Multimodal Hate Speech Detection in Multilingual Setting

Kyung Myung Ko
Department of Computer Science,
Purdue University
ko112@purdue.edu

Abstract

Multimedia content is prevalent and easily accessible. However, not all contents are appropriate for youth, as they may contain violence, toxicity, and discrimination. Exposing these contents without solid restriction to young viewers may negatively affect their behavior. Hence, the toxicity that reside within these content must be carefully analyzed. In this work, we investigate multimodal & multilingual hate speech dataset, MuTox. We replicate its baseline approach with the modified dataset that we obtained from scratch. We believe that this supplemental baseline work can be extended to close the gap for toxicity detection in low-resource languages.

1 Introduction

Toxicity is thoroughly spread out in the society. As people are able to connect each other easily, they are at the same time exposed to potential toxic and hate speech environment. With the rise of platforms that enables the massive communication between people, the solution for detection is further needed. Hate or abusive speech detection has been largely explored in the past few years (Röttger et al., 2020). The widely accessible multilingual datasets have been pushing this work even further (Jigsaw, 2019). However, the problem still remains as previous work focus solely on the content of the speech. To extend this question further, previous work has proposed what can be achieved with the multimodal speech and text dataset (Ghosh et al., 2021). Nonetheless, less emphasis have been put on effectively utilizing the multimodality text and speech. MuTox is composed of the speech and text multimodal dataset while also including the low-resource languages such as . incorporates both the multilinguality and the multimodality in the dataset (Costa-jussà et al., 2024). In this work, the authors released a speech and text multimodal hate speech dataset and propose a baseline evaluation of the binary hate speech classification. They first evaluate with the embedded-level representation of speech from the pre-trained model with the training objective in Machine Translation(MT) and the ASR component for Speech-To-Text conversion, where each are then fusioned with the text representation to perform multimodal binary classification. The ASR component outperformed the other in this experiment, but this seems obvious as the speech encoder was trained to achieve this objective. The question still remains as this misaligned objective in the pre-training step of the utilized model does not address the unique features that only exist in the speech modality.

There are two meaningful component in human speech that are relevant for classification: the textual content and the acoustic features. High-level syntactical representation of the textual content within the speech is a straightforward interpretation of speech and is widely accessible due to rapid development in Automatic Speech Recognition(ASR) task (Radford et al., 2023). On the other hand, low-level descriptors that exist within the waveform of the speech contain relevant information of prosodic cues that are not captured in static text. Recently, audio pre-trained models have been widely adopted to deal with audio classification tasks that require understanding of these descriptors. The CNN-based VGGish is often utilized as the baseline frozen encoder for various tasks regarding audio (Chen et al., 2020) and the rise of attention-based mechanisms have extended this development even further (Gong et al., 2021). Yet, recent work has shown that the low-level descriptors could also be captured via human-generated features to solve abusive speech detection problem (Spiesberger et al., 2023). openSmile is a open-source package for extracting features for speech processing and the Music Information Retrieval communities (Eyben et al., 2010). They are de-

signed for general audio-based research to affective computing. For example, the voice quality can be determined by HNR, Jitter, and Shimmer parameters and the tones with CHROMA, CENS, and CHROMA-based features. Indic languages are considered to be complex as variations of dialects are involved.

2 MuTox Baseline

We first obtained the raw MuTox dataset and replicate its baseline method, following the same training configuration and evaluation.

2.1 Dataset Analysis

MuTox dataset consists of utterances of 32 multilingual speech audio clips around 10 seconds. The original clips for the utterances have been separately extracted by the provided timestamps that align with the corresponding transcription. While the original dataset consists of 100K samples, the cleaning process reduced the total number of samples by half, as the majority of the clips contained expired URLs. The specifics of the cleaned dataset is shown in Table 1 & 2.

2.2 Model Architecture

The MuTox baseline architecture consists of 3 simple linear layers with the classification head on top for the binary classification task. Each speech and text data are separately processed by the pre-trained SONAR architecture (Duquenne et al., 2023). This architecture is built on top of NLLB 1B model for the Machine Translation(MT) objective and covers 200 languages (Costa-jussà et al., 2022).

2.3 Evaluation

The main metrics to evaluate the results were Area Under Curve(AUC) and Recall. We chose the top three high and low resource languages within our obtained dataset for the evaluation that had the fair amount of binary labels to compute the AUC metrics. We selected the best-performing model upon comparing the accuracy of the validation set while incurring early-stopping. The specifics are shown in Table 1.1 & 1.2 in Appendix.

2.4 Discussion

We performed cross validation for the low-resource samples. From the table, we analyze that the performance on the chosen languages that had fair amount of samples displays shows similar result as the baseline. We hypothesize that being able

lang	Original	Obtained
spa	16668	14708
eng	16535	13975
hun	2497	2207
ind	2484	2207
ces	2480	1498
cmn	2479	100
deu	2478	105
pes	2477	2109
pol	2472	1879
urd	2470	127
slk	2470	1823
deu	2478	105
rus	2468	99
tur	2466	112
vie	2462	105
fra	2461	115
deu	2478	105
fin	2461	2060
ben	2458	119
dan	2457	1986
ell	2457	1779
tgl	2454	124
swh	2430	111
nld	2429	104
heb	2417	2270
por	2415	128
arb	2399	115
hin	2331	87
ita	2308	119
est	2282	1980
bul	2249	2087
cat	2021	1923
Total	100935	54044

Table 1: Comparison of the original and the obtained dataset after the cleaning process.

	Original	T:NT	Obtained	T:NT
eng	2757	1:6.0	2291	1:6.1
spa	3315	1:5.0	2933	1:5.0
heb	100	1:24.2	88	1:25.8
pes	135	1:18.3	114	1:18.5
bul	317	1:7.1	267	1:7.8
por	364	1:6.6	22	1:5.8
urd	463	1:5.3	19	1:6.7
ita	327	1:7.1	21	1:5.7
Total	7778	1:5.8	5755	1:6.2

Table 2: Comparison of Toxic to Non-Toxic ratio in original vs obtained dataset for the selected language.

to keep the similar ratio of toxic to non-toxic data have contributed to this result. On the other hand, the languages with not enough samples did not express enough evidence for a fair evaluation with close to 0 in recall. We assume the lack of training samples led this simple model be ineffective to learn, bringing in the necessity of a more complex architecture and the aggregation of other datasets. During the training, we also faced limitation in regards to the computing environment. As the pre-trained speech encoders for each language that were necessary to encode the raw waveform each consisted of 10GB of storage space, we trained the model separately with each languages one after another, which may introduce bias in the model from the lack of variety of languages seen in each batch. We hope to make a comparison in the future regarding this training strategy and the regular way of exposing model with variety of languages in each batch.

lang	**Dev**	**DevTest**	$MuTox_{zs}$	$MuTox$
eng	824 125	1642 275	0.69	0.71
spa	865 155	1772 352	0.72	0.74
heb	666 27	223 9	0.74	0.75
pes	610 34	210 12	0.85	0.78
bul	572 78	198 29	0.82	0.83
por	36 6	15 6	0.60	0.81
urd	35 5	9 1	0.85	1.00
ita	25 4	18 5	0.52	0.46

Table 3: Area Under Curve

lang	**Dev**	**DevTest**	$MuTox_{zs}$	$MuTox$
eng	824 125	1642 275	0.34	0.40
spa	865 155	1772 352	0.39	0.41
heb	666 27	223 9	0.07	0.18
pes	610 34	210 12	0.31	0.31
bul	572 78	198 29	0.10	0.10
por	36 6	15 6	0.00	0.00
urd	35 5	9 1	0.00	0.00
ita	25 4	18 5	0.00	0.00

Table 4: Recall

3 Limitation

In this work, we analyze and demonstrate ways to cure toxicity detection in multilingual & multimodal setting. We believe that the proper alignment of the speech and the textual data will be able to mitigate the current concern. We also address our concern to the limitation of the dataset in the future. This may simply be solved with aggregation of more data for the low-resource languages and keeping the balance of the toxic to non-toxic ratio in some languages, following upon the original aggregated dataset proposed in the MuTox paper. We believe this work can be an extension to multimodal multilingual understanding and push it further.

4 Conclusion

In this work, we analyze and demonstrate ways to cure toxicity detection in multilingual & multimodal setting. We believe that the proper alignment of the speech and the textual data will be able to mitigate the current concern. We also address our concern to the limitation of the dataset in the future. This may simply be solved with aggregation of more data for the low-resource languages and keeping the balance of the toxic to non-toxic ratio in some languages, following upon the original aggregated dataset proposed in the MuTox paper. We believe this work can be an extension to multimodal multilingual understanding and push it further.

References

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. 2020. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Marta R Costa-jussà, Mariano Coria Meglioli, Pierre Andrews, David Dale, Prangthip Hansanti, Elahe Kalbassi, Alex Mourachko, Christophe Ropers, and Carleigh Wood. 2024. Mutox: Universal multilingual audio-based toxicity dataset and zero-shot detector. *arXiv preprint arXiv:2401.05060*.

Paul-Ambroise Duquenne, Holger Schwenk, and Benoît Sagot. 2023. Sentence-level multimodal and language-agnostic representations. *arXiv preprint arXiv:2308.11466*.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462.

Sreyan Ghosh, Samden Lepcha, Sahni Sakshi, Rajiv Ratn Shah, and Srinivasan Umesh. 2021. Detoxy:

A large-scale multimodal dataset for toxicity classification in spoken utterances. *arXiv preprint arXiv:2110.07592*.

Yuan Gong, Yu-An Chung, and James Glass. 2021. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*.

2019 Jigsaw. 2019. Toxic comment classification challenge:. *https://www.kaggle.com/c/jigsaw*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Paul Röttger, Bertram Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet B Pierrehumbert. 2020. Hatecheck: Functional tests for hate speech detection models. *arXiv preprint arXiv:2012.15606*.

Anika A Spiesberger, Andreas Triantafyllopoulos, Iosif Tsangko, and Björn W Schuller. 2023. Abusive speech detection in indic languages using acoustic features. In *Proc. INTERSPEECH*, volume 2023, pages 2683–2687.