, end-user, and multimedia content. Besides, it reduces time and cost related to experimental facilities, in-lab personnel, and traditional participant recruitment schemes (Garcia-Molina et al., 2016; Chai et al., 2019). However, crowdsourcing has several challenges that remain unresolved, such as 1) trustworthiness regarding the dataset, 2) unreliable annotation accuracy, 3) not supporting low-resource languages, and 4) a high number of malicious annotators who cheat with Google Translate to get more rewards (Wang et al., 2015).

Most Natural Language Processing (NLP) research projects focus on only 20 of the 7,000 languages in the world. African and Asian languages are among the low-resource languages still understudied (Magueresse et al., 2020; Wang et al., 2013). The potential reasons include the lack of online infrastructures such as payment methods, internet connection problems, a shortage of online native performers, and a lack of awareness of online jobs (Öhman, 2020). In previous studies, data annotation for Amharic hate speech has been conducted using in-lab facilities with a few personnel and limited user backgrounds (Abebaw et al., 2021; Mossie and Wang, 2018, 2020). This paper proposed a crowdsourcing annotation scheme for hate speech data collection and explored the challenges associated with low-resource languages in general and Amharic in particular. Furthermore, we have built a hate speech classification model and analyzed the results.

2 Related Work

Amharic is the working language of Ethiopia and the second-largest Semitic language family widely spoken next to Arabic. It uses the 'Fidäl' script, which originated from the Ge'ez alphabet (Salawu and Aseres, 2015; Gezmu et al., 2018). Amharic is a morphologically complex language and has 34 core characters, each having seven different varia-

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

tions to represent vowels. These days, with the ever-increasing increase of social media platforms and the spread of hateful content globally, hate speech studies have drawn the attention of researchers. Despite most studies being conducted for English, there is a rising attempt among the low-resource languages like Amharic and other African and Asian languages. (Yimam et al., 2021). Amharic hate speech studies are conducted by Mossie and Wang (2018, 2020); Abebaw et al. (2021) manually collecting their datasets from the Facebook pages of individuals and organizations. The annotations were done with a few annotators using in-lab annotation and achieved moderate kappa agreements greater than 0.50. Whereas studies by Mathew et al. (2021); Del Vigna12 et al. (2017); Ousidhoum et al. (2019) have employed MTurk crowdsourcing to annotate hate Speech data for English, Arabic, Germany, and French languages, and achieved slight-fair agreements of less than 0.25 except for Mathew et al. (2021), which is 0.46. Nowadays, crowdsourcing is getting more popular for data annotation due to its lower cost, higher speed, and diversity of opinions compared to in-lab annotations usually by experts and well-trained users (Drutsa et al., 2019). Amazon Mechanical Turk (MTurk), [2], Yandex Toloka [3], and Crowdflower [4] are among the top crowdsourcing platforms. However, performers at crowdsourcing marketplaces are non-professional and their annotation results are much noisier than those of expert annotations (Drutsa et al., 2019). MTurk is difficult to use from outside of the United States and Europe (Vaughan, 2017) since funding is difficult for researchers in developing nations (Öhman, 2020). We used Yandex Toloka, a rising crowdsourcing platform having more than 25K performers in more than 500 different projects every day (Garcia-Molina et al., 2016; Chai et al., 2019). Toloka is preferable for low-resource languages since it is cheap, supports annotations from developing countries, has a training facility for performers, and allows filtering of performers by language/country.

3 Data Collection and Model Building

A total of 5,400 tweets are selected from over 13 million tweets in the Ethiopian Twitter dataset repository, being collected since 2014 by Yimam et al. (2019) using seed keywords. The dataset consists of tweets collected over a one-month period in 5 different years, from 2018 to 2022, following controversial incidents that happened in Ethiopia and put the country in total violence. We used the PYCLD2[5] Python language identification library to select Amharic tweets. We have annotated each tweet by 3 different annotators using Yandex Toloka. We have used 50 control and 20 training tweets to control the annotation process and train annotators respectively. Each task presented to users contained 15 tweets and one of them was taken from control tweets using the smart mixing technique of Yandex Toloka.

We conducted two rounds of pilot annotations and used 400 tweets for both pilots. In the first pilot, 14 annotators participated and an agreement of 0.15 was achieved, while in the second pilot, 29 users participated in the annotation task where the majority were new users and achieved an agreement of 0.25. We examined the annotation results of both pilots manually using the control tweets and identified performers who probably use Google Translate while annotating, and blocked such users from participating in the main task. Moreover, we have sent personalized warning messages to 4 users as they seemed to annotate some tweets randomly. The possible challenges for our crowdsourcing can be the low prices per task, limited or no training for performers, lack of sufficient annotators for low-resource languages, and random annotators seeking more rewards.

For the main annotation, 5 pools were created where each pool contains 1000 tweets from 5 different years. In every pool, there were new users joining the task, while some others were banned from the project. Overall, 579 users from 27 different countries have participated in the task where 17 users are prohibited from the Toloka crowdsourcing system and 154 users from our project respectively. The majority of performers are only from three countries, namely 207 from Ethiopia, 197 from Pakistan, and 65 from the United States. Most of the performers participated in only a few tasks. We paid 0.1$ for each hit, which included 15 tweets, and one of them was a control tweet. We achieved an inter agreement of 0.34 for the overall dataset which seams good for crowdsourcing annotation approaches. The gold labels are determined based on majority voting for each tweet. For the tweets

[2] https://www.mturk.com/

[3] https://toloka.yandex.com/

[4] http://crowdflower.com/

[5] https://pypi.org/project/pycld2/

in which three annotators chose different labels, 801 out of 5,400 tweets were annotated by a fourth performer. However, for 134 tweets, performers chose the fourth label, which was excluded from the study. Finally, a total of 5,267 tweets were used for the experiments.

We have experimented with baseline classification tasks using 80:10:10 train, development, and test split. We employed the deep learning models Long-short-Term-Memory (LSTM) and Bidirectional-Long-Short-Term- Memory (BiLSTM), and achieved a similar F1-score of 0.44 for both models.

4 Conclusion and Recommendation

The paper presented a crowdsourcing-based hate speech data collection approach and studied the challenges where it is the first of its kind for the Amharic as far as our knowledge is concerned. We annotated 5,400 tweets into hate, offensive, normal, and unsure classes and reported 0.34 kappa agreement. The dataset can be a benchmark dataset for the crowdsourcing-based Amharic hate speech task. Classification outputs with deep learning models showed promising results for crowdsourcing-based hate speech studies. The dataset, the models, and the source code are publicly released under a permissive license to advance hate speech research in Amharic. In the future, we plan to build a semi-supervised 'active learning and distant supervision' approach to selecting hateful tweets employing a human-in-the-loop annotation approach. We also plan to explore and compare in-lab and crowdsourcing experiments combined with active learning setups.

5 Acknowledgment

This research is conducted with a research grant received from the Yandex Toloka crowdsourcing plat- form. We are pleased to thank Yandex Toloka for granting us the research fund for the data annotation. Besides, Toloka gave us opportunity to annotate our data on their platform.

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.

Chengliang Chai, Ju Fan, Guoliang Li, Jiannan Wang, and Yudian Zheng. 2019. Crowdsourcing database systems: Overview and challenges. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pages 2052–2055.

Fabio Del Vigna12, Andrea Cimino23, 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.

Alexey Drutsa, Viktoriya Farafonova, Valentina Fedorova, Olga Megorskaya, Evfrosiniya Zerminova, and Olga Zhilinskaya. 2019. Practice of efficient data collection via crowdsourcing at large-scale. *arXiv preprint arXiv:1912.04444*.

Hector Garcia-Molina, Manas Joglekar, Adam Marcus, Aditya Parameswaran, and Vasilis Verroios. 2016. Challenges in data crowdsourcing. *IEEE Transactions on Knowledge and Data Engineering*, 28(4):901–911.

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, New-Mexico, NM, USA.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *arXiv preprint arXiv:2006.07264*.

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. *Computer Science & Information Technology*, pages 41–55.

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

Emily Öhman. 2020. Challenges in annotation: annotator experiences from a crowdsourced emotion annotation task. In *Proceedings of the Digital Humanities in the Nordic Countries 5th Conference*, pages 293–301. CEUR Workshop Proceedings.

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.

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.

Jennifer Wortman Vaughan. 2017. Making better use of the crowd: How crowdsourcing can advance machine learning research. *J. Mach. Learn. Res.*, 18(1):7026–7071.

Aobo Wang, Cong Duy Vu Hoang, and Min-Yen Kan. 2013. Perspectives on crowdsourcing annotations for natural language processing. *Language resources and evaluation*, 47(1):9–31.

Zhenji Wang, Dan Tao, and Pingping Liu. 2015. Development and challenges of crowdsourcing quality of experience evaluation for multimedia. In *International Conference on Big Data Computing and Communications*, pages 444–452.

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 1–5, 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).