AUDIO VS. TEXT: IDENTIFY A POWERFUL MODALITY FOR EFFECTIVE HATE SPEECH DETECTION

Kirtilekha Bhesra[1], **Shivam Ashok Shukla**[2], **Akshay Agarwal**[2]
[1]C.V. Raman Global University Bhubaneshwar, [2]IISER, Bhopal
`20010097@cgu-odisha.ac.in,{shivams19, akagarwal}@iiserb.ac.in`

ABSTRACT

The boom in social media platforms has witnessed a significant jump in offensive, hate, and toxic languages. These toxic contents leave a significant effect on one's personality which can even lead to depression and can be in various forms including audio and text. **However, the primary concern is that the benchmark datasets such as Toxigen for hate speech detection, are only text-based**. *Therefore, in this research, for the first time, we have collected the audio-based hate speech dataset for unified security*. Utilizing both these modalities (text and audio), we have performed the benchmark study for audio and text hate detection and proposed a multimodal hate detection algorithm.

1 INTRODUCTION

Hate speech (Paz et al., 2020), characterized by offensive and harmful language targeted at specific groups, has become a significant concern in the digital age. Its online presence reduces the quality of digital communication and poses serious social and ethical challenges[1]. In response to this growing issue, researchers have devoted considerable effort to developing effective hate speech detection methods, primarily focusing on text-based content (MacAvaney et al., 2019). However, hate speech detection in audio data remains relatively unexplored, presenting unique challenges and opportunities. This research aims to bridge this gap and extend hate speech detection capabilities to audio and multiple modalities. Further, we have proposed a multimodal hate detection algorithm by combining the decision fusion of text and audio hate detection algorithms.

1.1 PROPOSED HATE AUDIO DATASET

To explore the potential of audio modality, we have generated hate and non-hate audio samples using the neural text-to-speech (TTS) [2] model. For that, we have used the 200 text samples of the ToxiGen dataset (Hartvigsen et al., 2022). Since hate speech might not belong to any specific demographic group, to cover a wide spectrum population, we have generated audio samples of different demographic entities varying in terms of age and gender. *In other words, the generated audio samples cover the voices of males, females, and children, and in total,* 600 *audio samples are generated*. We will release the dataset upon the acceptance of the paper.

2 HATE IDENTIFICATION

To benchmark the use of text and audio modality for hate information detection, we have performed an extensive set of experiments utilizing several machine learning classifiers including deep neural networks (DNN). The classifiers evaluated in this research are: (i) NuSVC, (ii) SVC, (iii) Logistic Regression (LR), (iv) Random Forest (RF), (v) SGD, (vi) Extreme Gradient Boosting (XGB), and (vii) 3 layer DNN. These classifiers are trained using default parameters in the Sklearn library (Pedregosa et al., 2011).

[1] https://news.un.org/en/story/2023/01/1132597

[2] https://voicemaker.in/

Table 1: Hate classification accuracy (%) using 'text' and 'audio' modalities.

Models	Text			Audio			
	enc-1	enc-2	enc-3	F-1	F-2	F-3	F-4
NuSVC	**77.0 ± 5.3**	76.0 ± 8.6	**77.5 ± 7.7**	**77.0 ± 7.3**	**79.0 ± 5.4**	**75.5 ± 5.8**	**77.0 ± 7.3**
SVC	74.0 ± 7.2	**76.5 ± 7.7**	74.5 ± 6.6	70.5 ± 3.7	74.5 ± 3.7	69.5 ± 5.1	70.5 ± 5.3
LR	71.0 ± 4.1	73.5 ± 5.6	71.0 ± 3.4	75.5 ± 3.7	76.5 ± 4.6	75.5 ± 4.5	75.5 ± 3.7
RF	69.5 ± 7.6	71.0 ± 8.5	70.0 ± 6.5	76.0 ± 7.8	71.0 ± 10.3	71.5 ± 9.3	75.5 ± 7.0
SGD	67.0 ± 4.3	72.5 ± 3.2	70.0 ± 4.2	**77.0 ± 7.0**	77.5 ± 5.7	74.0 ± 5.1	75.0 ± 4.5
XGB	62.5 ± 4.2	72.5 ± 7.1	65.5 ± 7.6	67.5 ± 7.2	68.0 ± 4.0	71.5 ± 5.8	71.5 ± 7.5
DNN	70.5 ± 6.2	75.5 ± 5.1	72.5 ± 5.7	76.0 ± 4.9	71.0 ± 3.4	71.5 ± 6.2	75.0 ± 7.1

Table 2: Average hate audio detection performance (%) of best-performing classifier (i.e., NuSVC) on different voices (M: man, W: woman, and C: child).

MFCC_Default			MFCC_13Coeff			MFCC_Hamming			MFCC_Balckman		
C	M	W	C	M	W	C	M	W	C	M	W
57.0	**77.0**	75.0	59.0	**79.0**	76.0	56.5	**75.5**	75.0	57.0	**77.0**	75.5

To ensure the generalizability of the performance, we have performed 5-fold cross-validation and every time the classifiers are trained on 4 folds and tested on the remaining 1 fold. The results are reported using average classification accuracy (%) along with standard deviation (%).

2.1 Results and Analysis of Hate Text Identification

To encode and extract the discriminating features of text samples, we have used multiple text encoders [3] namely (i) all-MiniLM-L6-v2 (**enc-1**), (ii) all-distilroberta-v1 (**enc-2**) (Sanh et al., 2019), and (iii) all-MiniLM-L12-v2 (**enc-3**), enc-1 and enc-2 are based on MiniLM (Wang et al., 2020b). These encoders are pre-trained on large-scale text datasets and effectively extract semantic information. The results of hate text detection are reported in Table 1. Encoder **enc-3**, with its 384-dimensional feature vectors, achieves the best average detection accuracy when paired with the NuSVC classifier surpassing each classifier including DNN for hate text identification. In contrast, **enc-1** and **enc-2** produce 768 and 384-dimensional feature vectors.

2.2 Results and Analysis of Hate Audio Identification

Similar to text encoders, to extract the discriminative features from the audio samples, we have utilized one of the popular methods namely, Mel-frequency cepstral coefficients (Ittichaichareon et al., 2012) (MFCCs). The four distinct MFCC extraction methods used in this research are: (i) default MFCC (**F-1**), (ii) 13-coefficient MFCCs (**F-2**), (iii) Hamming-windowed MFCCs (**F-3**), and (iv) Blackman-windowed MFCCs (**F-4**). To comprehensively evaluate the effectiveness of individual demographic entities, we trained hate detection models on each voice belonging to different genders and age groups. The hate audio detection performance of individual demographic voices is shown in Table 2. The analysis of hate audio detection can be described using the following: (i) analysis concerning demographic entity and (ii) effectiveness of feature encoders. It can be seen that the male voice outperforms the other types of voices in identifying hate audio. In terms of the effectiveness of the feature encoder, the 13-coefficient vector shows the best performance across each voice type.

Multimodal Hate Detection: *Further, to explore the possibility of combining the discriminating strength of audio and text, we have performed the decision fusion to boost the accuracy*. In this, we amalgamate the best-performing classifier, i.e. NuSVC trained on the most effective features (enc-3 & F-2) of these individual modalities. The proposed fusion shows an improved performance of $\mathbf{80.5}\% \pm \mathbf{4.7}\%$ compared to the best value of $\mathbf{79.0}\% \pm \mathbf{5.4}\%$ obtained using the audio.

3 Conclusion

Hate information identification and ban of that is an important topic for a secure and humble society. *To advance the research in this critical direction, we have proposed a novel audio hate dataset covering varying demographic entities for the first time*. The experiments performed using multiple text and audio encoders found that audio performs better than text. On top of that adult (man and woman) voices are found more effective than child voices. In the future, we aim to extend the audio dataset along with the development of a novel multimodal hate information detection algorithm.

[3] https://www.sbert.net/

URM Statement

The authors acknowledge that the key author of this work meets the URM criteria of the ICLR 2024 Tiny Papers Track.

References

Aymé Arango, Jorge Pérez, and Barbara Poblete. Hate speech detection is not as easy as you may think: A closer look at model validation. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*, pp. 45–54, 2019.

Fariha Tahosin Boishakhi, Ponkoj Chandra Shill, and Md Golam Rabiul Alam. Multi-modal hate speech detection using machine learning. In *2021 IEEE International Conference on Big Data (Big Data)*, pp. 4496–4499. IEEE, 2021.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. Hatebert: Retraining bert for abusive language detection in english. *arXiv preprint arXiv:2010.12472*, 2020.

Shammur Absar Chowdhury, Hamdy Mubarak, Ahmed Abdelali, Soon-gyo Jung, Bernard J Jansen, and Joni Salminen. A multi-platform arabic news comment dataset for offensive language detection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 6203–6212, 2020.

P Dhanalakshmi, S Palanivel, and Vennila Ramalingam. Classification of audio signals using svm and rbfnn. *Expert systems with applications*, 36(3):6069–6075, 2009.

P Dhanalakshmi, Sengottayan Palanivel, and Vennila Ramalingam. Classification of audio signals using aann and gmm. *Applied soft computing*, 11(1):716–723, 2011.

Goran Glavaš, Mladen Karan, and Ivan Vulić. Xhate-999: Analyzing and detecting abusive language across domains and languages. Association for Computational Linguistics, 2020.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *arXiv preprint arXiv:2203.09509*, 2022.

Michael Ibañez, Ranz Sapinit, Lloyd Antonie Reyes, Mohammed Hussien, Joseph Marvin Imperial, and Ramon Rodriguez. Audio-based hate speech classification from online short-form videos. In *2021 International Conference on Asian Language Processing (IALP)*, pp. 72–77. IEEE, 2021.

Chadawan Ittichaichareon, Siwat Suksri, and Thaweesak Yingthawornsuk. Speech recognition using mfcc. In *ICCGSM*, volume 9, 2012.

Mladen Karan and Jan Šnajder. Cross-domain detection of abusive language online. In *Proceedings of the 2nd workshop on abusive language online (ALW2)*, pp. 132–137, 2018.

Anna Koufakou, Endang Wahyu Pamungkas, Valerio Basile, Viviana Patti, et al. Hurtbert: Incorporating lexical features with bert for the detection of abusive language. In *Proceedings of the fourth workshop on online abuse and harms*, pp. 34–43. Association for Computational Linguistics, 2020.

D Pradeep Kumar, BJ Sowmya, KG Srinivasa, et al. A comparative study of classifiers for music genre classification based on feature extractors. In *2016 IEEE Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER)*, pp. 190–194. IEEE, 2016.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. Hate speech detection: Challenges and solutions. *PloS one*, 14(8):e0221152, 2019.

Johannes Skjeggestad Meyer and Björn Gambäck. A platform agnostic dual-strand hate speech detector. In *ACL 2019 The Third Workshop on Abusive Language Online Proceedings of the Workshop*. Association for Computational Linguistics, 2019.

Mainack Mondal, Leandro Araújo Silva, and Fabrício Benevenuto. A measurement study of hate speech in social media. In *Proceedings of the 28th ACM conference on hypertext and social media*, pp. 85–94, 2017.

Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. Hate speech detection and racial bias mitigation in social media based on bert model. *PloS one*, 15(8):e0237861, 2020.

Kadir Bulut Ozler, Kate Kenski, Steve Rains, Yotam Shmargad, Kevin Coe, and Steven Bethard. Fine-tuning for multi-domain and multi-label uncivil language detection. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pp. 28–33, 2020.

Endang Wahyu Pamungkas and Viviana Patti. Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon. In *Proceedings of the 57th annual meeting of the association for computational linguistics: Student research workshop*, pp. 363–370, 2019.

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

María Antonia Paz, Julio Montero-Díaz, and Alicia Moreno-Delgado. Hate speech: A systematized review. *Sage Open*, 10(4):2158244020973022, 2020.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *JMLR*, 12:2825–2830, 2011.

Joni Salminen, Maximilian Hopf, Shammur A Chowdhury, Soon-gyo Jung, Hind Almerekhi, and Bernard J Jansen. Developing an online hate classifier for multiple social media platforms. *Human-centric Computing and Information Sciences*, 10:1–34, 2020.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Karthikeyan Umapathy, Sridhar Krishnan, and Raveendra K Rao. Audio signal feature extraction and classification using local discriminant bases. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(4):1236–1246, 2007.

Kunze Wang, Dong Lu, Soyeon Caren Han, Siqu Long, and Josiah Poon. Detect all abuse! toward universal abusive language detection models. *arXiv preprint arXiv:2010.03776*, 2020a.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *NeurIPS*, 33: 5776–5788, 2020b.

Zeerak Waseem, James Thorne, and Joachim Bingel. Bridging the gaps: Multi task learning for domain transfer of hate speech detection. *Online harassment*, pp. 29–55, 2018.

Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg. Inducing a lexicon of abusive words–a feature-based approach. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1046–1056, 2018.

Ankit Yadav, Shubham Chandel, Sushant Chatufale, and Anil Bandhakavi. Lahm: Large annotated dataset for multi-domain and multilingual hate speech identification. *arXiv preprint arXiv:2304.00913*, 2023.

Lanqin Yuan, Tianyu Wang, Gabriela Ferraro, Hanna Suominen, and Marian-Andrei Rizoiu. Transfer learning for hate speech detection in social media. *arXiv preprint arXiv:1906.03829*, 2019.

A Related Work

In the literature, several hate speech detection algorithms have been developed. Similar to any classification problem, the algorithms developed for hate speech detection can be broadly categorized into two broad categories: (i) based on the utilization of handcrafted features along with traditional machine learning classifiers and (ii) use of deep learning architectures for hate speech detection. For example, Salminen et al. (Salminen et al., 2020) have utilized logistic regression, support vector machine, and gradient boosting classifiers on top of simple features such as the length of the comment, use of uppercase characters, and punctuation. Following a similar strategy, several other researchers (Chowdhury et al., 2020; Wiegand et al., 2018; Karan & Šnajder, 2018; Pamungkas & Patti, 2019; Pamungkas et al., 2020) have proposed traditional machine learning classifiers on the effective text features for hate speech detection. Apart from traditional machine learning classifiers, researchers have also used keyword-based filtering for the possible detection of hate speeches. For example, the sentence structure has been used to detect hate speeches (Mondal et al., 2017). While the above approaches are found effective they lack generalizability as the hate speeches have a wide sense of variability such as linguistic variability. Hence, recently several research efforts have been started utilizing deep learning architectures including convolutions neural networks (Meyer & Gambäck, 2019) (Wang et al., 2020a), LSTM (Wang et al., 2020a; Arango et al., 2019; Waseem et al., 2018; Yuan et al., 2019), and Transformers (Caselli et al., 2020; Koufakou et al., 2020; Glavaš et al., 2020; Mozafari et al., 2020; Ozler et al., 2020).

One of the major drawbacks of the above hate detection algorithms is that the systems are unimodal. It is well known that hate speeches are not limited to text-based social media content but have a large volume in terms of audio and video as well. Interestingly, the majority of the benchmark hate speech detection datasets are text-based which limits the development of a unified algorithm countering hate speeches. It is important to note that text-based speech datasets made a tremendous effort in developing an effective hate speech classifier. For example, ToxiGen (Hartvigsen et al., 2022) consists of 274k toxic and benign statements about 13 minority groups. Similarly, recently proposed LAHM (Yadav et al., 2023) is a multilingual such as English, and Hindi, and a multi-domain text hate speech dataset including abuse and racism. However, we assert that ignoring the other modalities in which hate is prominent can limit the universality of the detection algorithms and provide shallow protection to social media content. A few research efforts have also been started to use the audio modality for hate speech detection. For instance, Ibanez et al. (Ibañez et al., 2021) have used the MFCC audio features and Boishakhi et al. (Boishakhi et al., 2021) also used the text embedding vector along with MFCC audio features for the detection of hate speeches.

The above literature shows that limited work has been done on hate speech detection using audio modality or multimodal hate speech detection using audio and text. Further, no benchmark audio-based hate speech dataset exists similar to text-based datasets such as LAHM (Ibañez et al., 2021) consisting of demographic variations such as gender and age variations. Therefore, in this research, to tackle the existing limitations, we have not only proposed an audio-based hate speech detection dataset but also proposed an unimodal (audio and text separately) but also a multimodal hate speech detection algorithm.

Literature on acoustic and/or audio classification shows that the MFCC features are an effective medium in encoding the data. It is found in the classification of sound in multiple categories as compared to other features such as the Fourier transform and linear prediction analysis (Kumar et al., 2016; Dhanalakshmi et al., 2009; Umapathy et al., 2007; Dhanalakshmi et al., 2011).

B Proposed Hate Audio Dataset

In this paper, we have used the randomly extracted 200 text instances from Toxigen and generated the 600 voice samples of hate and non-hate classes. Out of these 600 audio samples, 200 samples belong to the adult male, 200 to the adult female, and the remaining 200 to the child demographic entity. Each audio sample ranges from a minimum of 3 sec to a maximum of 10 sec. The primary reason behind such a wide variety of audio data generation is that the production of hate data is not limited to any particular demographic group (male or female) or restricted to only adults (hence, the child category is selected). Audio samples are generated using the Neural TTS Speech-based mechanism and contain the voice of a native English speaker of a white (USA) ethnic background.

C IMPLEMENTATION DETAILS

We want to highlight that, we have mentioned that we have used the default parameters of the classifiers, encoders, and generation method have been used in this paper. For example, for SVM a default 'rbf' kernel of degree 3 has been used. Similarly for NuSVC, a default value of 0.5 for parameter nu is used along with the 'rbf' kernel of degree 3. These parameters are default in the sklearn libraries and default parameters against each classifier are used. In the future, we aim to tune these parameters along with the development of a novel multimodal deep attention classifier. Similarly for encoding the text and audio, pre-trained models are used and used as a feature extractor. Once the features are extracted, they are passed to the classifiers for hate detection.

To encode the audio data, four distinct approaches are used to extract Mel-Frequency Cepstral Coefficients (MFCCs) using the Librosa library[4], with varying hyperparameters. The first extractor, mfcc1, computes MFCCs with default Librosa settings. In contrast, mfcc2 specifically computes 13 MFCC coefficients, a common choice in many speech and audio processing tasks, offering a balance between capturing relevant features and computational efficiency. The third and fourth extractors, mfcc3 and mfcc4, differ in their window functions: mfcc3 uses a Hamming window, which minimizes the first side lobe, while mfcc4 employs a Blackman window, known for its high attenuation of side lobes. These window functions affect the time-frequency analysis of the audio signal, potentially capturing different audio characteristics.

[4] https://librosa.org/doc/main/generated/librosa.feature.mfcc.html