Comparative Analysis of Hate Speech Classification in Audio: A Study of Waveform vs. Spectrogram-based Approaches

Damian Kopp [1] Luca Engel [1] Nino Gerber [1]

Abstract

In this study, we compare hate speech classification in audio using waveform and spectrogram-based approaches. Our hypothesis suggests that detecting hate speech in audio files, generated by text-to-speech synthesis (TTS) of social media comments from the dataset ”Measuring-hate-speech”, is viable. We aim to determine if models using spectrograms outperform those using raw audio waveforms. To achieve this, we utilize the Coqui-TTS project for audio sample generation. We fine-tune two models, wav2vec2 and the audio spectrogram transformer (AST), using the generated samples. Additionally, we fine-tune a DistilBERT model trained on the non-synthesized text samples for comparison. Our evaluation on validation and test sets showcases the potential of audio-based hate speech classification. Overall, our findings underscore the effectiveness of audio-based hate speech classification methods and highlight a promising area for further research possibilities.

Keywords: hate speech classification, audio analysis, waveform, spectrogram, Coqui-TTS, wav2vec2, AST, DistilBERT.

1. Introduction

Audio data, initially represented as raw waveforms, captures amplitude samples at regular intervals over time. Converting these waveforms into spectrograms transforms the audio information into image form. Spectrograms visually depict frequency spectra evolving over time in a waveform [1]. This enables treating audio data akin to images, facilitating the application of image classifiers for classification.

Our hypothesis is that we can detect hate speech on audio files, where the audio files are generated by text-to-speech synthesis (TTS) of social media comments. More specifically, we wanted to find out if there is a significant difference in performance between models that use spectrograms as input and models that directly use audio wave forms. Existing models that use audio for classification tasks are mostly multi-modal models. So, they additionally consider other input modalities like image frames. We focused on single-modal audio models to find out if its possible to develop a robust classifier that only uses audio data as input. Lastly, we compare the results of our audio classification models to a text-based classification model using a DistilBERT model [2] fine-tuned on the same dataset.

2. Related Work

Looking for comparable projects, we found [3]. However, their work only focused on more traditional machine learning methods, like Logistic Regression, Random Forest, and Support Vector Machines. Also, the group extracted the audio from TikTok videos and exclusively worked with audio in Filipino language.

With wav2vec2 by Meta [4], we found a pretrained model that seemingly fitted our purpose. As a state-of-the-art model to learn meaningful representations from audio data, it seemed to be a good choice to finetune on.

In this paper [5], AST is introduced as the first convolution-free, purly attention-based model for audio classification. It achieves state-of-the-art results on various benchmarks. Therefore, we decided to fine-tune this model and compare its performance with the wav2vec2 model.

Concerning audio sample generation from text, we relied on the Coqui-TTS [6] project, an open-source Text-to-Speech (TTS) generator freely available for use. This project enables the training of fine-tuned synthesizer models as well as the utilization of pre-trained models for voice cloning or generating samples from existing voices.

3. Method

In our experiments, we fine-tuned both a pretrained wav2vec2 model that directly uses the raw wave forms, and the pretrained audio spectrogram transformer (AST) model as opposing model that uses spectrograms. As an additional comparison, we evaluated the performance of a text

[1] Group 45.

classifier by fine-tuning DistilBERT on our dataset. This shows how the the performance of different audio classification techniques compare to text classification. By using the same dataset for all three models, we are able to analyse and compare the performance of the three models.

The fine-tuning process of all three models made use of the Trainer class together with the TrainingArguments. Running fine-tuning with different hyperparameters hence comes down to adapting some input arguments of the TrainingArguments instance.

3.1. Dataset

The dataset utilized originates from UC Berkeley D-Lab and is denoted as ”Measuring Hate Speech” [7]. It encompasses 39,565 distinct samples, annotated by diverse annotators, sourced from a Twitter text corpus.

3.1.1. Preprocessing

During preprocessing, we extracted both the text and corresponding hate speech scores from each dataset entry. Within this dataset, we cataloged 25,255 supportive comments, 3,873 neutral comments, and 10,437 hateful comments. Subsequently, we curated a balanced dataset, segregating samples into two categories: non-hate for hate speech scores $<$ 0.5 and hate for hate speech scores $\geq$ 0.5. The spectrum of hate speech scores spans from a minimum of -8.34 to a maximum of 6.3. Our aim was to achieve a dataset distribution approximating a normal distribution, centered around the mean hate speech score of 0.5. Accordingly, 50% of the samples are hateful and the remaining are non-hateful, after the preprocessing step. The distribution is visually depicted in Figure 1.

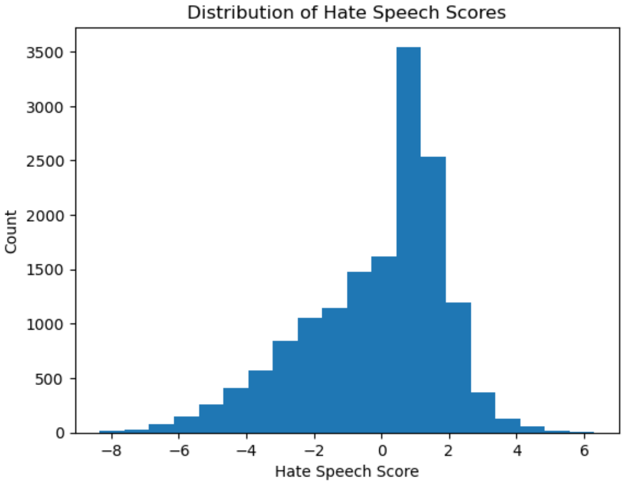

Figure 1. Distribution of hate speech scores after preprocessing

Furthermore, we prepared the samples for the generator and removed all emojis, hashtags, mentions, etc., which were present in the dataset.

3.1.2. Synthesis

Data synthesis was facilitated through the utilization of the Coqui-TTS project. Following rigorous experimentation with various models from the project, we ascertained that the ”Jenny” model yielded the most optimal results for our text corpus, providing clarity and coherence. Consequently, we leveraged the Jenny model to synthesize text samples, which manifested in clear and intelligible outputs. [8]

3.2. Wav2Vec2

The first experiment made use of the pretrained ”facebook/wav2vec2-base” model and feature extractor [4]. The reason for this choice is that this model seems sufficiently large to capture the most significant information, but still not too complex regarding training time. This trade-off should make training runs efficient and performant at the same time.

The pretrained model was built by Facebook AI to learn meaningful speech representations from audio inputs in a self-supervised manner. This model was used to build a hate speech classifier on top. [9]

3.3. AST

In addition to the waveform classification model, we decided to fine-tune a spectrogram-based classification model, AST, developed at MIT [5]. This model treats the spectrograms, which are visual representations of audio frequencies over time, as images and leverages techniques from image classification for this audio classification task. We chose the AST model as it provides an alternative perspective on audio classification compared to the Wav2Vec2 model. With the incorporation of this model, we aimed to explore how different representations of audio can impact the performance of hate speech classification models. Furthermore, training this model on the same dataset allows for a fair comparison and comprehensive analysis of the performances of these models. As there exist only fine-tuned versions of this model on Huggingface, we picked the already fine-tuned ”MIT/ast-finetuned-audioset-10-10-0.4593” model. [10]

3.4. DistilBERT

Additionally to the Audio classification models, we incorporated a fine-tuned DistilBERT model trained on the original text dataset. The reasoning behind this selection lies in its balance between model size and performance. As a distilled version of the BERT model, DistilBERT offers a computationally efficient model without a large accuracy compromise [2]. This makes it a suitable comparison for the fine-tuned AST and Wave2Vec2 models. Having fine-tuned DistilBERT on the same dataset that was synthesized for the audio-based classifiers allows for a precise comparison and analysis of the results. [11]

4. Validation

In the following tables, the performance metrics of the fine-tuned models on the validation and test set, respectively, are presented. For all three fine-tunings, the model yielding the the highest f1 score on the validation set across all trained epochs has been selected.

Model	Loss	Acc	Recall	Prec	F1
wav2vec2	0.656	0.622	0.785	0.599	0.680
AST	0.631	0.649	0.837	0.614	0.708
DistilBERT	0.998	0.774	0.812	0.753	0.781

Table 1. Performance Metrics on the Validation Set.

As becomes apparent in the following table, the models had a similar performance on the test set as on the validation set.

Model	Loss	Acc	Recall	Prec	F1
wav2vec2	0.660	0.619	0.782	0.594	0.676
AST	0.644	0.632	0.819	0.600	0.693
DistilBERT	1.064	0.754	0.793	0.741	0.766

Table 2. Performance Metrics on the Test Set.

For training the models, the following configurations were used. The base parameters were kept the same for all models. In table 4, the parameters that differed across the trainings are listed.

Parameter	Value
Optimizer	Adam
Lr scheduler type	linear
Lr scheduler warmup ratio	0.1
Weight Decay	0.005
Num epochs	10

Table 3. Training Base Configurations

The desicion to reduce the batch size and the number of gradient accumulation steps for the AST model had to be made based on the computational limitations:

Parameter	wave2vec2	AST	DistilBERT
Learning Rate	4e-5	3e-5	8e-5
Batch Size	32	16	32
GAS	4	2	4

Table 4. Training Individual Configurations. (GAS = Gradient Accumulation Steps)

Our fine-tuned AST model has a slightly better F1 score than the Wave2Vec model, outperforming it by 0.0172. There are multiple factors that could have contributed to this. One reason could be that the AST model has fewer parameters than the Wave2Vec2 model with 86M compared to 94M. This could have allowed the AST model to more quickly converge to a high performance on the dataset. However, it is imaginable that, with a bigger and more diverse dataset concerning voices, pitch, and background noise, the Wave2Vec2 might be more capable at classifying unseen data samples due to its size.

The fine-tuned DistilBERT model for text classification achieved the best performance with an F1 score of 0.7659. Then, the fine-tuned AST model with 0.6927 and, lastly, the Wave2Vec2 model with 0.6755. As DistilBERT is a distilled version of the state-of-the art 110 million parameter BERT model [12], it is to be expected that it outperforms the audio based classification models. Nevertheless, there is only a small performance decrease of the two audio classification models. This shows the potential

One limitation of the fine-tuned audio classification models is the training data. As it consists of a synthesized text dataset, the resemblance of the data to real human speech is limited. This could be reduced by creating a dataset with human voices, however, due to time limitations this would not have been feasible.

5. Conclusion

Our comparative analysis shows that audio classification is a viable approach for hate speech detection. We have shown that fine-tuning a the Wave2Vec2 and an AST model can achieve performances not far from a fine-tuned DistilBERT text classification performance on the same text-based dataset. With the relatively little amount of research having been done for audio-based hate speech classification, shown by the lack of purely audio-based hate speech classification datasets, our analysis highlights the potential of such models. Therefore, this area yields an interesting research field for further exploration of the classification performance on larger datasets and, more importantly, human speech based datasets.

References

[1] Wikipedia, “Spectrogram,” 2024.

[2] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter,” *arXiv*, vol. abs/1910.01108, 2019.

[3] M. Ibañez, R. Sapinit, L. A. Reyes, M. Hussien, J. M. Imperial, and R. Rodriguez, “Audio-based hate speech classification from online short-form videos.” 2021 International Conference on Asian Language Processing (IALP), 2021.

[4] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, “Wav2vec 2.0: Learning the structure of speech from raw audio,” *arXiv*, vol. abs/2006.11477, 2020.

[5] Y. Gong, Y. Chung, and J. R. Glass, “AST: audio spectrogram transformer,” *arXiv*, vol. abs/2104.01778, 2021.

[6] E. Gölge, J. Meyer, and K. Davis, “Github: coqui-ai,” 2024.

[7] C. J. Kennedy, G. Bacon, A. Sahn, and C. von Vacano, “Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application,” *arXiv*, vol. abs/2009.10277, 2020.

[8] N. Gerber, “Huggingface: Dl-project/hatespeech-synthesized-dataset,” 2024.

[9] D. Kopp, “Huggingface: Dl-project/hatespeech-wav2vec2,” 2024.

[10] L. Engel, “Huggingface: Dl-project/hatespeech-ast,” 2024.

[11] L. Engel, “Huggingface: Dl-project/hatespeech-distilbert,” 2024.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv*, vol. abs/1810.04805, 2019.