Hate Speech and Offensive Content Identification For Low-Resource Languages

Mohammadmostafa Rostamkhani and **Sauleh Eetemadi**
Iran University of Science and Technology at Tehran, Iran
mo_rostamkhani97@comp.iust.ac.ir, sauleh@iust.ac.ir

Abstract

In this paper, we address hate speech and offensive content detection for low-resource languages. To illustrate this, we selected Sinhala as our primary language example. We analyze the zero-shot performance of diverse models on the SOLD dataset and juxtapose it with ChatGPT's zero-shot performance. Surprisingly, our study indicates that, except for models fine-tuned on the SOLD dataset, ChatGPT consistently outperforms even those fine-tuned on Sinhala texts. These results underscore the remarkable zero-shot capabilities of ChatGPT. The research additionally conducts a comparison of multilingual models using translated text, as well as a combination of translated and original text, in contrast to source language-specific fine-tuned models on the original text. Our research highlights the potential advantages of fine-tuning models using both original and translated text, in contrast to solely using translated or original text. Furthermore, the results demonstrate the superior performance of fine-tuned source language models in hate speech detection.

1 Introduction

The proliferation of hate speech and offensive content on digital platforms has underscored the urgency of developing effective methods for their detection across languages. we can protect people from offensive content, detect offensive parts, and censor it. Different variety of methods have been used for hate speech detection tasks such as traditional classifiers (Thomas Davidson and Weber, 2017; Waseem and Hovy, 2016; MacAvaney and Frieder, 2019; Pamungkas et al., 2020), deep learning-based classifiers (Agrawal and Awekar, 2018; Badjatiya et al., 2017) or the combination of both approaches (Mossie and Wang, 2020). There is also some research for investigation on the importance of initial fine-tuning multilingual models on English hate speech and subsequently fine-tuning them with labeled data in the target language (Röttger et al., 2022). This paper addresses investigating hate speech identification in a low-resource Indo-Aryan language (Sinhala). The study encompasses classifying tweets as hate/offensive or not. We explore the zero-shot performance of various models, including ChatGPT. Additionally, we use three approaches for fine-tuning different models: 1) Leveraging translation services in combination with multilingual models, 2) Utilizing language models fine-tuned on the source languages, and 3) Fine-tuning multilingual models using a combination of both translated and original tweets.

2 Methodology

First, we investigate the zero-shot performance of various models. We employ ChatGPT, alongside Sinhala, multilingual, and English models, assessing their capabilities across diverse input text types. These include original Sinhala text, English translations of tweets using Google Translate API, and ChatGPT-generated translations of tweets. To achieve this, we divided the dataset into batches of 10 and requested ChatGPT to translate each batch while also identifying their labels for zero-shot evaluation.

In addition, we compare three different fine-tuning strategies. The first, uses a translation-based technique, employing the Google Translate API and ChatGPT to convert content in the source language into English, subsequently feeding translations to multilingual models (as an English text), An English-only model that is pre-trained on hate speech corpus. The second approach uses models fine-tuned on the source language. The third approach combines both translated and original text for fine-tuning multilingual models.

3 Experiments and Results

Our experiments reveal that, in terms of zero-shot performance, ChatGPT surpasses models except those specifically fine-tuned on the SOLD

Language	Model	Accuracy	Precision	Recall	F1-Score
Sinhala	ChatGPT	0.612	0.590	0.551	0.523
	xlm-roberta-base (Conneau et al., 2020)	0.588	0.413	0.496	0.376
	twitter-xlm-roberta-base-sentiment (Barbieri et al., 2022)	0.583	0.476	0.496	0.398
	SinhalaBERTo (keshan, 2021)	0.581	0.498	0.499	0.419
	Sinhala-roberta (d42kw01f, 2021)	0.542	0.514	0.513	0.511
	xlm-t-hasoc-hi (sinhala nlp, 2022a)	0.603	0.613	0.603	0.606
	xlm-t-hasoc-hi-sold-si (sinhala nlp, 2022b)	**0.834**	**0.833**	**0.834**	**0.832**
	xlm-t-sold-si (sinhala nlp, 2022c)	0.827	0.827	0.827	0.825
English (Google Translate)	xlm-roberta-base	0.587	0.473	0.498	0.386
	twitter-xlm-roberta-base-sentiment	0.496	0.517	0.517	0.496
	distilbert-base-uncased	0.594	0.547	0.500	0.373
	roberta-hate-speech-dynabench-r4-target (Bianchi et al., 2022)	0.586	0.544	0.527	0.501
English (ChatGPT)	xlm-roberta-base	0.570	0.482	0.493	0.427
	distilbert-base-uncased (Sanh et al., 2019)	0.561	0.517	0.513	0.498
	roberta-hate-speech-dynabench-r4-target	0.595	0.554	0.517	0.450

Table 1: Zero-shot performance of hate speech detection for Sinhala

Language	Model	Accuracy	Precision	Recall	F1-Score
English (Google Translate)	distilbert-base-uncased	0.641	0.623	0.613	0.614
	xlm-roberta-base	0.594	0.297	0.500	0.373
	twitter-xlm-roberta-base-sentiment	0.642	0.628	0.628	0.628
	roberta-hate-speech-dynabench-r4-target	0.657	0.642	0.622	0.623
Sinhala	SinhalaBERTo	0.823	0.822	0.808	0.813
	Sinhala-roberta	0.833	0.827	0.826	0.826
	xlm-roberta-base	0.823	0.820	0.811	0.814
	twitter-xlm-roberta-base-sentiment	0.812	0.832	0.783	0.793
	xlm-t-hasoc-hi	0.824	0.827	0.824	0.821
	xlm-t-hasoc-hi-sold-si	0.832	0.840	0.832	0.827
	xlm-t-sold-si	0.832	0.833	0.832	0.829
English + Sinhala	xlm-roberta-base	0.826	0.826	0.826	0.825
	twitter-xlm-roberta-base-sentiment	**0.843**	**0.842**	**0.843**	**0.842**

Table 2: Results of hate speech detection for Sinhala

dataset. In addition, utilizing models fine-tuned on the source languages outperformed the translation-based approach. This can be attributed to the preservation of linguistic nuances and contextual understanding inherent in language-specific models, as well as the absence of a proficient language model for correct and accurate translations also some issues present within the translated sentences. The fine-tuned models on source language have better performance in all cases for identifying hate speech and offensive content. The utilization of a blend of original and translated text demonstrates superior performance compared to both preceding methods. This could potentially be attributed to the cross-lingual knowledge transfer effect, where insights from one language positively impact the understanding of others.

4 Conclusion and Future Work

This study examines hate speech detection for low-resource languages. We choose Sinhala for this purpose. ChatGPT displays impressive zero-shot performance, surpassing models except those fine-tuned on the SOLD dataset. We compare translation-based multilingual models with language-specific fine-tuned models, highlighting the effectiveness of the latter. Furthermore, combining translated and original data improves hate speech detection results.

As part of our future work, we plan to extend our experiments to encompass additional languages.

References

Sweta Agrawal and Amit Awekar. 2018. Deep learning for detecting cyberbullying across multiple social media platforms. *CoRR*, abs/1801.06482.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta,

and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. *CoRR*, abs/1706.00188.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266, Marseille, France. European Language Resources Association.

Federico Bianchi, Stefanie HIlls, Patricia Rossini, Dirk Hovy, Rebekah Tromble, and Nava Tintarev. 2022. "it's not just hate": A multi-dimensional perspective on detecting harmful speech online. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8093–8099, Abu Dhabi, United Arab Emirates. 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. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

d42kw01f. 2021. Sinhala-roberta. https://huggingface.co/d42kw01f/Sinhala-RoBERTa. Accessed: 2023-09-10.

keshan. 2021. Sinhalaberto. https://huggingface.co/keshan/SinhalaBERTo. Accessed: 2023-09-10.

Yao H. R. Yang E. Russell K. Goharian N. MacAvaney, S. and O. Frieder. 2019. Hate speech detection: Challenges and solutions. *PloS one*.

Zewdie Mossie and Jenq-Haur Wang. 2020. Vulnerable community identification using hate speech detection on social media. *Information Processing Management*, 57(3):102087.

Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020. Misogyny detection in twitter: a multilingual and cross-domain study. *Information Processing Management*, 57(6):102360.

Paul Röttger, Debora Nozza, Federico Bianchi, and Dirk Hovy. 2022. Data-efficient strategies for expanding hate speech detection into under-resourced languages.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

sinhala nlp. 2022a. xlm-t-hasoc-hi. https://huggingface.co/sinhala-nlp/xlm-t-hasoc-hi-sold-si. Accessed: 2023-09-10.

sinhala nlp. 2022b. xlm-t-hasoc-hi-sold-si. https://huggingface.co/sinhala-nlp/xlm-t-hasoc-hi-sold-si. Accessed: 2023-09-10.

sinhala nlp. 2022c. xlm-t-sold-si. https://huggingface.co/sinhala-nlp/xlm-t-sold-si. Accessed: 2023-09-10.

Michael W. Macy Thomas Davidson, Dana Warmsley and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *CoRR*, abs/1703.04009.

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. Association for Computational Linguistics.